# Contents

1 **Introduction** 1-1
   1.1 Hypothesis Testing, Estimation, and Prediction .......................... 1-1
   1.2 Examples of Uses of Predictive Multivariable Modeling .................. 1-3
   1.3 Misunderstandings about Prediction vs. Classification .................. 1-5
   1.4 Planning for Modeling .............................................. 1-10
   1.5 Choice of the Model .................................................. 1-13
   1.6 Model uncertainty / Data-driven Model Specification .................... 1-14

2 **General Aspects of Fitting Regression Models** 2-1
   2.1 Notation for Multivariable Regression Models .......................... 2-5
   2.2 Model Formulations .................................................... 2-6
   2.3 Interpreting Model Parameters ....................................... 2-7
      2.3.1 Nominal Predictors .............................................. 2-7
      2.3.2 Interactions ....................................................... 2-8
      2.3.3 Example: Inference for a Simple Model .......................... 2-9
      2.3.4 Review of Composite (Chunk) Tests ............................. 2-11
   2.4 Relaxing Linearity Assumption for Continuous Predictors ............... 2-13
## CONTENTS

2.4.1 Avoiding Categorization ........................................ 2-13
2.4.2 Simple Nonlinear Terms ........................................ 2-18
2.4.3 Splines for Estimating Shape of Regression Function and Determining Predictor Transformations ..................... 2-18
2.4.4 Cubic Spline Functions ........................................... 2-21
2.4.5 Restricted Cubic Splines ......................................... 2-22
2.4.6 Choosing Number and Position of Knots ....................... 2-28
2.4.7 Nonparametric Regression ....................................... 2-30
2.4.8 Advantages of Regression Splines over Other Methods ...... 2-32

2.5 Recursive Partitioning: Tree-Based Models ....................... 2-34
2.5.1 New Directions in Predictive Modeling ......................... 2-35
2.5.2 Choosing Between Machine Learning and Statistical Modeling ................................................................. 2-38

2.6 Multiple Degree of Freedom Tests of Association ................. 2-41

2.7 Assessment of Model Fit ............................................ 2-43
2.7.1 Regression Assumptions ......................................... 2-43
2.7.2 Modeling and Testing Complex Interactions ................. 2-47
2.7.3 Fitting Ordinal Predictors ....................................... 2-52
2.7.4 Distributional Assumptions ..................................... 2-53

3 Missing Data .......................................................... 3-1

3.1 Types of Missing Data ............................................... 3-1
3.2 Prelude to Modeling .................................................. 3-2
3.3 Missing Values for Different Types of Response Variables .......... 3-3
3.4 Problems With Simple Alternatives to Imputation ............... 3-4
3.5 Strategies for Developing an Imputation Model ................. 3-7
  3.5.1 Interactions ........................................ 3-10
3.6 Single Conditional Mean Imputation .......................... 3-11
3.7 Predictive Mean Matching .................................. 3-12
3.8 Multiple Imputation ........................................ 3-12
3.9 Diagnostics .................................................. 3-17
3.10 Summary and Rough Guidelines ............................... 3-19
  3.10.1 Effective Sample Size ............................... 3-20
3.11 Bayesian Methods for Missing Data ........................... 3-22

4 Multivariable Modeling Strategies ................................. 4-1
  4.1 Prespecification of Predictor Complexity Without Later Simplification . 4-3
    4.1.1 Learning From a Saturated Model .................... 4-6
    4.1.2 Using Marginal Generalized Rank Correlations ........ 4-7
  4.2 Checking Assumptions of Multiple Predictors Simultaneously ........ 4-9
  4.3 Variable Selection ........................................ 4-10
    4.3.1 Maxwell’s Demon as an Analogy to Variable Selection . 4-17
  4.4 Overfitting and Limits on Number of Predictors ................ 4-19
  4.5 Shrinkage ............................................... 4-21
  4.6 Collinearity ............................................. 4-24
  4.7 Data Reduction ........................................... 4-26
4.7.1 Redundancy Analysis ........................................ 4-27
4.7.2 Variable Clustering ........................................... 4-28
4.7.3 Transformation and Scaling Variables Without Using Y .... 4-29
4.7.4 Simultaneous Transformation and Imputation ................. 4-31
4.7.5 Simple Scoring of Variable Clusters ......................... 4-35
4.7.6 Simplifying Cluster Scores .................................. 4-36
4.7.7 How Much Data Reduction Is Necessary? ..................... 4-36
4.8 Other Approaches to Predictive Modeling ....................... 4-39
4.9 Overly Influential Observations ................................. 4-39
4.10 Comparing Two Models ........................................ 4-42
4.11 Improving the Practice of Multivariable Prediction ............. 4-44
4.12 Summary: Possible Modeling Strategies ....................... 4-47
  4.12.1 Developing Predictive Models ............................. 4-47
  4.12.2 Developing Models for Effect Estimation ................. 4-49
  4.12.3 Developing Models for Hypothesis Testing ............... 4-50

5 Describing, Resampling, Validating, and Simplifying the Model 5-1
  5.1 Describing the Fitted Model .................................. 5-1
    5.1.1 Interpreting Effects .................................... 5-1
    5.1.2 Indexes of Model Performance ......................... 5-2
  5.2 The Bootstrap ................................................ 5-6
  5.3 Model Validation ............................................. 5-11
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3.1 Introduction</td>
<td>5-11</td>
</tr>
<tr>
<td>5.3.2 Which Quantities Should Be Used in Validation?</td>
<td>5-14</td>
</tr>
<tr>
<td>5.3.3 Data-Splitting</td>
<td>5-16</td>
</tr>
<tr>
<td>5.3.4 Improvements on Data-Splitting: Resampling</td>
<td>5-17</td>
</tr>
<tr>
<td>5.3.5 Validation Using the Bootstrap</td>
<td>5-18</td>
</tr>
<tr>
<td>5.4 Bootstrapping Ranks of Predictors</td>
<td>5-23</td>
</tr>
<tr>
<td>5.5 Simplifying the Final Model by Approximating It</td>
<td>5-25</td>
</tr>
<tr>
<td>5.5.1 Difficulties Using Full Models</td>
<td>5-25</td>
</tr>
<tr>
<td>5.5.2 Approximating the Full Model</td>
<td>5-25</td>
</tr>
<tr>
<td>5.6 How Do We Break Bad Habits?</td>
<td>5-27</td>
</tr>
<tr>
<td>6 R Software</td>
<td>6-1</td>
</tr>
<tr>
<td>6.1 The R Modeling Language</td>
<td>6-2</td>
</tr>
<tr>
<td>6.2 User-Contributed Functions</td>
<td>6-3</td>
</tr>
<tr>
<td>6.3 The \text{rms} Package</td>
<td>6-5</td>
</tr>
<tr>
<td>6.4 Other Functions</td>
<td>6-12</td>
</tr>
<tr>
<td>7 Modeling Longitudinal Responses using Generalized Least Squares</td>
<td>7-1</td>
</tr>
<tr>
<td>7.1 Notation</td>
<td>7-1</td>
</tr>
<tr>
<td>7.2 Model Specification for Effects on $E(Y)$</td>
<td>7-3</td>
</tr>
<tr>
<td>7.2.1 Common Basis Functions</td>
<td>7-3</td>
</tr>
<tr>
<td>7.2.2 Model for Mean Profile</td>
<td>7-3</td>
</tr>
<tr>
<td>7.2.3 Model Specification for Treatment Comparisons</td>
<td>7-4</td>
</tr>
</tbody>
</table>
7.3 Modeling Within-Subject Dependence ........................................... 7-6
7.4 Parameter Estimation Procedure ................................................. 7-10
7.5 Common Correlation Structures ................................................. 7-12
7.6 Checking Model Fit ................................................................. 7-14
7.7 R Software ........................................................................... 7-15
7.8 Case Study .............................................................. 7-16
  7.8.1 Graphical Exploration of Data ............................................... 7-16
  7.8.2 Using Generalized Least Squares .......................................... 7-20
  7.8.3 Bayesian Proportional Odds Random Effects Model ............. 7-29

8 Case Study in Data Reduction ..................................................... 8-1
  8.1 Data ............................................................................. 8-2
  8.2 How Many Parameters Can Be Estimated? .............................. 8-2
  8.3 Redundancy Analysis ............................................................. 8-2
  8.4 Variable Clustering ............................................................... 8-2
  8.5 Transformation and Single Imputation Using transcan ............ 8-2
  8.6 Data Reduction Using Principal Components ......................... 8-2
    8.6.1 Sparse Principal Components ......................................... 8-3
  8.7 Transformation Using Nonparametric Smoothers .................... 8-3

9 Maximum Likelihood Estimation ................................................. 9-1

10 Binary Logistic Regression ....................................................... 10-1
10.1 Model ................................................................. 10-2
  10.1.1 Model Assumptions and Interpretation of Parameters ....... 10-3
  10.1.2 Odds Ratio, Risk Ratio, and Risk Difference ................. 10-4
  10.1.3 Detailed Example ............................................ 10-5
  10.1.4 Design Formulations ........................................ 10-11

10.2 Estimation ......................................................... 10-13
  10.2.1 Maximum Likelihood Estimates .............................. 10-13
  10.2.2 Estimation of Odds Ratios and Probabilities ............... 10-13
  10.2.3 Minimum Sample Size Requirement ......................... 10-13

10.3 Test Statistics ..................................................... 10-16

10.4 Residuals .......................................................... 10-17

10.5 Assessment of Model Fit ......................................... 10-18

10.6 Collinearity ....................................................... 10-38

10.7 Overly Influential Observations ................................ 10-38

10.8 Quantifying Predictive Ability .................................. 10-38

10.9 Validating the Fitted Model ...................................... 10-40

10.10 Describing the Fitted Model .................................... 10-45

10.11 Bayesian Logistic Model Example ................................ 10-52

11 Case Study in Binary Logistic Regression, Model Selection and Approximation: Predicting Cause of Death 11-1

12 Logistic Model Case Study: Survival of Titanic Passengers 12-1
12.1 Descriptive Statistics ........................................ 12-2
12.2 Exploring Trends with Nonparametric Regression .............. 12-5
12.3 Binary Logistic Model with Casewise Deletion of Missing Values ..... 12-8
12.4 Examining Missing Data Patterns ................................ 12-13
12.5 Single Conditional Mean Imputation ................................ 12-16
12.6 Multiple Imputation ............................................. 12-20
12.7 Summarizing the Fitted Model ................................... 12-24
12.8 Bayesian Analysis .............................................. 12-27

13 Ordinal Logistic Regression .................................. 13-1

13.1 Background ..................................................... 13-1
13.2 Ordinality Assumption ......................................... 13-3
13.3 Proportional Odds Model ....................................... 13-4
13.3.1 Model ....................................................... 13-4
13.3.2 Assumptions and Interpretation of Parameters ............... 13-5
13.3.3 Estimation .................................................. 13-5
13.3.4 Residuals ................................................... 13-5
13.3.5 Assessment of Model Fit ................................... 13-6
13.3.6 Quantifying Predictive Ability ............................... 13-8
13.3.7 Describing the Model ....................................... 13-8
13.3.8 Validating the Fitted Model ................................. 13-9
13.3.9 R Functions ................................................ 13-9
13.4 Continuation Ratio Model .......................................... 13-11
  13.4.1 Model ......................................................... 13-11
  13.4.2 Assumptions and Interpretation of Parameters ............... 13-12
  13.4.3 Estimation ................................................... 13-12
  13.4.4 Residuals .................................................... 13-12
  13.4.5 Assessment of Model Fit .................................... 13-13
  13.4.6 Extended CR Model ......................................... 13-13
  13.4.7 Role of Penalization in Extended CR Model .................. 13-13
  13.4.8 Validating the Fitted Model ................................ 13-13
  13.4.9 R Functions ............................................... 13-13

14 Case Study in Ordinal Regression, Data Reduction, and Penalization 14-1

15 Regression Models for Continuous $Y$ and Case Study in Ordinal Regression 15-1
  15.1 Dataset and Descriptive Statistics ................................ 15-3
  15.2 The Linear Model ............................................... 15-7
    15.2.1 Checking Assumptions of OLS and Other Models .......... 15-7
  15.3 Quantile Regression ............................................. 15-11
  15.4 Ordinal Regression Models for Continuous $Y$ .................. 15-13
  15.5 Ordinal Regression Applied to HbA$_{1c}$ ....................... 15-19
    15.5.1 Checking Fit for Various Models Using Age ............... 15-19
    15.5.2 Examination of BMI ..................................... 15-24
# CONTENTS

16 Models Using Nonparametric Transformations of $X$ and $Y$ 16-1

17 Introduction to Survival Analysis 17-1

17.1 Background ....................................................... 17-1
17.2 Censoring, Delayed Entry, and Truncation ......................... 17-3
17.3 Notation, Survival, and Hazard Functions .......................... 17-4
17.4 Homogeneous Failure Time Distributions .......................... 17-8
17.5 Nonparametric Estimation of $S$ and $\Lambda$ ....................... 17-10
  17.5.1 Kaplan–Meier Estimator ..................................... 17-10
  17.5.2 Altschuler–Nelson Estimator ................................. 17-12
17.6 Analysis of Multiple Endpoints ................................. 17-13
  17.6.1 Competing Risks ............................................. 17-13
  17.6.2 Competing Dependent Risks ................................. 17-14
  17.6.3 State Transitions and Multiple Types of Nonfatal Events .... 17-17
  17.6.4 Joint Analysis of Time and Severity of an Event .......... 17-17
  17.6.5 Analysis of Multiple Events ................................. 17-18
17.7 $S$ Functions ...................................................... 17-20

18 Parametric Survival Models 18-1

18.1 Homogeneous Models (No Predictors) .......................... 18-1
  18.1.1 Specific Models ............................................. 18-2
  18.1.2 Estimation ................................................... 18-2
  18.1.3 Assessment of Model Fit ..................................... 18-4
## 18.2 Parametric Proportional Hazards Models

- **Model** ................................................. 18-5
- **Model Assumptions and Interpretation of Parameters** .................. 18-5
- **Hazard Ratio, Risk Ratio, and Risk Difference** ......................... 18-6
- **Specific Models** ....................................... 18-7
- **Assessment of Model Fit** ................................ 18-8

## 18.3 Accelerated Failure Time Models

- **Model** ................................................. 18-11
- **Model Assumptions and Interpretation of Parameters** .................. 18-11
- **Specific Models** ....................................... 18-12
- **Estimation** ............................................. 18-13
- **Residuals** ............................................. 18-13
- **Assessment of Model Fit** ................................ 18-13
- **Validating the Fitted Model** ................................ 18-21

## 18.4 R Functions ............................................. 18-23

## 19 Case Study in Parametric Survival Modeling and Model Approximation

- **Descriptive Statistics** .................................. 19-2
- **Checking Adequacy of Log-Normal Accelerated Failure Time Model** 19-7
- **Summarizing the Fitted Model** ................................ 19-14
- **Internal Validation of the Fitted Model Using the Bootstrap** ........ 19-18
- **Approximating the Full Model** ................................ 19-20
20 Cox Proportional Hazards Regression Model

20.1 Model ................................................................. 20-1

20.1.1 Preliminaries .................................................. 20-1
20.1.2 Model Definition ............................................. 20-2
20.1.3 Estimation of $\beta$ .......................................... 20-3
20.1.4 Model Assumptions and Interpretation of Parameters .... 20-3
20.1.5 Example ......................................................... 20-4
20.1.6 Design Formulations ........................................ 20-5
20.1.7 Extending the Model by Stratification .................... 20-6

20.2 Estimation of Survival Probability and Secondary Parameters ... 20-8

20.3 Sample Size Considerations ................................... 20-10

20.4 Test Statistics .................................................... 20-12

20.5 Residuals .......................................................... 20-13

20.6 Assessment of Model Fit ....................................... 20-14

20.6.1 Regression Assumptions ................................... 20-14
20.6.2 Proportional Hazards Assumption ......................... 20-19

20.7 What to Do When PH Fails .................................... 20-26

20.8 Collinearity ......................................................... 20-28

20.9 Overly Influential Observations ............................... 20-28

20.10 Quantifying Predictive Ability ................................ 20-28

20.11 Validating the Fitted Model .................................. 20-30

20.11.1 Validation of Model Calibration ......................... 20-30
CONTENTS

20.11.2 Validation of Discrimination and Other Statistical Indexes . . . 20-31
20.12 Describing the Fitted Model . . . . . . . . . . . . . . . . . . . . . . . 20-33
20.13 R Functions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 20-36
  20.13.1 Power and Sample Size Calculations, Hmisc Package . . . . . 20-36
  20.13.2 Cox Model using rms Package . . . . . . . . . . . . . . . . . . 20-36

21 Case Study in Cox Regression 21-1
  21.1 Choosing the Number of Parameters and Fitting the Model . . . . 21-1
  21.2 Checking Proportional Hazards . . . . . . . . . . . . . . . . . . . . 21-7
  21.3 Testing Interactions . . . . . . . . . . . . . . . . . . . . . . . . . . 21-9
  21.4 Describing Predictor Effects . . . . . . . . . . . . . . . . . . . . . . 21-10
  21.5 Validating the Model . . . . . . . . . . . . . . . . . . . . . . . . . . 21-11
  21.6 Presenting the Model . . . . . . . . . . . . . . . . . . . . . . . . . . 21-13

Bibliography 22-1

in the right margin indicates a hyperlink to a YouTube video related to the subject.

in the right margin is a hyperlink to an audio file elaborating on the notes. Red letters and numbers in the right margin are cues referred to within the audio recordings.

Rotated boxed blue text in the right margin at the start of a section represents the mnemonic key for linking to archival discussions about that section in vbiostatcourse.slack.com channel #rms. New discussions should be on datamethods and vandystats.zulipchat.com.

blog in the right margin is a link to a blog entry that further discusses the topic.
Course Philosophy

- Modeling is the endeavor to transform data into information and information into either prediction or evidence about the data generating mechanism.

- Models are usually the best descriptive statistics
  - adjust for one variable while displaying the association with $Y$ and another variable
  - descriptive statistics do not work in higher dimensions

- Satisfaction of model assumptions improves precision and increases statistical power
  - Be aware of assumptions, especially those mattering the most

- It is more productive to make a model fit step by step (e.g., transformation estimation) than to postulate a simple model and find out what went wrong
  - Model diagnostics are often not actionable
  - Changing the model in reaction to observed patterns ↑ uncertainty but is reflected by an apparent ↓ in uncertainty

- Graphical methods should be married to formal inference

---

*Thanks to Drew Levy for ideas that greatly improved this section.*
• Overfitting occurs frequently, so data reduction and model validation are important

• Software without multiple facilities for assessing and fixing model fit may only seem to be user-friendly

• Carefully fitting an improper model is better than badly fitting (and overfitting) a well-chosen one
  – E.g. small $N$ and overfitting vs. carefully formulated right hand side of model

• Methods which work for all types of regression models are the most valuable.

• In most research projects the cost of data collection far outweighs the cost of data analysis, so it is important to use the most efficient and accurate modeling techniques, to avoid categorizing continuous variables, and to not remove data from the estimation sample just to be able to validate the model.
  – A $100 analysis can make a $1,000,000 study worthless.

• The bootstrap is a breakthrough for statistical modeling and model validation.

• Bayesian modeling is ready for prime time.
  – Can incorporate non-data knowledge
– Provides full exact inferential tools even when penalizing $\beta$

– Rational way to account for model uncertainty

– Direct inference: evidence for all possible values of $\beta$

– More accurate way of dealing with missing data

• Using the data to guide the data analysis is almost as dangerous as not doing so.

• A good overall strategy is to decide how many degrees of freedom (i.e., number of regression parameters) can be “spent”, where they should be spent, to spend them with no regrets.

See the excellent text *Clinical Prediction Models* by Steyerberg [195].
Chapter 1

Introduction

1.1 Hypothesis Testing, Estimation, and Prediction

Even when only testing $H_0$ a model based approach has advantages:

- Permutation and rank tests not as useful for estimation

- Cannot readily be extended to cluster sampling or repeated measurements

- Models generalize tests
  - 2-sample $t$-test, ANOVA $\rightarrow$ multiple linear regression
  - Wilcoxon, Kruskal-Wallis, Spearman $\rightarrow$ proportional odds ordinal logistic model
– log-rank → Cox

• Models not only allow for multiplicity adjustment but for shrinkage of estimates

– Statisticians comfortable with $P$-value adjustment but fail to recognize that the difference between the most different treatments is badly biased

Statistical estimation is usually model-based

• Relative effect of increasing cholesterol from 200 to 250 mg/dl on hazard of death, holding other risk factors constant

• Adjustment depends on how other risk factors relate to hazard

• Usually interested in adjusted (partial) effects, not unadjusted (marginal or crude) effects
1.2 Examples of Uses of Predictive Multivariable Modeling

- Financial performance, consumer purchasing, loan pay-back
- Ecology
- Product life
- Employment discrimination
- Medicine, epidemiology, health services research
- Probability of diagnosis, time course of a disease
- Checking that a previously developed summary index (e.g., BMI) adequately summarizes its component variables
- Developing new summary indexes by how variables predict an outcome
- Comparing non-randomized treatments
- Getting the correct estimate of relative effects in randomized studies requires covariable adjustment if model is nonlinear
  - Crude odds ratios biased towards 1.0 if sample heterogeneous
• Estimating absolute treatment effect (e.g., risk difference)
  – Use e.g. difference in two predicted probabilities

• Cost-effectiveness ratios
  – incremental cost / incremental \textit{ABSOLUTE} benefit
  – most studies use avg. cost difference / avg. benefit, which may apply to no one
1.3 Misunderstandings about Prediction vs. Classification

- Many analysts desire to develop “classifiers” instead of predictions

- Outside of, for example, visual or sound pattern recognition, classification represents a premature decision

- See this blog for details

- Suppose that
  1. response variable is binary
  2. the two levels represent a sharp dichotomy with no gray zone (e.g., complete success vs. total failure with no possibility of a partial success)
  3. one is forced to assign (classify) future observations to only these two choices
  4. the cost of misclassification is the same for every future observation, and the ratio of the cost of a false positive to the cost of a false negative equals the (often hidden) ratio implied by the analyst’s classification rule

- Then classification is still suboptimal for driving the development of a predictive instrument as well as for hypothesis testing and estimation
• Classification and its associated classification accuracy measure—the proportion classified “correctly”—are very sensitive to the relative frequencies of the outcome variable. If a classifier is applied to another dataset with a different outcome prevalence, the classifier may no longer apply.

• Far better is to use the full information in the data to develop a probability model, then develop classification rules on the basis of estimated probabilities
  – ↑ power, ↑ precision, ↑ decision making

• Classification is more problematic if response variable is ordinal or continuous or the groups are not truly distinct (e.g., disease or no disease when severity of disease is on a continuum); dichotomizing it up front for the analysis is not appropriate
  – *minimum* loss of information (when dichotomization is at the median) is large
  – may require the sample size to increase many–fold to compensate for loss of information [73]

• Two-group classification represents artificial forced choice
  – best option may be “no choice, get more data”

• Unlike prediction (e.g., of absolute risk), classification implicitly uses utility (loss; cost of false positive or false nega-
tive) functions

- **Hidden problems:**
  - Utility function depends on variables not collected (subjects’ preferences) that are available only at the decision point
  - Assumes every subject has the same utility function
  - Assumes this function coincides with the analyst’s

- **Formal decision analysis uses**
  - optimum predictions using all available data
  - subject-specific utilities, which are often based on variables not predictive of the outcome

- **ROC analysis is misleading except for the special case of mass one-time group decision making with unknowable utilities**

See [218, 29, 77, 24, 70, 81].

---

To make an optimal decision you need to know all relevant data about an individual (used to estimate the probability of an outcome), and the utility (cost, loss function) of making each decision. Sensitivity and specificity do not provide this information. For example, if one estimated that the probability of a disease given age, sex, and symptoms is 0.1 and the “cost” of a false positive equaled the “cost” of a false negative, one would act as if the person does not have the disease. Given other utilities, one would make different decisions. If the utilities are unknown, one gives the best estimate of the probability of the outcome to the decision maker and let her incorporate her own unspoken utilities in making an optimum decision for her.

Besides the fact that cutoffs do not apply to individuals, only to groups, individual decision making does not utilize sensitivity and specificity. For an individual we can compute \( \text{Prob}(Y = 1 | X = x) \); we don’t care about \( \text{Prob}(Y = 1 | X > c) \), and an individual having \( X = x \) would be quite puzzled if she were given \( \text{Prob}(X > c | \text{future unknown } Y) \) when she already knows \( X = x \) so \( X \) is no longer a random variable.

Even when group decision making is needed, sensitivity and specificity can be bypassed. For mass marketing, for example, one can rank order individuals by the estimated probability of buying the product, to create a lift curve. This is then used to target the \( k \) most likely buyers where \( k \) is chosen to meet total program cost constraints.
Accuracy score used to drive model building should be a continuous score that utilizes all of the information in the data.

In summary:

- Classification is a forced choice — a decision.
- Decisions require knowledge of the cost or utility of making an incorrect decision.
- Predictions are made without knowledge of utilities.
- A prediction can lead to better decisions than classification. For example suppose that one has an estimate of the risk of an event, \( \hat{P} \). One might make a decision if \( \hat{P} < 0.10 \) or \( \hat{P} > 0.90 \) in some situations, even without knowledge of utilities. If on the other hand \( \hat{P} = 0.6 \) or the confidence interval for \( P \) is wide, one might
  - make no decision and instead opt to collect more data
  - make a tentative decision that is revisited later
  - make a decision using other considerations such as the infusion of new resources that allow targeting a larger number of potential customers in a marketing campaign

The Dichotomizing Motorist

- The speed limit is 60.
• I am going faster than the speed limit.

• Will I be caught?

An answer by a dichotomizer:

• Are you going faster than 70?

An answer from a better dichotomizer:

• If you are among other cars, are you going faster than 73?

• If you are exposed are you going faster than 67?

Better:

• How fast are you going and are you exposed?

Analogy to most medical diagnosis research in which +/- diagnosis is a false dichotomy of an underlying disease severity:

• The speed limit is moderately high.

• I am going fairly fast.

• Will I be caught?
Planning for Modeling

- Chance that predictive model will be used [175]
- Response definition, follow-up
- Variable definitions
- Observer variability
- Missing data
- Preference for continuous variables
- Subjects
- Sites

What can keep a sample of data from being appropriate for modeling:

1. Most important predictor or response variables not collected
2. Subjects in the dataset are ill-defined or not representative of the population to which inferences are needed
3. Data collection sites do not represent the population of sites
4. Key variables missing in large numbers of subjects
5. Data not missing at random
6. No operational definitions for key variables and/or measurement errors severe
7. No observer variability studies done

What else can go wrong in modeling?

1. The process generating the data is not stable.
2. The model is misspecified with regard to nonlinearities or interactions, or there are predictors missing.
3. The model is misspecified in terms of the transformation of the response variable or the model’s distributional assumptions.
4. The model contains discontinuities (e.g., by categorizing continuous predictors or fitting regression shapes with sudden changes) that can be gamed by users.
5. Correlations among subjects are not specified, or the correlation structure is misspecified, resulting in inefficient parameter estimates and overconfident inference.
6. The model is overfitted, resulting in predictions that are too extreme or positive associations that are false.
7. The user of the model relies on predictions obtained by extrapolating to combinations of predictor values well outside the range of the dataset used to develop the model.
8. Accurate and discriminating predictions can lead to behavior changes that make future predictions inaccurate.
Iezzoni [109] lists these dimensions to capture, for patient outcome studies:

1. age
2. sex
3. acute clinical stability
4. principal diagnosis
5. severity of principal diagnosis
6. extent and severity of comorbidities
7. physical functional status
8. psychological, cognitive, and psychosocial functioning
9. cultural, ethnic, and socioeconomic attributes and behaviors
10. health status and quality of life
11. patient attitudes and preferences for outcomes

General aspects to capture in the predictors:

1. baseline measurement of response variable
2. current status
3. trajectory as of time zero, or past levels of a key variable
4. variables explaining much of the variation in the response
5. more subtle predictors whose distributions strongly differ between levels of the key variable of interest in an observational study
1.5 Choice of the Model

- In biostatistics and epidemiology and most other areas we usually choose model empirically.

- Model must use data efficiently.

- Should model overall structure (e.g., acute vs. chronic).

- Robust models are better.

- Should have correct mathematical structure (e.g., constraints on probabilities).
1.6 Model uncertainty / Data-driven Model Specification

- Standard errors, C.L., $P$-values, $R^2$ wrong if computed as if the model pre-specified

- Stepwise variable selection is widely used and abused

- Bootstrap can be used to repeat all analysis steps to properly penalize variances, etc.

- Ye [238]: “generalized degrees of freedom” (GDF) for any “data mining” or model selection procedure based on least squares
  
  - Example: 20 candidate predictors, $n = 22$, forward step-wise, best 5-variable model: GDF=14.1
  
  - Example: CART, 10 candidate predictors, $n = 100$, 19 nodes: GDF=76

- See [140] for an approach involving adding noise to $Y$ to improve variable selection
Chapter 2

General Aspects of Fitting Regression Models

Regression modeling meets many analytic needs:

- Prediction, capitalizing on efficient estimation methods such as maximum likelihood and the predominant additivity in a variety of problems
  - E.g.: effects of age, smoking, and air quality add to predict lung capacity
  - When effects are predominantly additive, or when there aren’t too many interactions and one knows the likely interacting variables in advance, regression can beat machine learning techniques that assume interaction effects are likely to be as strong as main effects

- Separate effects of variables (especially exposure and treatment)
• Hypothesis testing

• Deep understanding of uncertainties associated with all model components
  – Simplest example: confidence interval for the slope of a predictor
  – Confidence intervals for predicted values; simultaneous confidence intervals for a series of predicted values

  * E.g.: confidence band for $Y$ over a series of values of $X$

**Alternative: Stratification**

• Cross-classify subjects on the basis of the $X$s, estimate a property of $Y$ for each stratum

• Only handles a small number of $X$s

• Does not handle continuous $X$

**Alternative: Single Trees (recursive partitioning/CART)**

• Interpretable because they are over-simplified and usually wrong

• Cannot separate effects
• Finds spurious interactions

• Require huge sample size

• Do not handle continuous $X$ effectively; results in very heterogeneous nodes because of incomplete conditioning

• Tree structure is unstable so insights are fragile

**Alternative: Machine Learning**

• E.g. random forests, bagging, boosting, support vector machines, neural networks, deep learning

• Allows for high-order interactions and does not require prespecification of interaction terms

• Almost automatic; can save analyst time and do the analysis in one step (long computing time)

• Uninterpretable black box

• Effects of individual predictors are not separable

• Interaction effects (e.g., differential treatment effect = precision medicine = personalized medicine) not available

• Because of not using prior information about dominance of additivity, can require 200 events per candidate predictor
when $Y$ is binary \[212\]

- Logistic regression may require 20 events per candidate predictor

- Can create a demand for “big data” where additive statistical models can work on moderate-size data

- See this article in *Harvard Business Review* for more about regression vs. complex methods
2.1 Notation for Multivariable Regression Models

- Weighted sum of a set of independent or predictor variables
- Interpret parameters and state assumptions by linearizing model with respect to regression coefficients
- Analysis of variance setups, interaction effects, nonlinear effects
- Examining the 2 regression assumptions

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>$Y$</strong></td>
<td>response (dependent) variable</td>
</tr>
<tr>
<td><strong>$X$</strong></td>
<td>$X_1, X_2, \ldots, X_p$ – list of predictors</td>
</tr>
<tr>
<td><strong>$\beta$</strong></td>
<td>$\beta_0, \beta_1, \ldots, \beta_p$ – regression coefficients</td>
</tr>
<tr>
<td><strong>$\beta_0$</strong></td>
<td>intercept parameter (optional)</td>
</tr>
<tr>
<td><strong>$\beta_1, \ldots, \beta_p$</strong></td>
<td>weights or regression coefficients</td>
</tr>
<tr>
<td><strong>$X\beta$</strong></td>
<td>$\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p, X_0 = 1$</td>
</tr>
</tbody>
</table>

Model: connection between $X$ and $Y$

$C(Y|X)$: property of distribution of $Y$ given $X$, e.g.

$C(Y|X) = E(Y|X)$ or $\text{Prob}\{Y = 1|X\}$. 
Model Formulations

General regression model

\[ C(Y|X) = g(X). \]

General linear regression model

\[ C(Y|X) = g(X\beta). \]

Examples

\[ C(Y|X) = E(Y|X) = X\beta, \]
\[ Y|X \sim N(X\beta, \sigma^2) \]
\[ C(Y|X) = \text{Prob}\{Y = 1|X\} = (1 + \exp(-X\beta))^{-1} \]

Linearize: \( h(C(Y|X)) = X\beta, h(u) = g^{-1}(u) \)

Example:

\[ C(Y|X) = \text{Prob}\{Y = 1|X\} = (1 + \exp(-X\beta))^{-1} \]
\[ h(u) = \text{logit}(u) = \log\left(\frac{u}{1-u}\right) \]
\[ h(C(Y|X)) = C''(Y|X) \text{ (link)} \]

General linear regression model:

\[ C''(Y|X) = X\beta. \]
2.3

Interpreting Model Parameters

Suppose that $X_j$ is linear and doesn’t interact with other $X$’s.

\[
C'(Y|X) = X\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \\
\beta_j = C'(Y|X_1, X_2, \ldots, X_j + 1, \ldots, X_p) \\
- C'(Y|X_1, X_2, \ldots, X_j, \ldots, X_p)
\]

Drop $'$ from $C'$ and assume $C(Y|X)$ is property of $Y$ that is linearly related to weighted sum of $X$’s.

2.3.1

Nominal Predictors

Nominal (polytomous) factor with $k$ levels: $k - 1$ dummy variables. E.g. $T = J, K, L, M$:

\[
C(Y|T = J) = \beta_0 \\
C(Y|T = K) = \beta_0 + \beta_1 \\
C(Y|T = L) = \beta_0 + \beta_2 \\
C(Y|T = M) = \beta_0 + \beta_3.
\]

\[
C(Y|T) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3,
\]

where

\[
X_1 = 1 \text{ if } T = K, \ 0 \text{ otherwise}
\]

Note that it is not necessary to “hold constant” all other variables to be able to interpret the effect of one predictor. It is sufficient to hold constant the weighted sum of all the variables other than $X_j$. And in many cases it is not physically possible to hold other variables constant while varying one, e.g., when a model contains $X$ and $X^2$ (David Hoaglin, personal communication).
\[ X_2 = 1 \text{ if } T = L, \ 0 \text{ otherwise} \]
\[ X_3 = 1 \text{ if } T = M, \ 0 \text{ otherwise}. \]

The test for any differences in the property \( C(Y) \) between treatments is \( H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \).

### 2.3.2 Interactions

\( X_1 \) and \( X_2 \), effect of \( X_1 \) on \( Y \) depends on level of \( X_2 \). One way to describe interaction is to add \( X_3 = X_1X_2 \) to model:

\[
C(Y|X) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2.
\]

\[
C(Y|X_1 + 1, X_2) \leftarrow C(Y|X_1, X_2)
\]
\[
= \beta_0 + \beta_1(X_1 + 1) + \beta_2X_2
\]
\[
+ \beta_3(X_1 + 1)X_2
\]
\[
= [\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2]
\]
\[
= \beta_1 + \beta_3X_2.
\]

One-unit increase in \( X_2 \) on \( C(Y|X) : \beta_2 + \beta_3X_1 \).

Worse interactions:

If \( X_1 \) is binary, the interaction may take the form of a difference in shape (and/or distribution) of \( X_2 \) vs. \( C(Y) \) depending on whether \( X_1 = 0 \) or \( X_1 = 1 \) (e.g. logarithm vs. square root).

This paper describes how interaction effects can be misleading.
Example: Inference for a Simple Model

Postulate the model \( C(Y|\text{age}, \text{sex}) = \beta_0 + \beta_1 \text{age} + \beta_2 (\text{sex} = f) + \beta_3 \text{age}(\text{sex} = f) \) where \( \text{sex} = f \) is a dummy indicator variable for \( \text{sex} = \text{female} \), i.e., the reference cell is \( \text{sex} = \text{male} \).

Model assumes

1. age is linearly related to \( C(Y) \) for males,
2. age is linearly related to \( C(Y) \) for females, and
3. interaction between age and sex is simple
4. whatever distribution, variance, and independence assumptions are appropriate for the model being considered.

Interpretations of parameters:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>( C(Y</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>( C(Y</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>( C(Y</td>
</tr>
<tr>
<td>( \beta_3 )</td>
<td>( C(Y</td>
</tr>
</tbody>
</table>

\( \beta_3 \) is the difference in slopes (female – male).

When a high-order effect such as an interaction effect is in the model, be sure to interpret low-order effects by finding out what makes the interaction effect ignorable. In our example, the interaction effect is zero when \( \text{age} = 0 \) or \( \text{sex} = \text{male} \).

\( ^b \)You can also think of the last part of the model as being \( \beta_3 X_3 \), where \( X_3 = \text{age} \times I[\text{sex} = f] \).
Hypotheses that are usually inappropriate:

1. $H_0 : \beta_1 = 0$: This tests whether age is associated with $Y$ for males
2. $H_0 : \beta_2 = 0$: This tests whether sex is associated with $Y$ for zero year olds

More useful hypotheses follow. For any hypothesis need to

- Write what is being tested
- Translate to parameters tested
- List the alternative hypothesis
- Not forget what the test is powered to detect
  - Test against nonzero slope has maximum power when linearity holds
  - If true relationship is monotonic, test for non-flatness will have some but not optimal power
  - Test against a quadratic (parabolic) shape will have some power to detect a logarithmic shape but not against a sine wave over many cycles
- Useful to write e.g. “$H_a : \text{age is associated with } C(Y)$, powered to detect a linear relationship”
### Most Useful Tests for Linear age × sex Model

<table>
<thead>
<tr>
<th>Null or Alternative Hypothesis</th>
<th>Mathematical Statement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effect of age is independent of sex or Effect of sex is independent of age or age and sex are additive age effects are parallel</td>
<td>$H_0 : \beta_3 = 0$</td>
</tr>
<tr>
<td>age interacts with sex age modifies effect of sex sex modifies effect of age sex and age are non-additive (synergistic)</td>
<td>$H_a : \beta_3 \neq 0$</td>
</tr>
<tr>
<td>age is not associated with $Y$ age is associated with $Y$ age is associated with $Y$ for either females or males</td>
<td>$H_0 : \beta_1 = \beta_3 = 0$ $H_a : \beta_1 \neq 0$ or $\beta_3 \neq 0$</td>
</tr>
<tr>
<td>sex is not associated with $Y$ sex is associated with $Y$ sex is associated with $Y$ for some value of age</td>
<td>$H_0 : \beta_2 = \beta_3 = 0$ $H_a : \beta_2 \neq 0$ or $\beta_3 \neq 0$</td>
</tr>
<tr>
<td>Neither age nor sex is associated with $Y$ Either age or sex is associated with $Y$</td>
<td>$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$</td>
</tr>
</tbody>
</table>

**Note:** The last test is called the global test of no association. If an interaction effect present, there is both an age and a sex effect. There can also be age or sex effects when the lines are parallel. The global test of association (test of total association) has 3 d.f. instead of 2 (age + sex) because it allows for unequal slopes.

### 2.3.4

**Review of Composite (Chunk) Tests**

- In the model
  \[ y \sim \text{age} + \text{sex} + \text{weight} + \text{waist} + \text{tricep} \]
we may want to jointly test the association between all body measurements and response, holding age and sex constant.

• This 3 d.f. test may be obtained two ways:
  
  – Remove the 3 variables and compute the change in $SSR$ or $SSE$
  
  – Test $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ using matrix algebra (e.g., `anova(fit, weight, waist, tricep)` if `fit` is a fit object created by the R `rms` package)
2.4

Relaxing Linearity Assumption for Continuous Predictors

2.4.1

Avoiding Categorization

Natura non facit saltus
(Nature does not make jumps)

Gottfried Wilhelm Leibniz

Lucy D’Agostino McGowan

- Relationships seldom linear except when predicting one variable from itself measured earlier

- Categorizing continuous predictors into intervals is a disaster; see references
  [180, 2, 104, 126, 4, 14, 71, 174, 201, 31]
  [145, 185, 5, 106, 149, 223, 73, 79, 46, 18] and Biostatistics for Biomedical Research, Chapter 18.

- Some problems caused by this approach:
1. Estimated values have reduced precision, and associated tests have reduced power

2. Categorization assumes relationship between predictor and response is flat within intervals; far less reasonable than a linearity assumption in most cases

3. To make a continuous predictor be more accurately modeled when categorization is used, multiple intervals are required

4. Because of sample size limitations in the very low and very high range of the variable, the outer intervals (e.g., outer quintiles) will be wide, resulting in significant heterogeneity of subjects within those intervals, and residual confounding

5. Categorization assumes that there is a discontinuity in response as interval boundaries are crossed. Other than the effect of time (e.g., an instant stock price drop after bad news), there are very few examples in which such discontinuities have been shown to exist.

6. Categorization only seems to yield interpretable estimates. E.g. odds ratio for stroke for persons with a systolic blood pressure $> 160$ mmHg compared to persons with a blood pressure $\leq 160$ mmHg $\rightarrow$ interpretation of OR depends on distribution of blood pressures in the sample (the proportion of subjects $> 170$, $> 180$, etc.). If blood pressure is modeled as a continuous variable (e.g., using a regression spline, quadratic, or linear effect) one can estimate the ratio of odds for exact settings of the predictor, e.g.,
the odds ratio for 200 mmHg compared to 120 mmHg.

7. Categorization does not condition on full information. When, for example, the risk of stroke is being assessed for a new subject with a known blood pressure (say 162 mmHg), the subject does not report to her physician “my blood pressure exceeds 160” but rather reports 162 mmHg. The risk for this subject will be much lower than that of a subject with a blood pressure of 200 mmHg.

8. If cutpoints are determined in a way that is not blinded to the response variable, calculation of $P$-values and confidence intervals requires special simulation techniques; ordinary inferential methods are completely invalid. E.g.: cutpoints chosen by trial and error utilizing $Y$, even informally $\rightarrow P$-values too small and CLs not accurate$^c$.

9. Categorization not blinded to $Y \rightarrow$ biased effect estimates $[4, 185]$

10. “Optimal” cutpoints do not replicate over studies. Hollander et al. $[106]$ state that “...the optimal cutpoint approach has disadvantages. One of these is that in almost every study where this method is applied, another cutpoint will emerge. This makes comparisons across studies extremely difficult or even impossible. Altman et al. point out this problem for studies of the prognostic relevance of the S-phase fraction in breast cancer published in the literature. They identified 19 different cutpoints used in the literature; some of them were solely used because they emerged as the 'optimal' cutpoint in a specific data set. In a meta-analysis on the relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients,

$^c$If a cutpoint is chosen that minimizes the $P$-value and the resulting $P$-value is 0.05, the true type I error can easily be above 0.5 $[106]$. 
12 studies were included with 12 different cutpoints. Interestingly, neither cathepsin-D nor the S-phase fraction are recommended to be used as prognostic markers in breast cancer in the recent update of the American Society of Clinical Oncology. Giannoni et al. [79] demonstrated that many claimed “optimal cutpoints” are just the observed median values in the sample, which happens to optimize statistical power for detecting a separation in outcomes.

11. Disagreements in cutpoints (which are bound to happen whenever one searches for things that do not exist) cause severe interpretation problems. One study may provide an odds ratio for comparing body mass index (BMI) $> 30$ with BMI $\leq 30$, another for comparing BMI $> 28$ with BMI $\leq 28$. Neither of these has a good definition and the two estimates are not comparable.

12. Cutpoints are arbitrary and manipulatable; cutpoints can be found that can result in both positive and negative associations [223].

13. If a confounder is adjusted for by categorization, there will be residual confounding that can be explained away by inclusion of the continuous form of the predictor in the model in addition to the categories.

- To summarize: The use of a (single) cutpoint $c$ makes many assumptions, including:

  1. Relationship between $X$ and $Y$ is discontinuous at $X = c$ and only $X = c$
  2. $c$ is correctly found as the cutpoint
  3. $X$ vs. $Y$ is flat to the left of $c$
4. $X$ vs. $Y$ is flat to the right of $c$

5. The choice of $c$ does not depend on the values of other predictors

Interactive demonstration of power loss of categorization vs. straight line and quadratic fits in OLS, with varying degree of nonlinearity and noise added to $X$ (must run in RStudio)

```r
require(Hmisc)
getRs('catgNoise.r')
```

Example of misleading results from creating intervals (here, deciles) of a continuous predictor. Final interval is extremely heterogeneous and is greatly influenced by very large glycohemoglobin values, creating the false impression of an inflection point at 5.9.

See this for excellent graphical examples of the harm of categorizing predictors, especially when using quantile groups.

---

4From NHANES III; Diabetes Care 32:1327-34; 2009 adapted from Diabetes Care 20:1183-1197; 1997.
2.4.2 Simple Nonlinear Terms

\[ C(Y|X_1) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2. \]

- \( H_0 : \) model is linear in \( X_1 \) vs. \( H_a : \) model is quadratic in \( X_1 \equiv H_0 : \beta_2 = 0. \)

- Test of linearity may be powerful if true model is not extremely non-parabolic

- Predictions not accurate in general as many phenomena are non-quadratic

- Can get more flexible fits by adding powers higher than 2

- But polynomials do not adequately fit logarithmic functions or “threshold” effects, and have unwanted peaks and valleys.

2.4.3 Splines for Estimating Shape of Regression Function and Determining Predictor Transformations

Draftsman’s spline : flexible strip of metal or rubber used to trace curves.

Spline Function : piecewise polynomial
Linear Spline Function: piecewise linear function

- Bilinear regression: model is $\beta_0 + \beta_1 X$ if $X \leq a$, $\beta_2 + \beta_3 X$ if $X > a$.

- Problem with this notation: two lines not constrained to join

- To force simple continuity: $\beta_0 + \beta_1 X + \beta_2 (X - a) \times I[X > a] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, where $X_2 = (X_1 - a) \times I[X_1 > a]$.

- Slope is $\beta_1$, $X \leq a$, $\beta_1 + \beta_2$, $X > a$.

- $\beta_2$ is the slope increment as you pass $a$

More generally: $X$-axis divided into intervals with endpoints $a$, $b$, $c$ (knots).

$$f(X) = \beta_0 + \beta_1 X + \beta_2 (X - a)_+ + \beta_3 (X - b)_+ + \beta_4 (X - c)_+,$$

where

$$(u)_+ = u, \ u > 0,$$

$$0, \ u \leq 0.$$
\[ f(X) = \beta_0 + \beta_1 X, \quad X \leq a \]
\[ = \beta_0 + \beta_1 X + \beta_2 (X - a) \quad a < X \leq b \]
\[ = \beta_0 + \beta_1 X + \beta_2 (X - a) + \beta_3 (X - b) \quad b < X \leq c \]
\[ = \beta_0 + \beta_1 X + \beta_2 (X - a) + \beta_3 (X - b) + \beta_4 (X - c) \quad c < X. \]

Figure 2.1: A linear spline function with knots at \( a = 1, b = 3, c = 5 \).

\[ C(Y|X) = f(X) = X\beta, \]

where \( X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \), and
\[ X_1 = X \quad X_2 = (X - a)_+ \]
\[ X_3 = (X - b)_+ \quad X_4 = (X - c)_+. \]

Overall linearity in \( X \) can be tested by testing \( H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \).
2.4.4 Cubic Spline Functions

Cubic splines are smooth at knots (function, first and second derivatives agree) — can’t see joins.

\[ f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 (X - a)^3_+ + \beta_5 (X - b)^3_+ + \beta_6 (X - c)^3_+ = X \beta \]

\[
X_1 = X \quad X_2 = X^2 \\
X_3 = X^3 \quad X_4 = (X - a)^3_+ \\
X_5 = (X - b)^3_+ \quad X_6 = (X - c)^3_+.
\]

\(k\) knots \(\rightarrow\) \(k + 3\) coefficients excluding intercept.

\(X^2\) and \(X^3\) terms must be included to allow nonlinearity when \(X < a\).

stats.stackexchange.com/questions/421964 has some useful descriptions of what happens at the knots, e.g.:

Knots are where different cubic polynomials are joined, and cubic splines force there to be three levels of continuity (the function, its slope, and its acceleration or second derivative (slope of the slope) do not change) at these points. At the knots the jolt (third derivative or rate of change of acceleration) is allowed to change suddenly, meaning the jolt is allowed to be
discontinuous at the knots. Between knots, jolt is constant.

The following graphs show the function and its first three derivatives (all further derivatives are zero) for the function given by 

\[ f(x) = x + x^2 + 2x^3 + 10(x - 0.25)^3 - 50(x - 0.5)^3 - 100(x - 0.75)^3 \]

for \( x \) going from 0 to 1, where there are three knots, at \( x = 0.25, 0.5, 0.75 \).

\[
x \leftarrow \text{seq}(0, 1, \text{length}=500)
\]
\[
x1 \leftarrow \text{pmax}(x - .25, 0)
\]
\[
x2 \leftarrow \text{pmax}(x - .50, 0)
\]
\[
x3 \leftarrow \text{pmax}(x - .75, 0)
\]
\[
b1 \leftarrow 1; b2 \leftarrow 1; b3 \leftarrow 2; b4 \leftarrow 10; b5 \leftarrow -50; b6 \leftarrow -100
\]
\[
y \leftarrow b1 \cdot x + b2 \cdot x^2 + b3 \cdot x^3 + b4 \cdot x1^3 + b5 \cdot x2^3 + b6 \cdot x3^3
\]
\[
y1 \leftarrow b1 + 2\cdot b2 \cdot x + 3\cdot b3 \cdot x^2 + 3\cdot b4 \cdot x1^2 + 3\cdot b5 \cdot x2^2 + 3\cdot b6 \cdot x3^2
\]
\[
y2 \leftarrow 2\cdot b2 + 6\cdot b3 \cdot x + 6\cdot b4 \cdot x1 + 6\cdot b5 \cdot x2 + 6\cdot b6 \cdot x3
\]
\[
y3 \leftarrow 6\cdot b3 + 6\cdot b4 \cdot (x1 > 0) + 6\cdot b5 \cdot (x2 > 0) + 6\cdot b6 \cdot (x3 > 0)
\]

\[
g \leftarrow \text{function()} \text{ abline(v=(1:3)/4, col=gray(.85))}
\]
\[
\text{plot(x, y, type=’l’, ylab=’’); g()}
\]
\[
\text{text(0, 1.5, ’Function’, adj=0)}
\]

\[
\text{plot(x, y1, type=’l’, ylab=’’); g()}
\]
\[
\text{text(0, -15, ’First Derivative: Slope
Rate of Change of Function’, adj=0)}
\]

\[
\text{plot(x, y2, type=’l’, ylab=’’); g()}
\]
\[
\text{text(0, -125, ’Second Derivative: Acceleration
Rate of Change of Slope’, adj=0)}
\]

\[
\text{plot(x, y3, type=’l’, ylab=’’); g()}
\]
\[
\text{text(0, -400, ’Third Derivative: Jolt
Rate of Change of Acceleration’, adj=0)}
\]

### 2.4.5

**Restricted Cubic Splines**

Stone and Koo [199]: cubic splines poorly behaved in tails. Constrain function to be linear in tails.

\( k + 3 \rightarrow k - 1 \) parameters [59].

To force linearity when \( X < a \): \( X^2 \) and \( X^3 \) terms must be omitted
Figure 2.2: A regular cubic spline function with three levels of continuity that prevent the human eye from detecting the knots. Also shown is the function’s first three derivatives. Knots are located at $x = 0.25, 0.5, 0.75$. For $x$ beyond the outer knots, the function is not restricted to be linear. Linearity would imply an acceleration of zero. Vertical lines are drawn at the knots.
To force linearity when \( X > \) last knot: last two \( \beta \)s are redundant, i.e., are just combinations of the other \( \beta \)s.

The restricted spline function with \( k \) knots \( t_1, \ldots, t_k \) is given by [59]

\[
f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{k-1} X_{k-1},
\]

where \( X_1 = X \) and for \( j = 1, \ldots, k-2, \)

\[
X_{j+1} = (X - t_j)^3_+ - (X - t_{k-1})^3_+ (t_k - t_j)/(t_k - t_{k-1}) + (X - t_k)^3_+ (t_{k-1} - t_j)/(t_k - t_{k-1}).
\]

\( X_j \) is linear in \( X \) for \( X \geq t_k \).

For numerical behavior and to put all basis functions for \( X \) on the same scale, \( \text{R Hmisc} \) and \( \text{rms} \) package functions by default divide the terms above by \( \tau = (t_k - t_1)^2 \).

```r
require(Hmisc)
x ← rcspline.eval(seq(0, 1,.01),
                   knots=seq(.05,.95,length=5), inclx=T)
xm ← x
x[xm > .0106] ← NA
matplot(x[,1], xm, type="l", ylim=c(0,.01),
        xlab=expression(X), ylab='', lty=1)
matplot(x[,1], x, type="l",
        xlab=expression(X), ylab='', lty=1)

x ← seq(0, 1, length=300)
for(nk in 3:6) {
  set.seed(nk)
  knots ← seq(.05, .95, length=nk)
  xx ← rcspline.eval(x, knots=knots, inclx=T)
  for(i in 1 : (nk - 1))
    xx[,i] ← (xx[,i] - min(xx[,i])) / (max(xx[,i]) - min(xx[,i]))
  for(i in 1 : 20) {
    beta ← 2*runif(nk-1) - 1
    xbeta ← xx %*% beta + 2 * runif(1) - 1
    xbeta ← (xbeta - min(xbeta)) / (max(xbeta) - min(xbeta))
  if(i == 1) {
    plot(x, xbeta, type="l", lty=1,
         xlab=expression(X), ylab='', bty="l")
    title(sub=paste(nk," knots"), adj=0, cex=.75)
  }
}
```
Figure 2.3: Restricted cubic spline component variables for $k = 5$ and knots at $X = .05,.275,.5,.725,$ and .95. Nonlinear basis functions are scaled by $\tau$. The left panel is a $y$–magnification of the right panel. Fitted functions such as those in Figure 2.4 will be linear combinations of these basis functions as long as knots are at the same locations used here.

Interactive demonstration of linear and cubic spline fitting, plus ordinary $4^{th}$ order polynomial. This can be run with RStudio or in an ordinary R session.

```r
for(j in 1 : nk)
  arrows(knots[j], .04, knots[j], -.03,
        angle=20, length=.07, lwd=1.5)
else lines(x, xbeta, col=i)

require(Hmisc)
getRs('demoSpline.r')                      # if using RStudio
getRs('demoSpline.r', put='source')       # if not
```

Paul Lambert’s excellent self-contained interactive demonstrations of continuity restrictions, cubic polynomial, linear spline, cubic spline, and restricted cubic spline fitting is at pclambert.net/interactivegraphs. Jordan Gauthier has another nice interactive demonstration at drjgauthier.shinyapps.io/spliny.
Figure 2.4: Some typical restricted cubic spline functions for $k = 3, 4, 5, 6$. The $y$-axis is $X\beta$. Arrows indicate knots. These curves were derived by randomly choosing values of $\beta$ subject to standard deviations of fitted functions being normalized.
Once $\beta_0, \ldots, \beta_{k-1}$ are estimated, the restricted cubic spline can be restated in the form

\[ f(X) = \beta_0 + \beta_1 X + \beta_2 (X - t_1)^3 + \beta_3 (X - t_2)^3 + \ldots + \beta_{k+1} (X - t_k)^3 \]

by dividing $\beta_2, \ldots, \beta_{k-1}$ by $\tau$ and computing

\[
\beta_k = \frac{[\beta_2(t_1 - t_k) + \beta_3(t_2 - t_k) + \ldots + \beta_{k-1}(t_{k-2} - t_k)]/(t_k - t_{k-1})}{(t_k - t_{k-1})} \\
\beta_{k+1} = \frac{[\beta_2(t_1 - t_{k-1}) + \beta_3(t_2 - t_{k-1}) + \ldots + \beta_{k-1}(t_{k-2} - t_{k-1})]}{(t_{k-1} - t_k)}. \\
\]

A test of linearity in $X$ can be obtained by testing

\[ H_0 : \beta_2 = \beta_3 = \ldots = \beta_{k-1} = 0. \]

Example: [186]
### Choosing Number and Position of Knots

- Knots are specified in advance in regression splines.
- Locations not important in most situations \([198, 64]\).
- Place knots where data exist — fixed quantiles of predictor’s marginal distribution.
- Fit depends more on choice of \(k\).
2.29

<table>
<thead>
<tr>
<th>k</th>
<th>Quantiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>.10 .5 .90</td>
</tr>
<tr>
<td>4</td>
<td>.05 .35 .65 .95</td>
</tr>
<tr>
<td>5</td>
<td>.05 .275 .5 .725 .95</td>
</tr>
<tr>
<td>6</td>
<td>.05 .23 .41 .59 .77 .95</td>
</tr>
<tr>
<td>7</td>
<td>.025 .1833 .3417 .5 .6583 .8167 .975</td>
</tr>
</tbody>
</table>

\( n < 100 \) – replace outer quantiles with 5th smallest and 5th largest \( X \) [199].

Choice of \( k \):

- Flexibility of fit vs. \( n \) and variance
- Usually \( k = 3, 4, 5 \). Often \( k = 4 \)
- Large \( n \) (e.g. \( n \geq 100 \)) – \( k = 5 \)
- Small \( n \) (< 30, say) – \( k = 3 \)
- Can use Akaike’s information criterion (AIC) [8, 214] to choose \( k \)
- This chooses \( k \) to maximize model likelihood ratio \( \chi^2 - 2k \).

See [83] for a comparison of restricted cubic splines, fractional polynomials, and penalized splines.
2.4.7 Nonparametric Regression

- Estimate tendency (mean or median) of $Y$ as a function of $X$
- Few assumptions
- Especially handy when there is a single $X$
- Plotted trend line may be the final result of the analysis
- Simplest smoother: moving average

<table>
<thead>
<tr>
<th>$X$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>5</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$</td>
<td>2.1</td>
<td>3.8</td>
<td>5.7</td>
<td>11.1</td>
<td>17.2</td>
</tr>
</tbody>
</table>

\[
\hat{E}(Y | X = 2) = \frac{2.1 + 3.8 + 5.7}{3} \\
\hat{E}(Y | X = \frac{2 + 3 + 5}{3}) = \frac{3.8 + 5.7 + 11.1}{3}
\]

- overlap OK
- problem in estimating $E(Y)$ at outer $X$-values
- estimates very sensitive to bin width

- Moving linear regression far superior to moving avg. (moving flat line)
• Cleveland’s [43] moving linear regression smoother loess (locally weighted least squares) is the most popular smoother. To estimate central tendency of $Y$ at $X = x$:
  – take all the data having $X$ values within a suitable interval about $x$ (default is $\frac{2}{3}$ of the data)
  – fit weighted least squares linear regression within this neighborhood
  – points near $x$ given the most weight\(^e\)
  – points near extremes of interval receive almost no weight
  – loess works much better at extremes of $X$ than moving avg.
  – provides an estimate at each observed $X$; other estimates obtained by linear interpolation
  – outlier rejection algorithm built-in

• loess works for binary $Y$ — just turn off outlier detection \(^A\)

• Other popular smoother: Friedman’s “super smoother”

• For loess or supsmu amount of smoothing can be controlled by analyst

\(^e\)Weight here means something different than regression coefficient. It means how much a point is emphasized in developing the regression coefficients.
Another alternative: smoothing splines\(^f\)

Smoother are very useful for estimating trends in residual plots

### 2.4.8

**Advantages of Regression Splines over Other Methods**

Regression splines have several advantages [94]:

- Parametric splines can be fitted using any existing regression program

- Regression coefficients estimated using standard techniques (ML or least squares), formal tests of no overall association, linearity, and additivity, confidence limits for the estimated regression function are derived by standard theory.

- The fitted function directly estimates transformation predictor should receive to yield linearity in \( C(Y|X) \).

- Even when a simple transformation is obvious, spline function can be used to represent the predictor in the final model (and the d.f. will be correct). Nonparametric methods do not yield a prediction equation.

- Extension to non-additive models.

---

\(^f\)These place knots at all the observed data points but penalize coefficient estimates towards smoothness.
burdensome computations.
2.5 Recursive Partitioning: Tree-Based Models

Breiman, Friedman, Olshen, and Stone [27]: CART (Classification and Regression Trees) — essentially model-free

Method:

- Find predictor so that best possible binary split has maximum value of some statistic for comparing 2 groups
- Within previously formed subsets, find best predictor and split maximizing criterion in the subset
- Proceed in like fashion until < k obs. remain to split
- Summarize Y for the terminal node (e.g., mean, modal category)
- Prune tree backward until it cross-validates as well as its “apparent” accuracy, or use shrinkage
Advantages/disadvantages of recursive partitioning:

- Does not require functional form for predictors
- Does not assume additivity — can identify complex interactions
- Can deal with missing data flexibly
- Interactions detected are frequently spurious
- Does not use continuous predictors effectively
- Penalty for overfitting in 3 directions
- Often tree doesn’t cross-validate optimally unless pruned back very conservatively
- Very useful in messy situations or those in which overfitting is not as problematic (confounder adjustment using propensity scores [47]; missing value imputation)

See [11].

2.5.1

New Directions in Predictive Modeling

The approaches recommended in this course are
• fitting fully pre-specified models without deletion of “insignificant” predictors

• using data reduction methods (masked to $Y$) to reduce the dimensionality of the predictors and then fitting the number of parameters the data’s information content can support

• use shrinkage (penalized estimation) to fit a large model without worrying about the sample size.

The data reduction approach can yield very interpretable, stable models, but there are many decisions to be made when using a two-stage (reduction/model fitting) approach, Newer approaches are evolving, including the following. These new approach handle continuous predictors well, unlike recursive partitioning.

• lasso (shrinkage using L1 norm favoring zero regression coefficients) [206, 197]

• elastic net (combination of L1 and L2 norms that handles the $p > n$ case better than the lasso) [244]

• adaptive lasso [241, 225]

• more flexible lasso to differentially penalize for variable selection and for regression coefficient estimation [173]

• group lasso to force selection of all or none of a group of
related variables (e.g., dummy variables representing a poly- 
tomous predictor)

• group lasso-like procedures that also allow for variables within 
a group to be removed [226]

• sparse-group lasso using L1 and L2 norms to achieve spare- 
ess on groups and within groups of variables [189]

• adaptive group lasso (Wang & Leng)

• Breiman’s nonnegative garrote [237]

• “preconditioning”, i.e., model simplification after developing 
a “black box” predictive model [156, 155]

• sparse principal components analysis to achieve parsimony 
in data reduction [234, 243, 132, 131]

• bagging, boosting, and random forests [99]

One problem prevents most of these methods from being ready 
for everyday use: they require scaling predictors before fitting 
the model. When a predictor is represented by nonlinear basis 
functions, the scaling recommendations in the literature are not 
sensible. There are also computational issues and difficulties 
obtaining hypothesis tests and confidence intervals.

When data reduction is not required, generalized additive mod-
Choosing Between Machine Learning and Statistical Modeling

- Statistical models allow for complexity (nonlinearity, interaction)
- Easy to allow every predictor to have nonlinear effect
- Easy to handle unlimited numbers of candidate predictors if assume additivity (e.g., using ridge regression, lasso, elastic net)
- Interactions should be pre-specified
- Machine learning is gaining attention but is oversold in some settings
- Researchers are under the mistaken impression that machine learning can be used on small samples

Considerations in Choosing One Approach over Another

A statistical model may be the better choice if

- Uncertainty is inherent and the signal:noise ratio is not large—
even with identical twins, one twin may get colon cancer and the other not; model tendencies instead of doing classification

- One could never have perfect training data, e.g., cannot repeatedly test one subject and have outcomes assessed without error

- One wants to isolate effects of a small number of variables

- Uncertainty in an overall prediction or the effect of a predictor is sought

- Additivity is the dominant way that predictors affect the outcome, or interactions are relatively small in number and can be pre-specified

- The sample size isn’t huge

- One wants to isolate (with a predominantly additive effect) the effects of “special” variables such as treatment or a risk factor

- One wants the entire model to be interpretable

Machine learning may be the better choice if

- The signal:noise ratio is large and the outcome being predicted doesn’t have a strong component of randomness; e.g.,
in visual pattern recognition an object must be an “E” or not an “E”

- The learning algorithm can be trained on an unlimited number of exact replications (e.g., 1000 repetitions of each letter in the alphabet or of a certain word to be translated to German)

- Overall prediction is the goal, without being able to succinctly describe the impact of any one variable (e.g., treatment)

- One is not very interested in estimating uncertainty in forecasts or in effects of select predictors

- Non-additivity is expected to be strong and can’t be isolated to a few pre-specified variables (e.g., in visual pattern recognition the letter “L” must have both a dominating vertical component and a dominating horizontal component)

- The sample size is huge [212]

- One does not need to isolate the effect of a special variable such as treatment

- One does not care that the model is a “black box”
2.6

Multiple Degree of Freedom Tests of Association

\[ C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2, \]

\[ H_0 : \beta_2 = \beta_3 = 0 \] with 2 d.f. to assess association between \( X_2 \) and outcome.

In the 5-knot restricted cubic spline model

\[ C(Y|X) = \beta_0 + \beta_1 X + \beta_2 X' + \beta_3 X'' + \beta_4 X''' , \]

\[ H_0 : \beta_1 = \ldots = \beta_4 = 0 \]

- Test of association: 4 d.f.
- Insignificant → dangerous to interpret plot
- What to do if 4 d.f. test insignificant, 3 d.f. test for linearity insig., 1 d.f. test sig. after delete nonlinear terms?

Grambsch and O’Brien [85] elegantly described the hazards of pretesting

- Studied quadratic regression
- Showed 2 d.f. test of association is nearly optimal even when regression is linear if nonlinearity entertained
Considered ordinary regression model
\[ E(Y|X) = \beta_0 + \beta_1X + \beta_2X^2 \]

Two ways to test association between \(X\) and \(Y\):

1. Fit quadratic model and test for linearity (\(H_0 : \beta_2 = 0\))

\(F\)-test for linearity significant at \(\alpha = 0.05\) level → report as the final test of association the 2 d.f. \(F\) test of \(H_0 : \beta_1 = \beta_2 = 0\)

2. If the test of linearity insignificant, refit without the quadratic term and final test of association is 1 d.f. test, \(H_0 : \beta_1 = 0|\beta_2 = 0\)

Showed that type I error > \(\alpha\)

Fairly accurate \(P\)-value obtained by instead testing against \(F\) with 2 d.f. even at second stage

Cause: are retaining the most significant part of \(F\)

**BUT** if test against 2 d.f. can only lose power when compared with original \(F\) for testing both \(\beta\)s

\(SSR\) from quadratic model > \(SSR\) from linear model
2.7 Assessment of Model Fit

2.7.1 Regression Assumptions

The general linear regression model is

\[ C(Y|X) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k. \]

Verify linearity and additivity. Special case:

\[ C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \]

where \( X_1 \) is binary and \( X_2 \) is continuous.

![Figure 2.5: Regression assumptions for one binary and one continuous predictor](image)

Methods for checking fit:

1. Fit simple linear additive model and check examine residual plots for patterns
• For OLS: box plots of $e$ stratified by $X_1$, scatterplots of $e$ vs. $X_2$ and $\hat{Y}$, with trend curves (want flat central tendency, constant variability)

• For normality, qqnorm plots of overall and stratified residuals

**Advantage**: Simplicity  
**Disadvantages**:

• Can only compute standard residuals for uncensored continuous response

• Subjective judgment of non-randomness

• Hard to handle interaction

• Hard to see patterns with large $n$ (trend lines help)

• Seeing patterns does not lead to corrective action

2. Scatterplot of $Y$ vs. $X_2$ using different symbols according to values of $X_1$

**Advantages**: Simplicity, can see interaction  
**Disadvantages**:

• Scatterplots cannot be drawn for binary, categorical, or censored $Y$

• Patterns difficult to see if relationships are weak or $n$ large

3. Stratify the sample by $X_1$ and quantile groups (e.g. deciles)
of $X_2$; estimate $C(Y|X_1, X_2)$ for each stratum

**Advantages**: Simplicity, can see interactions, handles censored $Y$ (if you are careful)

**Disadvantages**:

- Requires large $n$
- Does not use continuous var. effectively (no interpolation)
- Subgroup estimates have low precision
- Dependent on binning method

4. Separately for levels of $X_1$ fit a nonparametric smoother relating $X_2$ to $Y$

**Advantages**: All regression aspects of the model can be summarized efficiently with minimal assumptions

**Disadvantages**:

- Does not apply to censored $Y$
- Hard to deal with multiple predictors

5. Fit flexible nonlinear parametric model

**Advantages**:

- One framework for examining the model assumptions, fitting the model, drawing formal inference
- d.f. defined and all aspects of statistical inference “work as advertised”
Disadvantages:

- Complexity

- Generally difficult to allow for interactions when assessing patterns of effects

Confidence limits, formal inference can be problematic for methods 1-4.

Restricted cubic spline works well for method 5.

\[
\hat{C}(Y|X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_2' + \hat{\beta}_4 X_2'' \\
= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{f}(X_2),
\]

where

\[
\hat{f}(X_2) = \hat{\beta}_2 X_2 + \hat{\beta}_3 X_2' + \hat{\beta}_4 X_2'',
\]

\( \hat{f}(X_2) \) spline-estimated transformation of \( X_2 \).

- Plot \( \hat{f}(X_2) \) vs. \( X_2 \)

- \( n \) large → can fit separate functions by \( X_1 \)

- Test of linearity: \( H_0 : \beta_3 = \beta_4 = 0 \)

- Few good reasons to do the test other than to demonstrate that linearity is not a good default assumption

- Nonlinear → use transformation suggested by spline fit or keep spline terms
• Tentative transformation $g(X_2) \rightarrow$ check adequacy by expanding $g(X_2)$ in spline function and testing linearity

• Can find transformations by plotting $g(X_2)$ vs. $\hat{f}(X_2)$ for variety of $g$

• Multiple continuous predictors $\rightarrow$ expand each using spline

• Example: assess linearity of $X_2, X_3$

\[
C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X'_2 + \beta_4 X''_2 + \beta_5 X_3 + \beta_6 X'_3 + \beta_7 X''_3,
\]
Overall test of linearity $H_0 : \beta_3 = \beta_4 = \beta_6 = \beta_7 = 0$, with 4 d.f.

---

### 2.7.2

**Modeling and Testing Complex Interactions**

**Note:** Interactions will be misleading if main effects are not properly modeled [242].

Suppose $X_1$ binary or linear, $X_2$ continuous:

\[
C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X'_2 + \beta_4 X''_2 + \beta_5 X_1 X_2 + \beta_6 X_1 X'_2 + \beta_7 X_1 X''_2
\]
Simultaneous test of linearity and additivity: $H_0 : \beta_3 = \ldots = \beta_7 = 0$. 

 simsmodeling
• 2 continuous variables: could transform separately and form simple product

• **But** transformations depend on whether interaction terms adjusted for, so it is usually not possible to estimate transformations and interaction effects other than simultaneously

• Compromise: Fit interactions of the form $X_1 f(X_2)$ and $X_2 g(X_1)$:

$$C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_1' + \beta_3 X_1'' + \beta_4 X_2 + \beta_5 X_2' + \beta_6 X_2'' + \beta_7 X_1 X_2 + \beta_8 X_1 X_2' + \beta_9 X_1 X_2'' + \beta_{10} X_2 X_1' + \beta_{11} X_2 X_1''$$

• Test of additivity is $H_0 : \beta_7 = \beta_8 = \ldots = \beta_{11} = 0$ with 5 d.f.

• Test of lack of fit for the simple product interaction with $X_2$ is $H_0 : \beta_8 = \beta_9 = 0$

• Test of lack of fit for the simple product interaction with $X_1$ is $H_0 : \beta_{10} = \beta_{11} = 0$

General spline surface:

• Cover $X_1 \times X_2$ plane with grid and fit patch-wise cubic
polynomial in two variables

- Restrict to be of form \( aX_1 + bX_2 + cX_1X_2 \) in corners

- Uses all \((k - 1)^2\) cross-products of restricted cubic spline terms

- See Gray [86, 87, Section 3.2] for penalized splines allowing control of effective degrees of freedom. See Berhane et al. [19] for a good discussion of tensor splines.

Figure 2.6: Logistic regression estimate of probability of a hemorrhagic stroke for patients in the GUSTO-I trial given \(t\)-PA, using a tensor spline of two restricted cubic splines and penalization (shrinkage). Dark (cold color) regions are low risk, and bright (hot) regions are higher risk.

Figure 2.6 is particularly interesting because the literature had suggested (based on
approximately 24 strokes) that pulse pressure was the main cause of hemorrhagic stroke whereas this flexible modeling approach (based on approximately 230 strokes) suggests that mean arterial blood pressure (roughly a 45° line) is what is most important over a broad range of blood pressures. At the far right one can see that pulse pressure (axis perpendicular to 45° line) may have an impact although a non-monotonic one.

Other issues:

• \( Y \) non-censored (especially continuous) \( \rightarrow \) multi-dimensional scatterplot smoother [37]

• Interactions of order \( \geq 2 \): more trouble

• 2-way interactions among \( p \) predictors: pooled tests

• \( p \) tests each with \( p - 1 \) d.f.

Some types of interactions to pre-specify in clinical studies:

• Treatment \( \times \) severity of disease being treated

• Age \( \times \) risk factors

• Age \( \times \) type of disease

• Measurement \( \times \) state of a subject during measurement

• Race \( \times \) disease

• Calendar time \( \times \) treatment
• Quality $\times$ quantity of a symptom

• Measurement $\times$ amount of deterioration of the measurement

The last example is worth expanding as an example in model formulation. Consider the following study.

• A sample of patients seen over several years have a blood sample taken at time of hospitalization

• Blood samples are frozen

• Long after the last patient was sampled, the blood samples are thawed all in the same week and a blood analysis is done

• It is known that the quality of the blood analysis deteriorates roughly logarithmically by the age of the sample; blood measurements made on old samples are assumed to be less predictive of outcome

• This is reflected in an interaction between a function of sample age and the blood measurement $B^g$

• Patients were followed for an event, and the outcome variable of interest is the time from hospitalization to that event

---

For continuous $Y$, one might need to model the residual variance of $Y$ as increasing with sample age, in addition to modeling the mean function.
• To not assume a perfect logarithmic relationship for sample age on the effect of the blood measurement, a restricted cubic spline model with 3 default knots will be fitted for log sample age

• Sample age is assumed to not modify the effects of non-blood predictors patient age and sex

• Model may be specified the following way using the R `rms` package to fit a Cox proportional hazards model

• Test for nonlinearity of `sampleAge` tests the adequacy of assuming a plain logarithmic trend in sample age

```r
f ← cph(Surv(etime, event) ~ rcs(log(sampleAge), 3) * rcs(B, 4) +
        rcs(age, 5) * sex, data=mydata)
```

The $B \times sampleAge$ interaction effects have 6 d.f. and tests whether the sample deterioration affects the effect of $B$. By not assuming that $B$ has the same effect for old samples as for young samples, the investigator will be able to estimate the effect of $B$ on outcome when the blood analysis is ideal by inserting $sampleAge = 1$ day when requesting predicted values as a function of $B$.

### 2.7.3 Fitting Ordinal Predictors

• Small no. categories (3-4) → polytomous factor, dummy variables
• Design matrix for easy test of adequacy of initial codes → \( k \) original codes + \( k - 2 \) dummies

• More categories → score using data-driven trend. Later tests use \( k - 1 \) d.f. instead of 1 d.f.

• E.g., compute logit(mortality) vs. category

---

**2.7.4 Distributional Assumptions**

• Some models (e.g., logistic): all assumptions in \( C(Y|X) = X\beta \) (implicitly assuming no omitted variables!)

• Linear regression: \( Y \sim X\beta + \epsilon, \epsilon \sim n(0, \sigma^2) \)

• Examine distribution of residuals

• Some models (Weibull, Cox [52]):
  \( C(Y|X) = C(Y = y|X) = d(y) + X\beta \)
  \( C' = \log \text{hazard} \)

• Check form of \( d(y) \)

• Show \( d(y) \) does not interact with \( X \)
Chapter 3

Missing Data

3.1 Types of Missing Data

• Missing completely at random (MCAR)

• Missing at random (MAR)\(^a\)

• Informative missing
  (non-ignorable non-response)

See [61, 92, 1, 231, 33] for an introduction to missing data and imputation concepts.

\(^a\)“Although missing at random (MAR) is a non-testable assumption, it has been pointed out in the literature that we can get very close to MAR if we include enough variables in the imputation models” [92].
3.2 Prelude to Modeling

- Quantify extent of missing data
- Characterize types of subjects with missing data
- Find sets of variables missing on same subjects
3.3

Missing Values for Different Types of Response Variables

- Serial data with subjects dropping out (not covered in this course\(^b\))
- \(Y\) = time to event, follow-up curtailed: covered under survival analysis\(^c\)
- Often discard observations with completely missing \(Y\) but sometimes wasteful\(^d\)
- Characterize missings in \(Y\) before dropping obs.

\(^{b}\) Twist et al. [208] found instability in using multiple imputation of longitudinal data, and advantages of using instead full likelihood models.
\(^{c}\) White and Royston [230] provide a method for multiply imputing missing covariate values using censored survival time data.
\(^{d}\) \(Y\) is so valuable that if one is only missing a \(Y\) value, imputation is not worthwhile, and imputation of \(Y\) is not advised if MCAR or MAR.
3.4

Problems With Simple Alternatives to Imputation

Deletion of records—

- Badly biases parameter estimates when the probability of a case being incomplete is related to $Y$ and not just $X$ [137].

- Deletion because of a subset of $X$ being missing always results in inefficient estimates

- Deletion of records with missing $Y$ can result in biases [53] but is the preferred approach under MCAR\(^e\)

- However von Hippel [221] found advantages to a “use all variables to impute all variables then drop observations with missing $Y$” approach (but see [202])

- Lee and Carlin [130] suggest that observations missing on both $Y$ and on a predictor of major interest are not helpful

- Only discard obs. when
  - MCAR can be justified
  - Rarely missing predictor of overriding importance that can’t be imputed from other data

\(^e\)Multiple imputation of $Y$ in that case does not improve the analysis and assumes the imputation model is correct.
– Fraction of obs. with missings small and \( n \) is large

- No advantage of deletion except savings of analyst time

- Making up missing data better than throwing away real data

- See [118]

Adding extra categories of categorical predictors—

- Including missing data but adding a category ‘missing’ causes serious biases [1, 111, 209]

- Problem acute when values missing because subject too sick

- Difficult to interpret

- Fails even under MCAR [111, 1, 61, 211, 118]

- May be OK if values are “missing” because of “not applicable”

Likewise, serious problems are caused by setting missing continuous predictors to a constant (e.g., zero) and adding an indicator variable to try to estimate the effect of missing values.

Two examples from Donder et al.[61] using binary logistic regression, \( N = 500 \).

---

1E.g. you have a measure of marital happiness, dichotomized as high or low, but your sample contains some unmarried people. OK to have a 3-category variable with values high, low, and unmarried—Paul Allison, IMPUTE list, 4Jul09.
Results of 1000 Simulations With $\beta_1 = 1.0$ with MAR and Two Types of Imputation

<table>
<thead>
<tr>
<th>Imputation Method</th>
<th>$\hat{\beta}_1$</th>
<th>S.E.</th>
<th>Coverage of 0.90 C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single</td>
<td>0.989</td>
<td>0.09</td>
<td>0.64</td>
</tr>
<tr>
<td>Multiple</td>
<td>0.989</td>
<td>0.14</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Now consider a simulation with $\beta_1 = 1, \beta_2 = 0$, $X_2$ correlated with $X_1 (r = 0.75)$ but redundant in predicting $Y$, use missingness indicator when $X_1$ is MCAR in 0.4 of 500 subjects. This is also compared with grand mean fill-in imputation.

Results of 1000 Simulations Adding a Third Predictor Indicating Missing for $X_1$

<table>
<thead>
<tr>
<th>Imputation Method</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\beta}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indicator</td>
<td>0.55</td>
<td>0.51</td>
</tr>
<tr>
<td>Overall mean</td>
<td>0.55</td>
<td></td>
</tr>
</tbody>
</table>

In the incomplete observations the constant $X_1$ is uncorrelated with $X_2$. 
The goal of imputation is to preserve the information and meaning of the non-missing data.

There is a full Bayesian modeling alternative to all the methods presented below. The Bayesian approach requires more effort but has several advantages [69].

Exactly how are missing values estimated?

- Could ignore all other information — random or grand mean fill-in
- Can use external info not used in response model (e.g., zip code for income)
- Need to utilize reason for non-response if possible
- Use statistical model with sometimes-missing $X$ as response variable
- Model to estimate the missing values should include all variables that are either
  1. related to the missing data mechanism;
  2. have distributions that differ between subjects that have the target variable missing and those that have it mea-
sured;
3. associated with the sometimes-missing variable when it is not missing; or
4. included in the final response model [13, 92]

- Ignoring imputation results in biased $\hat{V}(\hat{\beta})$

- `transcan` function in Hmisc library: “optimal” transformations of all variables to make residuals more stable and to allow non-monotonic transformations

- `aregImpute` function in Hmisc: good approximation to full Bayesian multiple imputation procedure using the bootstrap

- `transcan` and `aregImpute` use the following for fitting imputation models:
  1. initialize NAs to median (mode for categoricals)
  2. expand all categorical predictors using dummy variables
  3. expand all continuous predictors using restricted cubic splines
  4. optionally optimally transform the variable being predicted by expanding it with restricted cubic splines and using the first canonical variate (multivariate regression) as the optimum transformation (maximizing $R^2$)
  5. one-dimensional scoring of categorical variables being predicted using canonical variates on dummy variables representing the categories (Fisher’s optimum scoring algo-
rithm); when imputing categories, solve for which category yields a score that is closest to the predicted score

- aregImpute and transcan work with `fit.mult.impute` to make final analysis of response variable relatively easy

- Predictive mean matching [137]: replace missing value with observed value of subject having closest predicted value to the predicted value of the subject with the `NA`. Key considerations are how to
  1. model the target when it is not `NA`
  2. match donors on predicted values
  3. avoid overuse of “good” donors to disallow excessive ties in imputed data
  4. account for all uncertainties

- Predictive model for each target uses any outcomes, all predictors in the final model other than the target, plus auxiliary variables not in the outcome model

- No distributional assumptions; nicely handles target variables with strange distributions [219]

- Predicted values need only be monotonically related to real predictive values
  - PMM can result in some donor observations being used repeatedly
– Causes lumpy distribution of imputed values

– Address by sampling from multinomial distribution, probabilities = scaled distance of all predicted values to predicted value ($y^*$) of observation needing imputing

– Tukey’s tricube function is a good weighting function (used in loess):
  \[ w_i = (1 - \min\left(d_i/s, 1\right))^3, \]
  \[ d_i = |\hat{y}_i - y^*| \]
  \[ s = 0.2 \times \text{mean}|\hat{y}_i - y^*| \] is a good default scale factor
  scale so that \( \sum w_i = 1 \)

• Recursive partitioning with surrogate splits — handles case where a predictor of a variable needing imputation is missing itself. But there are problems [163] even with completely random missingness.

• [231] discusses an alternative method based on choosing a donor observation at random from the \( q \) closest matches (\( q = 3 \), for example)

---

### Interactions

• When interactions are in the outcome model, oddly enough it may be better to treat interaction terms as “just another variable” and do unconstrained imputation of them [116]
3.6

Single Conditional Mean Imputation

- Can fill-in using unconditional mean or median if number of missings low and $X$ is unrelated to other $X$s

- Otherwise, first approximation to good imputation uses other $X$s to predict a missing $X$

- This is a single “best guess” conditional mean

- $\hat{X}_j = Z\hat{\theta}, Z = X_{-j}$ plus possibly auxiliary variables that precede $X_j$ in the causal chain that are not intended to be in the outcome model.
  Cannot include $Y$ in $Z$ without adding random errors to imputed values as done with multiple imputation (would steal info from $Y$)

- Recursive partitioning can sometimes be helpful for nonparametrically estimating conditional means
3.7 Predictive Mean Matching

3.8 Multiple Imputation

- Single imputation could use a random draw from the conditional distribution for an individual
  \[ \hat{X}_j = Z\hat{\theta} + \hat{\epsilon}, \quad Z = [\bar{X}_j, Y] \]  
  plus auxiliary variables
  \[ \hat{\epsilon} = n(0, \hat{\sigma}) \]
- bootstrap
- approximate Bayesian bootstrap [181, 92]: sample with replacement from sample with replacement of residuals

- Multiple imputations (M) with random draws
  - Draw sample of M residuals for each missing value to be imputed
  - Average M \( \hat{\beta} \)
  - In general can provide least biased estimates of \( \beta \)
  - Simple formula for imputation-corrected var(\( \hat{\beta} \))
    Function of average “apparent” variances and between-imputation variances of \( \hat{\beta} \)
CHAPTER 3. MISSING DATA

– Even when the $\chi^2$ distribution is a good approximation when data have no missing values, the $t$ or $F$ distributions are needed to have accurate $P$-values and confidence limits when there are missings [136, 176]

– **BUT** full multiple imputation needs to account for uncertainty in the imputation models by refitting these models for each of the $M$ draws

– transcan does not do that; aregImpute does

• Note that multiple imputation can and should use the response variable for imputing predictors [148]

• aregImpute algorithm [148]
  – Takes all aspects of uncertainty into account using the bootstrap
  
  – Different bootstrap resamples used for each imputation by fitting a flexible additive model on a sample with replacement from the original data
  
  – This model is used to predict all of the original missing and non-missing values for the target variable for the current imputation
  
  – Uses flexible parametric additive regression models to impute
There is an option to allow target variables to be optimally transformed, even non-monotonically (but this can overfit).

By default uses predictive mean matching for imputation; no residuals required (can also do more parametric regression imputation).

By default uses weighted PMM; many other matching options.

Uses by default van Buuren’s “Type 1” matching [33, Section 3.4.2] to capture the right amount of uncertainty by computing predicted values for missing values using a regression fit on the bootstrap sample, and finding donor observations by matching those predictions to predictions from potential donors using the regression fit from the original sample of complete observations.

When a predictor of the target variable is missing, it is first imputed from its last imputation when it was a target variable.

First 3 iterations of process are ignored (“burn-in”).

Compares favorably to R MICE approach.

Example:

```r
a ← aregImpute(~ age + sex + bp + death + heart.attack.before.death, data=mydata, n.impute=5)
```
f ← fit.mult.impute(death ~ rcs(age,3) + sex +
rcs(bp,5), lrm, a, data=mydata)

See Barzi and Woodward [13] for a nice review of multiple imputation with
detailed comparison of results (point estimates and confidence limits for
the effect of the sometimes-missing predictor) for various imputation meth-
ods. Barnes et al. [12] have a good overview of imputation methods and a
comparison of bias and confidence interval coverage for the methods when
applied to longitudinal data with a small number of subjects. Horton and
Kleinman [107] have a good review of several software packages for dealing
with missing data, and a comparison of them with aregImpute. Harel and
Zhou [92] provide a nice overview of multiple imputation and discuss some of
the available software. White and Carlin [229] studied bias of multiple impu-
tation vs. complete-case analysis. White et al. [231] provide much practical
guidance.

Caution: Methods can generate imputations having very reasonable distri-
butions but still not having the property that final response model regression
coefficients have nominal confidence interval coverage. It is worth checking
that imputations generate the correct collinearities among covariates.

- With MICE and aregImpute we are using the chained equation
  approach [231]

- Chained equations handles a wide variety of target variables
to be imputed and allows for multiple variables to be missing
on the same subject

- Iterative process cycles through all target variables to impute
  all missing values [213]
• Does not attempt to use the full Bayesian multivariate model for all target variables, making it more flexible and easy to use

• Possible to create improper imputations, e.g., imputing conflicting values for different target variables

• However, simulation studies [213] demonstrate very good performance of imputation based on chained equations
Diagnostics

- MCAR can be partially assessed by comparing distribution of non-missing $Y$ for those subjects with complete $X$ vs. those subjects having incomplete $X$ [137]

- Yucel and Zaslavsky [240] (see also [101])

- Interested in reasonableness of imputed values for a sometimes-missing predictor $X_j$

- Duplicate entire dataset

- In the duplicated observations set all non-missing values of $X_j$ to missing; let $w$ denote this set of observations set to missing

- Develop imputed values for the missing values of $X_j$

- In the observations in $w$ compare the distribution of imputed $X_j$ to the original values of $X_j$

- Bondarenko and Raghunathan [22] present a variety of useful diagnostics on the reasonableness of imputed values.
originally missing
all missing
impute

duplicate
3.10 Summary and Rough Guidelines

Table 3.1: Summary of Methods for Dealing with Missing Values

<table>
<thead>
<tr>
<th>Method</th>
<th>Deletion</th>
<th>Single</th>
<th>Multiple</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allows non-random missing</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Reduces sample size</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Apparent S.E. of $\hat{\beta}$ too low</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Increases real S.E. of $\hat{\beta}$</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>$\hat{\beta}$ biased</td>
<td>if not MCAR</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

The following contains crude guidelines. Simulation studies are needed to refine the recommendations. Here $f$ refers to the proportion of observations having any variables missing.

$f < 0.03$: It doesn’t matter very much how you impute missings or whether you adjust variance of regression coefficient estimates for having imputed data in this case. For continuous variables imputing missings with the median non-missing value is adequate; for categorical predictors the most frequent category can be used. Complete case analysis is also an option here. Multiple imputation may be needed to check that the simple approach “worked.”

$f \geq 0.03$: Use multiple imputation with number of imputations $g$ equal to $\max(5, 100f)$. Fewer imputations may be possible with very large sample sizes. See statisticalhorizons.com/how-many-imputations. Type 1 predictive mean matching is usually preferred, with weighted selection of donors. Account

---

*White et al. [231] recommend choosing $M$ so that the key inferential statistics are very reproducible should the imputation analysis be repeated. They suggest the use of 100$f$ imputations. See also [33, Section 2.7]. von Hippel [222] finds that the number of imputations should be quadratically increasing with the fraction of missing information.*
for imputation in estimating the covariance matrix for final parameter estimates. Use the $t$ distribution instead of the Gaussian distribution for tests and confidence intervals, if possible, using the estimated d.f. for the parameter estimates.

**Multiple predictors frequently missing:** More imputations may be required. Perform a “sensitivity to order” analysis by creating multiple imputations using different orderings of sometimes missing variables. It may be beneficial to initially sort variables so that the one with the most NAs will be imputed first.

Reason for missings more important than number of missing values.

Extreme amount of missing data does not prevent one from using multiple imputation, because alternatives are worse [110, 141].

---

### Effective Sample Size

It is useful to look at examples of effective sample sizes in the presence of missing data. If a sample of 1000 subjects contains various amounts and patterns of missings what size $n_c$ of a complete sample would have equivalent information for the intended purpose of the analysis?
1. A new marker was collected on a random sample of 200 of the subjects and one wants to estimate the added predictive value due to the marker: \( n_c = 200 \)

2. Height is missing on 100 subjects but we want to study association between BMI and outcome. Weight, sex, and waist circumference are available on all subjects: \( n_c = 980 \)

3. Each of 10 predictors is randomly missing on \( \frac{1}{10} \) of subjects, and the predictors are uncorrelated with each other and are each weakly related to the outcome: \( n_c = 500 \)

4. Same as previous but the predictors can somewhat be predicted from non-missing predictors: \( n_c = 750 \)

5. The outcome variable was not assessed on a random \( \frac{1}{5} \) of subjects: \( n_c = 800 \)

6. The outcome represents sensitive information, is missing on \( \frac{1}{2} \) of subjects, and we don’t know what made subjects respond to the question: \( n_c = 0 \) (serious selection bias)

7. One of the baseline variables was collected prospectively \( \frac{1}{2} \) of the time and for the other subjects it was retrospectively estimated only for subjects ultimately suffering a stroke and we don’t know which subjects had a stroke: \( n_c = 0 \) (study not worth doing)

8. The outcome variable was assessed by emailing the 1000 subjects, for which 800 responded, and we don’t know what made subjects respond: \( n_c = 0 \) (model will possibly be very biased—at least the intercept)
Bayesian Methods for Missing Data

- Multiple imputation developed as an approximation to a full Bayesian model

- Full Bayesian model treats missings as unknown parameters and provides exact inference and correct measures of uncertainty

- See this case study for an example

- The case study also shows how to do “posterior stacking” if you want to avoid having to specify a full model for missings, and instead use usual multiple imputations as described in this chapter
  
  - Run a multiple imputation algorithm

  - For each completed dataset run the Bayesian analysis and draw thousands of samples from the posterior distribution of the parameters

  - Pool all these posterior draws over all the multiple imputations and do posterior inference as usual with no special correction required

  - Made easy by the Hmisc package aregImpute function and the rms stackMI function as demonstrated in the Titanic
case study later in the notes.
Chapter 4

Multivariable Modeling Strategies

• “Spending d.f.”: examining or fitting parameters in models, or examining tables or graphs that utilize $Y$ to tell you how to model variables

• If wish to preserve statistical properties, can’t retrieve d.f. once they are “spent” (see Grambsch & O’Brien)

• If a scatterplot suggests linearity and you fit a linear model, how many d.f. did you actually spend (i.e., the d.f. that when put into a formula results in accurate confidence limits or $P$-values)?

• Decide number of d.f. that can be spent

• Decide where to spend them

• Spend them

• General references: [154, 196, 97, 80]
There are many choices to be made when deciding upon a global modeling strategy, including choice between:

- parametric and nonparametric procedures
- parsimony and complexity
- parsimony and good discrimination ability
- interpretable models and black boxes.
4.1 Prespecification of Predictor Complexity Without Later Simplification

- Rarely expect linearity
- Can’t always use graphs or other devices to choose transformation
- If select from among many transformations, results biased
- Need to allow flexible nonlinearity to potentially strong predictors not known to predict linearly
- Once decide a predictor is “in” can choose no. of parameters to devote to it using a general association index with $Y$
- Need a measure of “potential predictive punch”
- Measure needs to mask analyst to true form of regression to preserve statistical properties

Motivating examples:

```r
# Overfitting a flat relationship
require(rms)
set.seed(1)
x <- runif(1000)
y <- runif(1000, -0.5, 0.5)
dd <- datadist(x, y); options(datadist='dd')
par(mfrow=c(2,2), mar=c(2, 2, 3, 0.5))
pp <- function(actual) {
  yhat <- predict(f, data.frame(x=x))
}
```r
yreal ← actual(xs)
plot(0, 0, xlim=c(0,1),
     ylim=range(c(quantile(y, c(0.1, 0.9)), yhat, yreal)),
     type='n', axes=FALSE)
axis(1, labels=FALSE); axis(2, labels=FALSE)
lines(xs, yreal)
lines(xs, yhat, col='blue')
}
f ← ols(y ~ rcs(x, 5))
xs ← seq(0, 1, length=150)
pp(function(x) 0\times x)
title('Mild Error: \nOverfitting a Flat Relationship',
      cex=0.5)
y ← x + runif(1000, -0.5, 0.5)
f ← ols(y ~ rcs(x, 5))
pp(function(x) x)
title('Mild Error: \nOverfitting a Linear Relationship',
      cex=0.5)
y ← x^4 + runif(1000, -1, 1)
f ← ols(y ~ x)
pp(function(x) x^4)
title('Serious Error: \nUnderfitting a Steep Relationship',
      cex=0.5)
y ← -(x - 0.5)^2 + runif(1000, -0.2, 0.2)
f ← ols(y ~ x)
pp(function(x) -(x - 0.5)^2)
title('Tragic Error: \nMonotonic Fit to Non-Monotonic Relationship',
      cex=0.5)
```
Mild Error:
Overfitting a Flat Relationship

Mild Error:
Overfitting a Linear Relationship

Serious Error:
Underfitting a Steep Relationship

Tragic Error:
Monotonic Fit to Non-Monotonic Relationship

Table 4.1: Examples of Reducing the Number of Parameters

<table>
<thead>
<tr>
<th>Categorical predictor with $k$ levels</th>
<th>Collapse less frequent categories into “other”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous predictor represented as $k$-knot r.c. spline</td>
<td>Reduce $k$ to a number as low as 3, or 0 (linear)</td>
</tr>
</tbody>
</table>
Learning From a Saturated Model

When the effective sample size available is sufficiently large so that a saturated main effects model may be fitted, a good approach to gauging predictive potential is the following.

- Let all continuous predictors be represented as restricted cubic splines with $k$ knots, where $k$ is the maximum number of knots the analyst entertains for the current problem.

- Let all categorical predictors retain their original categories except for pooling of very low prevalence categories (e.g., ones containing $< 6$ observations).

- Fit this general main effects model.

- Compute the partial $\chi^2$ statistic for testing the association of each predictor with the response, adjusted for all other predictors. In the case of ordinary regression convert partial $F$ statistics to $\chi^2$ statistics or partial $R^2$ values.

- Make corrections for chance associations to “level the playing field” for predictors having greatly varying d.f., e.g., subtract the d.f. from the partial $\chi^2$ (the expected value of $\chi_p^2$ is $p$ under $H_0$).

- Make certain that tests of nonlinearity are not revealed as
this would bias the analyst.

- Sort the partial association statistics in descending order.

Commands in the \texttt{rms} package can be used to plot only what is needed. Here is an example for a logistic model.

\begin{verbatim}
f <- lrm(y ~ sex + race + rcs(age,5) + rcs(weight,5) + rcs(height,5) + rcs(blood.pressure,5))
plot(anova(f))
\end{verbatim}

### 4.1.2 Using Marginal Generalized Rank Correlations

When collinearities or confounding are not problematic, a quicker approach based on pairwise measures of association can be useful. This approach will not have numerical problems (e.g., singular covariance matrix) and is based on:

- 2 d.f. generalization of Spearman $\rho—R^2$ based on $\text{rank}(X)$ and $\text{rank}(X)^2$ vs. $\text{rank}(Y)$

- $\rho^2$ can detect U-shaped relationships

- For categorical $X$, $\rho^2$ is $R^2$ from dummy variables regressed against $\text{rank}(Y)$; this is tightly related to the Wilcoxon–Mann–Whitney–Kruskal–Wallis rank test for group differences\(^a\)

- Sort variables by descending order of $\rho^2$

\(^a\)This test statistic does not inform the analyst of \textit{which} groups are different from one another.
• Specify number of knots for continuous $X$, combine infrequent categories of categorical $X$ based on $\rho^2$

Allocating d.f. based on partial tests of association or sorting $\rho^2$ is a fair procedure because

• We already decided to keep variable in model no matter what $\rho^2$ or $\chi^2$ values are seen

• $\rho^2$ and $\chi^2$ do not reveal degree of nonlinearity; high value may be due solely to strong linear effect

• low $\rho^2$ or $\chi^2$ for a categorical variable might lead to collapsing the most disparate categories

Initial simulations show the procedure to be conservative. Note that one can move from simpler to more complex models but not the other way round
4.2 Checking Assumptions of Multiple Predictors Simultaneously

- Sometimes failure to adjust for other variables gives wrong transformation of an $X$, or wrong significance of interactions.
- Sometimes unwieldy to deal simultaneously with all predictors at each stage → assess regression assumptions separately for each predictor.
4.3 Variable Selection

- Series of potential predictors with no prior knowledge

- ↑ exploration → ↑ shrinkage (overfitting)

- Summary of problem: \( E(\hat{\beta}|\beta \text{ “significant”}) \neq \beta \) [39]

- Biased \( R^2, \hat{\beta}, \) standard errors, \( P \)-values too small

- \( F \) and \( \chi^2 \) statistics do not have the claimed distribution\(^b\) [85]

- Will result in residual confounding if use variable selection to find confounders [89]

- Derksen and Keselman [58] found that in stepwise analyses the final model represented noise 0.20-0.74 of time, final model usually contained \( \frac{1}{2} \) actual number of authentic predictors. Also:

  1. “The degree of correlation between the predictor variables affected the frequency with which authentic predictor variables found their way into the final model.

  2. The number of candidate predictor variables affected the number of noise variables that gained entry to the model.

\(^b\)Lockhart et al. [139] provide an example with \( n = 100 \) and 10 orthogonal predictors where all true \( \beta \)s are zero. The test statistic for the first variable to enter has type I error of 0.39 when the nominal \( \alpha \) is set to 0.05.
3. The size of the sample was of little practical importance in determining the number of authentic variables contained in the final model.

4. The population multiple coefficient of determination could be faithfully estimated by adopting a statistic that is adjusted by the total number of candidate predictor variables rather than the number of variables in the final model”.

- Global test with $p$ d.f. insignificant $\rightarrow$ stop

Simulation experiment, true $\sigma^2 = 6.25$, 8 candidate variables, 4 of them related to $Y$ in the population. Select best model using all possible subsets regression to maximize $R_{adj}^2$ (not usually recommended but gives variable selection more of a chance to work in this context).

Note: The audio was made using stepAIC with collinearities in predictors. The code below allows for several options. Here we use all possible subsets of predictors and force predictors to be uncorrelated, which is the easiest case for variable selection.

```r
require(MASS)
require(leaps)
sim ← function(n, sigma=2.5, method=c('stepaic', 'leaps'), pr=FALSE, prcor=FALSE, dataonly=FALSE) {
  method ← match.arg(method)
  if(uncorrelated) {
    x1 ← rnorm(n)
    x2 ← rnorm(n)
    x3 ← rnorm(n)
    x4 ← rnorm(n)
    x5 ← rnorm(n)
    x6 ← rnorm(n)
    x7 ← rnorm(n)
  }
```
CHAPTER 4. MULTIVARIABLE MODELING STRATEGIES

4-12

```r
x8 ← rnorm(n)
}
else {
  x1 ← rnorm(n)
x2 ← x1 + 2.0 * rnorm(n)  # was 0.5 * rnorm(n)
x3 ← rnorm(n)
x4 ← x3 + 1.5 * rnorm(n)
x5 ← x1 + rnorm(n)/1.3
x6 ← x2 + 2.25 * rnorm(n)  # was rnorm(n)/1.3
x7 ← x3 + x4 + 2.5 * rnorm(n)  # was + rnorm(n)
x8 ← x7 + 4.0 * rnorm(n)  # was + 0.5 * rnorm(n)
}
}
```

```r
z ← cbind(x1,x2,x3,x4,x5,x6,x7,x8)
if(prcor) return(round(cor(z), 2))
lp ← x1 + x2 + .5*x3 + .4*x7
y ← lp + sigma*rnorm(n)
if(dataonly) return(list(x=z, y=y))
if(method == 'leaps') {
  s ← summary(regsubsets(z, y))
  best ← which.max(s$adjr2)
  xvars ← s$which[best, -1]  # remove intercept
  ssr ← s$rss[best]
p ← sum(xvars)
x ← if(p == 0) 'none' else paste((1:8)[xvars], collapse='')
  if(pr) print(x)
  ssesw ← (n - 1) * var(y) - ssr
  s2s ← ssesw / (n - p - 1)
yhat ← if(p == 0) mean(y) else fitted(lm(y ~ z[, xvars]))
}
```

```r
f ← lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8)
if(method == 'stepaic') {
  g ← stepAIC(f, trace=0)
p ← g$rank - 1
  xs ← if(p == 0) 'none' else gsub('[ \+ x]', '', as.character(formula(g))[3])
  if(pr) print(xs)
  ssesw ← sum(resid(g)^2)
s2s ← ssesw/g$df.residual
  yhat ← fitted(g)
}
```

```r
# Set SSEsw / (n - gdf - 1) = true sigma^2
gdf ← n - 1 - ssesw / (sigma^2)
# Compute root mean squared error against true linear predictor
rmse.full ← sqrt(mean((fitted(f) - lp)^2))
rmse.step ← sqrt(mean((yhat - lp)^2))
list(stats=c(n=n, vratio=s2s/(sigma^2)),
  gdf=gdf, apparentdf=p, rmse.full=rmse.full, rmse.step=rmse.step),
  xselected=xs)
}
```

```r
rsim ← function(B, n, method=c('stepaic', 'leaps')) {
  method ← match.arg(method)
x ← character(B)
r ← matrix(NA, nrow=B, ncol=6)
for(i in 1:B) {
  w ← sim(n, method=method)
r[i,] ← w$stats
}
xs[i] ← w$xselected
}
colnames(r) ← names(w$stats)
s ← apply(r, 2, median)
p ← r[, 'apparentdf']
s['apparentdf'] ← mean(p)
print(round(s, 2))
print(table(p))
cat('Prob [correct model]= ', round(sum(xs == '1237')/B, 2), '
')

Show the correlation matrix being assumed for the Xs:

uncorrelated ← TRUE
sim(50000, prcor=TRUE)

Simulate to find the distribution of the number of variables selected, the proportion of simulations in which the true model \( M(X_1, X_2, X_3, X_7) \) was found, the median value of \( \hat{\sigma}^2/\sigma^2 \), the median effective d.f., and the mean number of apparent d.f., for varying sample sizes.

set.seed(11)
m ← 'leaps'  # all possible regressions stopping on R2adj
rsim(100, 20, method=m)  # actual model found twice out of 100

<table>
<thead>
<tr>
<th>n</th>
<th>vratio</th>
<th>gdf</th>
<th>apparentdf</th>
<th>rmse.full</th>
<th>rmse.step</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.00</td>
<td>0.94</td>
<td>5.32</td>
<td>4.10</td>
<td>1.62</td>
<td>1.58</td>
</tr>
</tbody>
</table>

P
1 2 3 4 5 6 7 8
3 14 18 22 27 11 4 1
Prob[correct model]= 0.02

rsim(100, 40, method=m)

<table>
<thead>
<tr>
<th>n</th>
<th>vratio</th>
<th>gdf</th>
<th>apparentdf</th>
<th>rmse.full</th>
<th>rmse.step</th>
</tr>
</thead>
<tbody>
<tr>
<td>40.00</td>
<td>0.61</td>
<td>17.89</td>
<td>4.38</td>
<td>1.21</td>
<td>1.24</td>
</tr>
</tbody>
</table>

P
2 3 4 5 6 7
As $n \uparrow$ the mean number of variables selected increased. The proportion of simulations in which the correct model was found increased from 0 to 0.53. $\sigma^2$ is underestimated when $n = 300$ by a factor of 0.42, resulting in the d.f. needed to de-bias $\hat{\sigma}^2$ being greater than $n$ when the apparent d.f. was only 5.16 on the average. Variable selection slightly increased closeness to the true $X\beta$.

If the simulations are re-run allowing for collinearities (unrelated=FALSE) one can expect variable selection to be even more problematic.

Variable selection methods [93]:

```
rsim(100, 150, method=m)
```

```
<table>
<thead>
<tr>
<th>n</th>
<th>vratio</th>
<th>gdf</th>
<th>apparentdf</th>
<th>rmse.full</th>
<th>rmse.step</th>
</tr>
</thead>
<tbody>
<tr>
<td>150.00</td>
<td>0.44</td>
<td>85.99</td>
<td>5.01</td>
<td>0.59</td>
<td>0.57</td>
</tr>
</tbody>
</table>
```

```
p 2 3 4 5 6 7 8
1 5 27 35 24 7 1
Prob[correct model]= 0.2
```

```
rsim(100, 300, method=m)
```

```
<table>
<thead>
<tr>
<th>n</th>
<th>vratio</th>
<th>gdf</th>
<th>apparentdf</th>
<th>rmse.full</th>
<th>rmse.step</th>
</tr>
</thead>
<tbody>
<tr>
<td>300.00</td>
<td>0.42</td>
<td>177.01</td>
<td>5.16</td>
<td>0.43</td>
<td>0.40</td>
</tr>
</tbody>
</table>
```

```
p 4 5 6 7 8
27 42 20 10 1
Prob[correct model]= 0.26
```

```
r simul(100, 2000)
```

```
<table>
<thead>
<tr>
<th>n</th>
<th>vratio</th>
<th>gdf</th>
<th>apparentdf</th>
<th>rmse.full</th>
<th>rmse.step</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000.00</td>
<td>1.00</td>
<td>6.43</td>
<td>4.58</td>
<td>0.17</td>
<td>0.15</td>
</tr>
</tbody>
</table>
```

```
p 4 5 6 7
53 37 9 1
Prob[correct model]= 0.53
```
• Forward selection, backward elimination

• Stopping rule: “residual $\chi^2$” with d.f. = no. candidates remaining at current step

• Test for significance or use Akaike’s information criterion (AIC [8]), here $\chi^2 - 2 \times d.f.$

• Better to use subject matter knowledge!

• No currently available stopping rule was developed for step-wise, only for comparing a limited number of pre-specified models [26, Section 1.3]

• Roecker [179] studied forward selection (FS), all possible subsets selection (APS), full fits

• APS more likely to select smaller, less accurate models than FS

• Neither as accurate as full model fit unless more than half of candidate variables redundant or unnecessary

• Step-down is usually better than forward [142] and can be used efficiently with maximum likelihood estimation [127]

• Fruitless to try different stepwise methods to look for agreement [233]
• Bootstrap can help decide between full and reduced model

• Full model fits gives meaningful confidence intervals with standard formulas, C.I. after stepwise does not \([3, 108, 26]\)

• Data reduction (grouping variables) can help

• Using the bootstrap to select important variables for inclusion in the final model \([183]\) is problematic \([9]\)

• It is not logical that a population regression coefficient would be exactly zero just because its estimate was “insignificant”

See also these articles:

• **Step away from stepwise** by Gary Smith

• **Five myths about variable selection** by Georg Heinze and Daniela Dunkler

• **Variable selection - A review and recommendations for the practicing statistician** by Georg Heinze, Christine Wallisch, Daniela Dunkler

• **Stopping stepwise** by Peter Flom
4.3.1 Maxwell’s Demon as an Analogy to Variable Selection

Some of the information in the data is spent on variable selection instead of using all information for estimation.

Model specification is preferred to model selection.

Information content of the data usually insufficient for reliable variable selection.

Maxwell imagines one container divided into two parts, A and B. Both parts are filled with the same gas at equal temperatures and placed next to each other. Observing the molecules on both sides, an imaginary demon guards a trapdoor between the two parts. When a faster-than-average molecule from A flies towards the trapdoor, the demon opens it, and the molecule will fly from A to B. Likewise, when a slower-than-average molecule from B flies towards the trapdoor, the demon will let it pass from B to A. The average speed of the molecules in B will have increased while in A they will have slowed down on average. Since average molecular speed corresponds to temperature, the temperature decreases in A and increases in B, contrary to the second law of thermodynamics.

Szilárd pointed out that a real-life Maxwell’s demon would need to have some means of measuring molecular speed, and that the act of acquiring information would require an
expenditure of energy. Since the demon and the gas are interacting, we must consider the total entropy of the gas and the demon combined. The expenditure of energy by the demon will cause an increase in the entropy of the demon, which will be larger than the lowering of the entropy of the gas.

Source: commons.wikimedia.org/wiki/File:YoungJamesClerkMaxwell.jpg
en.wikipedia.org/wiki/Maxwell’s_demon

Peter Ellis’ blog article contains excellent examples of issues discussed here but applied to time series modeling.
Overfitting and Limits on Number of Predictors

- Concerned with avoiding overfitting

- Assume typical problem in medicine, epidemiology, and the social sciences in which the signal:noise ratio is small (higher ratios allow for more aggressive modeling)

- \( p \) should be \( < \frac{m}{15} \) \([95, 96, 191, 158, 157, 220, 212]\)

- \( p = \) number of parameters in full model or number of candidate parameters in a stepwise analysis

- Derived from simulations to find minimum sample size so that apparent discrimination = validated discrimination

- Applies to typical signal:noise ratios found outside of tightly controlled experiments

- If true \( R^2 \) is high, many parameters can be estimated from smaller samples

- Ignores sample size needed just to estimate the intercept or, in semiparametric models, the underlying distribution function\(^c\)

\(^c\)The sample size needed for these is model-dependent
• Riley et al. [177, 178] have refined sample size estimation for continuous, binary, and time-to-event models to account for all of these issues.

• To just estimate $\sigma$ in a linear model with a multiplicative margin of error of 1.2 with 0.95 confidence requires $n = 70$

Table 4.2: Limiting Sample Sizes for Various Response Variables

<table>
<thead>
<tr>
<th>Type of Response Variable</th>
<th>Limiting Sample Size $m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>$n$ (total sample size)</td>
</tr>
<tr>
<td>Binary</td>
<td>$\min(n_1, n_2)$</td>
</tr>
<tr>
<td>Ordinal ($k$ categories)</td>
<td>$n - \frac{1}{n^2} \sum_{i=1}^{k} n_i^3$</td>
</tr>
<tr>
<td>Failure (survival) time</td>
<td>number of failures</td>
</tr>
</tbody>
</table>

$^a$If one considers the power of a two-sample binomial test compared with a Wilcoxon test if the response could be made continuous and the proportional odds assumption holds, the effective sample size for a binary response is $3n_1n_2/n \approx 3 \min(n_1, n_2)$ if $\frac{n_1}{n_2}$ is near 0 or 1 [232, Eq. 10, 15]. Here $n_1$ and $n_2$ are the marginal frequencies of the two response levels [157].

$^b$Based on the power of a proportional odds model two-sample test when the marginal cell sizes for the response are $n_1, \ldots, n_k$, compared with all cell sizes equal to unity (response is continuous) [232, Eq. 3]. If all cell sizes are equal, the relative efficiency of having $k$ response categories compared to a continuous response is $1 - \frac{1}{k}$ [232, Eq. 14], e.g., a 5-level response is almost as efficient as a continuous one if proportional odds holds across category cutoffs.

$^c$This is approximate, as the effective sample size may sometimes be boosted somewhat by censored observations, especially for non-proportional hazards methods such as Wilcoxon-type tests [17].

• Narrowly distributed predictor $\rightarrow$ even higher $n$

• $p$ includes all variables screened for association with response, including interactions

• Univariable screening (graphs, crosstabs, etc.) in no way reduces multiple comparison problems of model building [203]
4.5 Shrinkage

- Slope of calibration plot; regression to the mean
- Statistical estimation procedure — “pre-shrunken” models
- Aren’t regression coefficients OK because they’re unbiased?
- Problem is in how we use coefficient estimates
- Consider 20 samples of size $n = 50$ from $U(0, 1)$
- Compute group means and plot in ascending order
- Equivalent to fitting an intercept and 19 dummies using least squares
- Result generalizes to general problems in plotting $Y$ vs. $X\hat{\beta}$

```r
set.seed(123)
n ← 50
y ← runif(20*n)
group ← rep(1:20,each=n)
ybar ← tapply(y, group , mean)
ybar ← sort(ybar)
plot(1:20, ybar, type='n', axes=FALSE, ylim=c(.3,.7),
     xlab='Group', ylab='Group Mean')
lines(1:20, ybar)
points(1:20, ybar, pch=20, cex=.5)
axis(2)
axis(1, at=1:20, labels=FALSE)
for(j in 1:20) axis(1, at=j, labels=names(ybar)[j])
abline(h=.5, col=gray(.85))
```

- Prevent shrinkage by using pre-shrinkage
Figure 4.1: Sorted means from 20 samples of size 50 from a uniform [0, 1] distribution. The reference line at 0.5 depicts the true population value of all of the means.

- Spiegelhalter [194]: var. selection arbitrary, better prediction usually results from fitting all candidate variables and using shrinkage

- Shrinkage closer to that expected from full model fit than based on number of significant variables [50]

- Ridge regression [128, 214]

- Penalized MLE [217, 86, 98]

- Heuristic shrinkage parameter of van Houwelingen and le Cessie [214, Eq. 77]

\[
\hat{\gamma} = \frac{\text{model } \chi^2 - p}{\text{model } \chi^2},
\]
• OLS\textsuperscript{d}: $\hat{\gamma} = \frac{n-p-1}{n-1} R_{\text{adj}}^2 / R^2$

$R_{\text{adj}}^2 = 1 - (1 - R^2) \frac{n-1}{n-p-1}$

• $p$ close to no. candidate variables

• Copas [50, Eq. 8.5] adds 2 to numerator

---

\textsuperscript{4}An excellent discussion about such indexes may be found here.
4.6 Collinearity

- When at least 1 predictor can be predicted well from others

- Can be a blessing (data reduction, transformations)

  - $\uparrow$ s.e. of $\hat{\beta}$, $\downarrow$ power

- This is appropriate $\rightarrow$ asking too much of the data [40, Chap. 9]

- Variables compete in variable selection, chosen one arbitrary

- Does not affect joint influence of a set of highly correlated variables (use multiple d.f. tests)

- Does not at all affect predictions on model construction sample

- Does not affect predictions on new data [152, pp. 379-381] if
  1. Extreme extrapolation not attempted
  2. New data have same type of collinearities as original data

- Example: LDL and total cholesterol – problem only if more inconsistent in new data

- Example: age and $\text{age}^2$ – no problem
• One way to quantify for each predictor: variance inflation factors (VIF)

• General approach (maximum likelihood) — transform information matrix to correlation form, VIF = diagonal of inverse [56, 227]

• See Belsley [16, pp. 28-30] for problems with VIF

• Easy approach: SAS VARCLUS procedure [182], R varclus function, other clustering techniques: group highly correlated variables

• Can score each group (e.g., first principal component, \( PC_1 \) [55]); summary scores not collinear
4.7 Data Reduction

- Unless $n \gg p$, model unlikely to validate

- Data reduction: $\downarrow p$

- Use the literature to eliminate unimportant variables.

- Eliminate variables whose distributions are too narrow.

- Eliminate candidate predictors that are missing in a large number of subjects, especially if those same predictors are likely to be missing for future applications of the model.

- Use a statistical data reduction method such as incomplete principal components regression, nonlinear generalizations of principal components such as principal surfaces, sliced inverse regression, variable clustering, or ordinary cluster analysis on a measure of similarity between variables.

- Data reduction is completely masked to $Y$, which is precisely why it does not distort estimates, standard errors, $P$-values, or confidence limits

- Data reduction = unsupervised learning

- Example: dataset with 40 events and 60 candidate predictors
– Use variable clustering to group variables by correlation structure

– Use clinical knowledge to refine the clusters

– Keep age and severity of disease as separate predictors because of their strength

– For others create clusters: socioeconomic, risk factors/history, and physiologic function

– Summarize each cluster with its first principal component $PC_1$, i.e., the linear combination of characteristics that maximizes variance of the score across subjects subject to an overall constraint on the coefficients

– Fit outcome model with 5 predictors

**4.7.1 Redundancy Analysis**

- Remove variables that have poor distributions
  - E.g., categorical variables with fewer than 2 categories having at least 20 observations

- Use flexible additive parametric additive models to determine how well each variable can be predicted from the remaining variables
• Variables dropped in stepwise fashion, removing the most predictable variable at each step

• Remaining variables used to predict

• Process continues until no variable still in the list of predictors can be predicted with an $R^2$ or adjusted $R^2$ greater than a specified threshold or until dropping the variable with the highest $R^2$ (adjusted or ordinary) would cause a variable that was dropped earlier to no longer be predicted at the threshold from the now smaller list of predictors

• R function `redu`n in `Hmisc` package

• Related to principal variables [146] but faster

### 4.7.2 Variable Clustering

• Goal: Separate variables into groups
  – variables within group correlated with each other
  – variables not correlated with non-group members

• Score each dimension, stop trying to separate effects of factors measuring same phenomenon

• Variable clustering [182, 55] (oblique-rotation PC analysis)
→ separate variables so that first PC is representative of group

- Can also do hierarchical cluster analysis on similarity matrix based on squared Spearman or Pearson correlations, or more generally, Hoeffding’s $D$ [105].

- See [90] for a method related to variable clustering and sparse principal components.

- [41] implement many more variable clustering methods

Example: Figure 15.6

4.7.3 Transformation and Scaling Variables Without Using $Y$

- Reduce $p$ by estimating transformations using associations with other predictors

- Purely categorical predictors – correspondence analysis [129, 54, 42, 88, 147]

- Mixture of qualitative and continuous variables: qualitative principal components

- Maximum total variance (MTV) of Young, Takane, de Leeuw [239, 147]
1. Compute $PC_1$ of variables using correlation matrix
2. Use regression (with splines, dummies, etc.) to predict $PC_1$ from each $X$ — expand each $X_j$ and regress it separately on $PC_1$ to get working transformations
3. Recompute $PC_1$ on transformed $X$'s
4. Repeat 3-4 times until variation explained by $PC_1$ plateaus and transformations stabilize

- Maximum generalized variance (MGV) method of Sarle [123, pp. 1267-1268]
  1. Predict each variable from (current transformations of) all other variables
  2. For each variable, expand it into linear and nonlinear terms or dummies, compute first canonical variate
  3. For example, if there are only two variables $X_1$ and $X_2$ represented as quadratic polynomials, solve for $a, b, c, d$ such that $aX_1 + bX_2^2$ has maximum correlation with $cX_2 + dX_2^2$.
  4. Goal is to transform each var. so that it is most similar to predictions from other transformed variables
  5. Does not rely on PCs or variable clustering

- MTV (PC-based instead of canonical var.) and MGV implemented in SAS PROC PRINQUAL [123]
  1. Allows flexible transformations including monotonic splines
  2. Does not allow restricted cubic splines, so may be unstable unless monotonicity assumed
3. Allows simultaneous imputation but often yields wild estimates

### 4.7.4 Simultaneous Transformation and Imputation

**R transcan Function for Data Reduction & Imputation**

- Initialize missings to medians (or most frequent category)
- Initialize transformations to original variables
- Take each variable in turn as $Y$
- Exclude obs. missing on $Y$
- Expand $Y$ (spline or dummy variables)
- Score (transform $Y$) using first canonical variate
- Missing $Y \rightarrow$ predict canonical variate from $X$s
- The imputed values can optionally be shrunk to avoid over-fitting for small $n$ or large $p$
- Constrain imputed values to be in range of non-imputed ones
- Imputations on original scale
1. Continuous \( \rightarrow \) back-solve with linear interpolation
2. Categorical \( \rightarrow \) classification tree (most freq. cat.) or match to category whose canonical score is closest to one predicted

- Multiple imputation — bootstrap or approx. Bayesian boot.
  1. Sample residuals multiple times (default \( M = 5 \))
  2. Are on “optimally” transformed scale
  3. Back-transform
  4. `fit.mult.impute` works with `aregImpute` and `transcan` output to easily get imputation-corrected variances and avg. \( \hat{\beta} \)

- Option to insert constants as imputed values (ignored during transformation estimation); helpful when a lab value may be missing because the patient returned to normal

- Imputations and transformed values may be easily obtained for new data

- An R function `Function` will create a series of R functions that transform each predictor

- Example: \( n = 415 \) acutely ill patients
  1. Relate heart rate to mean arterial blood pressure
  2. Two blood pressures missing
  3. Heart rate not monotonically related to blood pressure
4. See Figures 4.2 and 4.3

```r
require(Hmisc)
getHdata(support)  # Get data frame from web site
heart.rate ← support$hrt
blood.pressure ← support$meanbp
blood.pressure[400:401]

Mean Arterial Blood Pressure Day 3
[1] 151 136

blood.pressure[400:401] ← NA  # Create two missings
d ← data.frame(heart.rate, blood.pressure)
par(pch=46)  # Figure 4.2
w ← transcan(~ heart.rate + blood.pressure, transformed=TRUE,
imputed=TRUE, show.na=TRUE, data=d)

Convergence criterion: 2.901 0.035
0.007
Convergence in 4 iterations
R² achieved in predicting each variable:

heart.rate  blood.pressure
0.259 0.259

Adjusted R²:

heart.rate  blood.pressure
0.254 0.253

w$imputed$blood.pressure

400 401
132.4057 109.7741

t ← w$transformed
spe ← round(c(spearman(heart.rate, blood.pressure),
spearman(t[,]'heart.rate'),
t[,]'blood.pressure'))), 2)

plot(heart.rate, blood.pressure)  # Figure 4.3
plot(t[,]'heart.rate'), t[,]'blood.pressure'),
xlab='Transformed hr', ylab='Transformed bp')

ACE (Alternating Conditional Expectation) of Breiman and Friedman [25]

1. Uses nonparametric “super smoother” [76]
2. Allows monotonicity constraints, categorical vars.
Figure 4.2: Transformations fitted using transcan. Tick marks indicate the two imputed values for blood pressure.

Figure 4.3: The lower left plot contains raw data (Spearman $\rho = -0.02$); the lower right is a scatterplot of the corresponding transformed values ($\rho = -0.13$). Data courtesy of the SUPPORT study [117].
3. Does not handle missing data

- These methods find *marginal* transformations

- Check adequacy of transformations using $Y$
  1. Graphical
  2. Nonparametric smoothers ($X$ vs. $Y$)
  3. Expand original variable using spline, test additional predictive information over original transformation

<table>
<thead>
<tr>
<th>Simple Scoring of Variable Clusters</th>
</tr>
</thead>
</table>

- Try to score groups of transformed variables with $PC_1$

- Reduces d.f. by pre-transforming var. and by combining multiple var.

- Later may want to break group apart, but delete all variables in groups whose summary scores do not add significant information

- Sometimes simplify cluster score by finding a subset of its constituent variables which predict it with high $R^2$.

Series of dichotomous variables:

- Construct $X_1 = 0-1$ according to whether any variables pos-
• Construct $X_2 = \text{number of positives}$

• Test whether original variables add to $X_1$ or $X_2$

---

### 4.7.6 Simplifying Cluster Scores

---

### 4.7.7 How Much Data Reduction Is Necessary?

Using Expected Shrinkage to Guide Data Reduction

• Fit full model with all candidates, $p$ d.f., LR likelihood ratio $\chi^2$

• Compute $\hat{\gamma}$

• If $< 0.9$, consider shrunken estimator from whole model, or data reduction (again not using $Y$)

• $q$ regression d.f. for reduced model

• Assume best case: discarded dimensions had no association with $Y$

• Expected loss in LR is $p - q$
• New shrinkage \[ LR - (p - q) - q \]/[LR - (p - q)]

• Solve for \( q \rightarrow q \leq (LR - p)/9 \)

• Under these assumptions, no hope unless original LR > \( p + 9 \)

• No \( \chi^2 \) lost by dimension reduction \( \rightarrow q \leq LR/10 \)

Example:

• Binary logistic model, 45 events on 150 subjects

• 10:1 rule \( \rightarrow \) analyze 4.5 d.f. total

• Analyst wishes to include age, sex, 10 others

• Not known if age linear or if age and sex additive

• 4 knots \( \rightarrow 3 + 1 + 1 \) d.f. for age and sex if restrict interaction to be linear

• Full model with 15 d.f. has LR=50

• Expected shrinkage factor \((50 - 15)/50 = 0.7\)

• LR > 15 + 9 = 24 \( \rightarrow \) reduction may help

• Reduction to \( q = (50 - 15)/9 \approx 4 \) d.f. necessary

• Have to assume age linear, reduce other 10 to 1 d.f.
- Separate hypothesis tests intended → use full model, adjust for multiple comparisons

<table>
<thead>
<tr>
<th>Goals</th>
<th>Reasons</th>
<th>Methods</th>
</tr>
</thead>
</table>
| Group predictors so that each group represents a single dimension that can be summarized with a single score | • ↓ d.f. arising from multiple predictors | Variable clustering  
• Subject matter knowledge  
• Group predictors to maximize proportion of variance explained by $PC_1$ of each group  
• Hierarchical clustering using a matrix of similarity measures between predictors |
| Transform predictors | • ↓ d.f. due to nonlinear and dummy variable components  
• Allows predictors to be optimally combined  
• Make $PC_1$ more reasonable summary  
• Use in customized model for imputing missing values on each predictor |  
• Maximum total variance on a group of related predictors  
• Canonical variates on the total set of predictors |
| Score a group of predictors | ↓ d.f. for group to unity |  
• $PC_1$  
• Simple point scores |
| Multiple dimensional scoring of all predictors | ↓ d.f. for all predictors combined |  
Principal components $1, 2, \ldots, k; k < p$ computed from all transformed predictors |
4.8 Other Approaches to Predictive Modeling

4.9 Overly Influential Observations

- Every observation should influence fit
- Major results should not rest on 1 or 2 obs.
- Overly infl. obs. → ↑ variance of predictions
- Also affects variable selection

Reasons for influence:

- Too few observations for complexity of model (see Sections 4.7, 4.3)
- Data transcription or entry errors
- Extreme values of a predictor
  1. Sometimes subject so atypical should remove from dataset
  2. Sometimes truncate measurements where data density ends
  3. Example:  \( n = 4000 \), 2000 deaths, white blood count range 500-100,000, .05,.95 quantiles=2755, 26700
4. Linear spline function fit
5. Sensitive to WBC $>$ 60000 ($n = 16$)
6. Predictions stable if truncate WBC to 40000 ($n = 46$ above 40000)

- Disagreements between predictors and response. Ignore unless extreme values or another explanation

- Example: $n = 8000$, one extreme predictor value not on straight line relationship with other $(X, Y) \rightarrow \chi^2 = 36$ for $H_0 :$ linearity

Statistical Measures:

- Leverage: capacity to be influential (not necessarily infl.)
  Diagonals of “hat matrix” $H = X(X'X)^{-1}X'$ — measures how an obs. predicts its own response [15]

- $h_{ii} > 2(p + 1)/n$ may signal a high leverage point [15]

- DFBETAS: change in $\hat{\beta}$ upon deletion of each obs, scaled by s.e.

- DFFIT: change in $X\hat{\beta}$ upon deletion of each obs

- DFFITS: DFFIT standardized by s.e. of $\hat{\beta}$

- Some classify obs as overly influential when $|DFFITS| > 2\sqrt{(p + 1)/(n - p - 1)}$ [15]
• Others examine entire distribution for “outliers”

• No substitute for careful examination of data [38, 193]

• Maximum likelihood estimation requires 1-step approximations
4.10

Comparing Two Models

- Level playing field (independent datasets, same no. candidate d.f., careful bootstrapping)

- Criteria:
  1. calibration
  2. discrimination
  3. face validity
  4. measurement errors in required predictors
  5. use of continuous predictors (which are usually better defined than categorical ones)
  6. omission of “insignificant” variables that nonetheless make sense as risk factors
  7. simplicity (though this is less important with the availability of computers)
  8. lack of fit for specific types of subjects

- Goal is to rank-order: ignore calibration

- Otherwise, dismiss a model having poor calibration

- Good calibration $\rightarrow$ compare discrimination (e.g., $R^2$ \textsuperscript{153}, model $\chi^2$, Somers’ $D_{xy}$, Spearman’s $\rho$, area under ROC curve)
• Worthwhile to compare models on a measure not used to optimize either model, e.g., mean absolute error, median absolute error if using OLS

• Rank measures may not give enough credit to extreme predictions → model $\chi^2$, $R^2$, examine extremes of distribution of $\hat{Y}$

• Examine differences in predicted values from the two models

• See [159, 162, 161, 160] for discussions and examples of low power for testing differences in ROC areas, and for other approaches.
4.11

**Improving the Practice of Multivariable Prediction**

See also Section 5.6.

Greenland [89] discusses many important points:

- Stepwise variable selection on confounders leaves important confounders uncontrolled

- Shrinkage is far superior to variable selection

- Variable selection does more damage to confidence interval widths than to point estimates

- Claims about unbiasedness of ordinary MLEs are misleading because they assume the model is correct and is the only model entertained

- “models need to be complex to capture uncertainty about the relations ... an honest uncertainty assessment requires parameters for all effects that we know may be present. This advice is implicit in an antiparsimony principle often attributed to L. J. Savage ‘All models should be as big as an elephant’ (see Draper, 1995)”

Greenland’s example of inadequate adjustment for confounders
as a result of using a bad modeling strategy:

- Case-control study of diet, food constituents, breast cancer
- 140 cases, 222 controls
- 35 food constituent intakes and 5 confounders
- Food intakes are correlated
- Traditional stepwise analysis not adjusting simultaneously for all foods consumed $\rightarrow$ 11 foods had $P < 0.05$
- Full model with all 35 foods competing $\rightarrow$ 2 had $P < 0.05$
- Rigorous simultaneous analysis (hierarchical random slopes model) penalizing estimates for the number of associations examined $\rightarrow$ no foods associated with breast cancer

Global Strategies

- Use a method known not to work well (e.g., stepwise variable selection without penalization; recursive partitioning), document how poorly the model performs (e.g. using the bootstrap), and use the model anyway
- Develop a black box model that performs poorly and is difficult to interpret (e.g., does not incorporate penalization)
• Develop a black box model that performs well and is difficult to interpret

• Develop interpretable approximations to the black box

• Develop an interpretable model (e.g. give priority to additive effects) that performs well and is likely to perform equally well on future data from the same stream

Preferred Strategy in a Nutshell

• Decide how many d.f. can be spent

• Decide where to spend them

• Spend them

• Don’t reconsider, especially if inference needed
4.12

Summary: Possible Modeling Strategies

4.12.1

Developing Predictive Models

1. Assemble accurate, pertinent data and lots of it, with wide distributions for $X$.

2. Formulate good hypotheses — specify relevant candidate predictors and possible interactions. Don’t use $Y$ to decide which $X$’s to include.

3. Characterize subjects with missing $Y$. Delete such subjects in rare circumstances [53]. For certain models it is effective to multiply impute $Y$.

4. Characterize and impute missing $X$. In most cases use multiple imputation based on $X$ and $Y$.

5. For each predictor specify complexity or degree of nonlinearity that should be allowed (more for important predictors or for large $n$) (Section 4.1)

6. Do data reduction if needed (pre-transformations, combinations), or use penalized estimation [98]

7. Use the entire sample in model development

8. Can do highly structured testing to simplify “initial” model
   (a) Test entire group of predictors with a single $P$-value
(b) Make each continuous predictor have same number of knots, and select the number that optimizes AIC
(c) Test the combined effects of all nonlinear terms with a single $P$-value

9. Make tests of linearity of effects in the model only to demonstrate to others that such effects are often statistically significant. Don’t remove individual insignificant effects from the model.

10. Check additivity assumptions by testing pre-specified interaction terms. Use a global test and either keep all or delete all interactions.

11. Check to see if there are overly-influential observations.

12. Check distributional assumptions and choose a different model if needed.

13. Do limited backwards step-down variable selection if parsimony is more important that accuracy [194]. But confidence limits, etc., must account for variable selection (e.g., bootstrap).

14. This is the “final” model.

15. Interpret the model graphically and by computing predicted values and appropriate test statistics. Compute pooled tests of association for collinear predictors.

16. Validate this model for calibration and discrimination ability, preferably using bootstrapping.
17. Shrink parameter estimates if there is overfitting but no further data reduction is desired (unless shrinkage built-in to estimation)

18. When missing values were imputed, adjust final variance-covariance matrix for imputation. Do this as early as possible because it will affect other findings.

19. When all steps of the modeling strategy can be automated, consider using Faraway’s method [72] to penalize for the randomness inherent in the multiple steps.

20. Develop simplifications to the final model as needed.

4.12.2 Developing Models for Effect Estimation

1. Less need for parsimony; even less need to remove insignificant variables from model (otherwise CLs too narrow)

2. Careful consideration of interactions; inclusion forces estimates to be conditional and raises variances

3. If variable of interest is mostly the one that is missing, multiple imputation less valuable

4. Complexity of main variable specified by prior beliefs, compromise between variance and bias

5. Don’t penalize terms for variable of interest

6. Model validation less necessary
Developing Models for Hypothesis Testing

1. Virtually same as previous strategy

2. Interactions require tests of effect by varying values of another variable, or “main effect + interaction” joint tests (e.g., is treatment effective for either sex, allowing effects to be different)

3. Validation may help quantify overadjustment
Chapter 5

Describing, Resampling, Validating, and Simplifying the Model

5.1 Describing the Fitted Model

5.1.1 Interpreting Effects

- Regression coefficients if 1 d.f. per factor, no interaction
- **Not** standardized regression coefficients
- Many programs print meaningless estimates such as effect of increasing age^2 by one unit, holding age constant
- Need to account for nonlinearity, interaction, and use meaningful ranges
- For monotonic relationships, estimate \( X \hat{\beta} \) at quartiles of continuous variables, separately for various levels of inter-
acting factors

• Subtract estimates, anti-log, e.g., to get inter-quartile-range odds or hazards ratios. Base C.L. on s.e. of difference. See Figure 19.10.

• Partial effect plot: Plot effect of each predictor on $X\beta$ or some transformation. See Figure 19.8. See also [114].

• Nomogram. See Figure 19.12.

• Use regression tree to approximate the full model

---

5.1.2 Indexes of Model Performance

Error Measures

• Central tendency of prediction errors
  – Mean absolute prediction error: mean $|Y - \hat{Y}|$
  – Mean squared prediction error
    * Binary $Y$: Brier score (quadratic proper scoring rule)
  – Logarithmic proper scoring rule (avg. log-likelihood)

• Discrimination measures
Pure discrimination: rank correlation of \((\hat{Y}, Y)\)

* Spearman \(\rho\), Kendall \(\tau\), Somers’ \(D_{xy}\)

* \(Y\) binary \(\rightarrow D_{xy} = 2 \times (C - \frac{1}{2})\)
  \(C = \) concordance probability = area under receiver operating characteristic curve \(\propto\) Wilcoxon-Mann-Whitney statistic

Mostly discrimination: \(R^2\)

* \(R^2_{adj}\) — overfitting corrected if model pre-specified

Brier score can be decomposed into discrimination and calibration components

Discrimination measures based on variation in \(\hat{Y}\)

* regression sum of squares

* \(g\)–index

Calibration measures

* calibration–in–the–large: average \(\hat{Y}\) vs. average \(Y\)

* high-resolution calibration curve (calibration–in–the–small). See Figure 12.7.

* calibration slope and intercept

* maximum absolute calibration error
– mean absolute calibration error
– 0.9 quantile of calibration error

See Van Calster et al. [210] for a nice discussion of different levels of calibration stringency and their relationship to likelihood of errors in decision making.

\( g\)-Index

- Based on Gini’s mean difference
  - mean over all possible \( i \neq j \) of \( |Z_i - Z_j| \)
  - interpretable, robust, highly efficient measure of variation

- \( g = \) Gini’s mean difference of \( X_i \hat{\beta} = \hat{Y} \)

- Example: \( Y = \) systolic blood pressure; \( g = 11 \text{mmHg} \) is typical difference in \( \hat{Y} \)

- Independent of censoring etc.

- For models in which anti-log of difference in \( \hat{Y} \) represent \( g \) meaningful ratios (odds ratios, hazard ratios, ratio of medians):
  \( g_r = \exp(g) \)

- For models in which \( \hat{Y} \) can be turned into a probability
estimate (e.g., logistic regression):

\[ g_p = \text{Gini’s mean difference of } \hat{P} \]

- These \( g \)-indexes represent e.g. “typical” odds ratios, “typical” risk differences

- Can define partial \( g \)
5.2 The Bootstrap

- If know population model, use simulation or analytic derivations to study behavior of statistical estimator

- Suppose \( Y \) has a cumulative dist. fctn. \( F(y) = \text{Prob}\{Y \leq y\} \)

- We have sample of size \( n \) from \( F(y) \), \( Y_1, Y_2, \ldots, Y_n \)

- Steps:
  1. Repeatedly simulate sample of size \( n \) from \( F \)
  2. Compute statistic of interest
  3. Study behavior over \( B \) repetitions

- Example: 1000 samples, 1000 sample medians, compute their sample variance

- \( F \) unknown \( \rightarrow \) estimate by empirical dist. fctn.

\[
F_n(y) = \frac{1}{n} \sum_{i=1}^{n} [Y_i \leq y].
\]

- Example: sample of size \( n = 30 \) from a normal distribution with mean 100 and SD 10

```r
set.seed(6)
x ← rnorm(30, 100, 20)
xs ← seq(50, 150, length=150)
```
\begin{verbatim}
cdf ← pnorm(xs, 100, 20)
plot(xs, cdf, type='l', ylim=c(0,1),
     xlab=expression(x),
     ylab=expression(paste("Prob[", X \le x, "]")))
lines(ecdf(x), cex=.5)
\end{verbatim}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{fig5_1.png}
\caption{Empirical and population cumulative distribution function}
\end{figure}

- $F_n$ corresponds to density function placing probability $\frac{1}{n}$ at each observed data point ($\frac{k}{n}$ if point duplicated $k$ times)

- Pretend that $F \equiv F_n$

- Sampling from $F_n \equiv$ sampling with replacement from observed data $Y_1, \ldots, Y_n$

- Large $n \rightarrow$ selects $1 - e^{-1} \approx 0.632$ of original data points in each bootstrap sample at least once

- Some observations not selected, others selected more than once
• Efron’s bootstrap → general-purpose technique for estimating properties of estimators without assuming or knowing distribution of data $F$

• Take $B$ samples of size $n$ with replacement, choose $B$ so that summary measure of individual statistics $\approx$ summary if $B = \infty$

• Bootstrap based on distribution of observed differences between a resampled parameter estimate and the original estimate telling us about the distribution of unobservable differences between the original estimate and the unknown parameter

Example: Data $(1, 5, 6, 7, 8, 9)$, obtain 0.80 confidence interval for population median, and estimate of population expected value of sample median (only to estimate the bias in the original estimate of the median).

```r
options(digits=3)
y ← c(2, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 19, 20, 21)
y ← c(1, 5, 6, 7, 8, 9)
set.seed(17)
n ← length(y)
n2 ← n/2
n21 ← n2 + 1
B ← 400
M ← double(B)
plot(0, 0, xlim=c(0, B), ylim=c(3, 9),
     xlab="Bootstrap Samples Used",
     ylab="Mean and 0.1, 0.9 Quantiles", type="n")
for (i in 1:B) {
  s ← sample(1:n, n, replace=T)
  x ← sort(y[s])
  m ← .5*(x[n2]+x[n21])
  M[i] ← m
  if(i ≤ 20) {
    w ← as.character(x)
    cat(w, "& ", sprintf("%.1f", m),
        if(i < 20) "\\n" else "\\hline\\n",
```
CHAPTER 5. DESCRIBING, RESAMPLING, VALIDATING, AND SIMPLIFYING THE MODEL

file = ' ~/doc/rms/validate/tab.tex ', append = i > 1)
}
points(i, mean(M[1:i]), pch = 46)
if(i ≥ 10) {
  q ← quantile(M[1:i], c(.1, .9))
  points(i, q[1], pch = 46, col = 'blue')
  points(i, q[2], pch = 46, col = 'blue')
}
}
table(M)

M
1 3 3.5 4 4.5 5 5.5 6 6.5 7 7.5 8 8.5 9
2 7 6 2 1 30 45 59 72 70 45 48 8 5

hist(M, nclass = length(unique(M)), xlab = '', main = '')

Figure 5.2: Estimating properties of sample median using the bootstrap

First 20 samples:
<table>
<thead>
<tr>
<th>Bootstrap Sample</th>
<th>Sample Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 5 5 7 8 9</td>
<td>6.0</td>
</tr>
<tr>
<td>1 1 5 7 9 9</td>
<td>6.0</td>
</tr>
<tr>
<td>6 7 7 8 9 9</td>
<td>7.5</td>
</tr>
<tr>
<td>1 1 5 6 8 9</td>
<td>5.5</td>
</tr>
<tr>
<td>1 6 7 7 8 8</td>
<td>7.0</td>
</tr>
<tr>
<td>1 5 6 8 8 9</td>
<td>7.0</td>
</tr>
<tr>
<td>1 6 8 8 9 9</td>
<td>8.0</td>
</tr>
<tr>
<td>5 5 6 7 8 9</td>
<td>6.5</td>
</tr>
<tr>
<td>1 5 6 7 7 8</td>
<td>6.5</td>
</tr>
<tr>
<td>1 5 6 8 9 9</td>
<td>7.0</td>
</tr>
<tr>
<td>1 5 7 7 8 9</td>
<td>7.0</td>
</tr>
<tr>
<td>1 5 6 6 7 8</td>
<td>6.0</td>
</tr>
<tr>
<td>1 6 6 7 8 9</td>
<td>6.5</td>
</tr>
<tr>
<td>5 6 7 7 8 9</td>
<td>7.0</td>
</tr>
<tr>
<td>1 5 6 8 8 8</td>
<td>7.0</td>
</tr>
<tr>
<td>1 1 6 6 7 8</td>
<td>6.0</td>
</tr>
<tr>
<td>5 5 5 8 8 9</td>
<td>6.5</td>
</tr>
<tr>
<td>5 6 6 6 7 7</td>
<td>6.0</td>
</tr>
<tr>
<td>1 5 7 9 9 9</td>
<td>8.0</td>
</tr>
<tr>
<td>1 1 5 5 5 7</td>
<td>5.0</td>
</tr>
</tbody>
</table>

- Histogram tells us whether we can assume normality for the bootstrap medians or need to use quantiles of medians to construct C.L.

- Need high $B$ for quantiles, low for variance (but see [23])

- See [66] for useful information about bootstrap confidence intervals and the latest R functions
5.3 Model Validation

5.3.1 Introduction

- External validation (best: another country at another time); also validates sampling, measurements\(^a\)

- Internal
  - apparent (evaluate fit on same data used to create fit)
  - data splitting
  - cross-validation
  - bootstrap: get overfitting-corrected accuracy index

- Best way to make model fit data well is to discard much of the data

- Predictions on another dataset will be inaccurate

- Need unbiased assessment of predictive accuracy

**Working definition of external validation**: Validation of a prediction tool on a sample that was not available at publi-

\(^a\)But in many cases it is better to combine data and include country or calendar time as a predictor.
cation time. **Alternate**: Validation of a prediction tool by an independent research team.

One suggested hierarchy of the quality of various validation methods is as follows, ordered from worst to best.

1. Attempting several validations (internal or external) and reporting only the one that “worked”
2. Reporting apparent performance on the training dataset (no validation)
3. Reporting predictive accuracy on an undersized independent test sample
4. Internal validation using data splitting where at least one of the training and test samples is not huge and the investigator is not aware of the arbitrariness of variable selection done on a single sample
5. Strong internal validation using 100 repeats of 10-fold cross-validation or several hundred bootstrap resamples, repeating all analysis steps involving $Y$ afresh at each re-sample and the arbitrariness of selected “important variables” is reported (if variable selection is used)
6. External validation on a large test sample, done by the original research team
7. Re-analysis by an independent research team using strong internal validation of the original dataset
8. External validation using new test data, done by an independent research team

9. External validation using new test data generated using different instruments/technology, done by an independent research team

Some points to consider:

- Unless both sample sizes are huge, external validation can be low precision

- External validation can be costly and slow and may result in disappointment that would have been revealed earlier with rigorous internal validation

- External validation is sometimes *gamed*; researchers disappointed in the validation sometimes ask for a “do over”; re-sampling validation is harder to game as long as all analytical steps using $Y$ are repeated each time.

- Instead of external validation to determine model applicability at a different time or place, and being disappointed if the model does not work in that setting, consider building a unified model containing time and place as predictors

- When the model was fully pre-specified, external validation tests *the model*
But when the model was fitted using machine learning, feature screening, variable selection, or model selection, the model developed using training data is usually only an example of a model, and the test sample validation could be called an example validation.

When resampling is used to repeat all modeling steps for each resample, rigorous internal validation tests the process used to develop the model and happens to also provide a high-precision estimate of the likely future performance of the “final” model developed using that process, properly penalizing for model uncertainty.

Resampling also reveals the volatility of the model selection process.

See BBR 10.11

Collins et al. [45] estimate that a typical sample size needed for externally validating a time-to-event model is 200 events.

5.3.2 Which Quantities Should Be Used in Validation?

- OLS: $R^2$ is one good measure for quantifying drop-off in predictive ability

- Example: $n = 10, p = 9$, apparent $R^2 = 1$ but $R^2$ will be close to zero on new subjects
• Example: $n = 20, p = 10$, apparent $R^2 = .9$, $R^2$ on new data $0.7$, $R^2_{adj} = 0.79$

• Adjusted $R^2$ solves much of the bias problem assuming $p$ in its formula is the largest number of parameters ever examined against $Y$

• Few other adjusted indexes exist

• Also need to validate models with phantom d.f.

• Cross-validation or bootstrap can provide unbiased estimate of any index; bootstrap has higher precision

• Two main types of quantities to validate
  1. Calibration or reliability: ability to make unbiased estimates of response ($\hat{Y}$ vs. $Y$)
  2. Discrimination: ability to separate responses
     OLS: $R^2$; $g$–index; binary logistic model: ROC area, equivalent to rank correlation between predicted probability of event and 0/1 event

• Unbiased validation nearly always necessary, to detect overfitting
5.3.3 Data-Splitting

- Split data into *training* and *test* sets

- Interesting to compare index of accuracy in training and test

- Freeze parameters from training

- Make sure you allow $R^2 = 1 - \frac{SSE}{SST}$ for test sample to be $< 0$

- Don’t compute ordinary $R^2$ on $X\hat{\beta}$ vs. $Y$; this allows for linear recalibration $aX\hat{\beta} + b$ vs. $Y$

- Test sample must be large enough to obtain very accurate assessment of accuracy

- Training sample is what’s left

- Example: overall sample $n = 300$, training sample $n = 200$, develop model, freeze $\hat{\beta}$, predict on test sample ($n = 100$), $R^2 = 1 - \frac{\sum(Y_i - X_i\hat{\beta})^2}{\sum(Y_i - \bar{Y})^2}$.

- Disadvantages of data splitting:
  1. Costly in $\downarrow n$ [179, 26]
  2. Requires *decision* to split at beginning of analysis
  3. Requires larger sample held out than cross-validation
4. Results vary if split again
5. Does not validate the final model (from recombined data)
6. Not helpful in getting CL corrected for var. selection
7. Nice summary of disadvantages:

5.3.4 Improvements on Data-Splitting: Resampling

- No sacrifice in sample size
- Work when modeling process automated
- Bootstrap excellent for studying arbitrariness of variable selection [183]. See P. 10-43.
- Cross-validation solves many problems of data splitting [214, 188, 236, 65]

- Example of $\times$-validation:
  1. Split data at random into 10 tenths
  2. Leave out $\frac{1}{10}$ of data at a time
  3. Develop model on $\frac{9}{10}$, including any variable selection, pre-testing, etc.
  4. Freeze coefficients, evaluate on $\frac{1}{10}$
  5. Average $R^2$ over 10 reps

- Drawbacks:
1. Choice of number of groups and repetitions
2. Doesn’t show full variability of var. selection
3. Does not validate full model
4. Lower precision than bootstrap
5. Need to do 50 repeats of 10-fold cross-validation to ensure adequate precision

• Randomization method
  1. Randomly permute \( Y \)
  2. Optimism = performance of fitted model compared to what expect by chance

---

### Validation Using the Bootstrap

• Estimate optimism of final whole sample fit without holding out data

• From original \( X \) and \( Y \) select sample of size \( n \) with replacement

• Derive model from bootstrap sample

• Apply to original sample

• Simple bootstrap uses average of indexes computed on original sample
• Estimated optimism = difference in indexes

• Repeat about $B = 100$ times, get average expected optimism

• Subtract average optimism from apparent index in final model

• Example: $n = 1000$, have developed a final model that is hopefully ready to publish. Call estimates from this final model $\hat{\beta}$.
  
  – final model has apparent $R^2 \ (R^2_{\text{app}}) = 0.4$

  – how inflated is $R^2_{\text{app}}$?

  – get resamples of size 1000 with replacement from original 1000

  – for each resample compute $R^2_{\text{boot}} =$ apparent $R^2$ in bootstrap sample

  – freeze these coefficients (call them $\hat{\beta}_{\text{boot}}$), apply to original (whole) sample $(X_{\text{orig}}, Y_{\text{orig}})$ to get $R^2_{\text{orig}} = R^2(X_{\text{orig}} \hat{\beta}_{\text{boot}}, Y_{\text{orig}})$

  – optimism = $R^2_{\text{boot}} - R^2_{\text{orig}}$

  – average over $B = 100$ optimisms to get $\overline{\text{optimism}}$

  – $R^2_{\text{overfitting corrected}} = R^2_{\text{app}} - \overline{\text{optimism}}$
• Example: See P. 10-41

• Is estimating unconditional (not conditional on $X$) distribution of $R^2$, etc. [72, p. 217]

• Conditional estimates would require assuming the model one is trying to validate

• Efron’s “.632” method may perform better (reduce bias further) for small $n$ [65], [67, p. 253], [68]

Bootstrap useful for assessing calibration in addition to discrimination:

• Fit $C(Y | X) = X \beta$ on bootstrap sample

• Re-fit $C(Y | X) = \gamma_0 + \gamma_1 X \hat{\beta}$ on same data

• $\hat{\gamma}_0 = 0$, $\hat{\gamma}_1 = 1$

• Test data (original dataset): re-estimate $\gamma_0$, $\gamma_1$

• $\hat{\gamma}_1 < 1$ if overfit, $\hat{\gamma}_0 > 0$ to compensate

• $\hat{\gamma}_1$ quantifies overfitting and useful for improving calibration [194]

• Use Efron’s method to estimate optimism in $(0, 1)$, estimate $(\gamma_0, \gamma_1)$ by subtracting optimism from $(0, 1)$
• See also Copas [49] and van Houwelingen and le Cessie [214, p. 1318]

See [75] for warnings about the bootstrap, and [65] for variations on the bootstrap to reduce bias.

Use bootstrap to choose between full and reduced models:

• Bootstrap estimate of accuracy for full model

• Repeat, using chosen stopping rule for each re-sample

• Full fit usually outperforms reduced model [194]

• Stepwise modeling often reduces optimism but this is not offset by loss of information from deleting marginal var.

<table>
<thead>
<tr>
<th>Method</th>
<th>Apparent Rank Correlation of Predicted vs. Observed</th>
<th>Over-Optimism</th>
<th>Bias-Corrected Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Model</td>
<td>0.50</td>
<td>0.06</td>
<td>0.44</td>
</tr>
<tr>
<td>Stepwise Model</td>
<td>0.47</td>
<td>0.05</td>
<td>0.42</td>
</tr>
</tbody>
</table>

In this example, stepwise modeling lost a possible $0.50 - 0.47 = 0.03$ predictive discrimination. The full model fit will especially be an improvement when

1. The stepwise selection deleted several variables which were almost significant.

2. These marginal variables have some real predictive value, even if it’s slight.
3. There is no small set of extremely dominant variables that would be easily found by stepwise selection.

Other issues:

- See [214] for many interesting ideas

- Faraway [72] shows how bootstrap is used to penalize for choosing transformations for $Y$, outlier and influence checking, variable selection, etc. simultaneously

- Brownstone [30, p. 74] feels that “theoretical statisticians have been unable to analyze the sampling properties of [usual multi-step modeling strategies] under realistic conditions” and concludes that the modeling strategy must be completely specified and then bootstrapped to get consistent estimates of variances and other sampling properties

- See Blettner and Sauerbrei [21] and Chatfield [39] for more interesting examples of problems resulting from data-driven analyses.
5.4 Bootstrapping Ranks of Predictors

- Order of importance of predictors not pre-specified
- Researcher interested in determining “winners” and “losers”
- Bootstrap useful in documenting the difficulty of this task
- Get confidence limits of the rank of each predictor in the scale of partial $\chi^2$ - d.f.

- Example using OLS

```r
# Use the plot method for anova, with pl=FALSE to suppress actual # plotting of chi-square - d.f. for each bootstrap repetition. # Rank the negative of the adjusted chi-squares so that a rank of # 1 is assigned to the highest. It is important to tell # plot.anova.rms not to sort the results, or every bootstrap # replication would have ranks of 1, 2, 3, ... for the stats.
require(rms)
n ← 300
set.seed(1)
d ← data.frame(x1=runif(n), x2=runif(n), x3=runif(n), x4=runif(n),
    x5=runif(n), x6=runif(n), x7=runif(n), x8=runif(n),
    x9=runif(n), x10=runif(n), x11=runif(n), x12=runif(n))
d$y ← with(d, 1*x1 + 2*x2 + 3*x3 + 4*x4 + 5*x5 + 6*x6 + 7*x7 +
    8*x8 + 9*x9 + 10*x10 + 11*x11 + 12*x12 + 9* rnorm(n))

f ← ols(y ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12, data=d)
B ← 1000
ranks ← matrix(NA, nrow=B, ncol=12)
rankvars ← function(fit)
    rank(plot(anova(fit), sort='none', pl=FALSE))
Rank ← rankvars(f)
for(i in 1:B) {
    j ← sample(1:n, n, TRUE)
    bootfit ← update(f, data=d, subset=j)
    ranks[i,] ← rankvars(bootfit)
}
lim ← t(apply(ranks, 2, quantile, probs=c(.025,.975)))
predictor ← factor(names(Rank), names(Rank))
w ← data.frame(predictor, Rank, lower=lim[,1], upper=lim[,2])
require(ggplot2)
ggplot(w, aes(x=predictor, y=Rank)) + geom_point() + coord_flip() +
```
Figure 5.3: Bootstrap percentile 0.95 confidence limits for ranks of predictors in an OLS model. Ranking is on the basis of partial $\chi^2$ minus d.f. Point estimates are original ranks.
5.5 Simplifying the Final Model by Approximating It

5.5.1 Difficulties Using Full Models

- Predictions are conditional on all variables, standard errors ↑ when predict for a low-frequency category
- Collinearity
- Can average predictions over categories to marginalize, ↓ s.e.

5.5.2 Approximating the Full Model

- Full model is gold standard
- Approximate it to any desired degree of accuracy
- If approx. with a tree, best c-v tree will have 1 obs./node
- Can use least squares to approx. model by predicting \( \hat{Y} = X \hat{\beta} \)
- When original model also fit using least squares, coef. of
approx. model against $\hat{Y} \equiv$ coef. of subset of variables fitted against $Y$ (as in stepwise)

- Model approximation still has some advantages
  1. Uses unbiased estimate of $\sigma$ from full fit
  2. Stopping rule less arbitrary
  3. Inheritance of shrinkage

- If estimates from full model are $\hat{\beta}$ and approx. model is based on a subset $T$ of predictors $X$, coef. of approx. model are $W\hat{\beta}$, where
  
  $W = (T'T)^{-1}T'X$

- Variance matrix of reduced coef.: $WVW'$
5.6 How Do We Break Bad Habits?

- Insist on validation of predictive models and discoveries

- Show collaborators that split-sample validation is not appropriate unless the number of subjects is huge
  - Split more than once and see volatile results
  - Calculate a confidence interval for the predictive accuracy in the test dataset and show that it is very wide

- Run simulation study with no real associations and show that associations are easy to find

- Analyze the collaborator’s data after randomly permuting the $Y$ vector and show some positive findings

- Show that alternative explanations are easy to posit
  - Importance of a risk factor may disappear if 5 “unimportant” risk factors are added back to the model
  - Omitted main effects can explain apparent interactions
  - *Uniqueness analysis*: attempt to predict the predicted values from a model derived by data torture from all of the features not used in the model
Chapter 6

R Software

R allows interaction spline functions, wide variety of predictor parameterizations, wide variety of models, unifying model formula language, model validation by resampling.

R is comprehensive:

• Easy to write R functions for new models $\rightarrow$ wide variety of modern regression models implemented (trees, nonparametric, ACE, AVAS, survival models for multiple events)

• Designs can be generated for any model $\rightarrow$ all handle “class” var, interactions, nonlinear expansions

• Single R objects (e.g., fit object) can be self-documenting $\rightarrow$ automatic hypothesis tests, predictions for new data

• Superior graphics

• Classes and generic functions
The R Modeling Language

R statistical modeling language:

<table>
<thead>
<tr>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>y ~ age + sex</code></td>
<td>age + sex main effects</td>
</tr>
<tr>
<td><code>y ~ age + sex + age:sex</code></td>
<td>add second-order interaction + all main effects</td>
</tr>
<tr>
<td><code>y ~ age*sex</code></td>
<td>second-order interaction + all main effects</td>
</tr>
<tr>
<td><code>y ~ (age + sex + pressure)^2</code></td>
<td>age + sex + pressure + age:sex + age:pressure...</td>
</tr>
<tr>
<td><code>y ~ (age + sex + pressure)^2 - sex:pressure</code></td>
<td>all main effects and all 2nd order interactions except sex:pressure</td>
</tr>
<tr>
<td><code>y ~ (age + race)*sex</code></td>
<td>age + race + sex + age:sex + race:sex</td>
</tr>
<tr>
<td><code>y ~ treatment*(age*race + age*sex)</code></td>
<td>no interact. with race, sex</td>
</tr>
<tr>
<td><code>sqrt(y) ~ sex*sqrt(age) + race</code></td>
<td>functions, with dummy variables generated if</td>
</tr>
<tr>
<td><code>sqrt(y) ~ sex*sqrt(age) + race</code></td>
<td># race is an R factor (classification) variable</td>
</tr>
<tr>
<td><code>y ~ sex + poly(age,2)</code></td>
<td>poly generates orthogonal polynomials</td>
</tr>
<tr>
<td><code>race.sex &lt;- interaction(race,sex)</code></td>
<td></td>
</tr>
<tr>
<td><code>y ~ age + race.sex</code></td>
<td>for when you want dummy variables for all combinations of the factors</td>
</tr>
</tbody>
</table>

The formula for a regression model is given to a modeling function, e.g.

```
lrm(y ~ rcs(x,4))
```

is read “use a logistic regression model to model y as a function of x, representing x by a restricted cubic spline with 4 default knots”\(^a\).

**update function:** re-fit model with changes in terms or data:

```
f <- lrm(y ~ rcs(x,4) + x2 + x3)
f2 <- update(f, subset=sex=="male")
f3 <- update(f, .~.-x2)          # remove x2 from model
f4 <- update(f, .~. + rcs(x5,5))# add rcs(x5,5) to model
f5 <- update(f, y2 ~ .)         # same terms, new response var.
```

\(^a\)lrm and rcs are in the rms package.
6.2 User-Contributed Functions

- R is high-level object-oriented language.
- R (UNIX, Linux, Mac, Windows)
- Multitude of user-contributed functions freely available
- International community of users

Some R functions:

- See Venables and Ripley
- Hierarchical clustering: `hclust`
- Principal components: `princomp`, `prcomp`
- Canonical correlation: `cancor`
- Nonparametric transform-both-sides additive models: `ace`, `avas`
- Parametric transform-both-sides additive models: `areg`, `areg.boot` (Hmisc package in R))
- Rank correlation methods: `rcorr`, `hoeffd`, `spearman2` (Hmisc)
- Variable clustering: `varclus (Hmisc)`

- Single imputation: `transcan (Hmisc)`

- Multiple imputation: `aregImpute (Hmisc)`

- Restricted cubic splines:
  `rcspline.eval (Hmisc)`

- Re-state restricted spline in simpler form:
  `rcspline.restate (Hmisc)`
6.3 The \texttt{rms} Package

- \texttt{datadist} function to compute predictor distribution summaries

\begin{verbatim}
\texttt{y \sim \text{sex} + \text{lsp}(\text{age},c(20,30,40,50,60)) +}
\texttt{sex }\%\text{i}\%\text{ lsp}(\text{age},c(20,30,40,50,60))
\end{verbatim}

E.g. restrict age $\times$ cholesterol interaction to be of form $AF(B) + BG(A)$:

\begin{verbatim}
\texttt{y \sim \text{lsp}(\text{age},30) + \text{rcs}(\text{cholesterol},4) +}
\texttt{\text{lsp}(\text{age},30) }\%\text{i}\%\text{ rcs}(\text{cholesterol},4)
\end{verbatim}

Special fitting functions by Harrell to simplify procedures described in these notes:

<table>
<thead>
<tr>
<th>Function</th>
<th>Purpose</th>
<th>Related R Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>ols</td>
<td>Ordinary least squares linear model</td>
<td>\texttt{lm}</td>
</tr>
<tr>
<td>lrm</td>
<td>Binary and ordinal logistic regression model</td>
<td>\texttt{glm}</td>
</tr>
<tr>
<td>blrm</td>
<td>Bayesian binary and ordinal logistic model</td>
<td></td>
</tr>
<tr>
<td>orm</td>
<td>Ordinal semi-parametric regression model for continuous $Y$ and several link functions</td>
<td>\texttt{polr, lrm}</td>
</tr>
<tr>
<td>psm</td>
<td>Accelerated failure time parametric survival models</td>
<td>\texttt{survreg}</td>
</tr>
<tr>
<td>cph</td>
<td>Cox proportional hazards regression</td>
<td>\texttt{coxph}</td>
</tr>
<tr>
<td>bj</td>
<td>Buckley-James censored least squares model</td>
<td>\texttt{survreg, lm}</td>
</tr>
<tr>
<td>Glm</td>
<td>\texttt{rms} version of \texttt{glm}</td>
<td>\texttt{glm}</td>
</tr>
<tr>
<td>Gls</td>
<td>\texttt{rms} version of \texttt{gls}</td>
<td>\texttt{gls (nlme package)}</td>
</tr>
<tr>
<td>Rq</td>
<td>\texttt{rms} version of \texttt{rq}</td>
<td>\texttt{rq (quantreg package)}</td>
</tr>
<tr>
<td>Function</td>
<td>Purpose</td>
<td>Related R Functions</td>
</tr>
<tr>
<td>----------</td>
<td>---------</td>
<td>---------------------</td>
</tr>
<tr>
<td>asis</td>
<td>No post-transformation (seldom used explicitly)</td>
<td>I</td>
</tr>
<tr>
<td>rcs</td>
<td>Restricted cubic splines</td>
<td>ns</td>
</tr>
<tr>
<td>pol</td>
<td>Polynomial using standard notation</td>
<td>poly</td>
</tr>
<tr>
<td>lsp</td>
<td>Linear spline</td>
<td></td>
</tr>
<tr>
<td>catg</td>
<td>Categorical predictor (seldom)</td>
<td>factor</td>
</tr>
<tr>
<td>scored</td>
<td>Ordinal categorical variables</td>
<td>ordered</td>
</tr>
<tr>
<td>matrix</td>
<td>Keep variables as group for anova and fastbw</td>
<td>matrix</td>
</tr>
<tr>
<td>strat</td>
<td>Non-modeled stratification factors</td>
<td>strata</td>
</tr>
<tr>
<td></td>
<td>(used for cph only)</td>
<td></td>
</tr>
</tbody>
</table>

Below notice that there are three graphic models implemented for depicting the effects of predictors in the fitted model: lattice graphics, a ggplot method using the ggplot2 package (which has an option to convert the result to plotly), and a direct plotly method. plotly is used to create somewhat interactive graphics with drill-down capability, and the rms package takes advantage of this capability. plotly graphics are best used with RStudio Rmarkdown html output.
<table>
<thead>
<tr>
<th>Function</th>
<th>Purpose</th>
<th>Related Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>print</td>
<td>Print parameters and statistics of fit</td>
<td></td>
</tr>
<tr>
<td>coef</td>
<td>Fitted regression coefficients</td>
<td></td>
</tr>
<tr>
<td>formula</td>
<td>Formula used in the fit</td>
<td></td>
</tr>
<tr>
<td>specs</td>
<td>Detailed specifications of fit</td>
<td></td>
</tr>
<tr>
<td>vcov</td>
<td>Fetch covariance matrix</td>
<td></td>
</tr>
<tr>
<td>logLik</td>
<td>Fetch maximized log-likelihood</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>Fetch AIC with option to put on chi-square basis</td>
<td></td>
</tr>
<tr>
<td>lrtest</td>
<td>Likelihood ratio test for two nested models</td>
<td></td>
</tr>
<tr>
<td>univarLR</td>
<td>Compute all univariable LR $\chi^2$</td>
<td></td>
</tr>
<tr>
<td>robcov</td>
<td>Robust covariance matrix estimates</td>
<td></td>
</tr>
<tr>
<td>bootcov</td>
<td>Bootstrap covariance matrix estimates</td>
<td></td>
</tr>
<tr>
<td>effective.df</td>
<td>Print effective d.f. for each type of variable</td>
<td></td>
</tr>
<tr>
<td>summary</td>
<td>Summary of effects of predictors</td>
<td></td>
</tr>
<tr>
<td>plot.summary</td>
<td>Plot continuously shaded confidence bars</td>
<td></td>
</tr>
<tr>
<td>anova</td>
<td>Wald tests of most meaningful hypotheses</td>
<td></td>
</tr>
<tr>
<td>plot.anova</td>
<td>Graphical depiction of anova</td>
<td></td>
</tr>
<tr>
<td>contrast</td>
<td>General contrasts, C.L., tests</td>
<td></td>
</tr>
<tr>
<td>gendata</td>
<td>Easily generate predictor combinations</td>
<td></td>
</tr>
<tr>
<td>predict</td>
<td>Obtain predicted values or design matrix</td>
<td></td>
</tr>
<tr>
<td>Predict</td>
<td>Obtain predicted values and confidence limits easily</td>
<td></td>
</tr>
<tr>
<td>plot.Predict</td>
<td>Plot the result of Predict using lattice</td>
<td></td>
</tr>
<tr>
<td>ggplot.Predict</td>
<td>Plot the result of Predict using ggplot2</td>
<td></td>
</tr>
<tr>
<td>plotp.Predict</td>
<td>Plot the result of Predict using plotly</td>
<td></td>
</tr>
<tr>
<td>fastbw</td>
<td>Fast backward step-down variable selection</td>
<td></td>
</tr>
<tr>
<td>residuals</td>
<td>(or resid) Residuals, influence stats from fit</td>
<td></td>
</tr>
<tr>
<td>sensuc</td>
<td>Sensitivity analysis for unmeasured confounder</td>
<td></td>
</tr>
<tr>
<td>which.influence</td>
<td>Which observations are overly influential</td>
<td>residuals</td>
</tr>
<tr>
<td>latex</td>
<td>( \text{\LaTeX} ) representation of fitted model</td>
<td></td>
</tr>
<tr>
<td>Function</td>
<td>Purpose</td>
<td>Related Functions</td>
</tr>
<tr>
<td>------------</td>
<td>----------------------------------------------</td>
<td>-------------------</td>
</tr>
<tr>
<td>Function</td>
<td>R function analytic representation of $X\beta$ from a fitted regression model</td>
<td>latex</td>
</tr>
<tr>
<td>Hazard</td>
<td>R function analytic representation of a fitted hazard function (for psm)</td>
<td></td>
</tr>
<tr>
<td>Survival</td>
<td>R function analytic representation of fitted survival function (for psm, cph)</td>
<td></td>
</tr>
<tr>
<td>Quantile</td>
<td>R function analytic representation of fitted function for quantiles of survival time (for psm, cph)</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>R function analytic representation of fitted function for mean survival time or for ordinal logistic nomogram</td>
<td>latex, plot</td>
</tr>
<tr>
<td>survest</td>
<td>Estimate survival probabilities (psm, cph)</td>
<td>survfit</td>
</tr>
<tr>
<td>survplot</td>
<td>Plot survival curves (psm, cph)</td>
<td>plot.survfit</td>
</tr>
<tr>
<td>survplotp</td>
<td>Plot survival curves with plotly features</td>
<td>survplot</td>
</tr>
<tr>
<td>validate</td>
<td>Validate indexes of model fit using resampling</td>
<td></td>
</tr>
<tr>
<td>val.prob</td>
<td>External validation of a probability model</td>
<td>lrm</td>
</tr>
<tr>
<td>val.surv</td>
<td>External validation of a survival model</td>
<td>calibrate</td>
</tr>
<tr>
<td>calibrate</td>
<td>Estimate calibration curve using resampling</td>
<td>val.prob</td>
</tr>
<tr>
<td>vif</td>
<td>Variance inflation factors for fitted model</td>
<td></td>
</tr>
<tr>
<td>naresid</td>
<td>Bring elements corresponding to missing data back into predictions and residuals</td>
<td></td>
</tr>
<tr>
<td>naprint</td>
<td>Print summary of missing values</td>
<td></td>
</tr>
<tr>
<td>impute</td>
<td>Impute missing values</td>
<td>aregImpute</td>
</tr>
</tbody>
</table>

**Functions related to Bayesian regression models:**

<table>
<thead>
<tr>
<th>Function</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>stackMI</td>
<td>Bayesian posterior stacking for multiple imputation</td>
</tr>
<tr>
<td>stanDx</td>
<td>Stan diagnostics on fit</td>
</tr>
<tr>
<td>stanDxplot</td>
<td>Trace plots to check posterior sampling convergence</td>
</tr>
<tr>
<td>PostF</td>
<td>Creates R function for computing posterior probabilities</td>
</tr>
<tr>
<td>plot.rmsb</td>
<td>Plot posterior densities, intervals, point summaries</td>
</tr>
<tr>
<td>compareBmods</td>
<td>Compare two models using LOO-cv</td>
</tr>
<tr>
<td>HPDint</td>
<td>Compute highest posterior density interval</td>
</tr>
<tr>
<td>distSym</td>
<td>Compute measure of symmetry of posterior distribution</td>
</tr>
</tbody>
</table>

An extensive overview of Bayesian capabilities of the `rms` package may be found at [hbiostat.org/R/rms/blrm.html](https://hbiostat.org/R/rms/blrm.html).

Global options `prType` and `grType` control printed and some graphical output, respectively as shown in example code below. The default is plain output and static graphics. If using `plotly` interactive graphics through `ggplot` or `plotp` or with `anova` or `summary` functions it is best to do so with RStudio `html` output or `html` notebooks. If using `html` output you must be producing
an html document or notebook. When setting `grType` to use \LaTeX{} or html it is highly recommended that you use the knitr package.

Example:

- **treat**: categorical variable with levels "a", "b", "c"

- **num.diseases**: ordinal variable, 0-4

- **age**: continuous
  Restricted cubic spline

- **cholesterol**: continuous
  (3 missings; use median)
  \( \log(\text{cholesterol}+10) \)

- **Allow treat \times cholesterol interaction**

- **Program to fit logistic model, test all effects in design, estimate effects (e.g. inter-quartile range odds ratios), plot estimated transformations**

```r
require(rms) # make new functions available
options(prType='latex') # print, summary, anova LaTeX output
options(grType='plotly') # plotly graphics for ggplot, anova, summary # default is 'base' for static graphics
ddist ← datadist(cholesterol, treat, num.diseases, age) # Could have used ddist ← datadist(data.frame.name)
options(datadist="ddist") # defines data dist. to rms
cholesterol ← impute(cholesterol)
fit ← lrm(y ~ treat + scored(num.diseases) + rcs(age) +
          log(cholesterol+10) + treat:log(cholesterol+10))
fit # outputs plain, LaTeX, or html markup
describe(y ~ treat + scored(num.diseases) + rcs(age))
# or use describe(formula(fit)) for all variables used in fit
```
# describe function (in Hmisc) gets simple statistics on variables
# fit ← robcov(fit)  # Would make all statistics that follow
# use a robust covariance matrix
# would need x=T, y=T in lrm()
specs(fit)  # Describe the design characteristics
anova(fit)  # Test these 2 by themselves
anova(fit, treat, cholesterol)  # Summarize anova graphically
plot(anova(fit))  # Print plain, LaTex, or html
summary(fit)  # Estimate effects using default ranges
plot(summary(fit))  # Graphical display of effects with C.I.
summary(fit, treat="b", age=60)  # Specify reference cell and adjustment val
summary(fit, age=c(50,70))  # Estimate effect of increasing age from 50 to 70
summary(fit, age=c(50,60,70))  # Increase age from 50 to 70, adjust to 60 when estimating effects of other factors

# If had not defined datadist, would have to define ranges for all var.

# Estimate and test treatment (b-a) effect averaged over 3 cholesterols
contrast(fit, list(treat='b', cholesterol=c(150,200,250)),
        list(treat='a', cholesterol=c(150,200,250)),
type='average')  # See the help file for contrast.rms for several examples of
# how to obtain joint tests of multiple contrasts and how to get
# double differences (interaction contrasts)

p ← predict(fit, age=seq(20,80,length=100), treat, conf.int=FALSE)  # Plot relationship between age and log
plot(p)  # or ggplot(p), plot(p)  # odds, separate curve for each treat,
plot(p, ~ age | treat)  # no C.I.
ggplot(p, groups=FALSE)  # Same but 2 panels
bplot(Predict(fit, age, cholesterol, np=50))  # 3-dimensional perspective plot for age,
# cholesterol, and log odds using default
# ranges for both variables
plot(Predict(fit, num.diseases, fun=function(x) 1/(1+exp(-x)), conf.int=.9),
ylab="Prob")  # Plot estimated probabilities instead of
# log odds (or use ggplot())
# can also use plotp() for plotly

# Again, if no datadist were defined, would have to tell plot all limits
logit ← predict(fit, expand.grid(treat="b", num.dis=1:3, age=c(20,40,60),
        cholesterol=seq(100,300,length=10)))  # Could also obtain list of predictor settings interactively}
logit ← predict(fit, gendata(fit, nobs=12))  # Since age doesn't interact with anything, we can quickly and
# interactively try various transformations of age, taking the spline
# function of age as the gold standard. We are seeking a linearizing
# transformation.

ag ← 10:80
logit ← predict(fit, expand.grid(treat="a", num.dis=0, age=ag,
        cholesterol=median(cholesterol)), type="terms")[,"age"]  # Note: if age interacted with anything, this would be the age
# "main effect" ignoring interaction terms
To examine interactions in a simpler way, you may want to group age into tertiles:

```r
age.tertile <- cut2(age, g=3) # For automatic ranges later, add age.tertile to datadist input
fit <- lrm(y ~ age.tertile * rcs(cholesterol))
```
6.4 Other Functions

- supsmu: Friedman’s “super smoother”
- lowess: Cleveland’s scatterplot smoother
- glm: generalized linear models (see Glm)
- gam: Generalized additive models
- rpart: Like original CART with surrogate splits for missings, censored data extension (Atkinson & Therneau)
- validate.rpart: in rms; validates recursive partitioning with respect to certain accuracy indexes
- loess: multi-dimensional scatterplot smoother

```r
f <- loess(y ~ age * pressure)
plot(f)  # cross-sectional plots
ages <- seq(20,70,length=40)
pressures <- seq(80,200,length=40)
pred <- predict(f, expand.grid(age=ages, pressure=pressures))
persp(ages, pressures, pred)  # 3-d plot
```
Chapter 7

Modeling Longitudinal Responses using Generalized Least Squares

7.1 Notation

• $N$ subjects

• Subject $i$ ($i = 1, 2, \ldots, N$) has $n_i$ responses measured at times $t_{i1}, t_{i2}, \ldots, t_{in_i}$

• Response at time $t$ for subject $i$: $Y_{it}$

• Subject $i$ has baseline covariates $X_i$

• Generally the response measured at time $t_{i1} = 0$ is a covariate in $X_i$ instead of being the first measured response $Y_{i0}$

• Time trend in response is modeled with $k$ parameters so that the time “main effect” has $k$ d.f.
- Let the basis functions modeling the time effect be $g_1(t), g_2(t), \ldots, g_k(t)$. 
7.2 Model Specification for Effects on $E(Y)$

7.2.1 Common Basis Functions

- $k$: dummy variables for $k + 1$ unique times (assumes no functional form for time but may spend many d.f.)
- $k = 1$ for linear time trend, $g_1(t) = t$
- $k$–order polynomial in $t$
- $k + 1$–knot restricted cubic spline (one linear term, $k - 1$ nonlinear terms)

7.2.2 Model for Mean Profile

- A model for mean time-response profile without interactions between time and any $X$:
  $E[Y_{it}|X_i] = X_i\beta + \gamma_1 g_1(t) + \gamma_2 g_2(t) + \ldots + \gamma_k g_k(t)$
- Model with interactions between time and some $X$’s: add product terms for desired interaction effects
- Example: To allow the mean time trend for subjects in group 1 (reference group) to be arbitrarily different from time trend
for subjects in group 2, have a dummy variable for group 2, a time “main effect” curve with \( k \) d.f. and all \( k \) products of these time components with the dummy variable for group 2

## 7.2.3 Model Specification for Treatment Comparisons

- In studies comparing two or more treatments, a response is often measured at baseline (pre-randomization).

- Analyst has the option to use this measurement as \( Y_{i0} \) or as part of \( X_i \).

- Jim Rochon (Rho, Inc., Chapel Hill NC) has the following comments about this:

  For RCTs, I draw a sharp line at the point when the intervention begins. The LHS [left hand side of the model equation] is reserved for something that is a response to treatment. Anything before this point can potentially be included as a covariate in the regression model. This includes the “baseline” value of the outcome variable. Indeed, the best predictor of the outcome at the end of the study is typically where the patient began at the beginning. It drinks up a lot of variability in the outcome; and, the effect of other covariates is typically mediated through this variable.

  I treat anything after the intervention begins as an outcome. In the western scientific method, an “effect” must follow the “cause” even if by a split second.

  Note that an RCT is different than a cohort study. In a cohort study, “Time 0” is not terribly meaningful. If we want to model, say, the trend over time, it would be legitimate, in my view, to include the “baseline” value on the LHS of that regression model.

  Now, even if the intervention, e.g., surgery, has an immediate effect, I would include still reserve the LHS for anything that might legitimately be considered as the response to the intervention. So, if we cleared a blocked artery and then measured the MABP, then that would still be included on the LHS.

  Now, it could well be that most of the therapeutic effect occurred by the time that the first repeated measure was taken, and then levels off. Then, a plot of the means would essentially be two parallel lines and the treatment effect is the distance between the lines, i.e., the difference in the intercepts.

  If the linear trend from baseline to Time 1 continues beyond Time 1, then the lines will have a common intercept but the slopes will diverge. Then, the treatment effect will the difference in slopes.

  One point to remember is that the estimated intercept is the value at time 0 that we predict from the set of repeated measures post randomization. In the first case above, the model will predict different intercepts even though randomization would suggest that they would start from the same place. This is because we were asleep at the switch and didn’t record the “action” from baseline to time 1. In the second case, the model will predict the same intercept values because the linear trend from baseline to time 1 was continued thereafter.

  More importantly, there are considerable benefits to including it as a covariate on the RHS. The baseline value tends to be the best predictor of the outcome post-randomization, and this maneuver increases the precision of
the estimated treatment effect. Additionally, any other prognostic factors correlated with the outcome variable will also be correlated with the baseline value of that outcome, and this has two important consequences. First, this greatly reduces the need to enter a large number of prognostic factors as covariates in the linear models. Their effect is already mediated through the baseline value of the outcome variable. Secondly, any imbalances across the treatment arms in important prognostic factors will induce an imbalance across the treatment arms in the baseline value of the outcome. Including the baseline value thereby reduces the need to enter these variables as covariates in the linear models.

Stephen Senn [187] states that temporally and logically, a “baseline cannot be a response to treatment”, so baseline and response cannot be modeled in an integrated framework.

...one should focus clearly on ‘outcomes’ as being the only values that can be influenced by treatment and examine critically any schemes that assume that these are linked in some rigid and deterministic view to ‘baseline’ values. An alternative tradition sees a baseline as being merely one of a number of measurements capable of improving predictions of outcomes and models it in this way.

The final reason that baseline cannot be modeled as the response at time zero is that many studies have inclusion/exclusion criteria that include cutoffs on the baseline variable. In other words, the baseline measurement comes from a truncated distribution. In general it is not appropriate to model the baseline with the same distributional shape as the follow-up measurements. Thus the approach recommended by Liang and Zeger [134] and Liu et al. [138] are problematic.

In addition to this, one of the paper’s conclusions that analysis of covariance is not appropriate if the population means of the baseline variable are not identical in the treatment groups is not correct [187]. See [115] for a rebuke of [138].
7.3 Modeling Within-Subject Dependence

- Random effects and mixed effects models have become very popular

- Disadvantages:
  - Induced correlation structure for $Y$ may be unrealistic
  - Numerically demanding
  - Require complex approximations for distributions of test statistics

- Extended linear model (with no random effects) is a logical extension of the univariate model (e.g., few statisticians use subject random effects for univariate $Y$)

- This was known as growth curve models and generalized least squares [169, 82] and was developed long before mixed effect models became popular

- Pinheiro and Bates (Section 5.1.2) state that "in some applications, one may wish to avoid incorporating random effects in the model to account for dependence among observations, choosing to use the within-group component $\Lambda_i$ to directly model variance-covariance structure of the response."
• We will assume that $Y_{it}|X_i$ has a multivariate normal distribution with mean given above and with variance-covariance matrix $V_i$, an $n_i \times n_i$ matrix that is a function of $t_{i1}, \ldots, t_{in_i}$.

• We further assume that the diagonals of $V_i$ are all equal.

• Procedure can be generalized to allow for heteroscedasticity over time or with respect to $X$ (e.g., males may be allowed to have a different variance than females).

• This extended linear model has the following assumptions:

  – all the assumptions of OLS at a single time point including correct modeling of predictor effects and univariate normality of responses conditional on $X$.

  – the distribution of two responses at two different times for the same subject, conditional on $X$, is bivariate normal with a specified correlation coefficient.

  – the joint distribution of all $n_i$ responses for the $i^{th}$ subject is multivariate normal with the given correlation pattern (which implies the previous two distributional assumptions).

  – responses from any times for any two different subjects are uncorrelated.
### What Methods To Use for Repeated Measurements / Serial Data?  

<table>
<thead>
<tr>
<th>Assumption</th>
<th>Repeated Measures ANOVA</th>
<th>GEE</th>
<th>Mixed Effects Model</th>
<th>GLS</th>
<th>LOCF</th>
<th>Summary Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assumes normality</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assumes independence of measurements within subject</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assumes a correlation structure</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| Requires same measurement times for all subjects | × | | | | | ?
| Does not allow smooth modeling of time to save d.f. | × | | | | | |
| Does not allow adjustment for baseline covariates | × | | | | | |
| Does not easily extend to non-continuous Y | × | | | | | |
| Loses information by not using intermediate measurements | | | | | | × |
| Does not allow widely varying # of observations per subject | × | × | × | | | × |
| Does not allow for subjects to have distinct trajectories | × | | | | | |
| Assumes subject-specific effects are Gaussian | | | | | | × |
| Badly biased if non-random dropouts | ? | × | × | | | |
| Biased in general | | | | | | |
| Harder to get tests & CLs | | | | | | × |
| Requires large # subjects/clusters | × | | | | | × |
| SEs are wrong | | | | | | | × |
| Assumptions are not verifiable in small samples | × | N/A | × | × | | × |
| Does not extend to complex settings such as time-dependent covariates and dynamic models | × | | | | | | ?

---

*a Thanks to Charles Berry, Brian Cade, Peter Flom, Bert Gunter, and Leena Choi for valuable input.

*b GEE: generalized estimating equations; GLS: generalized least squares; LOCF: last observation carried forward.

c E.g., compute within-subject slope, mean, or area under the curve over time. Assumes that the summary measure is an adequate summary of the time profile and assesses the relevant treatment effect.

d Unless one uses the Huynh-Feldt or Greenhouse-Geisser correction.

e For full efficiency, if using the working independence model.

f Or requires the user to specify one.

*g For full efficiency of regression coefficient estimates.

h Unless the last observation is missing.

i Or uses population averages.

j E.g., a model with a predictor that is a lagged value of the response variable.

k The cluster sandwich variance estimator used to estimate SEs in GEE does not perform well in this situation, and neither does the working independence model because it does not weight subjects properly.

l Unless one knows how to properly do a weighted analysis.

m Or uses high-dimensional integration to marginalize random effects, using complex approximations, and if using SAS, unintuitive d.f. for the various tests.

n Because there is no correct formula for SE of effects; ordinary SEs are not penalized for imputation and are too small.

o If correction not applied.
Gardiner et al. [78] compared several longitudinal data models, especially with regard to assumptions and how regression coefficients are estimated. Peters et al. [165] have an empirical study confirming that the “use all available data” approach of likelihood–based longitudinal models makes imputation of follow-up measurements unnecessary.
7.4 Parameter Estimation Procedure

- Generalized least squares

- Like weighted least squares but uses a covariance matrix that is not diagonal

- Each subject can have her own shape of $V_i$ due to each subject being measured at a different set of times

- Maximum likelihood

- Newton-Raphson or other trial-and-error methods used for estimating parameters

- For small number of subjects, advantages in using REML (restricted maximum likelihood) instead of ordinary MLE [60, Section 5.3], [168, Chapter 5], [82] (esp. to get more unbiased estimate of the covariance matrix)

- When imbalances are not severe, OLS fitted ignoring subject identifiers may be efficient
  
  - But OLS standard errors will be too small as they don’t take intra-cluster correlation into account

  - May be rectified by substituting covariance matrix estimated from Huber-White cluster sandwich estimator or
from cluster bootstrap

- When imbalances are severe and intra-subject correlations are strong, OLS is not expected to be efficient because it gives equal weight to each observation
  - a subject contributing two distant observations receives $\frac{1}{5}$ the weight of a subject having 10 tightly-spaced observations
7.5 Common Correlation Structures

- Usually restrict ourselves to *isotropic* correlation structures — correlation between responses within subject at two times depends only on a measure of distance between the two times, not the individual times.

- We simplify further and assume depends on $|t_1 - t_2|$.

- Can speak interchangeably of correlations of residuals within subjects or correlations between responses measured at different times on the same subject, conditional on covariates $X$.

- Assume that the correlation coefficient for $Y_{it_1}$ vs. $Y_{it_2}$ conditional on baseline covariates $X_i$ for subject $i$ is $h(|t_1 - t_2|, \rho)$, where $\rho$ is a vector (usually a scalar) set of fundamental correlation parameters.

- Some commonly used structures when times are continuous and are not equally spaced [168, Section 5.3.3] (nlme correlation function names are at the right if the structure is implemented in nlme):

  **Compound symmetry**: $h = \rho$ if $t_1 \neq t_2$, 1 if $t_1 = t_2$ (Essentially what two-way ANOVA assumes)  
  
  **Autoregressive-moving average lag 1**: $h = \rho^{(|t_1 - t_2|)} = \rho^s$
  
  where $s = |t_1 - t_2|$

  **Exponential**: $h = \exp(-s/\rho)$

  **Gaussian**: $h = \exp[-(s/\rho)^2]$

  **Linear**: $h = (1 - s/\rho)[s < \rho]$
Rational quadratic: \( h = 1 - \frac{(s/\rho)^2}{1 + (s/\rho)^2} \)  
Spherical: \( h = [1 - 1.5(s/\rho) + 0.5(s/\rho)^3][s < \rho] \)

Linear exponent AR(1): \( h = \rho^{d_{\min} + \delta d_{\min} \frac{s-d_{\min}}{d_{\max}-d_{\min}}} \), 1 if \( t_1 = t_2 \) [190]

The structures 3–7 use \( \rho \) as a scaling parameter, not as something restricted to be in \([0, 1]\).
7.6 Checking Model Fit

- Constant variance assumption: usual residual plots
- Normality assumption: usual qq residual plots
- Correlation pattern: **Variogram**
  - Estimate correlations of all possible pairs of residuals at different time points
  - Pool all estimates at same absolute difference in time \( s \)
  - Variogram is a plot with \( y = 1 - \hat{h}(s, \rho) \) vs. \( s \) on the \( x \)-axis
  - Superimpose the theoretical variogram assumed by the model
7.7 R Software

- Nonlinear mixed effects model package of Pinheiro & Bates

- For linear models, fitting functions are
  - `lme` for mixed effects models
  - `gls` for generalized least squares without random effects

- For this version the rms package has `Gls` so that many features of rms can be used:
  `anova`: all partial Wald tests, test of linearity, pooled tests
  `summary`: effect estimates (differences in \( \hat{Y} \)) and confidence limits, can be plotted
  `plot`, `ggplot`, `plotp`: continuous effect plots
  `nomogram`: nomogram
  `Function`: generate R function code for fitted model
  `latex`: \LaTeX{} representation of fitted model

In addition, `Gls` has a bootstrap option (hence you do not use `rms`'s `bootcov` for `Gls` fits).
To get regular `gls` functions named `anova` (for likelihood ratio tests, AIC, etc.) or `summary` use `anova.gls` or `summary.gls`

- `nlme` package has many graphics and fit-checking functions

- Several functions will be demonstrated in the case study
Consider the dataset in Table 6.9 of Davis [57, pp. 161-163] from a multicenter, randomized controlled trial of botulinum toxin type B (BotB) in patients with cervical dystonia from nine U.S. sites.

- Randomized to placebo \((N = 36)\), 5000 units of BotB \((N = 36)\), 10,000 units of BotB \((N = 37)\)

- Response variable: total score on Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS), measuring severity, pain, and disability of cervical dystonia (high scores mean more impairment)

- TWSTRS measured at baseline (week 0) and weeks 2, 4, 8, 12, 16 after treatment began

- Dataset cdystonia from web site

### 7.8.1 Graphical Exploration of Data

```r
require(rms)

options(prType='latex')  # for model print, summary, anova
gedata(cdystonia)
attach(cdystonia)

# Construct unique subject ID
uid <- with(cdystonia, factor(paste(site, id)))
```
# Tabulate patterns of subjects’ time points
```r
table(tapply(week, uid,
    function(w) paste(sort(unique(w)), collapse=' ')))
```

# Plot raw data, superposing subjects
```r
xl ← xlab('Week'); yl ← ylab('TWSTRS-total score')
ggplot(cdystonia, aes(x=week, y=twstrs, color=factor(id))) +
geom_line() + xl + yl + facet_grid(treat ~ site) +
guides(color=FALSE) # Fig. 7.1
```

---

**Figure 7.1**: Time profiles for individual subjects, stratified by study site and dose

---

# Show quartiles
```r
require(data.table)
```
```r
cdystonia ← data.table(cdystonia)
cdys ← cdystonia[, j=as.list(quantile(twstrs, (1 : 3)/4)),
    by = list(treat, week)]
cdys ← upData(cdys, rename=c('25%'='Q1', '50%'='Q2', '75%'='Q3'), print=FALSE)
ggplot(cdys, aes(x=week, y=Q2)) + x1 + yl + ylim(0, 70) +
```
geom_line() + facet_wrap(~ treat, nrow=2) +
geom_ribbon(aes(ymin=Q1, ymax=Q3), alpha=0.2)  # Fig. 7.2

Figure 7.2: Quartiles of TWSTRS stratified by dose

# Show means with bootstrap nonparametric CLs

cdys ← cdystonia[, j=as.list(smean.cl.boot(twstrs)),
by=list(treat, week)]

ggplot(cdys, aes(x=week, y=Mean)) + x1 + y1 + ylim(0, 70) +
geom_line() + facet_wrap(~ treat, nrow=2) +
geom_ribbon(aes(x=week, ymin=Lower, ymax=Upper), alpha=0.2)  # Fig. 7.3
Figure 7.3: Mean responses and nonparametric bootstrap 0.95 confidence limits for population means, stratified by dose
Model with $Y_{i0}$ as Baseline Covariate

```r
baseline ← subset(data.frame(cdystonia, uid), week == 0, -week)
baseline ← upData(baseline, rename=c(twstrs='twstrs0'), print=FALSE)
followup ← subset(data.frame(cdystonia, uid), week > 0, c(uid, week, twstrs))
rm(uid)
both ← merge(baseline, followup, by='uid')

dd ← datadist(both)
options(datadist='dd')
```

### 7.8.2 Using Generalized Least Squares

We stay with baseline adjustment and use a variety of correlation structures, with constant variance. Time is modeled as a restricted cubic spline with 3 knots, because there are only 3 unique interior values of `week`.

```r
require(nlme)

cp ← list(corCAR1, corExp, corCompSymm, corLin, corGaus, corSpher)
z ← vector('list', length(cp))
for(k in 1:length(cp)) {
  z[[k]] ← gls(twstrs ~ treat * rcs(week, 3) + rcs(twstrs0, 3) + rcs(age, 4) * sex, data=both, correlation=cp[[k]](form = ~week | uid))
}

anova(z[[1]], z[[2]], z[[3]], z[[4]], z[[5]], z[[6]])
```

<table>
<thead>
<tr>
<th>Model</th>
<th>df</th>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
</tr>
</thead>
<tbody>
<tr>
<td>z[[1]]</td>
<td>1</td>
<td>3553.906</td>
<td>3638.357</td>
<td>-1756.953</td>
</tr>
<tr>
<td>z[[2]]</td>
<td>2</td>
<td>3553.906</td>
<td>3638.357</td>
<td>-1756.953</td>
</tr>
<tr>
<td>z[[3]]</td>
<td>3</td>
<td>3587.974</td>
<td>3672.426</td>
<td>-1773.987</td>
</tr>
<tr>
<td>z[[4]]</td>
<td>4</td>
<td>3575.079</td>
<td>3659.531</td>
<td>-1767.540</td>
</tr>
<tr>
<td>z[[5]]</td>
<td>5</td>
<td>3621.081</td>
<td>3705.532</td>
<td>-1790.540</td>
</tr>
<tr>
<td>z[[6]]</td>
<td>6</td>
<td>3570.958</td>
<td>3655.409</td>
<td>-1765.479</td>
</tr>
</tbody>
</table>

AIC computed above is set up so that smaller values are best. From this the continuous-time AR1 and exponential structures
are tied for the best. For the remainder of the analysis use corCAR1, using Gls.

\[
a \leftarrow \text{Gls}(\text{twstrs} \sim \text{treat} \times \text{rcs(week, 3)} + \text{rcs(twstrs0, 3)} + \text{rcs(age, 4)} \times \text{sex}, \text{data=both, correlation=corCAR1(form=~week | uid)})
\]

Generalized Least Squares Fit by REML

\[
\text{Gls}(\text{model} = \text{twstrs} \sim \text{treat} \times \text{rcs(week, 3)} + \text{rcs(twstrs0, 3)} + \text{rcs(age, 4)} \times \text{sex}, \text{data=both, correlation=corCAR1(form=~week | uid)})
\]

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Log-restricted-likelihood</th>
<th>Clusters</th>
<th>Model d.f.</th>
<th>g</th>
<th>σ</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>522</td>
<td>-1756.95</td>
<td>108</td>
<td>17</td>
<td>11.334</td>
<td>8.5917</td>
<td>504</td>
</tr>
</tbody>
</table>

|       | \(\hat{\beta}\) | S.E. | \(t\) | \(Pr(>|t|)\) |
|-------|------------------|------|-------|-----------------|
| Intercept    | -0.3093          | 11.8804 | -0.03 | 0.9792          |
| treat=5000U  | 0.4344           | 2.5962 | 0.17  | 0.8672          |
| treat=Placebo| 7.1433           | 2.6133 | 2.73  | 0.0065          |
| week          | 0.2879           | 0.2973 | 0.97  | 0.3334          |
| week'         | 0.7313           | 0.3078 | 2.38  | 0.0179          |
| twstrs0       | 0.8071           | 0.1449 | 5.57  | <0.0001         |
| twstrs0'      | 0.2129           | 0.1795 | 1.19  | 0.2360          |
| age           | 0.1178           | 0.2346 | -0.50 | 0.6158          |
| age'          | 0.6968           | 0.6484 | 1.07  | 0.2830          |
| age''         | -3.4018          | 2.5599 | -1.33 | 0.1845          |
| sex=M         | 24.2802          | 18.6208 | 1.30  | 0.1929          |
| treat=5000U \times week | 0.0745       | 0.4221 | 0.18  | 0.8599          |
| treat=Placebo \times week | -0.1256       | 0.4243 | -0.30 | 0.7674          |
| treat=5000U \times week' | -0.4389       | 0.4363 | -1.01 | 0.3149          |
| treat=Placebo \times week' | -0.6459       | 0.4381 | -1.47 | 0.1411          |
| age \times sex=M | -0.5846       | 0.4447 | -1.31 | 0.1892          |
| age' \times sex=M | 1.4652        | 1.2388 | 1.18  | 0.2375          |
| age'' \times sex=M | -4.0338       | 4.8123 | -0.84 | 0.4023          |

Correlation Structure: Continuous AR(1)
Formula: ~week | uid
Parameter estimate(s):
  Phi
0.8666689
\[ \hat{\rho} = 0.8672, \] the estimate of the correlation between two measurements taken one week apart on the same subject. The estimated correlation for measurements 10 weeks apart is \[ 0.8672^{10} = 0.24. \]

\[ v \leftarrow \text{Variogram}(a, \text{form} = \sim \text{week} | \text{uid}) \]
\[ \text{plot}(v) \] \# Figure 7.4

Figure 7.4: Variogram, with assumed correlation pattern superimposed

Check constant variance and normality assumptions:

\[ \text{anova}(a) \]
Figure 7.5: Three residual plots to check for absence of trends in central tendency and in variability. Upper right panel shows the baseline score on the $x$-axis. Bottom left panel shows the mean $\pm 2 \times$ SD. Bottom right panel is the QQ plot for checking normality of residuals from the GLS fit.
Wald Statistics for twstrs

<table>
<thead>
<tr>
<th></th>
<th>( \chi^2 )</th>
<th>d.f.</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>treat (Factor+Higher Order Factors)</td>
<td>22.11</td>
<td>6</td>
<td>0.0012</td>
</tr>
<tr>
<td>All Interactions</td>
<td>14.94</td>
<td>4</td>
<td>0.0048</td>
</tr>
<tr>
<td>week (Factor+Higher Order Factors)</td>
<td>77.27</td>
<td>6</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>14.94</td>
<td>4</td>
<td>0.0048</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>6.61</td>
<td>3</td>
<td>0.0852</td>
</tr>
<tr>
<td>twstrs0</td>
<td>233.83</td>
<td>2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>1.41</td>
<td>1</td>
<td>0.2354</td>
</tr>
<tr>
<td>age (Factor+Higher Order Factors)</td>
<td>9.68</td>
<td>6</td>
<td>0.1388</td>
</tr>
<tr>
<td>All Interactions</td>
<td>4.86</td>
<td>3</td>
<td>0.1826</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>7.59</td>
<td>4</td>
<td>0.1077</td>
</tr>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>5.67</td>
<td>4</td>
<td>0.2252</td>
</tr>
<tr>
<td>All Interactions</td>
<td>4.86</td>
<td>3</td>
<td>0.1826</td>
</tr>
<tr>
<td>treat × week (Factor+Higher Order Factors)</td>
<td>14.94</td>
<td>4</td>
<td>0.0048</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>2.27</td>
<td>2</td>
<td>0.3208</td>
</tr>
<tr>
<td>Nonlinear Interaction : ( f(A,B) ) vs. ( AB )</td>
<td>2.27</td>
<td>2</td>
<td>0.3208</td>
</tr>
<tr>
<td>age × sex (Factor+Higher Order Factors)</td>
<td>4.86</td>
<td>3</td>
<td>0.1826</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>3.76</td>
<td>2</td>
<td>0.1526</td>
</tr>
<tr>
<td>Nonlinear Interaction : ( f(A,B) ) vs. ( AB )</td>
<td>3.76</td>
<td>2</td>
<td>0.1526</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>15.03</td>
<td>8</td>
<td>0.0586</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>19.75</td>
<td>7</td>
<td>0.0061</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>28.54</td>
<td>11</td>
<td>0.0027</td>
</tr>
<tr>
<td>TOTAL</td>
<td>322.98</td>
<td>17</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

plot(anova(a))             # Figure 7.6

ylm ← ylim(25, 60)
p1 ← ggplot(Predict(a, week, treat, conf.int=FALSE),
            adj.subtitle=FALSE, legend.position='top') + ylm
p2 ← ggplot(Predict(a, twstrs0), adj.subtitle=FALSE) + ylm
p3 ← ggplot(Predict(a, age, sex), adj.subtitle=FALSE,
             legend.position='top') + ylm
gridExtra::grid.arrange(p1, p2, p3, ncol=2)     # Figure 7.7

summary(a)                  # Shows for week 8

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>High</th>
<th>( \Delta )</th>
<th>Effect</th>
<th>S.E.</th>
<th>Lower 0.95</th>
<th>Upper 0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>week</td>
<td>4</td>
<td>12</td>
<td>8</td>
<td>6.69100</td>
<td>1.10570</td>
<td>4.5238</td>
<td>8.8582</td>
</tr>
<tr>
<td>twstrs0</td>
<td>39</td>
<td>53</td>
<td>14</td>
<td>13.55100</td>
<td>0.88618</td>
<td>11.8140</td>
<td>15.2880</td>
</tr>
<tr>
<td>age</td>
<td>46</td>
<td>65</td>
<td>19</td>
<td>2.50270</td>
<td>2.05140</td>
<td>-1.5179</td>
<td>6.5234</td>
</tr>
<tr>
<td>treat — 5000U:10000U</td>
<td>1</td>
<td>2</td>
<td></td>
<td>0.59167</td>
<td>1.99830</td>
<td>-3.3249</td>
<td>4.5083</td>
</tr>
<tr>
<td>treat — Placebo:10000U</td>
<td>1</td>
<td>3</td>
<td></td>
<td>5.49300</td>
<td>2.00430</td>
<td>1.5647</td>
<td>9.4212</td>
</tr>
<tr>
<td>sex — M:F</td>
<td>1</td>
<td>2</td>
<td></td>
<td>-1.08500</td>
<td>1.77860</td>
<td>-4.5711</td>
<td>2.4011</td>
</tr>
</tbody>
</table>
Figure 7.6: Results of \texttt{anova.rms} from generalized least squares fit with continuous time AR1 correlation structure.

Figure 7.7: Estimated effects of time, baseline TWSTRS, age, and sex.
# To get results for week 8 for a different reference group
# for treatment, use e.g. summary(a, week=4, treat='Placebo')

# Compare low dose with placebo, separately at each time
k1 ← contrast(a, list(week=c(2,4,8,12,16), treat='5000 U'),
              list(week=c(2,4,8,12,16), treat='Placebo'))
options(width=80)
print(k1, digits=3)

week twstrs0 age sex Contrast S.E. Lower Upper Z Pr(>|z|)
1 2 46 56 F -6.31 2.10 -10.43 -2.186 -3.00 0.0027
2 4 46 56 F -5.91 1.82 -9.47 -2.349 -3.25 0.0011
3 8 46 56 F -4.90 2.01 -8.85 -0.953 -2.43 0.0150
4* 12 46 56 F -3.07 1.75 -6.49 0.361 -1.75 0.0795
5* 16 46 56 F -1.02 2.10 -5.14 3.092 -0.49 0.6260

Redundant contrasts are denoted by *
Confidence intervals are 0.95 individual intervals

# Compare high dose with placebo
k2 ← contrast(a, list(week=c(2,4,8,12,16), treat='10000 U'),
              list(week=c(2,4,8,12,16), treat='Placebo'))
print(k2, digits=3)

week twstrs0 age sex Contrast S.E. Lower Upper Z Pr(>|z|)
1 2 46 56 F -6.89 2.07 -10.96 -2.83 -3.32 0.0009
2 4 46 56 F -6.64 1.79 -10.15 -3.13 -3.70 0.0002
3 8 46 56 F -5.49 2.00 -9.42 -1.56 -2.74 0.0061
4* 12 46 56 F -1.76 1.74 -5.17 1.65 -1.01 0.3109
5* 16 46 56 F 2.62 2.09 -1.47 6.71 1.25 0.2099

Redundant contrasts are denoted by *
Confidence intervals are 0.95 individual intervals

k1 ← as.data.frame(k1[c('week', 'Contrast', 'Lower', 'Upper')])
p1 ← ggplot(k1, aes(x=week, y=Contrast)) + geom_point() +
     geom_line() + ylab('Low Dose - Placebo') +
     geom_errorbar(aes(ymin=Lower, ymax=Upper), width=0)
k2 ← as.data.frame(k2[c('week', 'Contrast', 'Lower', 'Upper')])
p2 ← ggplot(k2, aes(x=week, y=Contrast)) + geom_point() +
     geom_line() + ylab('High Dose - Placebo') +
     geom_errorbar(aes(ymin=Lower, ymax=Upper), width=0)
gridExtra::grid.arrange(p1, p2, ncol=2)  # Figure 7.8

Although multiple d.f. tests such as total treatment effects or
treatment × time interaction tests are comprehensive, their in-
creased degrees of freedom can dilute power. In a treatment
comparison, treatment contrasts at the last time point (single
d.f. tests) are often of major interest. Such contrasts are informed by all the measurements made by all subjects (up until dropout times) when a smooth time trend is assumed.

```r
n <- nomogram(a, age=c(seq(20, 80, by=10), 85))
plot(n, cex.axis=.55, cex.var=.8, lmgp=.25)  # Figure 7.9
```
Figure 7.9: Nomogram from GLS fit. Second axis is the baseline score.
7.8.3 Bayesian Proportional Odds Random Effects Model

- Develop a $y$-transformation invariant longitudinal model
- Proportional odds model with no grouping of TWSTRS scores
- Bayesian random effects model
- Random effects Gaussian with exponential prior distribution for its SD, with mean 1.0
- Compound symmetry correlation structure
- Demonstrates a large amount of patient-to-patient intercept variability

```r
options(stancompiled='~/R/stan', mc.cores=parallel::detectCores())
bpo ← blrm(twstrs ~ treat * rcs(week, 3) + rcs(twstrs0, 3) +
          rcs(age, 4) * sex + cluster(uid), data=both, file='bpo.rds')
# file = means that after the first time the model is run, it will not
# be re-run unless the data, fitting options, or underlying Stan code change
stanDx(bpo)
```

Iterations: 2000 on each of 4 chains, with 4000 posterior distribution samples saved

For each parameter, n_eff is a crude measure of effective sample size
and Rhat is the potential scale reduction factor on split chains
(at convergence, Rhat=1)

<table>
<thead>
<tr>
<th>$y \geq 7$</th>
<th>n_eff</th>
<th>Rhat</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y \geq 9$</td>
<td>2040</td>
<td>1.001</td>
</tr>
<tr>
<td>$y \geq 10$</td>
<td>1823</td>
<td>1.000</td>
</tr>
<tr>
<td>$y \geq 11$</td>
<td>1229</td>
<td>1.002</td>
</tr>
<tr>
<td>$y \geq 13$</td>
<td>1060</td>
<td>1.001</td>
</tr>
<tr>
<td>$y \geq 14$</td>
<td>993</td>
<td>1.001</td>
</tr>
<tr>
<td>$y \geq 15$</td>
<td>896</td>
<td>1.001</td>
</tr>
<tr>
<td>y &gt;= 16</td>
<td>801 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 17</td>
<td>730 1.000</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 18</td>
<td>687 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 19</td>
<td>638 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 20</td>
<td>611 1.002</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 21</td>
<td>593 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 22</td>
<td>582 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 23</td>
<td>563 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 24</td>
<td>549 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 25</td>
<td>524 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 26</td>
<td>499 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 27</td>
<td>486 1.002</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 28</td>
<td>474 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 29</td>
<td>474 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 30</td>
<td>466 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 31</td>
<td>460 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 32</td>
<td>457 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 33</td>
<td>461 1.002</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 34</td>
<td>466 1.001</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 35</td>
<td>456 1.002</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 36</td>
<td>460 1.002</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 37</td>
<td>463 1.002</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 38</td>
<td>474 1.003</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 39</td>
<td>481 1.002</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 40</td>
<td>494 1.002</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 41</td>
<td>494 1.003</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 42</td>
<td>514 1.003</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 43</td>
<td>521 1.003</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 44</td>
<td>550 1.004</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 45</td>
<td>564 1.004</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 46</td>
<td>580 1.004</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 47</td>
<td>614 1.004</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 48</td>
<td>638 1.004</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 49</td>
<td>669 1.003</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 50</td>
<td>704 1.005</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 51</td>
<td>819 1.005</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 52</td>
<td>858 1.005</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 53</td>
<td>928 1.005</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 54</td>
<td>1011 1.004</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 55</td>
<td>1063 1.004</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 56</td>
<td>1130 1.004</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 57</td>
<td>1197 1.003</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 58</td>
<td>1260 1.003</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 59</td>
<td>1394 1.003</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 60</td>
<td>1486 1.003</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 61</td>
<td>1495 1.005</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 62</td>
<td>1533 1.004</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 63</td>
<td>1604 1.005</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 64</td>
<td>1674 1.005</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 65</td>
<td>2167 1.004</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 66</td>
<td>2266 1.004</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 67</td>
<td>2412 1.003</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 68</td>
<td>2493 1.003</td>
<td></td>
</tr>
<tr>
<td>y &gt;= 71</td>
<td>2799 1.003</td>
<td></td>
</tr>
<tr>
<td>treat = 5000U</td>
<td>777 1.003</td>
<td></td>
</tr>
<tr>
<td>treat = Placebo</td>
<td>797 1.003</td>
<td></td>
</tr>
<tr>
<td>week</td>
<td>2286 1.001</td>
<td></td>
</tr>
</tbody>
</table>
Bayesian Logistic Regression Model

```r
blrm(formula = twstrs ~ treat * rcs(week, 3) + rcs(twstrs0, 3) +
      rcs(age, 4) * sex + cluster(uid), data = both, file = "bpo.rds")
```

Mixed Calibration/ Discrimination Indexes

<table>
<thead>
<tr>
<th>Obs</th>
<th>Mixed Calibration/ Discrimination Indexes</th>
<th>Discrimination Indexes</th>
<th>Rank Discrim. Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOO log L -1745.6±23.74</td>
<td>g: 3.841 [3.324, 4.3]</td>
<td>C: 0.793 [0.786, 0.799]</td>
</tr>
<tr>
<td></td>
<td>LOO IC 3491.2±47.48</td>
<td>g_p: 0.435 [0.42, 0.449]</td>
<td>D_{xy}: 0.585 [0.571, 0.599]</td>
</tr>
<tr>
<td></td>
<td>Effective p 178.38±7.9</td>
<td>EV: 0.593 [0.552, 0.644]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B: 0.149 [0.139, 0.16]</td>
<td>v: 11.484 [8.542, 14.317]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>vp: 0.148 [0.137, 0.16]</td>
<td></td>
</tr>
<tr>
<td>sigma</td>
<td>1.8879 [1.541, 2.2594]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>y&lt;7</th>
<th>Mean β</th>
<th>Median β</th>
<th>S.E.</th>
<th>Lower</th>
<th>Upper</th>
<th>Pr(β &gt; 0)</th>
<th>Symmetry</th>
</tr>
</thead>
<tbody>
<tr>
<td>y&lt;9</td>
<td>-1.8339</td>
<td>-1.8370</td>
<td>4.2976</td>
<td>-9.9353</td>
<td>7.1692</td>
<td>0.3265</td>
<td>1.01</td>
</tr>
<tr>
<td>y&lt;10</td>
<td>-2.8233</td>
<td>-2.8895</td>
<td>4.2096</td>
<td>-11.2088</td>
<td>5.3319</td>
<td>0.2502</td>
<td>1.02</td>
</tr>
<tr>
<td>y&lt;11</td>
<td>-3.9806</td>
<td>-4.0789</td>
<td>4.1573</td>
<td>-11.9557</td>
<td>4.2070</td>
<td>0.1705</td>
<td>1.03</td>
</tr>
<tr>
<td>y&lt;12</td>
<td>-4.4177</td>
<td>-4.5345</td>
<td>4.1537</td>
<td>-12.9680</td>
<td>3.2245</td>
<td>0.1440</td>
<td>1.03</td>
</tr>
<tr>
<td>y&lt;13</td>
<td>-4.6167</td>
<td>-4.7443</td>
<td>4.1482</td>
<td>-12.2323</td>
<td>3.9488</td>
<td>0.1328</td>
<td>1.03</td>
</tr>
<tr>
<td>y&lt;14</td>
<td>-4.9748</td>
<td>-5.0854</td>
<td>4.1427</td>
<td>-12.4926</td>
<td>3.5744</td>
<td>0.1172</td>
<td>1.04</td>
</tr>
<tr>
<td>y&lt;15</td>
<td>-5.2833</td>
<td>-5.3604</td>
<td>4.1391</td>
<td>-12.8419</td>
<td>3.1908</td>
<td>0.1022</td>
<td>1.03</td>
</tr>
<tr>
<td>y&lt;16</td>
<td>-5.6706</td>
<td>-5.7772</td>
<td>4.1339</td>
<td>-13.1391</td>
<td>2.9395</td>
<td>0.0860</td>
<td>1.04</td>
</tr>
<tr>
<td>y&lt;17</td>
<td>-6.4756</td>
<td>-6.5635</td>
<td>4.1274</td>
<td>-14.4598</td>
<td>1.5527</td>
<td>0.0578</td>
<td>1.03</td>
</tr>
<tr>
<td>y&lt;18</td>
<td>-6.7335</td>
<td>-6.8183</td>
<td>4.1278</td>
<td>-14.4950</td>
<td>1.4811</td>
<td>0.0520</td>
<td>1.03</td>
</tr>
<tr>
<td>y&lt;19</td>
<td>-7.0189</td>
<td>-7.1225</td>
<td>4.1307</td>
<td>-15.3432</td>
<td>0.6939</td>
<td>0.0442</td>
<td>1.04</td>
</tr>
<tr>
<td>y&lt;20</td>
<td>-7.2130</td>
<td>-7.3241</td>
<td>4.1321</td>
<td>-15.5414</td>
<td>0.4529</td>
<td>0.0398</td>
<td>1.03</td>
</tr>
<tr>
<td>y&lt;21</td>
<td>-7.3942</td>
<td>-7.4941</td>
<td>4.1312</td>
<td>-15.6961</td>
<td>0.3093</td>
<td>0.0362</td>
<td>1.04</td>
</tr>
<tr>
<td>y&lt;22</td>
<td>-7.8073</td>
<td>-7.9052</td>
<td>4.1320</td>
<td>-16.1170</td>
<td>-0.0241</td>
<td>0.297</td>
<td>1.03</td>
</tr>
<tr>
<td>y&lt;23</td>
<td>-8.0745</td>
<td>-8.1467</td>
<td>4.1318</td>
<td>-16.1530</td>
<td>-0.0594</td>
<td>0.2670</td>
<td>1.04</td>
</tr>
<tr>
<td>y&lt;24</td>
<td>-8.3592</td>
<td>-8.4095</td>
<td>4.1316</td>
<td>-16.5031</td>
<td>-0.3911</td>
<td>0.0235</td>
<td>1.04</td>
</tr>
<tr>
<td>y&lt;25</td>
<td>-8.6194</td>
<td>-8.6829</td>
<td>4.1315</td>
<td>-16.7479</td>
<td>-0.6811</td>
<td>0.0210</td>
<td>1.03</td>
</tr>
<tr>
<td>y&lt;26</td>
<td>-9.0128</td>
<td>-9.0882</td>
<td>4.1359</td>
<td>-17.0816</td>
<td>-0.9606</td>
<td>0.0170</td>
<td>1.04</td>
</tr>
<tr>
<td>y&lt;27</td>
<td>-9.3039</td>
<td>-9.3882</td>
<td>4.1376</td>
<td>-17.4863</td>
<td>-1.3867</td>
<td>0.0138</td>
<td>1.03</td>
</tr>
<tr>
<td>y&lt;28</td>
<td>-9.5464</td>
<td>-9.6180</td>
<td>4.1388</td>
<td>-17.6929</td>
<td>-1.5988</td>
<td>0.0112</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 29</td>
<td>Mean β</td>
<td>Median β</td>
<td>S.E.</td>
<td>Lower</td>
<td>Upper</td>
<td>Pr(β &gt; 0)</td>
<td>Symmetry</td>
</tr>
<tr>
<td>--------</td>
<td>---------</td>
<td>----------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>-----------</td>
<td>----------</td>
</tr>
<tr>
<td>-9.7813</td>
<td>-9.8544</td>
<td>4.1401</td>
<td>-17.9507</td>
<td>-1.8841</td>
<td>0.0102</td>
<td>1.03</td>
<td></td>
</tr>
<tr>
<td>y ≥ 30</td>
<td>-10.0834</td>
<td>-10.1615</td>
<td>4.1408</td>
<td>-18.3938</td>
<td>-2.3286</td>
<td>0.0085</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 31</td>
<td>-10.3770</td>
<td>-10.4721</td>
<td>4.1420</td>
<td>-18.5219</td>
<td>-2.5135</td>
<td>0.0070</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 32</td>
<td>-10.4931</td>
<td>-10.5894</td>
<td>4.1417</td>
<td>-18.6187</td>
<td>-2.5744</td>
<td>0.0068</td>
<td>1.05</td>
</tr>
<tr>
<td>y ≥ 33</td>
<td>-10.8594</td>
<td>-10.9362</td>
<td>4.1459</td>
<td>-18.9980</td>
<td>-2.9566</td>
<td>0.0053</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 34</td>
<td>-11.1697</td>
<td>-11.2474</td>
<td>4.1472</td>
<td>-19.3911</td>
<td>-3.2846</td>
<td>0.0045</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 36</td>
<td>-11.6397</td>
<td>-11.7070</td>
<td>4.1478</td>
<td>-19.5184</td>
<td>-3.4532</td>
<td>0.0025</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 37</td>
<td>-11.9155</td>
<td>-11.9935</td>
<td>4.1494</td>
<td>-19.9564</td>
<td>-3.8344</td>
<td>0.0022</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 38</td>
<td>-12.1432</td>
<td>-12.2174</td>
<td>4.1498</td>
<td>-20.1489</td>
<td>-4.0637</td>
<td>0.0018</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 39</td>
<td>-12.3892</td>
<td>-12.4709</td>
<td>4.1528</td>
<td>-20.4753</td>
<td>-4.3654</td>
<td>0.0010</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 40</td>
<td>-12.5750</td>
<td>-12.6619</td>
<td>4.1546</td>
<td>-20.7773</td>
<td>-4.6331</td>
<td>0.0010</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 41</td>
<td>-12.7589</td>
<td>-12.8548</td>
<td>4.1569</td>
<td>-20.7494</td>
<td>-4.6136</td>
<td>0.0010</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 43</td>
<td>-13.3138</td>
<td>-13.4103</td>
<td>4.1602</td>
<td>-21.4720</td>
<td>-5.2685</td>
<td>0.0008</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 44</td>
<td>-13.6518</td>
<td>-13.7490</td>
<td>4.1624</td>
<td>-21.7085</td>
<td>-5.4876</td>
<td>0.0008</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 45</td>
<td>-13.9704</td>
<td>-14.0880</td>
<td>4.1631</td>
<td>-21.9898</td>
<td>-5.8527</td>
<td>0.0008</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 46</td>
<td>-14.2720</td>
<td>-14.3627</td>
<td>4.1644</td>
<td>-22.3300</td>
<td>-6.1606</td>
<td>0.0008</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 47</td>
<td>-14.6911</td>
<td>-14.7788</td>
<td>4.1690</td>
<td>-22.7546</td>
<td>-6.5777</td>
<td>0.0003</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 48</td>
<td>-14.9815</td>
<td>-15.0652</td>
<td>4.1687</td>
<td>-23.0143</td>
<td>-6.8499</td>
<td>0.0003</td>
<td>1.04</td>
</tr>
</tbody>
</table>
### Table 7.1: Model Coefficients for y ≥ 49

<table>
<thead>
<tr>
<th></th>
<th>Mean $\hat{\beta}$</th>
<th>Median $\hat{\beta}$</th>
<th>S.E.</th>
<th>Lower</th>
<th>Upper</th>
<th>Pr($\hat{\beta} &gt; 0$)</th>
<th>Symmetry</th>
</tr>
</thead>
<tbody>
<tr>
<td>y ≥ 49</td>
<td>-15.3561</td>
<td>-15.4516</td>
<td>4.1709</td>
<td>-23.3881</td>
<td>-7.2513</td>
<td>0.0003</td>
<td>1.04</td>
</tr>
<tr>
<td>y ≥ 50</td>
<td>-15.6742</td>
<td>-15.7710</td>
<td>4.1720</td>
<td>-23.6612</td>
<td>-7.5099</td>
<td>0.0003</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 51</td>
<td>-16.2033</td>
<td>-16.2828</td>
<td>4.1750</td>
<td>-24.1261</td>
<td>-7.9737</td>
<td>0.0003</td>
<td>1.02</td>
</tr>
<tr>
<td>y ≥ 52</td>
<td>-16.5657</td>
<td>-16.6517</td>
<td>4.1780</td>
<td>-24.5407</td>
<td>-8.3475</td>
<td>0.0003</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 53</td>
<td>-17.0093</td>
<td>-17.0916</td>
<td>4.1801</td>
<td>-24.9680</td>
<td>-8.7594</td>
<td>0.0003</td>
<td>1.02</td>
</tr>
<tr>
<td>y ≥ 54</td>
<td>-17.5094</td>
<td>-17.5848</td>
<td>4.1814</td>
<td>-25.3941</td>
<td>-9.1859</td>
<td>0.0000</td>
<td>1.03</td>
</tr>
<tr>
<td>y ≥ 55</td>
<td>-17.9251</td>
<td>-18.0089</td>
<td>4.1890</td>
<td>-26.0294</td>
<td>-9.7594</td>
<td>0.0000</td>
<td>1.02</td>
</tr>
<tr>
<td>y ≥ 56</td>
<td>-18.1785</td>
<td>-18.2806</td>
<td>4.1931</td>
<td>-26.5999</td>
<td>-10.2886</td>
<td>0.0000</td>
<td>1.02</td>
</tr>
<tr>
<td>y ≥ 57</td>
<td>-18.6475</td>
<td>-18.7216</td>
<td>4.1954</td>
<td>-26.7549</td>
<td>-10.4311</td>
<td>0.0000</td>
<td>1.02</td>
</tr>
<tr>
<td>y ≥ 58</td>
<td>-19.2056</td>
<td>-19.2750</td>
<td>4.1977</td>
<td>-26.8995</td>
<td>-10.5137</td>
<td>0.0000</td>
<td>1.00</td>
</tr>
<tr>
<td>y ≥ 61</td>
<td>-20.5875</td>
<td>-20.6712</td>
<td>4.2131</td>
<td>-28.9822</td>
<td>-12.5520</td>
<td>0.0000</td>
<td>0.99</td>
</tr>
<tr>
<td>y ≥ 62</td>
<td>-20.9643</td>
<td>-21.0283</td>
<td>4.2251</td>
<td>-29.2904</td>
<td>-12.8757</td>
<td>0.0000</td>
<td>1.00</td>
</tr>
<tr>
<td>y ≥ 63</td>
<td>-21.3694</td>
<td>-21.4600</td>
<td>4.2380</td>
<td>-29.6267</td>
<td>-13.2462</td>
<td>0.0000</td>
<td>1.00</td>
</tr>
<tr>
<td>y ≥ 64</td>
<td>-21.5122</td>
<td>-21.6191</td>
<td>4.2311</td>
<td>-29.5315</td>
<td>-13.0426</td>
<td>0.0000</td>
<td>1.01</td>
</tr>
<tr>
<td>y ≥ 65</td>
<td>-22.3988</td>
<td>-22.3237</td>
<td>4.2425</td>
<td>-30.2928</td>
<td>-13.7575</td>
<td>0.0000</td>
<td>1.01</td>
</tr>
<tr>
<td>y ≥ 66</td>
<td>-22.6169</td>
<td>-22.7128</td>
<td>4.2512</td>
<td>-30.6187</td>
<td>-13.9764</td>
<td>0.0000</td>
<td>1.00</td>
</tr>
<tr>
<td>y ≥ 67</td>
<td>-23.0283</td>
<td>-23.0856</td>
<td>4.2563</td>
<td>-31.2660</td>
<td>-14.6301</td>
<td>0.0001</td>
<td>1.02</td>
</tr>
<tr>
<td>y ≥ 68</td>
<td>-23.7955</td>
<td>-23.8681</td>
<td>4.2835</td>
<td>-32.1459</td>
<td>-15.2872</td>
<td>0.0000</td>
<td>1.00</td>
</tr>
<tr>
<td>y ≥ 69</td>
<td>-24.6800</td>
<td>-24.7032</td>
<td>4.3020</td>
<td>-32.6177</td>
<td>-15.6389</td>
<td>0.0000</td>
<td>1.02</td>
</tr>
</tbody>
</table>

#### treat=5000U
- 0.1091 0.1179 0.7183 -1.2523 1.5316 0.5665 1.02

#### treat=Placebo
- 2.4094 2.4065 0.7540 0.8709 3.8075 0.9992 1.01

#### week
- 0.1230 0.1234 0.0798 -0.0298 0.2824 0.9428 1.03

#### week'
- 0.1907 0.1911 0.0876 0.0240 0.3636 0.9840 0.98

#### twstrs0
- 0.2289 0.2294 0.0504 0.1266 0.3217 1.0000 1.01

#### twstrs0'
- 0.1296 0.1288 0.0620 0.0060 0.2490 0.9845 1.02

#### age
- 0.0168 -0.0187 0.0821 -0.1792 0.1457 0.4080 1.03

#### age'
- 0.1947 0.2003 0.2239 -0.2424 0.6355 0.8065 0.96

#### age''
- 0.0408 0.0203 1.6624 -3.3156 3.1535 0.5030 1.02

<table>
<thead>
<tr>
<th></th>
<th>Mean $\hat{\beta}$</th>
<th>Median $\hat{\beta}$</th>
<th>S.E.</th>
<th>Lower</th>
<th>Upper</th>
<th>Pr($\hat{\beta} &gt; 0$)</th>
<th>Symmetry</th>
</tr>
</thead>
<tbody>
<tr>
<td>age x sex=M</td>
<td>4.7997</td>
<td>4.9317</td>
<td>6.5916</td>
<td>-8.6780</td>
<td>17.0193</td>
<td>0.7685</td>
<td>1.01</td>
</tr>
<tr>
<td>treat=5000U \times week</td>
<td>0.0511</td>
<td>0.0525</td>
<td>0.1089</td>
<td>-0.1526</td>
<td>0.2780</td>
<td>0.6905</td>
<td>1.02</td>
</tr>
<tr>
<td>treat=Placebo \times week</td>
<td>-0.0562</td>
<td>-0.0573</td>
<td>0.1123</td>
<td>-0.2719</td>
<td>0.1630</td>
<td>0.3050</td>
<td>0.99</td>
</tr>
<tr>
<td>treat=5000U \times week'</td>
<td>-0.1621</td>
<td>-0.1632</td>
<td>0.1189</td>
<td>-0.3999</td>
<td>0.0745</td>
<td>0.0848</td>
<td>1.04</td>
</tr>
<tr>
<td>treat=Placebo \times week'</td>
<td>-0.1375</td>
<td>-0.1363</td>
<td>0.1236</td>
<td>-0.3728</td>
<td>0.1066</td>
<td>0.1295</td>
<td>0.99</td>
</tr>
<tr>
<td>age x sex=M</td>
<td>-0.1061</td>
<td>-0.1107</td>
<td>0.1566</td>
<td>-0.4711</td>
<td>0.1954</td>
<td>0.2488</td>
<td>1.03</td>
</tr>
<tr>
<td>age' x sex=M</td>
<td>0.1499</td>
<td>0.1602</td>
<td>0.4295</td>
<td>-0.6931</td>
<td>0.9703</td>
<td>0.6425</td>
<td>0.97</td>
</tr>
<tr>
<td>age'' x sex=M</td>
<td>0.0408</td>
<td>0.0203</td>
<td>1.6624</td>
<td>-3.3156</td>
<td>3.1535</td>
<td>0.5030</td>
<td>1.02</td>
</tr>
</tbody>
</table>

a ← anova(bpo)
a

Relative Explained Variation for twstrs. Approximate total model Wald $\chi^2$ used in denominators of REV:254.4 [206.6, 325.7].
- Show the final graphic (high dose:placebo contrast as function of time)

- Intervals are 0.95 highest posterior density intervals

- \( y \)-axis: log-odds ratio
```r
wks <- c(2, 4, 8, 12, 16)
k <- contrast(bpo, list(week=wks, treat='10000 U'),
              list(week=wks, treat='Placebo'),
              cnames=paste('Week', wks))
k
```

<table>
<thead>
<tr>
<th>week</th>
<th>Contrast</th>
<th>S.E.</th>
<th>Lower</th>
<th>Upper</th>
<th>Pr(Contrast &gt;0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Week 2</td>
<td>-2.2969155</td>
<td>0.6271798</td>
<td>-3.587160</td>
<td>-1.1438202</td>
</tr>
<tr>
<td>2</td>
<td>Week 4</td>
<td>-2.1844434</td>
<td>0.5649108</td>
<td>-3.235373</td>
<td>-1.0540233</td>
</tr>
<tr>
<td>3</td>
<td>Week 8</td>
<td>-1.8219799</td>
<td>0.6177201</td>
<td>-2.981526</td>
<td>-0.5511903</td>
</tr>
<tr>
<td>4*</td>
<td>Week 12</td>
<td>-0.9094399</td>
<td>0.5610337</td>
<td>-1.957164</td>
<td>0.020536</td>
</tr>
<tr>
<td>5*</td>
<td>Week 16</td>
<td>0.1406193</td>
<td>0.6420881</td>
<td>-1.135811</td>
<td>1.3502284</td>
</tr>
</tbody>
</table>

Redundant contrasts are denoted by *

Intervals are 0.95 highest posterior density intervals
Contrast is the posterior mean

```r
plot(k)
```

```r
k <- as.data.frame(k[c('week', 'Contrast', 'Lower', 'Upper')])
ggplot(k, aes(x=week, y=Contrast)) + geom_point() + geom_line() + ylab('High Dose - Placebo') + geom_errorbar(aes(ymin=Lower, ymax=Upper), width=0)
```
For each posterior draw compute the difference in means and get an exact (to within simulation error) 0.95 highest posterior density intervals for these differences.

```r
M ← Mean(bpo) # create R function that computes mean Y from X*beta
k ← contrast(bpo, list(week=wks, treat='10000U'),
              list(week=wks, treat='Placebo'),
              fun=M, cnames=paste('Week', wks))
plot(k, which='diff') + theme(legend.position='bottom')
```

```r
f ← function(x) {
    hpd ← HPDint(x, prob=0.95) # is in rms
    ...}
```
```r
r ← c(mean(x), median(x), hpd)
names(r) ← c('Mean', 'Median', 'Lower', 'Upper')
r
w ← as.data.frame(t(apply(k$estaa - k$estab, 2, f)))
week ← as.numeric(sub('Week ', '', rownames(w)))

ggplot(w, aes(x=week, y=Mean)) + geom_point() +
geom_line() + ylab('High Dose - Placebo') +
geom_errorbar(aes(ymin=Lower, ymax=Upper), width=0) +
scale_y_continuous(breaks=c(-8, -4, 0, 4))
```

![Graph showing the longitudinal response of high dose compared to placebo over weeks.]
Chapter 8

**Case Study in Data Reduction**

See Chapter 8 in the text. Links to narrations are on the next page.
CHAPTER 8. CASE STUDY IN DATA REDUCTION

8.1 Data

8.2 How Many Parameters Can Be Estimated?

8.3 Redundancy Analysis

8.4 Variable Clustering

8.5 Transformation and Single Imputation Using transcan

8.6 Data Reduction Using Principal Components

Dotted blue line in Fig. 8.5 should be at 3958.
8.6.1 Sparse Principal Components

8.7 Transformation Using Nonparametric Smoothers
Chapter 9

Maximum Likelihood Estimation

See Chapter 9 in the book.
Chapter 10

Binary Logistic Regression

• $Y = 0, 1$

• Time of event not important

• $\ln C(Y|X) C$ is $\text{Prob}\{Y = 1\}$

• $g(u)$ is $\frac{1}{1+e^{-u}}$
10.1 Model

\[ \text{Prob}\{Y = 1|X\} = \left[1 + \exp(-X\beta)\right]^{-1}. \]
\[ P = \left[1 + \exp(-x)\right]^{-1} \]

\[ B = \frac{P}{1-P} \]
\[ O = \frac{O}{1+O} \]
\[ X_\beta = \log \frac{P}{1-P} \]
\[ e^{X_\beta} = O \]
Model Assumptions and Interpretation of Parameters

\[
\text{logit}\{Y = 1|X}\} = \logit(P) = \log[P/(1 - P)] = X\beta,
\]

- Increase \(X_j\) by \(d\) \(\rightarrow\) increase odds \(Y = 1\) by \(\exp(\beta_j d)\),
  increase log odds by \(\beta_j d\).

- If there is only one predictor \(X\) and that predictor is binary, the model can be written
  \[
  \begin{align*}
  \text{logit}\{Y = 1|X = 0\} & = \beta_0 \\
  \text{logit}\{Y = 1|X = 1\} & = \beta_0 + \beta_1.
  \end{align*}
  \]

- One continuous predictor:
  \[
  \text{logit}\{Y = 1|X\} = \beta_0 + \beta_1 X,
  \]

- Two treatments (indicated by \(X_1 = 0\) or \(1\)) and one continuous covariable \((X_2)\).
  \[
  \begin{align*}
  \text{logit}\{Y = 1|X\} & = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \\
  \text{logit}\{Y = 1|X_1 = 0, X_2\} & = \beta_0 + \beta_2 X_2 \\
  \text{logit}\{Y = 1|X_1 = 1, X_2\} & = \beta_0 + \beta_1 + \beta_2 X_2.
  \end{align*}
  \]
10.1.2 Odds Ratio, Risk Ratio, and Risk Difference

- Odds ratio capable of being constant

- Ex: risk factor doubles odds of disease

<table>
<thead>
<tr>
<th>Without Risk Factor</th>
<th>With Risk Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probability</td>
<td>Odds</td>
</tr>
<tr>
<td>.2</td>
<td>.25</td>
</tr>
<tr>
<td>.5</td>
<td>1</td>
</tr>
<tr>
<td>.8</td>
<td>4</td>
</tr>
<tr>
<td>.9</td>
<td>9</td>
</tr>
<tr>
<td>.98</td>
<td>49</td>
</tr>
</tbody>
</table>

```
plot(0, 0, type="n", xlab="Risk for Subject Without Risk Factor", ylab="Increase in Risk", xlim=c(0,1), ylim=c(0,.6)) # Figure 10.2
i <- 0
or <- c(1.1,1.25,1.5,1.75,2,3,4,5,10)
for(h in or) {
  i <- i + 1
  p <- seq(.0001, .9999, length=200)
  logit <- log(p/(1 - p)) # same as qlogis(p)
  logit <- logit + log(h) # modify by odds ratio
  p2 <- 1/(1 + exp(-logit)) # same as plogis(logit)
  d <- p2 - p
  lines(p, d, lty=i)
  maxd <- max(d)
  smax <- p[d==maxd]
  text(smax, maxd + .02, format(h), cex=.6)
}
```

Let $X_1$ be a binary risk factor and let $A = \{X_2, \ldots, X_p\}$ be the other factors. Then the estimate of $\text{Prob}\{Y = 1|X_1 = 1, A\} - \text{Prob}\{Y = 1|X_1 = 0, A\}$ is

$$
\frac{1}{1 + \exp \left[ -\hat{\beta}_0 - \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p \right]}
- \frac{1}{1 + \exp \left[ -\hat{\beta}_0 - \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p \right]}
$$
CHAPTER 10. BINARY LOGISTIC REGRESSION

10.1.3 Detailed Example

Figure 10.2: Absolute benefit as a function of risk of the event in a control subject and the relative effect (odds ratio) of the risk factor. The odds ratios are given for each curve.

\[
= \frac{1}{1 + \left(\frac{1-R}{R}\right)\exp(-\beta_1)} - \hat{R},
\]

where \( R = \text{Prob}[Y = 1|X_1 = 0, A] \).

- Risk ratio is \( \frac{1+e^{-X_2\beta}}{1+e^{-X_1\beta}} \)

- Does not simplify like odds ratio, which is \( \frac{e^{X_1\beta}}{e^{X_2\beta}} = e^{(X_1-X_2)\beta} \)

<table>
<thead>
<tr>
<th>Females Age</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>37 39 39 42 47 48 48 52 53 55 56 57 58 59 60 64 65 68 70</td>
<td>0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Males Age</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>34 38 40 40 41 43 43 44 46 47 48 48 50 50 52 55 60 61 61</td>
<td>1 1 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
</tbody>
</table>

require(rms)
gethdata(sex.age.response)
d ← sex.age.response
dd ← datadist(d); options(datadist='dd')
f ← lrm(response ~ sex + age, data=d)
fasr ← f  # Save for later
w ← function(...) with(d, {
m ← sex=='male'
f ← sex=='female'
lpoints(age[f], response[f], pch=1)
lpoints(age[m], response[m], pch=2)
af ← cut2(age, c(45,55), levels.mean=TRUE)
prop ← tapply(response, list(af, sex), mean, na.rm=TRUE)
agem ← as.numeric(row.names(prop))
lpoints(agem, prop[,'female'], pch=4, cex=1.3, col='green')
lpoints(agem, prop[,'male'], pch=5, cex=1.3, col='green')
x ← rep(62, 4); y ← seq(.25,.1, length=4)
lpoints(x, y, pch=c(1, 2, 4, 5), col=rep(c('blue','green'), each=2))
ltex(x+5, y, c('F Observed','M Observed', 'F Proportion','M Proportion'), cex=.8)
}

plot(Predict(f, age=seq(34, 70, length=200), sex, fun=plogis), ylab='Pr[response]', ylim=c(-.02, 1.02), addpanel=w)
ltex ← function(fit) latex(fit, inline=TRUE, columns=54, file='', after='$.', digits=3, size='Ssize', before='$X\hat{\beta}= ')
ltex(f)

\[ X\hat{\beta} = -9.84 + 3.49[\text{male}] + 0.158 \text{age.} \]

<table>
<thead>
<tr>
<th>sex</th>
<th>response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td></td>
</tr>
<tr>
<td>Row Pct</td>
<td>0 1 Total</td>
</tr>
<tr>
<td>F</td>
<td>14 6 20</td>
</tr>
<tr>
<td></td>
<td>70.00 30.00</td>
</tr>
<tr>
<td>M</td>
<td>6 14 20</td>
</tr>
<tr>
<td></td>
<td>30.00 70.00</td>
</tr>
<tr>
<td>Total</td>
<td>20 20 40</td>
</tr>
</tbody>
</table>

M:F odds ratio = (14/6)/(6/14) = 5.44, log=1.695

sex × response

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi Square</td>
<td>1</td>
<td>6.400</td>
<td>0.011</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>6.583</td>
<td>0.010</td>
</tr>
</tbody>
</table>
Figure 10.3: Data, subgroup proportions, and fitted logistic model, with 0.95 pointwise confidence bands.
Log likelihood \((\beta_1 = 0)\) : -27.727
Log likelihood (max) : -24.435
\( LR \chi^2(H_0 : \beta_1 = 0) \) : -2(-27.727 - -24.435) = 6.584

Next, consider the relationship between age and response, ignoring sex.

The estimate of \( \beta_1 \) is in rough agreement with that obtained from the frequency table. The 55+:<45 log odds ratio is .875, and since the respective mean ages in the 55+ and <45 age groups are 61.1 and 40.2, an estimate of the log odds ratio increase per year is \(.875/(61.1–40.2)=.875/20.9=.042\).
Log likelihood ($\beta_1 = 0$) : -27.727
Log likelihood (max) : -26.511
LR $\chi^2(H_0 : \beta_1 = 0)$ : -2(-27.727 - -26.511) = 2.432

(Compare 2.432 with the Wald statistic 2.28.)

Next we consider the simultaneous association of age and sex with response.

A logistic model for relating sex and age simultaneously to response is given below.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std Err</th>
<th>Wald $\chi^2$</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>-9.843</td>
<td>3.676</td>
<td>7.171</td>
<td></td>
</tr>
<tr>
<td>$\beta_1$ (sex)</td>
<td>3.490</td>
<td>1.199</td>
<td>8.469</td>
<td>0.004</td>
</tr>
<tr>
<td>$\beta_2$ (age)</td>
<td>0.158</td>
<td>0.062</td>
<td>6.576</td>
<td>0.010</td>
</tr>
</tbody>
</table>

Likelihood ratio tests are obtained from the information below.
Log likelihood \((\beta_1 = 0, \beta_2 = 0)\) : -27.727 
Log likelihood (max) : -19.458 
Log likelihood \((\beta_1 = 0)\) : -26.511 
Log likelihood \((\beta_2 = 0)\) : -24.435 
LR \(\chi^2\) \((H_0 : \beta_1 = \beta_2 = 0)\) : -2(-27.727- -19.458) = 16.538 
LR \(\chi^2\) \((H_0 : \beta_1 = 0)\) \(\text{sex}|\text{age}\) : -2(-26.511- -19.458) = 14.106 
LR \(\chi^2\) \((H_0 : \beta_2 = 0)\) \(\text{age}|\text{sex}\) : -2(-24.435- -19.458) = 9.954 

The 14.1 should be compared with the Wald statistic of 8.47, and 9.954 should be compared with 6.58. The fitted logistic model is plotted separately for females and males in Figure 10.3. The fitted model is 

\[
\logit\{\text{Response} = 1|\text{sex}, \text{age}\} = -9.84 + 3.49 \times \text{sex} + 0.158 \times \text{age},
\]

where as before sex=0 for females, 1 for males. For example, for a 40 year old female, the predicted logit is \(-9.84 + 0.158(40) = -3.52\). The predicted probability of a response is 

\[
\frac{1}{1 + \exp(3.52)} = 0.029.
\]

For a 40 year old male, the predicted logit is \(-9.84 + 3.49 + 0.158(40) = -0.03\), with a probability of 0.492.

---

### 10.1.4 Design Formulations

- Can do ANOVA using \(k - 1\) dummies for a \(k\)-level predictor
- Can get same \(\chi^2\) statistics as from a contingency table
- Can go farther: covariable adjustment
• Simultaneous comparison of multiple variables between two groups: Turn problem backwards to predict group from all the dependent variables

• This is more robust than a parametric multivariate test

• Propensity scores for adjusting for nonrandom treatment selection: Predict treatment from all baseline variables

• Adjusting for the predicted probability of getting a treatment adjusts adequately for confounding from all of the variables

• In a randomized study, using logistic model to adjust for covariables, even with perfect balance, will improve the treatment effect estimate
10.2 Estimation

10.2.1 Maximum Likelihood Estimates

Like binomial case but $P$s vary; $\hat{\beta}$ computed by trial and error using an iterative maximization technique

10.2.2 Estimation of Odds Ratios and Probabilities

\[
\hat{P}_i = \left[1 + \exp(-X_i\hat{\beta})\right]^{-1}.
\]

\[
\{1 + \exp[-(X_i\hat{\beta} \pm zs)]\}^{-1}.
\]

10.2.3 Minimum Sample Size Requirement

- Simplest case: no covariates, only an intercept
- Consider margin of error of 0.1 in estimating $\theta = \text{Prob}[Y = 1]$ with 0.95 confidence
- Worst case: $\theta = \frac{1}{2}$
- Requires $n = 96$ observations

\[a\] The general formula for the sample size required to achieve a margin of error of $\delta$ in estimating a true probability of $\theta$ at the 0.95 confidence
• Single binary predictor with prevalence $\frac{1}{2}$: need $n = 96$ for each value of $X$

• For margin of error of $\pm 0.05$, $n = 384$ is required (if true probabilities near 0.5 are possible); $n = 246$ required if true probabilities are only known not to be in $[0.2, 0.8]$.

• Single continuous predictor $X$ having a normal distribution with mean zero and standard deviation $\sigma$, with true $P = \frac{1}{1+\exp(-X)}$ so that the expected number of events is $\frac{n}{2}$.

  Compute mean of $\max_{X \in [-1.5, 1.5]} |P - \hat{P}|$ over 1000 simulations for varying $n$ and $\sigma$.

```
sigmas ← c(.5, .75, 1, 1.25, 1.5, 1.75, 2, 2.5, 3, 4)
ns ← seq(25, 300, by=25)
sim ← 1000
xs ← seq(-1.5, 1.5, length=200)
pactual ← plogis(xs)

dn ← list(sigma=format(sigmas), n=format(ns))
maxerr ← N1 ← array(NA, c(length(sigmas), length(ns), dn))
require(rms)

i ← 0
for(s in sigmas) {
  i ← i + 1
  j ← 0
  for(n in ns) {
    j ← j + 1
    n1 ← maxe ← 0
    for(k in 1:nsim) {
      x ← rnorm(n, 0, s)
      P ← plogis(x)
      y ← ifelse(runif(n) ≤ P, 1, 0)
      n1 ← n1 + sum(y)
      beta ← lrm.fit(x, y)$coefficients
      maxe ← maxe + max(abs(phat - pactual))
    }
    n1 ← n1/nsim
    maxe ← maxe/nsim
    maxerr[i,j] ← maxe
  }
}
```

level is $n = \left(\frac{1.96}{2}\right)^2 \times \theta(1-\theta)$. Set $\theta = \frac{1}{2}$ for the worst case.

An average absolute error of 0.05 corresponds roughly to a 0.95 confidence interval margin of error of 0.1.
\begin{verbatim}
N1[i,j] ← n1
}
}
xrange ← range(xs)
simerr ← llist(N1, maxerr, sigmas, ns, nsim, xrange)

maxe ← reShape(maxerr)

# Figure 10.4
xYplot(maxerr ~ n, groups=sigma, data=maxe,
       ylab=expression(paste('Average Maximum ',
                           abs(hat(P) - P))),
       type='l', lty=rep(1:2, 5), label.curve=FALSE,
       abline=list(h=c(.15, .1, .05), col=gray(.85)))
Key(.8, .68, other=list(cex=.7,
            title=expression(~~~~~~~~~sigma)))
\end{verbatim}

Figure 10.4: Simulated expected maximum error in estimating probabilities for \( x \in [-1.5, 1.5] \) with a single normally distributed \( X \) with mean zero
10.3 Test Statistics

- Likelihood ratio test best
- Score test second best (score $\chi^2 \equiv$ Pearson $\chi^2$)
- Wald test may misbehave but is quick
10.4 Residuals

Partial residuals (to check predictor transformations)

\[ r_{ij} = \hat{\beta}_j X_{ij} + \frac{Y_i - \hat{P}_i}{\hat{P}_i(1 - \hat{P}_i)}, \]
10.5 Assessment of Model Fit

\[
\logit\{Y = 1|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2,
\]

\[X_1 = 1\]

\[X_1 = 0\]

\[X_2\]

Figure 10.5: Logistic regression assumptions for one binary and one continuous predictor

getHdata(acath)
acath$sex ← factor(acath$sex, 0:1, c('male','female'))
dd ← datadist(acath); options(datadist='dd')
f ← lrm(sigdz ∼ rcs(age, 4) * sex, data=acath)

w ← function(...) 
with(acath, {
  plsmo(age, sigdz, group=sex, fun=qlogis, lty='dotted',
  add=TRUE, grid=TRUE)
  af ← cut2(age, g=10, levels.mean=TRUE)
  prop ← qlogis(tapply(sigdz, list(af, sex), mean,
    na.rm=TRUE))
  agem ← as.numeric(row.names(prop))
  lpoints(agem, prop[,,'female'], pch=4, col='green')
  lpoints(agem, prop[,,'male'], pch=2, col='green')
})  # Figure 10.6
plot(Predict(f, age, sex), ylim=c(-2,4), addpanel=w,
  label.curve=list(offset=unit(0.5, 'cm')))

- Can verify by plotting stratified proportions
CHAPTER 10. BINARY LOGISTIC REGRESSION

Figure 10.6: Logit proportions of significant coronary artery disease by sex and deciles of age for n=3504 patients, with spline fits (smooth curves). Spline fits are for $k = 4$ knots at age= 36, 48, 56, and 68 years, and interaction between age and sex is allowed. Shaded bands are pointwise 0.95 confidence limits for predicted log odds. Smooth nonparametric estimates are shown as dotted curves. Data courtesy of the Duke Cardiovascular Disease Databank.

- $\hat{P} = \frac{\text{number of events}}{\text{stratum size}}$
- $\hat{O} = \frac{\hat{P}}{1-\hat{P}}$
- Plot $\log \hat{O}$ (scale on which linearity is assumed)
- Stratified estimates are noisy
- 1 or 2 $X$s $\rightarrow$ nonparametric smoother
- `plsmo` function makes it easy to use loess to compute logits of nonparametric estimates (fun=qlogis)
- General: restricted cubic spline expansion of one or more predictors

$$\logit\{Y = 1|X\} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_2' + \hat{\beta}_4 X_2''$$
\[
\hat{\beta}_0 + \hat{\beta}_1 X_1 + f(X_2),
\]

\[
\text{logit}\{Y = 1 | X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2' + \beta_4 X_2'' + \beta_5 X_1 X_2 + \beta_6 X_1 X_2' + \beta_7 X_1 X_2''
\]

```r
lr <- function(formula) {
  f <- lrm(formula, data=acath)
  stats <- f$stats[c('Model L.R.', 'd.f. ')]
  cat('L.R. Chi-square: ', round(stats[1],1), 'd.f.: ', stats[2], 'n')
  f
}
a <- lr(sigdz ~ sex + age)

L.R. Chi-square: 766  d.f.: 2

b <- lr(sigdz ~ sex * age)

L.R. Chi-square: 768.2  d.f.: 3

c <- lr(sigdz ~ sex + rcs(age,4))

L.R. Chi-square: 769.4  d.f.: 4

d <- lr(sigdz ~ sex * rcs(age,4))

L.R. Chi-square: 782.5  d.f.: 7

lrtest(a, b)

Model 1: sigdz ~ sex + age
Model 2: sigdz ~ sex * age

L.R. Chisq  d.f.  P
2.1964146 1.0000000 0.1383322

lrtest(a, c)

Model 1: sigdz ~ sex + age
Model 2: sigdz ~ sex + rcs(age, 4)

L.R. Chisq  d.f.  P
3.4502500 2.0000000 0.1781508

lrtest(a, d)
```
Model 1: sigdz $\sim$ sex + age
Model 2: sigdz $\sim$ sex * rcs(age, 4)

L.R. Chisq        d.f.    P
16.547036344  5.000000000  0.005444012

lrtest(b, d)

Model 1: sigdz $\sim$ sex * age
Model 2: sigdz $\sim$ sex * rcs(age, 4)

L.R. Chisq        d.f.    P
14.350621767  4.000000000  0.006256138

lrtest(c, d)

Model 1: sigdz $\sim$ sex + rcs(age, 4)
Model 2: sigdz $\sim$ sex * rcs(age, 4)

L.R. Chisq        d.f.    P
13.096786352  3.000000000  0.004431906

<table>
<thead>
<tr>
<th>Model / Hypothesis</th>
<th>Likelihood Ratio</th>
<th>d.f.</th>
<th>$P$</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>a: sex, age (linear, no interaction)</td>
<td>766.0</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b: sex, age, age $\times$ sex</td>
<td>768.2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c: sex, spline in age</td>
<td>769.4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d: sex, spline in age, interaction</td>
<td>782.5</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$H_0$: no age $\times$ sex interaction given linearity</td>
<td>2.2</td>
<td>1</td>
<td>.14</td>
<td>($b - a$)</td>
</tr>
<tr>
<td>$H_0$: age linear $</td>
<td>$ no interaction</td>
<td>3.4</td>
<td>2</td>
<td>.18</td>
</tr>
<tr>
<td>$H_0$: age linear, no interaction</td>
<td>16.6</td>
<td>5</td>
<td>.005</td>
<td>($d - a$)</td>
</tr>
<tr>
<td>$H_0$: age linear, product form interaction</td>
<td>14.4</td>
<td>4</td>
<td>.006</td>
<td>($d - b$)</td>
</tr>
<tr>
<td>$H_0$: no interaction, allowing for nonlinearity in age</td>
<td>13.1</td>
<td>3</td>
<td>.004</td>
<td>($d - c$)</td>
</tr>
</tbody>
</table>

- Example of finding transform. of a single continuous predictor

- Duration of symptoms vs. odds of severe coronary disease
• Look at AIC to find best # knots for the money

<table>
<thead>
<tr>
<th>k</th>
<th>Model $\chi^2$</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>99.23</td>
<td>97.23</td>
</tr>
<tr>
<td>3</td>
<td>112.69</td>
<td>108.69</td>
</tr>
<tr>
<td>4</td>
<td>121.30</td>
<td>115.30</td>
</tr>
<tr>
<td>5</td>
<td>123.51</td>
<td>115.51</td>
</tr>
<tr>
<td>6</td>
<td>124.41</td>
<td>114.41</td>
</tr>
</tbody>
</table>

```
dz ← subset(acath, sigdz==1)
dd ← datadist(dz)

f ← lrm(tvd1m ~ rcs(cad.dur, 5), data=dz)
w ← function(...)
   with(dz, {
      plsmo(cad.dur, tvd1m, fun=qlogis, add=TRUE,
            grid=TRUE, lty='dotted')
x ← cut2(cad.dur, g=15, levels.mean=TRUE)
prop ← qlogis(tapply(tvd1m, x, mean, na.rm=TRUE))
xm ← as.numeric(names(prop))
lpoints(xm, prop, pch=2, col='green')
})  # Figure 10.7
plot(Predict(f, cad.dur), addpanel=w)
```

Figure 10.7: Estimated relationship between duration of symptoms and the log odds of severe coronary artery disease for $k = 5$. Knots are marked with arrows. Solid line is spline fit; dotted line is a nonparametric loess estimate.
Figure 10.8: Fitted linear logistic model in $\log_{10}(\text{duration}+1)$, with subgroup estimates using groups of 150 patients. Fitted equation is $\text{logit}(\text{tvdlm}) = -0.9809 + 0.7122 \log_{10}(\text{months} + 1)$. 
CHAPTER 10. BINARY LOGISTIC REGRESSION

Modeling Interaction Surfaces

- Sample of 2258 pts\(^c\)
- Predict significant coronary disease
- For now stratify age into tertiles to examine interactions simply
- Model has 2 dummies for age, sex, age × sex, 4-knot restricted cubic spline in cholesterol, age tertile × cholesterol

```r
acath <- transform(acath,
    cholesterol = cholest,
    age.tertile = cut2(age,g=3),
    sx = as.integer(acath$sex) - 1)
# sz for loess, need to code as numeric
dd <- datadist(acath); options(datadist='dd')

# First model stratifies age into tertiles to get more empirical estimates of age × cholesterol interaction
f <- lrm(sigdz ~ age.tertile*(sex + rcs(cholesterol,4)),
         data=acath)
```

Logistic Regression Model

```r
lrm(formula = sigdz ~ age.tertile * (sex + rcs(cholesterol, 4)),
     data = acath)
```

Frequencies of Missing Values Due to Each Variable

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>sigdz</th>
<th>age.tertile</th>
<th>sex</th>
<th>cholesterol</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2258</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1246</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Model Likelihood Ratio Test</th>
<th>Discrimination Indexes</th>
<th>Rank Discrim. Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>LR (\chi^2) = 533.52</td>
<td>(R^2) = 0.291</td>
<td>(C) = 0.780</td>
</tr>
<tr>
<td>0</td>
<td>d.f. = 14</td>
<td>(g) = 1.316</td>
<td>(D_{xy}) = 0.560</td>
</tr>
<tr>
<td>1</td>
<td>(\text{Pr}(&gt;\chi^2) &lt; 0.0001)</td>
<td>(g_r) = 3.729</td>
<td>(\gamma) = 0.560</td>
</tr>
<tr>
<td>max (\frac{</td>
<td>\partial \log L}{\partial \beta})</td>
<td>(2 \times 10^{-8})</td>
<td>(g_p) = 0.252</td>
</tr>
<tr>
<td>Brier</td>
<td></td>
<td>0.173</td>
<td></td>
</tr>
</tbody>
</table>

\(^c\)Many patients had missing cholesterol.
|                  | $\hat{\beta}$ | S.E. | Wald Z | Pr($>|Z|)$ |
|------------------|---------------|------|--------|-----------|
| Intercept        | -0.4155       | 1.0987 | -0.38  | 0.7053    |
| age.tertile=[49,58) | 0.8781       | 1.7337 | 0.51  | 0.6125    |
| age.tertile=[58,82] | 4.7861       | 1.8143 | 2.64  | 0.0083    |
| sex=female       | -1.6123       | 0.1751 | -9.21 | <0.0001   |
| cholesterol      | 0.0029        | 0.0060 | 0.48  | 0.6347    |
| cholesterol'     | 0.0384        | 0.0242 | 1.59  | 0.1126    |
| cholesterol"     | -0.1148       | 0.0768 | -1.49 | 0.1350    |
| age.tertile=[49,58) $\times$ sex=female | -0.7900 | 0.2537 | -3.11 | 0.0018    |
| age.tertile=[58,82] $\times$ cholesterol | 0.0011 | 0.0095 | 0.11  | 0.9093    |
| age.tertile=[58,82] $\times$ cholesterol' | -0.0158 | 0.0099 | -1.59 | 0.1111    |
| age.tertile=[58,82] $\times$ cholesterol" | -0.0183 | 0.0365 | -0.50 | 0.6162    |
| age.tertile=[49,58) $\times$ cholesterol | 0.0011 | 0.0095 | 0.11  | 0.9093    |
| age.tertile=[58,82] $\times$ cholesterol' | 0.0127 | 0.0406 | 0.31  | 0.7550    |
| age.tertile=[58,82] $\times$ cholesterol" | 0.0582 | 0.1301 | -0.07 | 0.9436    |

\[
\beta = -0.415 + 0.878[\text{age.tertile} \in [49,58)) + 4.79[\text{age.tertile} \in [58,82)) - 1.61[\text{female}] + 0.00287[\text{cholesterol}] + 1.52\times10^{-6}(\text{cholesterol} - 160)^3_+ - 4.53\times10^{-6}(\text{cholesterol} - 208)^3_+ + 3.44\times10^{-6}(\text{cholesterol} - 243)^3_+ - 4.28\times10^{-7}(\text{cholesterol} - 319)^3_+ + 0.00108[\text{cholesterol}] - 7.23\times10^{-7}(\text{cholesterol} - 160)^3_+ + 2.3\times10^{-6}(\text{cholesterol} - 208)^3_+ - 1.84\times10^{-6}(\text{cholesterol} - 243)^3_+ + 2.69\times10^{-7}(\text{cholesterol} - 319)^3_+].
\]

\[
\text{plot(Predict(f, cholesterol, age.tertile),}
\]

**Crudely categorizing age into tertiles**

<table>
<thead>
<tr>
<th></th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>age.tertile (Factor+Higher Order Factors)</td>
<td>120.74</td>
<td>10</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>21.87</td>
<td>8</td>
<td>0.0052</td>
</tr>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>329.54</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>9.78</td>
<td>2</td>
<td>0.0075</td>
</tr>
<tr>
<td>cholesterol (Factor+Higher Order Factors)</td>
<td>93.75</td>
<td>9</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>10.03</td>
<td>6</td>
<td>0.1235</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>9.96</td>
<td>6</td>
<td>0.1263</td>
</tr>
<tr>
<td>age.tertile $\times$ sex (Factor+Higher Order Factors)</td>
<td>9.78</td>
<td>2</td>
<td>0.0075</td>
</tr>
<tr>
<td>age.tertile $\times$ cholesterol (Factor+Higher Order Factors)</td>
<td>10.03</td>
<td>6</td>
<td>0.1235</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>2.62</td>
<td>4</td>
<td>0.6237</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>2.62</td>
<td>4</td>
<td>0.6237</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>9.96</td>
<td>6</td>
<td>0.1263</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>21.87</td>
<td>8</td>
<td>0.0052</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>29.67</td>
<td>10</td>
<td>0.0010</td>
</tr>
<tr>
<td>TOTAL</td>
<td>410.75</td>
<td>14</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
• Now model age as continuous predictor

• Start with nonparametric surface using \( Y = 0/1 \)

```r
# Re-do model with continuous age
f ← loess(sigdz ~ age * (sx + cholesterol), data=acath, 
           parametric="sx", drop.square="sx")
ages ← seq(25, 75, length=40)
chols ← seq(100, 400, length=40)
g ← expand.grid(cholesterol=chols, age=ages, sx=0)
# drop sex dimension of grids since held to 1 value
p ← drop(predict(f, g))
(p[p < 0.001] ← 0.001;
p[p > 0.999] ← 0.999)
zl ← c(-3, 6)  # Figure 10.10
wireframe(qlogis(p) ~ cholesterol*age,
         xlab=list(rot=30), ylab=list(rot=-40),
         zlab=list(label='log odds', rot=90), zlim=zl,
         scales = list(arrows = FALSE), data=g)
```

• Next try parametric fit using linear spline in age, chol. (3 knots each), all product terms. For all the remaining 3-d plots we limit plotting to points that are supported by at least 5 subjects beyond those cholesterol/age combinations

```r
f ← lrm(sigdz ~ lsp(age,c(46,52,59)) *
        (sex + lsp(cholesterol,c(196,224,259)))),
data=acath)
ltx(f)
```
Figure 10.10: Local regression fit for the logit of the probability of significant coronary disease vs. age and cholesterol for males, based on the loess function.
\[ X\hat{\beta} = -1.83 + 0.0232\text{age} + 0.0759(\text{age} - 46)_+ - 0.0025(\text{age} - 52)_+ + 2.27(\text{age} - 59)_+ + 3.02[\text{female}] - 0.0177\text{cholesterol} + 0.114(\text{cholesterol} - 196)_+ - 0.131(\text{cholesterol} - 224)_+ + 0.0651(\text{cholesterol} - 259)_+ + [\text{female}][-0.112\text{age} + 0.0852(\text{age} - 46)_+ - 0.0302(\text{age} - 52)_+ + 0.176(\text{age} - 59)_+] + \text{age} \times 0.00577(\text{cholesterol} - 0.00286(\text{cholesterol} - 196)_+ + 0.0382(\text{cholesterol} - 224)_+ - 0.00205(\text{cholesterol} - 259)_+ + (\text{age} - 46)_+[-0.00936\text{cholesterol} + 0.00643(\text{cholesterol} - 196)_+ - 0.0115(\text{cholesterol} - 224)_+ + 0.00756(\text{cholesterol} - 259)_+] + (\text{age} - 52)_+ [0.000433\text{cholesterol} - 0.0037(\text{cholesterol} - 196)_+ + 0.00815(\text{cholesterol} - 224)_+ - 0.00715(\text{cholesterol} - 259)_+] + (\text{age} - 59)_+[-0.0124\text{cholesterol} + 0.015(\text{cholesterol} - 196)_+ - 0.0067(\text{cholesterol} - 224)_+ + 0.00752(\text{cholesterol} - 259)_+]. \]

\[
\text{print(anova}(f), \text{caption} = '\text{Linear spline surface}', \text{size} = '\text{smaller}')
\]

### Linear spline surface

<table>
<thead>
<tr>
<th>Model</th>
<th>(\chi^2)</th>
<th>d.f.</th>
<th>(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>age (Factor+Higher Order Factors)</td>
<td>164.17</td>
<td>24</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>42.28</td>
<td>20</td>
<td>0.0025</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>25.21</td>
<td>18</td>
<td>0.1192</td>
</tr>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>343.80</td>
<td>5</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>23.90</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>cholesterol (Factor+Higher Order Factors)</td>
<td>100.13</td>
<td>20</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>16.27</td>
<td>16</td>
<td>0.4341</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>16.35</td>
<td>15</td>
<td>0.3595</td>
</tr>
<tr>
<td>age \times sex (Factor+Higher Order Factors)</td>
<td>23.90</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>12.97</td>
<td>3</td>
<td>0.0047</td>
</tr>
<tr>
<td>Nonlinear Interaction : (f(A,B)) vs. (AB)</td>
<td>12.97</td>
<td>3</td>
<td>0.0047</td>
</tr>
<tr>
<td>age \times cholesterol (Factor+Higher Order Factors)</td>
<td>16.27</td>
<td>16</td>
<td>0.4341</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>11.45</td>
<td>15</td>
<td>0.7204</td>
</tr>
<tr>
<td>Nonlinear Interaction : (f(A,B)) vs. (AB)</td>
<td>11.45</td>
<td>15</td>
<td>0.7204</td>
</tr>
<tr>
<td>(f(A,B)) vs. (Af(B) + Bg(A))</td>
<td>9.38</td>
<td>9</td>
<td>0.4033</td>
</tr>
<tr>
<td>Nonlinear Interaction in age vs. (Af(B))</td>
<td>9.99</td>
<td>12</td>
<td>0.6167</td>
</tr>
<tr>
<td>Nonlinear Interaction in cholesterol vs. (Bg(A))</td>
<td>10.75</td>
<td>12</td>
<td>0.5503</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>33.22</td>
<td>24</td>
<td>0.0995</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>42.28</td>
<td>20</td>
<td>0.0025</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>49.03</td>
<td>26</td>
<td>0.0041</td>
</tr>
<tr>
<td>TOTAL</td>
<td>449.26</td>
<td>29</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

```r
perim \leftarrow \text{with}(\text{acath}, \text{perimeter}(\text{cholesterol}, \text{age}, \text{xinc}=20, \text{n}=5))

zl \leftarrow c(-2, 4)  \quad \# \text{Figure 10.11}

\text{bplot}(\text{Predict}(f, \text{cholesterol}, \text{age}, \text{np}=40), \text{perim}=\text{perim}, \text{lfun}=\text{wireframe}, \text{zlim}=\text{zl}, \text{adjsubtitle}=\text{FALSE})
```

- Next try smooth spline surface, include all cross-products!

```r
f \leftarrow \text{lrm}(\text{sigdz} \sim \text{rcs(\text{age},4)}*(\text{sex} + \text{rcs(\text{cholesterol},4))), \text{data}=\text{acath}, \text{tol}=1e-11)

\text{ltx}(f)
```

\[ X\hat{\beta} = -6.41 + 0.166\text{age} - 0.00067(\text{age} - 36)_+^3 + 0.00543(\text{age} - 48)_+^3 - 0.00727(\text{age} - 56)_+^3 + 0.00251(\text{age} - 68)_+^3 + 2.87[\text{female}] + 0.00979(\text{cholesterol} + 1.96 \times 10^{-6}(\text{cholesterol} - 160)_+^3 - 7.16 \times 10^{-6}(\text{cholesterol} - 208)_+^3 + 6.35 \times 10^{-6}(\text{cholesterol} - 243)_+^3 - 1.16 \times 10^{-6}(\text{cholesterol} - 319)_+^3 + [\text{female}][-0.109\text{age} + 7.52 \times 10^{-5}(\text{age} - 36)_+^2 + 0.00015(\text{age} - 48)_+^3 - 0.00045(\text{age} -
```
Figure 10.11: Linear spline surface for males, with knots for age at 46, 52, 59 and knots for cholesterol at 196, 224, and 259 (quartiles).
\[
56)^3 + 0.00225(\text{age} - 68)^3 + \alpha \cdot 0.00028(\text{cholesterol} - 160)^3 + 3.03 \times 10^{-5}(\text{cholesterol} - 208)^3 - 4.99 \times 10^{-8}(\text{cholesterol} - 243)^3 + 1.69 \times 10^{-8}(\text{cholesterol} - 319)^3 + \alpha \cdot 0.00341(\text{cholesterol} - 4.02 \times 10^{-7}(\text{cholesterol} - 160)^3 + 9.71 \times 10^{-7}(\text{cholesterol} - 208)^3 - 5.79 \times 10^{-7}(\text{cholesterol} - 243)^3 + 8.79 \times 10^{-9}(\text{cholesterol} - 319)^3 + \alpha \cdot 0.029(\text{cholesterol} + 3.04 \times 10^{-6}(\text{cholesterol} - 160)^3 - 7.34 \times 10^{-6}(\text{cholesterol} - 208)^3 + 4.36 \times 10^{-6}(\text{cholesterol} - 243)^3 - 5.82 \times 10^{-8}(\text{cholesterol} - 319)^3).
\]

print(anova(f), caption='Cubic spline surface',
      size='smaller')

Cubic spline surface

<table>
<thead>
<tr>
<th></th>
<th>(\chi^2)</th>
<th>d.f.</th>
<th>(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>age (Factor+Higher Order Factors)</td>
<td>165.23</td>
<td>15</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>37.32</td>
<td>12</td>
<td>0.0002</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>21.01</td>
<td>10</td>
<td>0.0210</td>
</tr>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>343.67</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>23.31</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>cholesterol (Factor+Higher Order Factors)</td>
<td>97.50</td>
<td>12</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>12.95</td>
<td>9</td>
<td>0.1649</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>13.62</td>
<td>8</td>
<td>0.0923</td>
</tr>
<tr>
<td>age (\times) sex (Factor+Higher Order Factors)</td>
<td>23.31</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>13.37</td>
<td>2</td>
<td>0.0013</td>
</tr>
<tr>
<td>Nonlinear Interaction : (f(A,B)) vs. (AB)</td>
<td>13.37</td>
<td>2</td>
<td>0.0013</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>12.95</td>
<td>9</td>
<td>0.1649</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>7.27</td>
<td>8</td>
<td>0.5078</td>
</tr>
<tr>
<td>Nonlinear Interaction : (f(A,B)) vs. (AB)</td>
<td>7.27</td>
<td>8</td>
<td>0.5078</td>
</tr>
<tr>
<td>(f(A,B)) vs. (Af(B) + Bg(A))</td>
<td>5.41</td>
<td>4</td>
<td>0.2480</td>
</tr>
<tr>
<td>Nonlinear Interaction in age vs. (Af(B))</td>
<td>6.44</td>
<td>6</td>
<td>0.3753</td>
</tr>
<tr>
<td>Nonlinear Interaction in cholesterol vs. (Bg(A))</td>
<td>6.27</td>
<td>6</td>
<td>0.3931</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>29.22</td>
<td>14</td>
<td>0.0097</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>37.32</td>
<td>12</td>
<td>0.0002</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>45.41</td>
<td>16</td>
<td>0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>450.88</td>
<td>19</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

# Figure 10.12:
bplot(Predict(f, cholesterol, age, np=40), perim=perim, lfun=wireframe, zlim=zl, adj.subtitle=FALSE)

- Now restrict surface by excluding doubly nonlinear terms

```r
f ← lrm(sigdz ~ sex*rcs(age,4) + rcs(cholesterol,4) + 
    rcs(age,4) %ia% rcs(cholesterol,4), data=acath)
print(anova(f), size='smaller', caption='Singly nonlinear cubic spline surface')
```
Figure 10.12: Restricted cubic spline surface in two variables, each with $k = 4$ knots
Singly nonlinear cubic spline surface

<table>
<thead>
<tr>
<th></th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>343.42</td>
<td>4</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>All Interactions</td>
<td>24.05</td>
<td>3</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>age (Factor+Higher Order Factors)</td>
<td>169.35</td>
<td>11</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>All Interactions</td>
<td>34.80</td>
<td>8</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>16.55</td>
<td>6</td>
<td>0.0111</td>
</tr>
<tr>
<td>cholesterol (Factor+Higher Order Factors)</td>
<td>93.62</td>
<td>8</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>All Interactions</td>
<td>10.83</td>
<td>5</td>
<td>0.0548</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>10.87</td>
<td>4</td>
<td>0.0281</td>
</tr>
<tr>
<td>age × cholesterol (Factor+Higher Order Factors)</td>
<td>10.83</td>
<td>5</td>
<td>0.0548</td>
</tr>
<tr>
<td>All Interactions</td>
<td>3.12</td>
<td>4</td>
<td>0.5372</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>1.60</td>
<td>2</td>
<td>0.4996</td>
</tr>
<tr>
<td>Nonlinear Interaction in age vs. A(B)</td>
<td>1.64</td>
<td>2</td>
<td>0.4000</td>
</tr>
<tr>
<td>sex × age (Factor+Higher Order Factors)</td>
<td>24.05</td>
<td>3</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>All Interactions</td>
<td>13.58</td>
<td>2</td>
<td>0.0011</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>13.58</td>
<td>2</td>
<td>0.0011</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>27.89</td>
<td>10</td>
<td>0.0019</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>34.80</td>
<td>8</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>45.45</td>
<td>12</td>
<td>$&lt;0.0001$</td>
</tr>
<tr>
<td>TOTAL</td>
<td>453.10</td>
<td>15</td>
<td>$&lt;0.0001$</td>
</tr>
</tbody>
</table>

# Figure 10.13:

```r
bplot(Predict(f, cholesterol, age, np=40), perim=perim, 
  lfun=wireframe, zlim=zl, adj.subtitle=FALSE)
ltx(f)
```

\[
\hat{X} \beta = -7.2 + 2.96 [\text{female}] + 0.164 \text{age} + 7.23 \times 10^{-5} (\text{age} - 36)^3 + 0.000106 (\text{age} - 48)^3 - 1.63 \times 10^{-5} (\text{age} - 56)^3 + 4.99 \times 10^{-5} (\text{age} - 68)^3 + 0.0148 \text{cholesterol} + 1.21 \times 10^{-6} (\text{cholesterol} - 160)^3 - 5.5 \times 10^{-6} (\text{cholesterol} - 208)^3 + 5.5 \times 10^{-6} (\text{cholesterol} - 243)^3 - 1.21 \times 10^{-6} (\text{cholesterol} - 319)^3 + 0.00029 \text{cholesterol} + 9.28 \times 10^{-9} (\text{cholesterol} - 160)^3 + 1.7 \times 10^{-8} (\text{cholesterol} - 208)^3 - 4.43 \times 10^{-8} (\text{cholesterol} - 243)^3 - 1.79 \times 10^{-8} (\text{cholesterol} - 319)^3 + \text{cholesterol} [2.3 \times 10^{-7} (\text{age} - 36)^3 + 4.21 \times 10^{-7} (\text{age} - 48)^3 - 1.31 \times 10^{-6} (\text{age} - 56)^3 + 6.64 \times 10^{-7} (\text{age} - 68)^3 + [\text{female}][-0.111 \text{age} + 8.03 \times 10^{-5} (\text{age} - 36)^3 + 0.000135 (\text{age} - 48)^3 - 0.00044 (\text{age} - 56)^3 + 0.000224 (\text{age} - 68)^3].
\]

- Finally restrict the interaction to be a simple product

```r
f ← lrm(sigdz ~ rcs(age,4)*sex + rcs(cholesterol,4) + 
  age %ia% cholesterol, data=acath)
print(anova(f), caption='Linear interaction surface', 
  size='smaller')
```
Figure 10.13: Restricted cubic spline fit with age × spline(cholesterol) and cholesterol × spline(age)
Linear interaction surface

<table>
<thead>
<tr>
<th></th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>age (Factor+Higher Order Factors)</td>
<td>167.83</td>
<td>7</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>31.03</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>14.58</td>
<td>4</td>
<td>0.0057</td>
</tr>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>345.88</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>22.30</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>cholesterol (Factor+Higher Order Factors)</td>
<td>89.37</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>7.99</td>
<td>1</td>
<td>0.0047</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>10.65</td>
<td>2</td>
<td>0.0049</td>
</tr>
<tr>
<td>age × cholesterol (Factor+Higher Order Factors)</td>
<td>7.99</td>
<td>1</td>
<td>0.0047</td>
</tr>
<tr>
<td>age × sex (Factor+Higher Order Factors)</td>
<td>22.30</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>12.06</td>
<td>2</td>
<td>0.0024</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>12.06</td>
<td>2</td>
<td>0.0024</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>25.72</td>
<td>6</td>
<td>0.0003</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>31.03</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>43.59</td>
<td>8</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>452.75</td>
<td>11</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

The Wald test for age × cholesterol interaction yields $\chi^2 = 7.99$ with 1 d.f., $p = .005$.

- See how well this simple interaction model compares with initial model using 2 dummies for age

- Request predictions to be made at mean age within tertiles
Figure 10.14: Spline fit with nonlinear effects of cholesterol and age and a simple product interaction.

Figure 10.15: Predictions from linear interaction model with mean age in tertiles indicated.
• Using residuals for “duration of symptoms” example

```r
f <- lrm(tvdlm ~ cad.dur, data=dz, x=TRUE, y=TRUE)
resid(f, "partial", pl="loess", xlim=c(0,250), ylim=c(-3,3))
scatid(dz$cad.dur)
log.cad.dur <- log10(dz$cad.dur + 1)
f <- lrm(tvdlm ~ log.cad.dur, data=dz, x=TRUE, y=TRUE)
resid(f, "partial", pl="loess", ylim=c(-3,3))
scatid(log.cad.dur)  # Figure 10.16
```

![Graphs showing partial residuals for duration and log10(duration+1). Data density shown at top of each plot.](image)

Figure 10.16: Partial residuals for duration and log10(duration+1). Data density shown at top of each plot.

• Relative merits of strat., nonparametric, splines for checking fit
## Hosmer-Lemeshow Test

Hosmer-Lemeshow test is a commonly used test of goodness-of-fit of a binary logistic model. It compares the proportion of events with mean predicted probability within deciles of \( \hat{P} \):

- **Arbitrary** (number of groups, how to form groups)
- **Low power** (too many d.f.)
- **Does not reveal the culprits**

### Table of Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Choice Required</th>
<th>Assumes Additivity</th>
<th>Uses Ordering of ( X )</th>
<th>Low Variance</th>
<th>Good Resolution on ( X )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratification Intervals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoother on ( X_1 ) stratifying on ( X_2 )</td>
<td>Bandwidth</td>
<td>( \times )</td>
<td>(not on ( X_2 ))</td>
<td>( \times )</td>
<td>( \times ) ( X_1 )</td>
</tr>
<tr>
<td>Smooth partial residual plot</td>
<td>Bandwidth</td>
<td>( \times )</td>
<td>( \times )</td>
<td>( \times )</td>
<td>( \times ) ( X_1 )</td>
</tr>
<tr>
<td>Spline model for all ( X )'s</td>
<td>Knots</td>
<td>( \times )</td>
<td>( \times )</td>
<td>( \times )</td>
<td>( \times ) ( X_1 )</td>
</tr>
</tbody>
</table>

### Notes

- A new omnibus test based on SSE has more power and requires no grouping; still does not lead to corrective action.
- Any omnibus test lacks power against specific alternatives such as nonlinearity or interaction.
Collinearity

Overly Influential Observations

Quantifying Predictive Ability

• Generalized $R^2$: equals ordinary $R^2$ in normal case:

$$R^2_N = \frac{1 - \exp(-LR/n)}{1 - \exp(-L_0/n)},$$

• Brier score (calibration + discrimination):

$$B = \frac{1}{n} \sum_{i=1}^{n} (\hat{P}_i - Y_i)^2,$$

• $c$ = “concordance probability” = ROC area
  – Related to Wilcoxon-Mann-Whitney stat and Somers’ $D_{xy}$

$$D_{xy} = 2(c - .5).$$

  – Good pure index of predictive discrimination for a single model

  – Not useful for comparing two models [48, 162]

\[\text{But see [160].}\]
“Coefficient of discrimination” [207]: average $\hat{P}$ when $Y = 1$ minus average $\hat{P}$ when $Y = 0$
  - Has many advantages. Tjur shows how it ties in with sum of squares–based $R^2$ measures.

“Percent classified correctly” has lots of problems
  - improper scoring rule; optimizing it will lead to incorrect model
  - arbitrary, insensitive, uses a strange loss (utility function)
Validating the Fitted Model

- Possible indexes \[10\]
  - Accuracy of \(\hat{P}\): calibration
    Plot \(\frac{1}{1+e^{-X_{new}\hat{\beta}_{old}}}\) against estimated prob. that \(Y = 1\) on new data
  - Discrimination: \(C\) or \(D_{xy}\)
  - \(R^2\) or \(B\)

- Use bootstrap to estimate calibration equation
  \[P_c = \text{Prob}\{Y = 1|X\hat{\beta}\} = [1 + \exp-(\gamma_0 + \gamma_1 X\hat{\beta})]^{-1},\]
  \[E_{max}(a, b) = \max_{a \leq \hat{P} \leq b} |\hat{P} - \hat{P}_c|,\]

- Bootstrap validation of age-sex-response data, 150 samples

- 2 predictors forced into every model

\[
d \leftarrow \text{sex.age.response}
dd \leftarrow \text{datadist}(d); \text{options(datadist='dd')}
f \leftarrow \text{lrm(response ~ sex + age, data=d, x=TRUE, y=TRUE)}
\text{set.seed}(3) \# \text{for reproducibility}
v1 \leftarrow \text{validate}(f, B=150)
\]

\[
\text{latex(v1,}
\caption='Bootstrap Validation, 2 Predictors Without Stepdown',
\text{insert.bottom='\\label{pg:lrm-sex-age-response-boot}',
\text{digits=2, size='Ssize', file=')}
}
### Bootstrap Validation, 2 Predictors Without Stepdown

<table>
<thead>
<tr>
<th>Index</th>
<th>Original Sample</th>
<th>Training Sample</th>
<th>Test Sample</th>
<th>Optimism</th>
<th>Corrected Index</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{xy}$</td>
<td>0.70</td>
<td>0.70</td>
<td>0.67</td>
<td>0.03</td>
<td>0.66</td>
<td>150</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.45</td>
<td>0.48</td>
<td>0.43</td>
<td>0.05</td>
<td>0.40</td>
<td>150</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.04</td>
<td>150</td>
<td></td>
</tr>
<tr>
<td>Slope</td>
<td>1.00</td>
<td>1.00</td>
<td>0.92</td>
<td>0.08</td>
<td>0.92</td>
<td>150</td>
</tr>
<tr>
<td>$E_{max}$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>0.03</td>
<td>150</td>
<td></td>
</tr>
<tr>
<td>$D$</td>
<td>0.39</td>
<td>0.43</td>
<td>0.36</td>
<td>0.07</td>
<td>0.32</td>
<td>150</td>
</tr>
<tr>
<td>$U$</td>
<td>−0.05</td>
<td>−0.05</td>
<td>0.02</td>
<td>−0.07</td>
<td>0.02</td>
<td>150</td>
</tr>
<tr>
<td>$Q$</td>
<td>0.44</td>
<td>0.48</td>
<td>0.34</td>
<td>0.14</td>
<td>0.30</td>
<td>150</td>
</tr>
<tr>
<td>$B$</td>
<td>0.16</td>
<td>0.15</td>
<td>0.18</td>
<td>−0.03</td>
<td>0.19</td>
<td>150</td>
</tr>
<tr>
<td>$g$</td>
<td>2.10</td>
<td>2.38</td>
<td>1.97</td>
<td>0.41</td>
<td>1.70</td>
<td>150</td>
</tr>
<tr>
<td>$g_p$</td>
<td>0.35</td>
<td>0.35</td>
<td>0.34</td>
<td>0.01</td>
<td>0.34</td>
<td>150</td>
</tr>
</tbody>
</table>

- Allow for step-down at each re-sample
- Use individual tests at $\alpha = 0.10$
- Both age and sex selected in 137 of 150, neither in 3 samples

```r
v2 ← validate(f, B=150, bw=TRUE,
  rule='p', sls=.1, type='individual')
latex(v2,
  caption='Bootstrap Validation, 2 Predictors with Stepdown',
  digits=2, B=15, file='', size='Ssize')
```

### Bootstrap Validation, 2 Predictors with Stepdown

<table>
<thead>
<tr>
<th>Index</th>
<th>Original Sample</th>
<th>Training Sample</th>
<th>Test Sample</th>
<th>Optimism</th>
<th>Corrected Index</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{xy}$</td>
<td>0.70</td>
<td>0.71</td>
<td>0.65</td>
<td>0.07</td>
<td>0.63</td>
<td>150</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.45</td>
<td>0.50</td>
<td>0.41</td>
<td>0.09</td>
<td>0.36</td>
<td>150</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>−0.01</td>
<td>0.01</td>
<td>150</td>
</tr>
<tr>
<td>Slope</td>
<td>1.00</td>
<td>1.00</td>
<td>0.83</td>
<td>0.17</td>
<td>0.83</td>
<td>150</td>
</tr>
<tr>
<td>$E_{max}$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
<td>150</td>
</tr>
<tr>
<td>$D$</td>
<td>0.39</td>
<td>0.46</td>
<td>0.35</td>
<td>0.11</td>
<td>0.28</td>
<td>150</td>
</tr>
<tr>
<td>$U$</td>
<td>−0.05</td>
<td>−0.05</td>
<td>0.05</td>
<td>−0.10</td>
<td>0.05</td>
<td>150</td>
</tr>
<tr>
<td>$Q$</td>
<td>0.44</td>
<td>0.51</td>
<td>0.29</td>
<td>0.21</td>
<td>0.22</td>
<td>150</td>
</tr>
<tr>
<td>$B$</td>
<td>0.16</td>
<td>0.14</td>
<td>0.18</td>
<td>−0.04</td>
<td>0.20</td>
<td>150</td>
</tr>
<tr>
<td>$g$</td>
<td>2.10</td>
<td>2.60</td>
<td>1.90</td>
<td>0.70</td>
<td>1.40</td>
<td>150</td>
</tr>
<tr>
<td>$g_p$</td>
<td>0.35</td>
<td>0.35</td>
<td>0.33</td>
<td>0.02</td>
<td>0.33</td>
<td>150</td>
</tr>
</tbody>
</table>
### Factors Retained in Backwards Elimination

First 15 Resamples

<table>
<thead>
<tr>
<th>sex</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
<tr>
<td>•</td>
<td></td>
</tr>
</tbody>
</table>

### Frequencies of Numbers of Factors Retained

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>11</td>
<td>138</td>
</tr>
</tbody>
</table>

• Try adding 5 noise candidate variables

```r
set.seed(133)
n ← nrow(d)
x1 ← runif(n)
x2 ← runif(n)
x3 ← runif(n)
x4 ← runif(n)
x5 ← runif(n)
f ← lrm(response ~ age + sex + x1 + x2 + x3 + x4 + x5,
          data=d, x=TRUE, y=TRUE)
v3 ← validate(f, B=150, bw=TRUE,
              rule='p', sls=.1, type='individual')
k ← attr(v3, 'kept')
# Compute number of x1-x5 selected
nx ← apply(k[,3:7], 1, sum)
# Get selections of age and sex
v ← colnames(k)
as ← apply(k[,1:2], 1,
        function(x) paste(v[1:2][x], collapse=', '))
table(paste(as, ', ', nx, ', Xs'))
```

<table>
<thead>
<tr>
<th></th>
<th>0 Xs</th>
<th>1 Xs age, sex</th>
<th>0 Xs age, sex</th>
<th>1 Xs age, sex</th>
<th>2 Xs age, sex</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
<td>4</td>
<td>30</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>age, sex</td>
<td>3 Xs</td>
<td>sex 0 Xs sex</td>
<td>1 Xs sex</td>
<td>2 Xs sex</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>9</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
latex(v3, 
caption='Bootstrap Validation with 5 Noise Variables and Stepdown',
digits=2, B=15, size='Ssize', file='')

Bootstrap Validation with 5 Noise Variables and Stepdown

<table>
<thead>
<tr>
<th>Index</th>
<th>Original Sample</th>
<th>Training Sample</th>
<th>Test Sample</th>
<th>Optimism</th>
<th>Corrected Index</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{xy}$</td>
<td>0.70</td>
<td>0.47</td>
<td>0.38</td>
<td>0.09</td>
<td>0.61</td>
<td>136</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.45</td>
<td>0.34</td>
<td>0.23</td>
<td>0.11</td>
<td>0.34</td>
<td>136</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>−0.04</td>
<td>0.04</td>
<td>136</td>
</tr>
<tr>
<td>Slope</td>
<td>1.00</td>
<td>1.00</td>
<td>0.77</td>
<td>0.23</td>
<td>0.77</td>
<td>136</td>
</tr>
<tr>
<td>$E_{\text{max}}$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>136</td>
</tr>
<tr>
<td>$D$</td>
<td>0.39</td>
<td>0.31</td>
<td>0.18</td>
<td>0.13</td>
<td>0.26</td>
<td>136</td>
</tr>
<tr>
<td>$U$</td>
<td>−0.05</td>
<td>−0.05</td>
<td>0.06</td>
<td>−0.11</td>
<td>0.06</td>
<td>136</td>
</tr>
<tr>
<td>$Q$</td>
<td>0.44</td>
<td>0.36</td>
<td>0.12</td>
<td>0.24</td>
<td>0.20</td>
<td>136</td>
</tr>
<tr>
<td>$B$</td>
<td>0.16</td>
<td>0.18</td>
<td>0.21</td>
<td>−0.04</td>
<td>0.20</td>
<td>136</td>
</tr>
<tr>
<td>$g$</td>
<td>2.10</td>
<td>1.81</td>
<td>1.06</td>
<td>0.75</td>
<td>1.35</td>
<td>136</td>
</tr>
<tr>
<td>$g_p$</td>
<td>0.35</td>
<td>0.24</td>
<td>0.19</td>
<td>0.04</td>
<td>0.31</td>
<td>136</td>
</tr>
</tbody>
</table>

Factors Retained in Backwards Elimination
First 15 Resamples

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x4</th>
<th>x5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Frequencies of Numbers of Factors Retained

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>13</td>
<td>33</td>
<td>27</td>
<td>8</td>
<td>5</td>
</tr>
</tbody>
</table>

- Repeat but force age and sex to be in all models

v4 ← validate(f, B=150, bw=TRUE, rule='p', sls=.1, 
type='individual', force=1:2)
\texttt{ap4} \leftarrow \texttt{round(v4[,}'index.orig', 2)}
\texttt{bc4} \leftarrow \texttt{round(v4[,}'index.corrected', 2)}

\texttt{latex(v4,}
\texttt{caption='Bootstrap Validation with 5 Noise Variables and Stepdown,}
\texttt{Forced Inclusion of age and sex',}
\texttt{digits=2, B=15, size='Ssize')}

\begin{table}[h]
\centering
\begin{tabular}{lccccccc}
\hline
\multicolumn{1}{l|}{Index} & \multicolumn{3}{c}{Original Sample} & \multicolumn{2}{c}{Training Sample} & \multicolumn{2}{c}{Test Sample} \\
\hline
$D_{xy}$ & 0.70 & 0.76 & 0.66 & 0.10 & 0.60 & 130 \\
$R^2$ & 0.45 & 0.54 & 0.41 & 0.12 & 0.33 & 130 \\
Intercept & 0.00 & 0.00 & 0.06 & -0.06 & 0.06 & 130 \\
Slope & 1.00 & 1.00 & 0.76 & 0.24 & 0.76 & 130 \\
$E_{max}$ & 0.00 & 0.00 & 0.07 & 0.07 & 0.07 & 130 \\
$D$ & 0.39 & 0.50 & 0.35 & 0.15 & 0.24 & 130 \\
$U$ & -0.05 & -0.05 & 0.08 & -0.13 & 0.08 & 130 \\
$Q$ & 0.44 & 0.55 & 0.27 & 0.28 & 0.16 & 130 \\
$B$ & 0.16 & 0.14 & 0.18 & -0.04 & 0.21 & 130 \\
g & 2.10 & 2.75 & 1.89 & 0.86 & 1.25 & 130 \\
g_p & 0.35 & 0.37 & 0.33 & 0.04 & 0.31 & 130 \\
\hline
\end{tabular}
\caption{Bootstrap Validation with 5 Noise Variables and Stepdown, Forced Inclusion of age and sex}
\end{table}

Factors Retained in Backwards Elimination
\begin{center}
First 15 Resamples
\end{center}
\begin{align*}
\texttt{age} & \cdot \cdot \\
\texttt{sex} & \cdot \cdot \\
\texttt{x1} & \cdot \\
\texttt{x2} & \cdot \\
\texttt{x3} & \cdot \\
\texttt{x4} & \cdot \\
\texttt{x5} & \cdot \\
\end{align*}

\begin{tabular}{lcccc}
\hline
\multicolumn{1}{l|}{2} & \multicolumn{1}{l|}{3} & \multicolumn{1}{l|}{4} & \multicolumn{1}{l|}{5} & \multicolumn{1}{l|}{6} \\
\hline
88 & 29 & 10 & 1 & 2 \\
\end{tabular}

Frequencies of Numbers of Factors Retained
Describing the Fitted Model

\[
s \leftarrow \text{summary(f.linia)} \\
\text{print(s, size='Ssize')} \\
\]

<table>
<thead>
<tr>
<th></th>
<th>Low</th>
<th>High</th>
<th>Effect</th>
<th>S.E.</th>
<th>Lower 0.95</th>
<th>Upper 0.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>46</td>
<td>59</td>
<td>13</td>
<td>0.9063</td>
<td>0.5460</td>
<td>1.2665</td>
</tr>
<tr>
<td>\text{Odds Ratio}</td>
<td>2.47510</td>
<td>1.7264</td>
<td>3.5486</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cholesterol</td>
<td>196</td>
<td>259</td>
<td>63</td>
<td>0.7548</td>
<td>0.4874</td>
<td>1.0222</td>
</tr>
<tr>
<td>\text{Odds Ratio}</td>
<td>2.12720</td>
<td>1.6281</td>
<td>2.7792</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>sex — female:male</td>
<td>1</td>
<td>2</td>
<td>-2.43</td>
<td>0.1484</td>
<td>-2.7206</td>
<td>-2.1389</td>
</tr>
<tr>
<td>\text{Odds Ratio}</td>
<td>0.08806</td>
<td>0.0658</td>
<td>0.1177</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{plot(s)} \quad \# \text{Figure 10.17}
\]

\[
\text{Odds Ratio}
\]

\[
\begin{align*}
\text{age} & \quad 59:46 \\
\text{cholesterol} & \quad 259:196 \\
\text{sex} \quad \text{female:male}
\end{align*}
\]

Adjusted to: age=52, sex=male, cholesterol=224.5

Figure 10.17: Odds ratios and confidence bars, using quartiles of age and cholesterol for assessing their effects on the odds of coronary disease.
Figure 10.18: Linear spline fit for probability of bacterial vs. viral meningitis as a function of age at onset [192]. Points are simple proportions by age quantile groups.
Figure 10.19: (A) Relationship between myocardium at risk and ventricular fibrillation, based on the individual best fit equations for animals anesthetized with pentobarbital and α-chloralose. The amount of myocardium at risk at which 0.5 of the animals are expected to fibrillate (MAR_{50}) is shown for each anesthetic group. (B) Relationship between myocardium at risk and ventricular fibrillation, based on equations derived from the single slope estimate. Note that the MAR_{50} describes the overall relationship between myocardium at risk and outcome when either the individual best fit slope or the single slope estimate is used. The shift of the curve to the right during α-chloralose anesthesia is well described by the shift in MAR_{50}. Test for interaction had P=0.10. Reprinted by permission, NRC Research Press.
Figure 10.20: A nomogram for estimating the likelihood of significant coronary artery disease (CAD) in women. ECG = electrocardiographic; MI = myocardial infarction [171]. Reprinted from American Journal of Medicine, Vol 75, Pryor DB et al., “Estimating the likelihood of significant coronary artery disease”, p. 778, Copyright 1983, with permission from Excerpta Medica, Inc.
### Figure 10.21: Nomogram for estimating probability of bacterial (ABM) vs. viral (AVM) meningitis.

Step 1, place ruler on reading lines for patient’s age and month of presentation and mark intersection with line A; step 2, place ruler on values for glucose ratio and total polymorphonuclear leukocyte (PMN) count in cerebrospinal fluid and mark intersection with line B; step 3, use ruler to join marks on lines A and B, then read off the probability of ABM vs. AVM [192].
# Draw a nomogram that shows examples of confidence intervals

```r
nom <- nomogram(f.linia, cholesterol=seq(150, 400, by=50),
    interact=list(age=seq(30, 70, by=10)),
    lp.at=seq(-2, 3.5, by=.5),
    conf.int=TRUE, conf.lp="all",
    fun=function(x)1/(1+exp(-x)), # or plogis
    funlabel="Probability of CAD",
    fun.at=c(seq(.1, .9, by=.1), .95, .99)
)
```

Figure 10.22

```r
plot(nom, col.grid = gray(c(0.8, 0.95)),
    varname.label=FALSE, ia.space=1, xfrac=.46, lmgp=.2)
```
Figure 10.22: Nomogram relating age, sex, and cholesterol to the log odds and to the probability of significant coronary artery disease. Select one axis corresponding to sex and to age $\in \{30, 40, 50, 60, 70\}$. There was linear interaction between age and sex and between age and cholesterol. $0.70$ and $0.90$ confidence intervals are shown ($0.90$ in gray). Note that for the 'Linear Predictor' scale there are various lengths of confidence intervals near the same value of $X\hat{\beta}$, demonstrating that the standard error of $X\hat{\beta}$ depends on the individual $X$ values. Also note that confidence intervals corresponding to smaller patient groups (e.g., females) are wider.
10.11 Bayesian Logistic Model Example

This re-analysis of data in Section 10.1.3 was done by Nathan James, Dinesh Karki, and Elizabeth McNeer. Ryan Jarrett added the section on bivariate confidence regions. This Bayesian analysis uses the R \texttt{brms} package \cite{Bates2017} front-end to the Stan Bayesian modeling system \cite{Stan1,Stan2}.

```r
require(brms)
require(cluster)
require(MASS)
set.seed(42)
```

Fit the frequentist model using \texttt{lrm} and the Bayesian model using \texttt{brm} in the \texttt{brms} package. For the Bayesian model, the intercept prior was a Student's $t$-distribution with 3 degrees of freedom and the age and sex parameters were given mean zero priors with standard deviations computed to achieve specified tail prior probabilities. Four MCMC chains with 5000 iterations were used with a warm-up of 2500 iterations each resulting in 10000 draws from the posterior distribution.

```r
dd ← datadist(sex.age.response)
options(datadist = 'dd')

# Frequentist model
fit_lrm ← lrm(response ~ sex + age, data=sex.age.response)

# Bayesian model
# Distribute chains across cpu cores:
options(mc.cores=parallel::detectCores())

# Set priors
# Solve for SD such that sex effect has only a 0.025 chance of
# being above 5 (or being below -5)
s1 ← 5 / qnorm(0.975)

# Solve for SD such that 10-year age effect has only 0.025 chance
```
The model summaries for the frequentist and Bayesian models are shown below, with posterior means computed as Bayesian “point estimates.” The parameter estimates are similar for the two approaches. The frequentist 0.95 confidence interval for the age parameter is 0.037 - 0.279 while the Bayesian 0.95 credible interval is 0.051 - 0.270. Similarly, the 0.95 confidence interval for sex is 1.139 - 5.840 and the corresponding Bayesian 0.95 credible interval is 1.377 - 5.336.

Logistic Regression Model

\[
\text{lr}(\text{formula} = \text{response} \sim \text{sex} + \text{age}, \text{data} = \text{sex.age.response})
\]
The figure shows the posterior draws for the \( \text{age} \) and \( \text{sex} \) parameters as well as the trace of the 4 MCMC chains for each parameter and the bivariate posterior distribution. The posterior distributions of each parameter are roughly mound shaped.
and the overlap between chains in the trace plots indicates good convergence. The bivariate density plot indicates moderate correlation between the age and sex parameters.

```
# display posterior densities for age and sex parameters
plot(fit_brms, c("age","sex"), combo=c("dens","trace","hex"))
```

A plot of the marginal effects from the Bayesian model reveals the same pattern as Figure 10.3.

```
# Marginal effects plot
plot(marginal_effects(fit_brms, "age:sex"))
```

```
# Frequentist
# variance-covariance for sex and age parameters
sex_age_vcov ← vcov(fit_lrm)[2:3,2:3]

# Sampling based parameter estimate correlation coefficient
f_cc ← sex_age_vcov[1,2] / (sqrt(sex_age_vcov[1,1]) * sqrt(sex_age_vcov[2,2]))
```
Using the code in the block above, we calculate the frequentist sampling-based parameter estimate correlation coefficient is 0.75 while the linear correlation between the posterior draws for the age and sex parameters is 0.65. Both models indicate a comparable amount of correlation between the parameters, though in difference senses (sampling data vs. sampling posterior distribution of parameters).

The posterior probability that sex has a positive relationship with hospital death is estimated as \( \operatorname{Prob}(\beta_{sex} > 0) = 0.9999 \) while the posterior probability that age has a positive relationship with hospital death is \( \operatorname{Prob}(\beta_{age} > 0) = 0.9991 \) and the probability of both events is \( \operatorname{Prob}(\beta_{sex} > 0 \cap \beta_{age} > 0) = 0.999. \) Even using somewhat skeptical priors centered around 0, male gender
and increasing age are highly likely to be associated with the response.

As seen above, the MCMC algorithm used by \texttt{brms} provides us with samples from the joint posterior distribution of $\beta_{age}$ and $\beta_{sex}$. Unlike frequentist intervals which require the log-likelihood to be approximately quadratic in form, there are no such restrictions placed on the posterior distribution, as it will always be proportional to the product of the likelihood density and the prior, regardless of the likelihood function that is used. In this specific example, we notice that the bivariate density is somewhat skewed – a characteristic that would likely lead to unequal tail coverage probabilities if a symmetric confidence interval is used.

\begin{verbatim}
  ggplot(post_samps, aes(x=b_sexmale, y = b_age)) +
  geom_hex() +
  theme(legend.position="none")
\end{verbatim}

Create a 0.95 bivariate credible interval for the joint distribution of age and sex. Any number of intervals could be drawn, as any region that covers 0.95 of the posterior density could be
accurately be called a 0.95 credible interval. Commonly used: maximum a-posteriori probability (MAP) interval, which seeks to find the region that holds 0.95 of the density, while also having the smallest area. In a 1-dimensional setting, this would translate into having the shortest interval length, and therefore the most precise estimate. The figure below shows the point estimate calculated by brms as well as the corresponding MAP interval.

```r
# Calculate MAP interval
# Code from http://www.sumsar.net/blog/2014/11/
# how-to-summarize-a-2d-posterior-using-a-highest-density-ellipse/
samples ← as.matrix(post_samps)
coverage = 0.95
fit ← cov.mve(samples, quantile.used = round(nrow(samples) * coverage))
points_in_ellipse ← samples[fit$best,]
ellipse_boundary ← predict(ellipsoidhull(points_in_ellipse))
map ← data.frame(ellipse_boundary)
names(map) ← c("y","x")

ggplot(post_samps, aes(x=b_sexmale, y = b_age)) +
  geom_hex() +
  geom_polygon(data = map, aes(x=x,y=y), color = "grey", alpha = 0) +
  geom_point(aes(x = fixef(fit_brms)[,1][2], y = fixef(fit_brms)[,1][3]), color = "grey") +
  theme(legend.position="none")
```

In the above figure, the point estimate does not appear quite
at the point of highest density. This is because \texttt{brms} estimates the posterior mean, rather than the posterior mode. You have the full posterior density, so you can calculate whatever you’d like if you don’t want the mean.

See how the MAP interval compares to a confidence ellipse that we would calculate using a frequentist approach.

```r
# Function takes in variance-covariance matrix (D), point estimates (d), # and a level of significance (alpha)
EllipseDF ← function(D, d, alpha = 0.05) {
  delta = sqrt(eigen(D)$values)
  # Root eigenvalues correspond to the half-lengths of the ellipse
  V = eigen(D)$vectors
  # Eigenvectors give the axes of the confidence ellipse
  R = sqrt(qchisq(1-alpha, df = length(delta)))
  # Scaling factor to get to 0.95 confidence
  a = R*delta[1] # scale the ellipse axes
  b = R*delta[2]
  t ← seq(0, 2*pi, length.out=200)
  # Generate radian measures from 0 to 2pi
  points.proj = V %*% t(cbind(a * cos(t), b * sin(t)))
  # Transform circle into ellipse
  return(data.frame(x = (points.proj)[1,] + d[1],
                    y = (points.proj)[2,] + d[2]))
}
D ← vcov(fit_lrm)[-1,-1]
beta ← coef(fit_lrm)[-1]
ci_ellipse ← EllipseDF(D, beta, alpha = 0.05)

ggplot(post_samps, aes(x=b_sexmale, y = b_age)) +
  geom_hex() +
  geom_polygon(data = map, aes(x=x,y=y), color = "grey", alpha = 0) +
  geom_polygon(data = ci_ellipse, aes(x = x,y = y), color = "red", alpha = 0) +
  geom_point(aes(x = fixef(fit_brms)[,1][2], y = fixef(fit_brms)[,1][3]), color = "grey") +
  geom_point(aes(x = coef(fit_lrm)[2], y = coef(fit_lrm)[3]), color = "red") +
  theme(legend.position="none")
```
Much of the region covered by the confidence ellipse (in red) is shared by the MAP ellipse, but the point estimate appears to target the posterior mode and constructs a symmetric confidence region about it, while the MAP region is not symmetric about the posterior mean\(^e\). As a result, even in this simple 2-parameter logistic regression model, the confidence ellipse is likely to have problematic asymmetric coverage of the true parameter.

**Note:** Most of these results can be easily obtained using the \texttt{rms} package in conjunction with the \texttt{rstan} package, as shown in the Titanic case study.

---

\(^e\)To be fair, there’s no requirement saying that confidence intervals/regions must be symmetric, they just usually are computed that way.
Chapter 11

Case Study in Binary Logistic Regression, Model Selection and Approximation: Predicting Cause of Death

See new Chapter 11 in book.
Chapter 12

Logistic Model Case Study: Survival of Titanic Passengers

Data source: The Titanic Passenger List edited by Michael A. Findlay, originally published in Eaton & Haas (1994) Titanic: Triumph and Tragedy, Patrick Stephens Ltd, and expanded with the help of the Internet community. The original html files were obtained from Philip Hind (1999) (http://atschool.eduweb.co.uk/phind). The dataset was compiled and interpreted by Thomas Cason. It is available in Rand spreadsheet formats from hbiostat.org/data under the name titanic3.
### Descriptive Statistics

```R
require(rms)

options(prType='latex') # for print, summary, anova
getHdata(titanic3) # get dataset from web site
# List of names of variables to analyze
v ← c('pclass','survived','age','sex','sibsp','parch')
t3 ← titanic3[, v]
units(t3$age) ← 'years'
latex(describe(t3), file='')
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>missing</th>
<th>distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>pclass</td>
<td>1309</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>survived</td>
<td>1309</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>age</td>
<td>1046</td>
<td>263</td>
<td>98</td>
</tr>
<tr>
<td>sex</td>
<td>1309</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>sibsp</td>
<td>1309</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>parch</td>
<td>1309</td>
<td>0</td>
<td>8</td>
</tr>
<tr>
<td>/dd</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Survived

<table>
<thead>
<tr>
<th>n</th>
<th>missing</th>
<th>distinct</th>
<th>Info</th>
<th>Sum</th>
<th>Mean</th>
<th>Gmd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1309</td>
<td>0</td>
<td>2</td>
<td>0.708</td>
<td>500</td>
<td>0.382</td>
<td>0.4725</td>
</tr>
</tbody>
</table>

#### Age [years]

<table>
<thead>
<tr>
<th>n</th>
<th>missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1046</td>
<td>263</td>
<td>98</td>
<td>0.999</td>
<td>29.88</td>
<td>16.06</td>
</tr>
</tbody>
</table>

#### Sex

<table>
<thead>
<tr>
<th>n</th>
<th>missing</th>
<th>distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>1309</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

#### Sibsp: Number of Siblings/Spouses Aboard

<table>
<thead>
<tr>
<th>n</th>
<th>missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1309</td>
<td>0</td>
<td>7</td>
<td>0.67</td>
<td>0.4989</td>
<td>0.777</td>
</tr>
</tbody>
</table>

#### Parch: Number of Parents/Children Aboard

<table>
<thead>
<tr>
<th>n</th>
<th>missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
</tr>
</thead>
<tbody>
<tr>
<td>1309</td>
<td>0</td>
<td>8</td>
<td>0.549</td>
<td>0.385</td>
<td>0.6375</td>
</tr>
</tbody>
</table>

### Figure 12.1

```R
dd ← datadist(t3)
# describe distributions of variables to rms
options(datadist='dd')
s ← summary(survived ~ age + sex + pclass +
            cut2(sibsp,0:3) + cut2(parch,0:3), data=t3)
plot(s, main='', subtitles=FALSE)  # Figure 12.1
```
Show 4-way relationships after collapsing levels. Suppress estimates based on < 25 passengers.

```r
# Figure 12.2:
A tn <- transform(t3,  
agec = ifelse(age < 21, 'child', 'adult'),  
sibsp = ifelse(sibsp == 0, 'no sib/sp', 'sib/sp'),  
parch = ifelse(parch == 0, 'no par/child', 'par/child'))  
g <- function(y) if(length(y) < 25) NA else mean(y)  
s <- with(tn, summarize(survived,  
  llist(agec, sex, pclass, sibsp, parch), g))  
# llist, summarize in Hmisc package

ggplot(subset(s, agec != 'NA'),  
aes(x=survived, y=pclass, shape=sex)) +  
  geom_point() + facet_grid(agec ~ sibsp * parch) +  
  xlab('Proportion Surviving') + ylab('Passenger Class') +  
  scale_x_continuous(breaks=c(0, .5, 1))
```

Figure 12.1: Univariable summaries of Titanic survival
Figure 12.2: Multi-way summary of Titanic survival
12.2 Exploring Trends with Nonparametric Regression

# Figure 12.3
b ← scale_size_discrete(range=c(.1, .85))

y1 ← ylab(NULL)
p1 ← ggplot(t3, aes(x=age, y=survived)) +
    histSpikeg(survived ~ age, lowess=TRUE, data=t3) +
    ylim(0,1) + yl
p2 ← ggplot(t3, aes(x=age, y=survived, color=sex)) +
    histSpikeg(survived ~ age + sex, lowess=TRUE, data=t3) +
    ylim(0,1) + yl
p3 ← ggplot(t3, aes(x=age, y=survived, size=pclass)) +
    histSpikeg(survived ~ age + pclass, lowess=TRUE, data=t3) +
    b + ylim(0,1) + yl
p4 ← ggplot(t3, aes(x=age, y=survived, color=sex, size=pclass)) +
    histSpikeg(survived ~ age + sex + pclass, lowess=TRUE, data=t3) +
    b + ylim(0,1) + yl
gridExtra::grid.arrange(p1, p2, p3, p4, ncol=2)  # combine 4

# Figure 12.4

top ← theme(legend.position='top')
p1 ← ggplot(t3, aes(x=age, y=survived, color=cut2(sibsp, 0:2)) +
    stat_plsmo() + b + ylim(0,1) + yl + top +
    scale_color_discrete(name='siblings/spouses')
p2 ← ggplot(t3, aes(x=age, y=survived, color=cut2(parch, 0:2)) +
    stat_plsmo() + b + ylim(0,1) + yl + top +
    scale_color_discrete(name='parents/children')
ggridExtra::grid.arrange(p1, p2, ncol=2)
Figure 12.3: Nonparametric regression (loess) estimates of the relationship between age and the probability of surviving the Titanic, with tick marks depicting the age distribution. The top left panel shows unstratified estimates of the probability of survival. Other panels show nonparametric estimates by various stratifications.
Figure 12.4: Relationship between age and survival stratified by the number of siblings or spouses on board (left panel) or by the number of parents or children of the passenger on board (right panel).
### Table 12.1: Wald Statistics for survived

<table>
<thead>
<tr>
<th>Factor</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>187.15</td>
<td>15</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>59.74</td>
<td>14</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>pclass (Factor+Higher Order Factors)</td>
<td>100.10</td>
<td>20</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>46.51</td>
<td>18</td>
<td>0.0003</td>
</tr>
<tr>
<td>age (Factor+Higher Order Factors)</td>
<td>56.20</td>
<td>32</td>
<td>0.0052</td>
</tr>
<tr>
<td>All Interactions</td>
<td>34.57</td>
<td>28</td>
<td>0.1826</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>28.66</td>
<td>24</td>
<td>0.2331</td>
</tr>
<tr>
<td>sibsp (Factor+Higher Order Factors)</td>
<td>19.67</td>
<td>5</td>
<td>0.0014</td>
</tr>
<tr>
<td>All Interactions</td>
<td>12.13</td>
<td>4</td>
<td>0.0164</td>
</tr>
<tr>
<td>parch (Factor+Higher Order Factors)</td>
<td>3.51</td>
<td>5</td>
<td>0.6217</td>
</tr>
<tr>
<td>All Interactions</td>
<td>3.51</td>
<td>4</td>
<td>0.4761</td>
</tr>
<tr>
<td>sex $\times$ pclass (Factor+Higher Order Factors)</td>
<td>42.43</td>
<td>10</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>sex $\times$ age (Factor+Higher Order Factors)</td>
<td>15.89</td>
<td>12</td>
<td>0.1962</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>14.47</td>
<td>9</td>
<td>0.1066</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>4.17</td>
<td>3</td>
<td>0.2441</td>
</tr>
<tr>
<td>pclass $\times$ age (Factor+Higher Order Factors)</td>
<td>13.47</td>
<td>16</td>
<td>0.6385</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>12.92</td>
<td>12</td>
<td>0.3749</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>6.88</td>
<td>6</td>
<td>0.3324</td>
</tr>
<tr>
<td>age $\times$ sibsp (Factor+Higher Order Factors)</td>
<td>12.13</td>
<td>4</td>
<td>0.0164</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>1.76</td>
<td>3</td>
<td>0.6235</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>1.76</td>
<td>3</td>
<td>0.6235</td>
</tr>
<tr>
<td>age $\times$ parch (Factor+Higher Order Factors)</td>
<td>3.51</td>
<td>4</td>
<td>0.4761</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>1.80</td>
<td>3</td>
<td>0.6147</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>1.80</td>
<td>3</td>
<td>0.6147</td>
</tr>
<tr>
<td>sex $\times$ pclass $\times$ age (Factor+Higher Order Factors)</td>
<td>8.34</td>
<td>8</td>
<td>0.4006</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>7.74</td>
<td>6</td>
<td>0.2581</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>28.66</td>
<td>24</td>
<td>0.2331</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>75.61</td>
<td>30</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>79.49</td>
<td>33</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>241.93</td>
<td>39</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

## 12.3

### Binary Logistic Model with Casewise Deletion of Missing Values

First fit a model that is saturated with respect to age, sex, pclass. Insufficient variation in sibsp, parch to fit complex interactions or nonlinearities.

```r
f1 ← lrm(survived ~ sex*pclass*rcs(age,5) + rcs(age,5)*(sibsp + parch), data=t3) # Table 12.1
print(anova(f1), table.env=TRUE, label='titanic-anova3', size='small')
```

3-way interactions, parch clearly insignificant, so drop

```r
f ← lrm(survived ~ (sex + pclass + rcs(age,5))^2 + rcs(age,5)*sibsp, data=t3)
```
Logistic Regression Model

```r
print(f)
```

### Logistic Regression Model

```r
lrm(formula = survived ~ (sex + pclass + rcs(age, 5))^2 + rcs(age, 5) * sibsp, data = t3)
```

#### Frequencies of Missing Values Due to Each Variable

<table>
<thead>
<tr>
<th></th>
<th>survived</th>
<th>sex</th>
<th>pclass</th>
<th>age</th>
<th>sibsp</th>
</tr>
</thead>
<tbody>
<tr>
<td>obs</td>
<td>1046</td>
<td>0</td>
<td>0</td>
<td>263</td>
<td>0</td>
</tr>
</tbody>
</table>

#### Model Likelihood and Discrimination Indexes

<table>
<thead>
<tr>
<th></th>
<th>LR $\chi^2$</th>
<th>d.f.</th>
<th>Discrimination Indexes</th>
<th>Rank Discrim. Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>553.87</td>
<td>26</td>
<td>$R^2$ 0.555</td>
<td>$C$ 0.878</td>
</tr>
<tr>
<td>max $</td>
<td>\frac{\partial \log L}{\partial \beta}</td>
<td>$</td>
<td>$6 \times 10^{-6}$</td>
<td>$g_2$ 2.427</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$g_r$ 11.325</th>
<th>$\gamma$ 0.758</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$g_p$ 0.365</td>
<td>$\tau_a$ 0.366</td>
</tr>
<tr>
<td></td>
<td>Brier 0.130</td>
<td></td>
</tr>
</tbody>
</table>

#### Estimated Coefficients

|                      | $\hat{\beta}$ | S.E. | Wald $Z$ | Pr($>|Z|)$ |
|----------------------|----------------|------|----------|------------|
| Intercept            | 3.3075         | 1.8427 | 1.79  | 0.0727     |
| sex=male             | -1.1478        | 1.0878 | -1.06  | 0.2914     |
| pclass=2nd           | 6.7309         | 3.9617 | 1.70   | 0.0893     |
| pclass=3rd           | -1.6437        | 1.8299 | -0.90  | 0.3691     |
| age                  | 0.0886         | 0.1346 | 0.66   | 0.5102     |
| age'                 | -0.7410        | 0.6513 | -1.14  | 0.2552     |
| age''                | 4.9264         | 4.0047 | 1.23   | 0.2186     |
| age'''               | -6.6129        | 5.4100 | -1.22  | 0.2216     |
| sibsp                | -1.0446        | 0.3441 | -3.04  | 0.0024     |
| sex=male $\times$ pclass=2nd | -0.7682 | 0.7083 | -1.08  | 0.2781     |
| sex=male $\times$ pclass=3rd | 2.1520  | 0.6214 | 3.46   | 0.0005     |
| sex=male $\times$ age | -0.2191       | 0.0722 | -3.04  | 0.0024     |
| sex=male $\times$ age' | 1.0842        | 0.3886 | 2.79   | 0.0053     |
| sex=male $\times$ age'' | -6.5578      | 2.6511 | -2.47  | 0.0134     |
| sex=male $\times$ age''' | 8.3716        | 3.8532 | 2.17   | 0.0298     |
| pclass=2nd $\times$ age | -0.5446       | 0.2653 | -2.05  | 0.0401     |
| pclass=3rd $\times$ age | -0.1634       | 0.1308 | -1.25  | 0.2118     |
| pclass=2nd $\times$ age' | 1.9156        | 1.0189 | 1.88   | 0.0601     |
| pclass=3rd $\times$ age' | 0.8205        | 0.6091 | 1.35   | 0.1780     |
| pclass=2nd $\times$ age'' | -8.9545       | 5.5027 | -1.63  | 0.1037     |
| pclass=3rd $\times$ age'' | -5.4276       | 3.6475 | -1.49  | 0.1367     |
| pclass=2nd $\times$ age''' | 9.3926        | 6.9559 | 1.35   | 0.1769     |
| pclass=3rd $\times$ age''' | 7.5403        | 4.8519 | 1.55   | 0.1202     |
| age $\times$ sibsp   | 0.0357         | 0.0340 | 1.05   | 0.2933     |
Table 12.2: Wald Statistics for survived

<table>
<thead>
<tr>
<th></th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>199.42</td>
<td>7</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>56.14</td>
<td>6</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>pclass (Factor+Higher Order Factors)</td>
<td>108.73</td>
<td>12</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>42.83</td>
<td>10</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>age (Factor+Higher Order Factors)</td>
<td>47.04</td>
<td>20</td>
<td>0.0006</td>
</tr>
<tr>
<td>All Interactions</td>
<td>24.51</td>
<td>16</td>
<td>0.0789</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>22.72</td>
<td>15</td>
<td>0.0902</td>
</tr>
<tr>
<td>sibsp (Factor+Higher Order Factors)</td>
<td>19.95</td>
<td>5</td>
<td>0.0013</td>
</tr>
<tr>
<td>All Interactions</td>
<td>10.99</td>
<td>4</td>
<td>0.0267</td>
</tr>
<tr>
<td>sex × pclass (Factor+Higher Order Factors)</td>
<td>35.40</td>
<td>2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>sex × age (Factor+Higher Order Factors)</td>
<td>10.08</td>
<td>4</td>
<td>0.0391</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>8.17</td>
<td>3</td>
<td>0.0426</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A,B)$ vs. $AB$</td>
<td>8.17</td>
<td>3</td>
<td>0.0426</td>
</tr>
<tr>
<td>pclass × age (Factor+Higher Order Factors)</td>
<td>6.86</td>
<td>8</td>
<td>0.5516</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>6.11</td>
<td>6</td>
<td>0.4113</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A,B)$ vs. $AB$</td>
<td>6.11</td>
<td>6</td>
<td>0.4113</td>
</tr>
<tr>
<td>age × sibsp (Factor+Higher Order Factors)</td>
<td>10.99</td>
<td>4</td>
<td>0.0267</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>1.81</td>
<td>3</td>
<td>0.6134</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A,B)$ vs. $AB$</td>
<td>1.81</td>
<td>3</td>
<td>0.6134</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>22.72</td>
<td>15</td>
<td>0.0902</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>67.58</td>
<td>18</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>70.68</td>
<td>21</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>253.18</td>
<td>26</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

|                           | $\hat{\beta}$ | S.E. | Wald $Z$  | $Pr(>|Z|)$ |
|---------------------------|----------------|------|-----------|------------|
| age’ × sibsp              | -0.0467        | 0.2213 | -0.21     | 0.8330     |
| age” × sibsp              | 0.5574         | 1.6680 | 0.33      | 0.7382     |
| age”’ × sibsp             | -1.1937        | 2.5711 | -0.46     | 0.6425     |

print(anova(f), table.env=TRUE, label='titanic-anova2', size='small') #12.2

Show the many effects of predictors.

B

```r
p ← Predict(f, age, sex, pclass, sibsp=0, fun=plogis)
ggplot(p)  # F ig. 12.5

ggplot(Predict(f, sibsp, age=c(10,15,20,50), conf.int=FALSE))
## F igure 12.6
```

Note that children having many siblings apparently had lower survival. Married adults had slightly higher survival than unmarried ones.

Validate the model using the bootstrap to check overfitting. Ignoring two very insignificant pooled tests.
Figure 12.5: Effects of predictors on probability of survival of Titanic passengers, estimated for zero siblings or spouses

Figure 12.6: Effect of number of siblings and spouses on the log odds of surviving, for third class males
f ← update(f, x=TRUE, y=TRUE)
# x=TRUE, y=TRUE adds raw data to fit object so can bootstrap
set.seed(131)  # so can replicate re-samples
latex(validate(f, B=200), digits=2, size='Ssize')

<table>
<thead>
<tr>
<th>Index</th>
<th>Original Sample</th>
<th>Training Sample</th>
<th>Test Sample</th>
<th>Optimism</th>
<th>Corrected Index</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{xy}$</td>
<td>0.76</td>
<td>0.77</td>
<td>0.74</td>
<td>0.03</td>
<td>0.72</td>
<td>200</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.55</td>
<td>0.58</td>
<td>0.53</td>
<td>0.05</td>
<td>0.50</td>
<td>200</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.08</td>
<td>0.08</td>
<td>-0.08</td>
<td>200</td>
</tr>
<tr>
<td>Slope</td>
<td>1.00</td>
<td>1.00</td>
<td>0.86</td>
<td>0.14</td>
<td>0.86</td>
<td>200</td>
</tr>
<tr>
<td>$E_{max}$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
<td>200</td>
</tr>
<tr>
<td>$D$</td>
<td>0.53</td>
<td>0.56</td>
<td>0.49</td>
<td>0.06</td>
<td>0.46</td>
<td>200</td>
</tr>
<tr>
<td>$U$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>-0.01</td>
<td>0.01</td>
<td>200</td>
</tr>
<tr>
<td>$Q$</td>
<td>0.53</td>
<td>0.56</td>
<td>0.49</td>
<td>0.07</td>
<td>0.46</td>
<td>200</td>
</tr>
<tr>
<td>$B$</td>
<td>0.13</td>
<td>0.13</td>
<td>0.13</td>
<td>-0.01</td>
<td>0.14</td>
<td>200</td>
</tr>
<tr>
<td>$g$</td>
<td>2.43</td>
<td>2.78</td>
<td>2.40</td>
<td>0.38</td>
<td>2.04</td>
<td>200</td>
</tr>
<tr>
<td>$g_p$</td>
<td>0.37</td>
<td>0.37</td>
<td>0.35</td>
<td>0.02</td>
<td>0.35</td>
<td>200</td>
</tr>
</tbody>
</table>

cal ← calibrate(f, B=200)  # Figure 12.7
plot(cal, subtitles=FALSE)

n=1046  Mean absolute error=0.011  Mean squared error=0.00016
0.9 Quantile of absolute error=0.018

Figure 12.7: Bootstrap overfitting-corrected loess nonparametric calibration curve for casewise deletion model

But moderate problem with missing data
### 12.4 Examining Missing Data Patterns

```r
na.patterns ← naclus(titanic3)
require(rpart)  # Recursive partitioning package

who.na ← rpart(is.na(age) ~ sex + pclass + survived +
            sibsp + parch, data=titanic3, minbucket=15)
naplot(na.patterns, 'na per var')
plot(who.na, margin=.1); text(who.na) # Figure 12.8
plot(na.patterns)
```

![Fraction of NAs in each Variable](image)

**Figure 12.8:** Patterns of missing data. Upper left panel shows the fraction of observations missing on each predictor. Lower panel depicts a hierarchical cluster analysis of missingness combinations. The similarity measure shown on the Y-axis is the fraction of observations for which both variables are missing. Right panel shows the result of recursive partitioning for predicting \( \text{is.na(age)} \). The `rpart` function found only strong patterns according to passenger class.

```r
plot(summary(is.na(age) ~ sex + pclass + survived +
            sibsp + parch, data=t3)) # Figure 12.9

m ← lrm(is.na(age) ~ sex * pclass + survived + sibsp + parch,
            data=t3)
print(m, needspace='3.5in')
```
Logistic Regression Model

\[
\text{lrn(formula = is.na(age) ~ sex * pclass + survived + sibsp + parch, data = t3)}
\]

<table>
<thead>
<tr>
<th>Obs</th>
<th>Model Likelihood Ratio Test</th>
<th>Discrimination Indexes</th>
<th>Rank Discrim. Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>FALSE 1046</td>
<td>LR $\chi^2$ 114.99 d.f. 8</td>
<td>$R^2$ 0.133 $g$ 1.015</td>
<td>$C$ 0.703</td>
</tr>
<tr>
<td>TRUE 263</td>
<td>Pr($&gt; \chi^2$) &lt;0.0001</td>
<td>$g_r$ 2.759 $\gamma$ 0.451</td>
<td>$D_{xy}$ 0.406</td>
</tr>
<tr>
<td>max $</td>
<td>\frac{\partial \log L}{\partial \beta}</td>
<td>5 \times 10^{-6}$</td>
<td></td>
</tr>
<tr>
<td>Overall 1309</td>
<td>Brier 0.148</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
**Table 12.3: Wald Statistics for \( \text{is.na(age)} \)**

<table>
<thead>
<tr>
<th></th>
<th>( \chi^2 )</th>
<th>d.f.</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>5.61</td>
<td>3</td>
<td>0.1324</td>
</tr>
<tr>
<td>All Interactions</td>
<td>5.58</td>
<td>2</td>
<td>0.0614</td>
</tr>
<tr>
<td>pclass (Factor+Higher Order Factors)</td>
<td>68.43</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>5.58</td>
<td>2</td>
<td>0.0614</td>
</tr>
<tr>
<td>survived</td>
<td>0.98</td>
<td>1</td>
<td>0.3232</td>
</tr>
<tr>
<td>sibsp</td>
<td>0.35</td>
<td>1</td>
<td>0.5548</td>
</tr>
<tr>
<td>parch</td>
<td>7.92</td>
<td>1</td>
<td>0.0049</td>
</tr>
<tr>
<td>sex ( \times ) pclass (Factor+Higher Order Factors)</td>
<td>5.58</td>
<td>2</td>
<td>0.0614</td>
</tr>
<tr>
<td>TOTAL</td>
<td>82.90</td>
<td>8</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

| \( \hat{\beta} \)       | S.E.         | Wald Z | Pr(\( > |Z| \)) |
|--------------------------|--------------|--------|---------------|
| Intercept                | -2.2030      | -6.05  | <0.0001       |
| sex=male                 | 0.6440       | 1.63   | 0.1033        |
| pclass=2nd               | -1.0079      | -1.51  | 0.1300        |
| pclass=3rd               | 1.6124       | 4.48   | <0.0001       |
| survived                 | -0.1806      | -0.99  | 0.3232        |
| sibsp                    | 0.0435       | 0.59   | 0.5548        |
| parch                    | -0.3526      | -2.81  | 0.0049        |
| sex=male \( \times \) pclass=2nd | 0.1347       | 0.18   | 0.8583        |
| sex=male \( \times \) pclass=3rd | -0.8563      | -2.03  | 0.0422        |

```r
print(aov(m), table.env=TRUE, label='titanic-anova.na')) # Table 12.3
```

**pclass** and **parch** are the important predictors of missing age.
12.5

Single Conditional Mean Imputation

First try: conditional mean imputation
Default spline transformation for age caused distribution of imputed values to be much different from non-imputed ones; constrain to linear

```
xtrans ← transcan(~ I(age) + sex + pclass + sibsp + parch, 
                 imputed=TRUE, pl=FALSE, pr=FALSE, data=t3)
```

```
summary(xtrans)
```

```
transcan(x = ~I(age) + sex + pclass + sibsp + parch, imputed = TRUE, 
         pr = FALSE, pl = FALSE, data = t3)
```

Iterations: 5

$R^2$ achieved in predicting each variable:

- **age**: 0.264
- **sex**: 0.076
- **pclass**: 0.242
- **sibsp**: 0.249
- **parch**: 0.291

Adjusted $R^2$:

- **age**: 0.260
- **sex**: 0.073
- **pclass**: 0.239
- **sibsp**: 0.245
- **parch**: 0.288

Coefficients of canonical variates for predicting each (row) variable

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>sex</th>
<th>pclass</th>
<th>sibsp</th>
<th>parch</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.92</td>
<td>6.05</td>
<td>-2.02</td>
<td>-2.65</td>
<td></td>
</tr>
<tr>
<td>sex</td>
<td>0.03</td>
<td>-0.56</td>
<td>-0.01</td>
<td>-0.75</td>
<td></td>
</tr>
<tr>
<td>pclass</td>
<td>0.08</td>
<td>-0.26</td>
<td>0.03</td>
<td>0.28</td>
<td></td>
</tr>
<tr>
<td>sibsp</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.03</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>parch</td>
<td>-0.03</td>
<td>-0.30</td>
<td>0.23</td>
<td>0.75</td>
<td></td>
</tr>
</tbody>
</table>

Summary of imputed values

<table>
<thead>
<tr>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>263</td>
</tr>
<tr>
<td>.25</td>
</tr>
<tr>
<td>26.17</td>
</tr>
</tbody>
</table>

lowest: 9.82894 11.75710 13.22440 15.15250 17.28300
highest: 33.24650 34.73840 38.63790 40.83950 42.76770

Starting estimates for imputed values:
# Look at mean imputed values by sex, pclass and observed means
# age.i is age, filled in with conditional mean estimates
age.i ← with(t3, impute(xtrans, age, data=t3))
i ← is.imputed(age.i)
with(t3, tapply(age.i[i], list(sex[i], pclass[i]), mean))

with(t3, tapply(age, list(sex, pclass), mean, na.rm=TRUE))

dd ← datadist(dd, age.i)
f.si ← lrm(survived ~ (sex + pclass + rcs(age.i, 5))^2 +
rcs(age.i, 5)*sibsp, data=t3)
print(f.si, coefs=FALSE)

Logistic Regression Model

lrm(formula = survived ~ (sex + pclass + rcs(age.i, 5))^2 + rcs(age.i, 5) * sibsp, data = t3)

<table>
<thead>
<tr>
<th></th>
<th>Model Likelihood Ratio Test</th>
<th>Discrimination Indexes</th>
<th>Rank Discrim. Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs</td>
<td>LR $\chi^2$</td>
<td>640.85</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>809</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>500</td>
<td></td>
</tr>
<tr>
<td>max</td>
<td>$\frac{\partial \log L}{\partial \beta}$</td>
<td>0.0004</td>
<td>$g_r$</td>
</tr>
</tbody>
</table>

p1 ← Predict(f, age, pclass, sex, sibsp=0, fun=plogis)
p2 ← Predict(f.si, age.i, pclass, sex, sibsp=0, fun=plogis)
p ← rbind('Casewise Deletion'=p1, 'Single Imputation'=p2, rename=c(age.i='age'))
ggplot(p, groups='sex', ylab='Probability of Surviving')
# Figure 12.10

print(anova(f.si), table.env=TRUE, label='titanic-anova.si') # Table 12.4
Figure 12.10: Predicted probability of survival for males from fit using casewise deletion (bottom) and single conditional mean imputation (top). sibsp is set to zero for these predicted values.
Table 12.4: Wald Statistics for survived

<table>
<thead>
<tr>
<th></th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>245.39</td>
<td>7</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>52.85</td>
<td>6</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>pclass (Factor+Higher Order Factors)</td>
<td>112.07</td>
<td>12</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>36.79</td>
<td>10</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>age.i (Factor+Higher Order Factors)</td>
<td>49.32</td>
<td>20</td>
<td>0.0003</td>
</tr>
<tr>
<td>All Interactions</td>
<td>25.62</td>
<td>16</td>
<td>0.0595</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>19.71</td>
<td>15</td>
<td>0.1835</td>
</tr>
<tr>
<td>sibsp (Factor+Higher Order Factors)</td>
<td>22.02</td>
<td>5</td>
<td>0.0005</td>
</tr>
<tr>
<td>All Interactions</td>
<td>12.28</td>
<td>4</td>
<td>0.0154</td>
</tr>
<tr>
<td>sex × pclass (Factor+Higher Order Factors)</td>
<td>30.29</td>
<td>2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>sex × age.i (Factor+Higher Order Factors)</td>
<td>8.91</td>
<td>4</td>
<td>0.0633</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>5.62</td>
<td>3</td>
<td>0.1319</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>5.62</td>
<td>3</td>
<td>0.1319</td>
</tr>
<tr>
<td>pclass × age.i (Factor+Higher Order Factors)</td>
<td>6.05</td>
<td>8</td>
<td>0.6421</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>5.44</td>
<td>6</td>
<td>0.4888</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>5.44</td>
<td>6</td>
<td>0.4888</td>
</tr>
<tr>
<td>age.i × sibsp (Factor+Higher Order Factors)</td>
<td>12.28</td>
<td>4</td>
<td>0.0154</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>2.05</td>
<td>3</td>
<td>0.5614</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>2.05</td>
<td>3</td>
<td>0.5614</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>19.71</td>
<td>15</td>
<td>0.1835</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>67.00</td>
<td>18</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>69.53</td>
<td>21</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>305.74</td>
<td>26</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
12.6

Multiple Imputation

The following uses `aregImpute` with predictive mean matching. By default, `aregImpute` does not transform `age` when it is being predicted from the other variables. Four knots are used to transform `age` when used to impute other variables (not needed here as no other missings were present). Since the fraction of observations with missing age is $\frac{263}{1309} = 0.2$ we use 20 imputations.

```r
set.seed(17)  # so can reproduce random aspects
mi ← aregImpute(~ age + sex + pclass + sibsp + parch + survived, 
                 data=t3, n.impute=20, nk=4, pr=FALSE)
```

Multiple Imputation using Bootstrap and PMM

```r
aregImpute(formula = ~ age + sex + pclass + sibsp + parch + survived, 
            data = t3, n.impute = 20, nk = 4, pr = FALSE)
```

<table>
<thead>
<tr>
<th>n: 1309</th>
<th>p: 6</th>
<th>Imputations: 20</th>
<th>nk: 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of NAs:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>age</td>
<td>sex</td>
<td>pclass</td>
<td>sibsp</td>
</tr>
<tr>
<td>263</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>type</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>s 1</td>
</tr>
<tr>
<td>sex</td>
<td>c 1</td>
</tr>
<tr>
<td>pclass</td>
<td>c 2</td>
</tr>
<tr>
<td>sibsp</td>
<td>s 3</td>
</tr>
<tr>
<td>parch</td>
<td>s 3</td>
</tr>
<tr>
<td>survived</td>
<td>l 1</td>
</tr>
</tbody>
</table>

Transformation of Target Variables Forced to be Linear

R-squares for Predicting Non-Missing Values for Each Variable Using Last Imputations of Predictors

```r
age 0.373
```

# Print the first 10 imputations for the first 10 passengers
# having missing age

mi$imputed$age[1:10, 1:10]

<p>| | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>41</td>
<td>47.0</td>
<td>24</td>
<td>44</td>
<td>60.0</td>
<td>47</td>
<td>28.0</td>
<td>29</td>
<td>49</td>
<td>17</td>
</tr>
<tr>
<td>38</td>
<td>53</td>
<td>44.0</td>
<td>76</td>
<td>59</td>
<td>35.0</td>
<td>39</td>
<td>16.0</td>
<td>54</td>
<td>19</td>
<td>29</td>
</tr>
<tr>
<td>41</td>
<td>45</td>
<td>46.0</td>
<td>28</td>
<td>40</td>
<td>50.0</td>
<td>61</td>
<td>19.0</td>
<td>63</td>
<td>18</td>
<td>61</td>
</tr>
<tr>
<td>47</td>
<td>31</td>
<td>28.5</td>
<td>33</td>
<td>35</td>
<td>61.0</td>
<td>55</td>
<td>45.5</td>
<td>38</td>
<td>41</td>
<td>47</td>
</tr>
<tr>
<td>60</td>
<td>35</td>
<td>40.0</td>
<td>49</td>
<td>41</td>
<td>27.0</td>
<td>36</td>
<td>51.0</td>
<td>2</td>
<td>33</td>
<td>27</td>
</tr>
<tr>
<td>70</td>
<td>30</td>
<td>30.0</td>
<td>16</td>
<td>53</td>
<td>56.0</td>
<td>70</td>
<td>17.0</td>
<td>38</td>
<td>45</td>
<td>51</td>
</tr>
<tr>
<td>71</td>
<td>55</td>
<td>36.0</td>
<td>36</td>
<td>42</td>
<td>42.0</td>
<td>33</td>
<td>65.0</td>
<td>46</td>
<td>39</td>
<td>57</td>
</tr>
<tr>
<td>75</td>
<td>24</td>
<td>36.0</td>
<td>47</td>
<td>49</td>
<td>45.5</td>
<td>47</td>
<td>47.0</td>
<td>38</td>
<td>55</td>
<td>56</td>
</tr>
<tr>
<td>81</td>
<td>60</td>
<td>45.0</td>
<td>46</td>
<td>28</td>
<td>55.0</td>
<td>42</td>
<td>45.0</td>
<td>61</td>
<td>33</td>
<td>45</td>
</tr>
<tr>
<td>107</td>
<td>46</td>
<td>29.0</td>
<td>40</td>
<td>58</td>
<td>71.0</td>
<td>58</td>
<td>47.0</td>
<td>63</td>
<td>61</td>
<td>56</td>
</tr>
</tbody>
</table>

Show the distribution of imputed (black) and actual ages (gray).

plot(mi)
Ecdf(t3$age, add=TRUE, col = 'gray', lwd = 2, subtitles = FALSE)

Figure 12.11: Distributions of imputed and actual ages for the Titanic dataset. Imputed values are in black and actual ages in gray.

Fit logistic models for 20 completed datasets and print the ratio of imputation-corrected variances to average ordinary variances

f.mi ← fit.mult.impute(
  survived ~ (sex + pclass + rcs(age,5))^2 +
  rcs(age,5)*sibsp,
  lrm, mi, data=t3, pr=FALSE)
print(anova(f.mi), table.env=TRUE, label='titanic-anova.mi', size='small')  # Table 12.5

The Wald $\chi^2$ for age is reduced by accounting for imputation
Table 12.5: Wald Statistics for survived

<table>
<thead>
<tr>
<th>Term</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>237.81</td>
<td>7</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>53.44</td>
<td>6</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>pclass (Factor+Higher Order Factors)</td>
<td>113.77</td>
<td>12</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>38.60</td>
<td>10</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>age (Factor+Higher Order Factors)</td>
<td>49.97</td>
<td>20</td>
<td>0.0002</td>
</tr>
<tr>
<td>All Interactions</td>
<td>26.00</td>
<td>16</td>
<td>0.0540</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>23.03</td>
<td>15</td>
<td>0.0835</td>
</tr>
<tr>
<td>sibsp (Factor+Higher Order Factors)</td>
<td>25.08</td>
<td>5</td>
<td>0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>13.42</td>
<td>4</td>
<td>0.0094</td>
</tr>
<tr>
<td>sex × pclass (Factor+Higher Order Factors)</td>
<td>32.70</td>
<td>2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>sex × age (Factor+Higher Order Factors)</td>
<td>10.54</td>
<td>4</td>
<td>0.0322</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>8.40</td>
<td>3</td>
<td>0.0384</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A,B)$ vs. AB</td>
<td>8.40</td>
<td>3</td>
<td>0.0384</td>
</tr>
<tr>
<td>pclass × age (Factor+Higher Order Factors)</td>
<td>5.53</td>
<td>8</td>
<td>0.6996</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>4.67</td>
<td>6</td>
<td>0.5870</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A,B)$ vs. AB</td>
<td>4.67</td>
<td>6</td>
<td>0.5870</td>
</tr>
<tr>
<td>age × sibsp (Factor+Higher Order Factors)</td>
<td>13.42</td>
<td>4</td>
<td>0.0094</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>2.11</td>
<td>3</td>
<td>0.5492</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A,B)$ vs. AB</td>
<td>2.11</td>
<td>3</td>
<td>0.5492</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>23.03</td>
<td>15</td>
<td>0.0835</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>66.42</td>
<td>18</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>69.10</td>
<td>21</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>294.26</td>
<td>26</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

but is increased by using patterns of association with survival status to impute missing age.

Show estimated effects of age by classes.

```r
p1 <- Predict(f.si, age.i, pclass, sex, sibsp=0, fun=plogis)
p2 <- Predict(f.mi, age, pclass, sex, sibsp=0, fun=plogis)
p <- rbind('Single Imputation'=p1, 'Multiple Imputation'=p2, rename=c(age.i='age'))
ggplot(p, groups='sex', ylab='Probability of Surviving')
# Figure 12.12
```
Figure 12.12: Predicted probability of survival for males from fit using single conditional mean imputation again (top) and multiple 
random draw imputation (bottom). Both sets of predictions are for sibsp=0.
12.7 Summarizing the Fitted Model

Show odds ratios for changes in predictor values

```r
# Get predicted values for certain types of passengers
s <- summary(f.mi, age=c(1,30), sibsp=0:1)
# override default ranges for 3 variables
plot(s, log=TRUE, main='')  # Figure 12.13

phat <- predict(f.mi, 
    combos <- expand.grid(age=c(2,21,50), sex=levels(t3$sex), 
        pclass=levels(t3$pclass), 
        sibsp=0), type='fitted')
# Can also use: Predict(f.mi, age=c(2,21,50), sex, pclass, 
# sibsp=0, fun=plogis)$yhat
options(digits=1)
data.frame(combos, phat)
```

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>pclass</th>
<th>sibsp</th>
<th>phat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>female</td>
<td>1st</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>female</td>
<td>1st</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>female</td>
<td>1st</td>
<td>0.97</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>male</td>
<td>1st</td>
<td>0.89</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>male</td>
<td>1st</td>
<td>0.47</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>male</td>
<td>1st</td>
<td>0.26</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>female</td>
<td>2nd</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>21</td>
<td>female</td>
<td>2nd</td>
<td>0.90</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>female</td>
<td>2nd</td>
<td>0.81</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>male</td>
<td>2nd</td>
<td>1.00</td>
</tr>
<tr>
<td>11</td>
<td>21</td>
<td>male</td>
<td>2nd</td>
<td>0.08</td>
</tr>
<tr>
<td>12</td>
<td>50</td>
<td>male</td>
<td>2nd</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Figure 12.13: Odds ratios for some predictor settings
We can also get predicted values by creating an S function that will evaluate the model on demand.

```
pred.logit ← Function(f.mi)
# Note: if don't define sibsp to pred.logit, defaults to 0
# normally just type the function name to see its body
latex(pred.logit, file='', type='Sinput', size='small',
width.cutoff=49)
```

```
pred.logit ← function (sex = "male", pclass = "3rd", age = 28,
sibsp = 0)
{
  3.373079 - 1.0484795 * (sex == "male") + 5.8078168 *
  (pclass == "2nd") - 1.4370771 * (pclass ==
  "3rd") + 0.078347318 * age - 0.00027150053 *
  pmax(age = 6, 0)^3 + 0.0017093284 * pmax(age =
  21, 0)^3 + 0.00010126373 * pmax(age = 36, 0)^3 - 7.5314668e-05 *
  pmax(age = 56, 0)^3 - 1.799235 * sibsp +
  (sex == "male") * (-0.47754081 * (pclass ==
  "2nd") + 2.0665924 * (pclass == "3rd")) +
  (sex == "male") * (-0.21884197 * age + 0.00042463444 *
  pmax(age = 6, 0)^3 - 0.0023860246 * pmax(age =
  21, 0)^3 + 0.0030996682 * pmax(age = 28,
  0)^3 - 0.0012255784 * pmax(age = 36, 0)^3 +
  8.7300463e-05 * pmax(age = 56, 0)^3) +
  (pclass == "2nd") * (-0.47647131 * age + 0.00068483 *
  pmax(age = 6, 0)^3 - 0.0029990417 * pmax(age =
  21, 0)^3 + 0.0031221255 * pmax(age = 28,
  0)^3 - 0.00083472782 * pmax(age = 36,
  0)^3 + 2.6813959e-05 * pmax(age = 56,
  0)^3) + (pclass == "3rd") * (-0.16335774 *
  age + 0.0030986546 * pmax(age = 6, 0)^3 -
  0.0018174716 * age + 2.0590588e-05 *
  pmax(age = 21, 0)^3 + 0.00083472782 * pmax(age =
  28, 0)^3 + 0.0023860246 * pmax(age = 36,
  0)^3 + 9.8357307e-05 * pmax(age = 56, 0)^3)
}
```

```
# Run the newly created function
plogis(pred.logit(age=c(2,21,50), sex='male', pclass='3rd'))
```

```
[1] 0.912648 0.134219 0.050343
```
A nomogram could be used to obtain predicted values manually, but this is not feasible when so many interaction terms are present.
12.8 Bayesian Analysis

- Repeat the multiple imputation-based approach but using a Bayesian binary logistic model.

- Using default `blrm` function normal priors on regression coefficients with zero mean and large SD making the priors almost flat.

- `blrm` uses the `rstan` package that provides the full power of Stan to R.

- Could use smaller SDs to get penalized estimates.

- Using 4 independent Markov chain Hamiltonian posterior sampling procedures each with 1000 burn-in iterations that are discarded, and 1000 “real” iterations for a total of 4000 posterior sample draws.

- Use the first 10 multiple imputations already developed above (object `mi`), running the Bayesian procedure separately for 10 completed datasets.

- Merely have to stack the posterior draws into one giant sample to account for imputation and get correct posterior distribution.

# Use all available CPU cores. Each chain will be run on its
Bayesian Logistic Regression Model

\[
\text{stackMI(formula = survived} \sim (\text{sex} + \text{pclass} + \text{rcs(age, 5)})^2 + \\
\text{rcs(age, 5)} \times \text{sibsp,} \\
\text{fitter = blrm, xtrans = mi, data = t3, n.impute = 10, refresh = 25})
\]

<table>
<thead>
<tr>
<th></th>
<th>Mixed Calibration/ Discrimination Indexes</th>
<th>Discrimination Indexes</th>
<th>Rank Discrim. Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOO log L -573.79±23.41</td>
<td>2.54 [2.229, 2.933]</td>
<td>C 0.868 [0.863, 0.873]</td>
</tr>
<tr>
<td></td>
<td>LOO IC 1147.58±46.83</td>
<td>0.358 [0.342, 0.374]</td>
<td>Dxy 0.736 [0.726, 0.746]</td>
</tr>
<tr>
<td></td>
<td>Effective p 34.97±4.96</td>
<td>EV 0.464 [0.425, 0.507]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>B 0.134 [0.132, 0.136]</td>
<td>v 5.9 [3.906, 8.215]</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>vp 0.109 [0.1, 0.119]</td>
<td></td>
</tr>
<tr>
<td>Obs</td>
<td>1309</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>809</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Draws</td>
<td>40000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chains</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Imputations</td>
<td>10</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p</td>
<td>26</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Mean $\hat{\beta}$</th>
<th>Median $\hat{\beta}$</th>
<th>S.E.</th>
<th>Lower</th>
<th>Upper</th>
<th>Pr($\hat{\beta} &gt; 0$)</th>
<th>Symmetry</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.2225</td>
<td>4.0273</td>
<td>2.0970</td>
<td>0.4260</td>
<td>8.5633</td>
<td>0.9914</td>
<td>1.31</td>
</tr>
<tr>
<td>sex=male</td>
<td>-1.1510</td>
<td>-1.1419</td>
<td>1.0628</td>
<td>-3.2847</td>
<td>0.9010</td>
<td>0.1364</td>
<td>0.96</td>
</tr>
<tr>
<td>pclass=2nd</td>
<td>6.8808</td>
<td>6.4263</td>
<td>4.2477</td>
<td>-0.6601</td>
<td>15.9072</td>
<td>0.9693</td>
<td>1.33</td>
</tr>
<tr>
<td>pclass=3rd</td>
<td>-2.1324</td>
<td>-1.9398</td>
<td>2.0892</td>
<td>-6.3605</td>
<td>1.7612</td>
<td>0.1415</td>
<td>0.77</td>
</tr>
<tr>
<td>age</td>
<td>0.0388</td>
<td>0.0491</td>
<td>0.1456</td>
<td>-0.2545</td>
<td>0.3178</td>
<td>0.6345</td>
<td>0.82</td>
</tr>
<tr>
<td>age’</td>
<td>-0.5732</td>
<td>-0.5893</td>
<td>0.7114</td>
<td>-1.9683</td>
<td>0.8299</td>
<td>0.2057</td>
<td>1.05</td>
</tr>
<tr>
<td>age”</td>
<td>3.8672</td>
<td>3.8878</td>
<td>4.0961</td>
<td>-4.1161</td>
<td>11.9727</td>
<td>0.8280</td>
<td>1.00</td>
</tr>
<tr>
<td>age”’</td>
<td>-5.4720</td>
<td>-5.4669</td>
<td>5.6795</td>
<td>-16.8849</td>
<td>5.3680</td>
<td>0.1676</td>
<td>0.99</td>
</tr>
<tr>
<td>sibsp</td>
<td>-1.2769</td>
<td>-1.2611</td>
<td>0.3490</td>
<td>-1.9597</td>
<td>-0.5984</td>
<td>0.0000</td>
<td>0.88</td>
</tr>
<tr>
<td>sex=male × pclass=2nd</td>
<td>-0.5118</td>
<td>-0.5247</td>
<td>0.7325</td>
<td>-1.8819</td>
<td>0.9701</td>
<td>0.2392</td>
<td>1.05</td>
</tr>
<tr>
<td>sex=male × pclass=3rd</td>
<td>2.2253</td>
<td>2.2022</td>
<td>0.6309</td>
<td>1.0169</td>
<td>3.4720</td>
<td>1.0000</td>
<td>1.13</td>
</tr>
<tr>
<td>sex=male × age</td>
<td>-0.2227</td>
<td>-0.2226</td>
<td>0.0674</td>
<td>-0.3578</td>
<td>-0.0938</td>
<td>0.0006</td>
<td>0.99</td>
</tr>
<tr>
<td>sex=male × age’</td>
<td>1.0600</td>
<td>1.0574</td>
<td>0.3905</td>
<td>0.2804</td>
<td>1.8159</td>
<td>0.9964</td>
<td>1.02</td>
</tr>
<tr>
<td>sex=male × age”</td>
<td>-5.8017</td>
<td>-5.7881</td>
<td>2.5897</td>
<td>-10.9156</td>
<td>-0.7616</td>
<td>0.0120</td>
<td>0.98</td>
</tr>
<tr>
<td>sex=male × age”’</td>
<td>7.3004</td>
<td>7.2669</td>
<td>3.9604</td>
<td>-0.1971</td>
<td>15.3274</td>
<td>0.9684</td>
<td>1.02</td>
</tr>
<tr>
<td>pclass=2nd × age</td>
<td>-0.5509</td>
<td>-0.5253</td>
<td>0.2762</td>
<td>-1.1229</td>
<td>-0.0437</td>
<td>0.0114</td>
<td>0.78</td>
</tr>
<tr>
<td>pclass=3rd × age</td>
<td>-0.1314</td>
<td>-0.1418</td>
<td>0.1436</td>
<td>-0.4068</td>
<td>0.1555</td>
<td>0.1694</td>
<td>1.21</td>
</tr>
<tr>
<td>pclass=2nd × age’</td>
<td>1.9861</td>
<td>1.9285</td>
<td>1.1041</td>
<td>-0.1511</td>
<td>4.1750</td>
<td>0.9730</td>
<td>1.15</td>
</tr>
<tr>
<td>pclass=3rd × age’</td>
<td>0.6838</td>
<td>0.6999</td>
<td>0.6798</td>
<td>-0.6634</td>
<td>2.0090</td>
<td>0.8448</td>
<td>0.93</td>
</tr>
<tr>
<td>pclass=2nd × age”</td>
<td>-8.6655</td>
<td>-8.4927</td>
<td>5.7150</td>
<td>-20.2487</td>
<td>2.1550</td>
<td>0.0583</td>
<td>0.93</td>
</tr>
<tr>
<td>pclass=3rd × age”’</td>
<td>-4.1869</td>
<td>-4.2309</td>
<td>3.8214</td>
<td>-11.7962</td>
<td>3.2183</td>
<td>0.1360</td>
<td>1.03</td>
</tr>
</tbody>
</table>
• Note that fit indexes have HPD uncertainty intervals

• Everthing above accounts for imputation

• Look at diagnostics

\texttt{stanDx(bt)}

Diagnostics for each of 10 imputations

Iterations: 2000 on each of 4 chains, with 4000 posterior distribution samples saved

For each parameter, n\textsubscript{eff} is a crude measure of effective sample size and Rhat is the potential scale reduction factor on split chains (at convergence, Rhat=1)

<table>
<thead>
<tr>
<th>Imputation 1</th>
<th>n\textsubscript{eff}</th>
<th>Rhat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2374</td>
<td>1.000</td>
</tr>
<tr>
<td>sex=\text{male}</td>
<td>4513</td>
<td>1.000</td>
</tr>
<tr>
<td>pclass=2nd</td>
<td>1125</td>
<td>1.003</td>
</tr>
<tr>
<td>pclass=3rd</td>
<td>4437</td>
<td>1.000</td>
</tr>
<tr>
<td>age</td>
<td>4295</td>
<td>0.999</td>
</tr>
<tr>
<td>age'</td>
<td>4452</td>
<td>1.000</td>
</tr>
<tr>
<td>age''</td>
<td>4444</td>
<td>1.001</td>
</tr>
<tr>
<td>age''''</td>
<td>4339</td>
<td>1.001</td>
</tr>
<tr>
<td>sibsp</td>
<td>4119</td>
<td>1.000</td>
</tr>
<tr>
<td>sex=\text{male} * pclass=2nd</td>
<td>2884</td>
<td>1.000</td>
</tr>
<tr>
<td>sex=\text{male} * pclass=3rd</td>
<td>2945</td>
<td>1.000</td>
</tr>
<tr>
<td>sex=\text{male} * age</td>
<td>5237</td>
<td>0.999</td>
</tr>
<tr>
<td>sex=\text{male} * age'</td>
<td>4804</td>
<td>1.000</td>
</tr>
<tr>
<td>sex=\text{male} * age''</td>
<td>4445</td>
<td>1.000</td>
</tr>
<tr>
<td>sex=\text{male} * age''''</td>
<td>4122</td>
<td>1.000</td>
</tr>
<tr>
<td>pclass=2nd * age</td>
<td>1177</td>
<td>1.003</td>
</tr>
<tr>
<td>pclass=3rd * age</td>
<td>4469</td>
<td>1.000</td>
</tr>
<tr>
<td>pclass=2nd * age'</td>
<td>1438</td>
<td>1.002</td>
</tr>
<tr>
<td>pclass=3rd * age'</td>
<td>4626</td>
<td>1.000</td>
</tr>
<tr>
<td>pclass=2nd * age''</td>
<td>1831</td>
<td>1.002</td>
</tr>
<tr>
<td>pclass=3rd * age''</td>
<td>4592</td>
<td>1.001</td>
</tr>
<tr>
<td>pclass=2nd * age''''</td>
<td>2366</td>
<td>1.001</td>
</tr>
<tr>
<td>pclass=3rd * age''''</td>
<td>4474</td>
<td>1.001</td>
</tr>
<tr>
<td>Imputation 1: age * sibsp</td>
<td>3744 1.000</td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------</td>
<td></td>
</tr>
<tr>
<td>Imputation 1: age’ * sibsp</td>
<td>3949 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 1: age’’ * sibsp</td>
<td>4404 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 1: age’’’ * sibsp</td>
<td>4704 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 2: Intercept</td>
<td>2462 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 2: sex=male</td>
<td>4021 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 2: pclass=2nd</td>
<td>1464 1.005</td>
<td></td>
</tr>
<tr>
<td>Imputation 2: pclass=3rd</td>
<td>4330 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 2: age</td>
<td>4671 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 2: age’</td>
<td>4635 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 2: age’’</td>
<td>4378 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 2: age’’’</td>
<td>4204 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 2: age * sibsp</td>
<td>3653 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 2: age’ * sibsp</td>
<td>3276 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 2: age’’ * sibsp</td>
<td>3651 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 2: age’’’ * sibsp</td>
<td>4088 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: Intercept</td>
<td>2448 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: sex=male</td>
<td>4862 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: pclass=2nd</td>
<td>788 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: pclass=3rd</td>
<td>4099 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: age</td>
<td>3916 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: age’</td>
<td>3934 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: age’’</td>
<td>4045 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: age’’’</td>
<td>4192 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: sibsp</td>
<td>4364 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: sex=male * pclass=2nd</td>
<td>3287 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: sex=male * pclass=3rd</td>
<td>3165 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: sex=male * age</td>
<td>5165 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: sex=male * age’</td>
<td>4807 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: sex=male * age’’</td>
<td>4904 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: sex=male * age’’’</td>
<td>5100 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: pclass=2nd * age</td>
<td>855 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: pclass=3rd * age</td>
<td>4276 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: pclass=2nd * age’</td>
<td>1221 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: pclass=3rd * age’</td>
<td>4226 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: pclass=2nd * age’’</td>
<td>1681 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: pclass=3rd * age’’</td>
<td>4224 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: pclass=2nd * age’’’</td>
<td>2357 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: pclass=3rd * age’’’</td>
<td>4233 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: age * sibsp</td>
<td>3576 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: age’ * sibsp</td>
<td>3079 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: age’’ * sibsp</td>
<td>3243 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 3: age``` * sibsp</td>
<td>3468 1.001</td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>-----------</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: Intercept</td>
<td>2144 1.002</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: sex=male</td>
<td>4442 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: pclass=2nd</td>
<td>882 1.003</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: pclass=3rd</td>
<td>4370 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: age</td>
<td>4099 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: age'</td>
<td>4195 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: age''</td>
<td>4278 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: age'''</td>
<td>4320 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: sibsp</td>
<td>4706 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: sex=male * pclass=2nd</td>
<td>3184 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: sex=male * pclass=3rd</td>
<td>3110 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: sex=male * age</td>
<td>4900 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: sex=male * age'</td>
<td>4795 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: sex=male * age''</td>
<td>4639 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: sex=male * age'''</td>
<td>4549 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: pclass=2nd * age</td>
<td>922 1.003</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: pclass=3rd * age</td>
<td>4300 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: pclass=2nd * age'</td>
<td>1119 1.003</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: pclass=3rd * age''</td>
<td>4393 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: pclass=2nd * age'''</td>
<td>1429 1.002</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: pclass=3rd * age''</td>
<td>4456 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: pclass=2nd * age'''</td>
<td>2005 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: pclass=3rd * age'''</td>
<td>4493 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: age * sibsp</td>
<td>3823 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: age' * sibsp</td>
<td>3265 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: age'' * sibsp</td>
<td>3515 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 4: age''' * sibsp</td>
<td>3987 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: Intercept</td>
<td>1987 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: sex=male</td>
<td>4095 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: pclass=2nd</td>
<td>1019 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: pclass=3rd</td>
<td>4714 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: age</td>
<td>3537 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: age'</td>
<td>3211 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: age''</td>
<td>3354 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: age'''</td>
<td>3561 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: sibsp</td>
<td>4209 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: sex=male * pclass=2nd</td>
<td>2508 1.002</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: sex=male * pclass=3rd</td>
<td>2595 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: sex=male * age</td>
<td>5372 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: sex=male * age'</td>
<td>4857 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: sex=male * age''</td>
<td>4787 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: sex=male * age'''</td>
<td>4642 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: pclass=2nd * age</td>
<td>1051 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: pclass=3rd * age</td>
<td>4230 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: pclass=2nd * age'</td>
<td>1252 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: pclass=3rd * age''</td>
<td>3881 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: pclass=2nd * age'''</td>
<td>1551 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: pclass=3rd * age''</td>
<td>3942 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: pclass=2nd * age'''</td>
<td>1949 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: pclass=3rd * age'''</td>
<td>3908 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: age * sibsp</td>
<td>3121 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: age' * sibsp</td>
<td>2784 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: age'' * sibsp</td>
<td>3161 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 5: age''' * sibsp</td>
<td>3489 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: Intercept</td>
<td>2218 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: sex=male</td>
<td>4144 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: pclass=2nd</td>
<td>1313 1.000</td>
<td></td>
</tr>
<tr>
<td>--------------------------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: pclass=3rd</td>
<td>3497 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: age</td>
<td>3831 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: age'</td>
<td>4121 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: age''</td>
<td>4131 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: age'''</td>
<td>4139 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: sibsp</td>
<td>4135 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: sex=male * pclass=2nd</td>
<td>3066 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: sex=male * pclass=3rd</td>
<td>2779 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: sex=male * age</td>
<td>5091 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: sex=male * age'</td>
<td>4873 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: sex=male * age''</td>
<td>4618 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: sex=male * age'''</td>
<td>4427 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: pclass=2nd * age</td>
<td>1382 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: pclass=3rd * age</td>
<td>3898 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: pclass=2nd * age'</td>
<td>1763 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: pclass=3rd * age'</td>
<td>4223 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: pclass=2nd * age''</td>
<td>2331 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: pclass=3rd * age''</td>
<td>4298 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: pclass=2nd * age''''</td>
<td>3011 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: pclass=3rd * age''''</td>
<td>4286 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: age * sibsp</td>
<td>4167 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: age' * sibsp</td>
<td>4372 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: age'' * sibsp</td>
<td>4621 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 6: age''' * sibsp</td>
<td>4795 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: Intercept</td>
<td>1934 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: sex=male</td>
<td>4205 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: pclass=2nd</td>
<td>893 1.008</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: pclass=3rd</td>
<td>3493 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: age</td>
<td>3162 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: age'</td>
<td>3549 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: age''</td>
<td>3922 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: age'''</td>
<td>4192 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: sibsp</td>
<td>3274 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: sex=male * pclass=2nd</td>
<td>2941 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: sex=male * pclass=3rd</td>
<td>2659 1.002</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: sex=male * age</td>
<td>5372 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: sex=male * age'</td>
<td>5071 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: sex=male * age''</td>
<td>4841 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: sex=male * age'''</td>
<td>4693 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: pclass=2nd * age</td>
<td>960 1.007</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: pclass=3rd * age</td>
<td>3617 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: pclass=2nd * age'</td>
<td>1274 1.006</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: pclass=3rd * age'</td>
<td>3952 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: pclass=2nd * age''</td>
<td>1770 1.005</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: pclass=3rd * age''</td>
<td>4272 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: pclass=2nd * age''''</td>
<td>2291 1.004</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: pclass=3rd * age''''</td>
<td>4361 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: age * sibsp</td>
<td>3020 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: age' * sibsp</td>
<td>3392 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: age'' * sibsp</td>
<td>3857 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 7: age''' * sibsp</td>
<td>4202 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: Intercept</td>
<td>2236 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: sex=male</td>
<td>4113 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: pclass=2nd</td>
<td>1209 1.002</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: pclass=3rd</td>
<td>3708 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: age</td>
<td>3643 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: age'</td>
<td>3956 1.000</td>
<td></td>
</tr>
<tr>
<td>-----------------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: age''</td>
<td>4111 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: age'''</td>
<td>4162 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: sibsp</td>
<td>4601 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: sex=male * pclass=2nd</td>
<td>3118 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: sex=male * pclass=3rd</td>
<td>2835 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: sex=male * age</td>
<td>5532 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: sex=male * age'</td>
<td>5351 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: sex=male * age''</td>
<td>5202 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: sex=male * age'''</td>
<td>4937 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: pclass=2nd * age</td>
<td>1254 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: pclass=3rd * age</td>
<td>3965 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: pclass=2nd * age'</td>
<td>1547 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: pclass=3rd * age'</td>
<td>4161 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: pclass=2nd * age''</td>
<td>1969 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: pclass=3rd * age''</td>
<td>4373 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: pclass=2nd * age'''</td>
<td>2517 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: pclass=3rd * age'''</td>
<td>4520 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: age * sibsp</td>
<td>3995 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: age' * sibsp</td>
<td>3738 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: age'' * sibsp</td>
<td>4014 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 8: age''' * sibsp</td>
<td>4262 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: Intercept</td>
<td>2274 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: sex=male</td>
<td>4400 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: pclass=2nd</td>
<td>1152 1.002</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: pclass=3rd</td>
<td>3877 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: age</td>
<td>3776 1.002</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: age'</td>
<td>3966 1.002</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: age''</td>
<td>4069 1.002</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: age'''</td>
<td>4211 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: sibsp</td>
<td>4510 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: sex=male * pclass=2nd</td>
<td>3337 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: sex=male * pclass=3rd</td>
<td>2993 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: sex=male * age</td>
<td>5176 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: sex=male * age'</td>
<td>5356 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: sex=male * age''</td>
<td>5298 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: sex=male * age'''</td>
<td>5218 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: pclass=2nd * age</td>
<td>1254 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: pclass=3rd * age</td>
<td>4053 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: pclass=2nd * age'</td>
<td>1536 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: pclass=3rd * age'</td>
<td>4191 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: pclass=2nd * age''</td>
<td>1999 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: pclass=3rd * age''</td>
<td>4344 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: pclass=2nd * age'''</td>
<td>2634 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: pclass=3rd * age'''</td>
<td>4429 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: age * sibsp</td>
<td>3931 1.002</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: age' * sibsp</td>
<td>3780 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: age'' * sibsp</td>
<td>4169 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 9: age''' * sibsp</td>
<td>4521 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: Intercept</td>
<td>2976 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: sex=male</td>
<td>4422 0.999</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: pclass=2nd</td>
<td>2643 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: pclass=3rd</td>
<td>4434 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: age</td>
<td>3912 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: age'</td>
<td>3409 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: age''</td>
<td>3333 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: age'''</td>
<td>3546 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 10:</td>
<td>sibsp 5241 1.000</td>
<td></td>
</tr>
<tr>
<td>-----------------------------</td>
<td>------------------</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: sex=male * pclass=2nd</td>
<td>2613 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: sex=male * pclass=3rd</td>
<td>2249 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: sex=male * age</td>
<td>6345 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: sex=male * age'</td>
<td>5310 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: sex=male * age''</td>
<td>4868 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: sex=male * age'''</td>
<td>4637 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: pclass=2nd * age</td>
<td>2758 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: pclass=3rd * age</td>
<td>4418 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: pclass=2nd * age'</td>
<td>2801 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: pclass=3rd * age'</td>
<td>3929 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: pclass=2nd * age''</td>
<td>3105 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: pclass=3rd * age''</td>
<td>3804 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: pclass=2nd * age'''</td>
<td>3857 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: pclass=3rd * age'''</td>
<td>3956 1.001</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: age * sibsp</td>
<td>3534 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: age' * sibsp</td>
<td>3038 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: age'' * sibsp</td>
<td>3457 1.000</td>
<td></td>
</tr>
<tr>
<td>Imputation 10: age''' * sibsp</td>
<td>3849 1.000</td>
<td></td>
</tr>
</tbody>
</table>

# Look at convergence of only 2 parameters

```r
stanDxplot(bt, c('sex=males', 'pclass=3rd', 'age'), rev=TRUE)
```
- Difficult to see but there are 40 traces (10 imputations × 4 chains)
- Diagnostics look good; posterior samples can be trusted
- Plot posterior densities for select parameters
 Also shows the 10 densities before stacking

\[
\text{plot(bt, c('sex=male', 'pclass=3rd', 'age'), nrow=2)}
\]

- Plot partial effect plots with 0.95 highest posterior density intervals

\[
p \leftarrow \text{Predict(bt, age, sex, pclass, sibsp=0, fun=plogis)}
\]
\[
\text{ggplot(p)}
\]
• Compute approximate measure of explained outcome variation for predictors

```r
plot(anova(bt))
```
• Contrast second class males and females, both at 5 years and 30 years of age, all other things being equal

• Compute 0.95 HPD interval for the contrast and a joint uncertainty region

• Compute \( P(\text{both contrasts} < 0) \), both \(-2\), and \( P(\text{either one} < 0) \)

```r
k ← contrast(bt, list(sex='male', age=c(5, 30), pclass='2nd'),
              list(sex='female', age=c(5, 30), pclass='2nd'),
              cnames = c('age 5 M-F', 'age 30 M-F'))

   age Contrast     S.E.    Lower    Upper  Pr(Contrast>0)
1    age 5 M-F    5  -2.7761 0.77920  -4.2387  -1.1877   2e-04
2    age 30 M-F  30  -4.1500 0.51818  -5.1883  -3.1573  0e+00

Intervals are 0.95 highest posterior density intervals
Contrast is the posterior mean
```

```r
plot(k)
```

```
plot(k, bivar=TRUE)  # assumes an ellipse
plot(k, bivar=TRUE, bivarmethod='kernel')  # doesn't
P ← PostF(k, pr=TRUE)
```
Contrast names: age 5 M-F, age 30 M-F

\[ P\{\text{age 5 M-F} < 0 \& \text{age 30 M-F} < 0\} \]

[1] 0.9998

\[ P\{\text{age 5 M-F} < -2 \& \text{age 30 M-F} < -2\} \]

[1] 0.83955

\[ P\{\text{age 5 M-F} < 0 \mid \text{age 30 M-F} < 0\} \]

[1] 1

---

**R Software Used**

<table>
<thead>
<tr>
<th>Package</th>
<th>Purpose</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hmisc</td>
<td>Miscellaneous functions</td>
<td>summary, plsmo, naclus, llist, latex, summarize, Dotplot, describe</td>
</tr>
<tr>
<td>Hmisc</td>
<td>Imputation</td>
<td>transcan, impute, fit.mult.impute, aregImpute, stackMI</td>
</tr>
<tr>
<td>rms</td>
<td>Modeling</td>
<td>datadist, lrm, blrm, rcs, plot, summary, nomogram, Function, anova</td>
</tr>
<tr>
<td></td>
<td>Model presentation</td>
<td>Predict, summary, contrast</td>
</tr>
<tr>
<td></td>
<td>Estimation</td>
<td>validate, calibrate</td>
</tr>
<tr>
<td></td>
<td>Model validation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Misc. Bayesian</td>
<td>stanDx, stanDxplot, plot</td>
</tr>
<tr>
<td>rpart*</td>
<td>Recursive partitioning</td>
<td>rpart</td>
</tr>
</tbody>
</table>

*Written by Atkinson & Therneau*
Chapter 13

Ordinal Logistic Regression

13.1 Background

- Levels of $Y$ are ordered; no spacing assumed

- If no model assumed, one can still assess association between $X$ and $Y$

- Example: $Y = 0, 1, 2$ corresponds to no event, heart attack, death. Test of association between race (3 levels) and outcome (3 levels) can be obtained from a $2 \times 2$ d.f. $\chi^2$ test for a contingency table

- If willing to assuming an ordering of $Y$ and a model, can test for association using $2 \times 1$ d.f.

- Proportional odds model: generalization of Wilcoxon-Mann-Whitney-Kruskal-Wallis-Spearman
• Can have \( n \) categories for \( n \) observations!

• Continuation ratio model: discrete proportional hazards model
## Ordinality Assumption

- Assume $X$ is linearly related to some appropriate log odds.
- Estimate mean $X \mid Y$ with and without assuming the model holds.
13.3 Proportional Odds Model

13.3.1 Model

- Walker & Duncan [224] — most popular ordinal response model

- For convenience $Y = 0, 1, 2, \ldots, k$

$$ \Pr[Y \geq j | X] = \frac{1}{1 + \exp[-(\alpha_j + X\beta)]}, $$

where $j = 1, 2, \ldots, k$.

- $\alpha_j$ is the logit of $\text{Prob}[Y \geq j]$ when all $X$s are zero

- $\text{Odds}[Y \geq j | X] = \exp(\alpha_j + X\beta)$

- $\text{Odds}[Y \geq j | X_m = a + 1] / \text{Odds}[Y \geq j | X_m = a] = e^{\beta_m}$

- Same odds ratio $e^{\beta_m}$ for any $j = 1, 2, \ldots, k$

- $\text{Odds}[Y \geq j | X] / \text{Odds}[Y \geq v | X] = \frac{e^{\alpha_j + X\beta}}{e^{\alpha_v + X\beta}} = e^{\alpha_j - \alpha_v}$

- $\text{Odds}[Y \geq j | X] = constant \times \text{Odds}[Y \geq v | X]$

- Assumes OR for 1 unit increase in age is the same when considering the probability of death as when considering the
probability of death or heart attack

- PO model only uses ranks of $Y$; same $\hat{\beta}$s if transform $Y$; is robust to outliers

### 13.3.2 Assumptions and Interpretation of Parameters

### 13.3.3 Estimation

### 13.3.4 Residuals

- Construct binary events $Y \geq j, j = 1, 2, \ldots, k$ and use corresponding predicted probabilities

$$
\hat{P}_{ij} = \frac{1}{1 + \exp[-(\hat{\alpha}_j + X_i\hat{\beta})]},
$$

- Score residual for subject $i$ predictor $m$:

$$
U_{im} = X_{im}(Y_i \geq j) - \hat{P}_{ij},
$$

- For each column of $U$ plot mean $\bar{U}_m$ and C.L. against $Y$

- Partial residuals are more useful as they can also estimate
covariable transformations [125, 44]:

\[ r_{im} = \hat{\beta}_m X_{im} + \frac{Y_i - \hat{P}_i}{\hat{P}_i(1 - \hat{P}_i)}, \]

where

\[ \hat{P}_i = \frac{1}{1 + \exp[-(\alpha + X_i\hat{\beta})]}. \]

- Smooth \( r_{im} \) vs. \( X_{im} \) to estimate how \( X_m \) relates to the log relative odds that \( Y = 1 \mid X_m \)

- For ordinal \( Y \) compute binary model partial res. for all cut-offs \( j \):

\[ r_{im} = \hat{\beta}_m X_{im} + \frac{[Y_i \geq j] - \hat{P}_{ij}}{\hat{P}_{ij}(1 - \hat{P}_{ij})}, \]

Li and Shepherd[133] have a residual for ordinal models that serves for the entire range of \( Y \) without the need to consider cutoffs. Their residual is useful for checking functional form of predictors but not the proportional odds assumption.

---

**Assessment of Model Fit**

- Section 13.2

- Stratified proportions \( Y \geq j, j = 1, 2, \ldots, k \), since \( \logit(Y \geq j \mid X) - \logit(Y \geq i \mid X) = \alpha_j - \alpha_i \), for any constant \( X \)

```r
require(Hmisc)
```
When $Y$ is continuous or almost continuous and $X$ is discrete, the PO model assumes that the logit of the cumulative distribution function of $Y$ is parallel across categories of $X$. The corresponding, more rigid, assumptions of the ordinary linear model (here, parametric ANOVA) are parallelism and linearity if the normal inverse cumulative distribution function across categories of $X$. As an example consider the web site’s diabetes dataset, where we consider the distribution of log glycohemoglobin across subjects’ body frames.
Figure 13.2: Transformed empirical cumulative distribution functions stratified by body frame in the diabetes dataset. Left panel: checking all assumptions of the parametric ANOVA. Right panel: checking all assumptions of the PO model (here, Kruskal–Wallis test).

13.3.6 Quantifying Predictive Ability

13.3.7 Describing the Model

For PO models there are four and sometimes five types of relevant predictions:

1. logit\[Y \geq j | X]\], i.e., the linear predictor
2. Prob\[Y \geq j | X]\]
CHAPTER 13. ORDINAL LOGISTIC REGRESSION

3. \( \text{Prob}[Y = j|X] \)
4. Quantiles of \( Y|X \) (e.g., the median\(^a\))
5. \( E(Y|X) \) if \( Y \) is interval scaled.

Graphics:

1. Partial effect plot (prob. scale or mean)
2. Odds ratio chart
3. Nomogram (possibly including the mean)

---

Validating the Fitted Model

---

R Functions

The \texttt{rms} package’s \texttt{lrm} and \texttt{orm} functions fit the PO model directly, assuming that the levels of the response variable (e.g., the levels of a factor variable) are listed in the proper order. \texttt{predict} computes all types of estimates except for quantiles. \texttt{orm} allows for more link functions than the logistic and is intended to efficiently handle hundreds of intercepts as happens when \( Y \) is continuous.

The R functions \texttt{popower} and \texttt{posamsize} (in the \texttt{Hmisc} package) compute power and sample size estimates for ordinal responses using the proportional odds model.

\(^a\)If \( Y \) does not have very many levels, the median will be a discontinuous function of \( X \) and may not be satisfactory.
The function `plot.xmean.ordinaly` in `rms` computes and graphs the quantities described in Section 13.2. It plots simple $Y$-stratified means overlaid with $\hat{E}(X|Y = j)$, with $j$ on the $x$-axis. The $\hat{E}$s are computed for both PO and continuation ratio ordinal logistic models.

The `Hmisc` package’s `summary.formula` function is also useful for assessing the PO assumption.

Generic `rms` functions such as `validate`, `calibrate`, and `nomogram` work with PO model fits from `lrm` as long as the analyst specifies which intercept(s) to use.

`rms` has a special function generator `Mean` for constructing an easy-to-use function for getting the predicted mean $Y$ from a PO model. This is handy with `plot` and `nomogram`. If the fit has been run through `bootcov`, it is easy to use the `Predict` function to estimate bootstrap confidence limits for predicted means.
Continuation Ratio Model

Model

Unlike the PO model, which is based on cumulative probabilities, the continuation ratio (CR) model is based on conditional probabilities. The (forward) CR model \([74, 7, 20]\) is stated as follows for \(Y = 0, \ldots, k\):

\[
\Pr(Y = j | Y \geq j, X) = \frac{1}{1 + \exp[-(\theta_j + X\gamma)]}
\]

\[
\logit(Y = 0 | Y \geq 0, X) = \logit(Y = 0 | X) = \theta_0 + X\gamma
\]

\[
\logit(Y = 1 | Y \geq 1, X) = \theta_1 + X\gamma
\]

\[
\logit(Y = k - 1 | Y \geq k - 1, X) = \theta_{k-1} + X\gamma.
\]

The CR model has been said to be likely to fit ordinal responses when subjects have to “pass through” one category to get to the next. The CR model is a discrete version of the Cox proportional hazards model. The discrete hazard function is defined as \(\Pr(Y = j | Y \geq j)\).

Advantage of CR model: easy to allow unequal slopes across
Assumptions and Interpretation of Parameters

Estimation

Residuals

To check CR model assumptions, binary logistic model partial residuals are again valuable. We separately fit a sequence of binary logistic models using a series of binary events and the corresponding applicable (increasingly small) subsets of subjects, and plot smoothed partial residuals against $X$ for all of the binary events. Parallelism in these plots indicates that the CR model’s constant $\gamma$ assumptions are satisfied.
Assessment of Model Fit

Extended CR Model

Role of Penalization in Extended CR Model

Validating the Fitted Model

R Functions

The `cr.setup` function in `rms` returns a list of vectors useful in constructing a dataset used to trick a binary logistic function such as `lrm` into fitting CR models.
Chapter 14

Case Study in Ordinal Regression, Data Reduction, and Penalization

See new Chapter 14 in book.
Chapter 15

Regression Models for Continuous $Y$ and Case Study in Ordinal Regression

This chapter concerns univariate continuous $Y$. There are many multivariable models for predicting such response variables.

- linear models with assumed normal residuals, fitted with ordinary least squares
- generalized linear models and other parametric models based on special distributions such as the gamma
- generalized additive models (GAMs)
- generalization of GAMs to also nonparametrically transform $Y$
- quantile regression (see Section 15.3)
- other robust regression models that, like quantile regres-
sion, use an objective different from minimizing the sum of squared errors [215]

- semiparametric models based on the ranks of $Y$, such as the Cox proportional hazards model and the proportional odds ordinal logistic model

- cumulative probability models (often called *cumulative link models*) which are semiparametric models from a wider class of families than the logistic

Semiparametric models that treat $Y$ as ordinal but not interval-scaled have many advantages including robustness and freedom of distributional assumptions for $Y$ conditional on any given set of predictors.

Advantages are demonstrated in a case study of a cumulative probability ordinal model. Some of the results are compared to quantile regression and OLS. Many of the methods used in the case study also apply to ordinary linear models.
15.1 Dataset and Descriptive Statistics

- Diabetes Mellitus (DM) type II (adult onset diabetes) is strongly associated with obesity

- Primary laboratory test for diabetes: glycosylated hemoglobin (HbA$_{1c}$), also called glycated hemoglobin, glycohemoglobin, or hemoglobin $A_{1c}$.

- HbA$_{1c}$ reflects average blood glucose for the preceding 60 to 90 days

- HbA$_{1c} > 7.0$ usually taken as a positive diagnosis of diabetes

- Goal of analysis:
  - better understand effects of body size measurements on risk of DM
  - enhance screening for DM

- Best way to develop a model for DM screening is not to fit binary logistic model with HbA$_{1c} > 7$ as the response variable

  - All cutpoints are arbitrary; no justification for any putative cut
– HbA$_{1c}$ = 6.9, 7.1 = 10
– Larger standard errors of $\hat{\beta}$, lower power, wider confidence bands
– Better: predict continuous HbA$_{1c}$ using continuous response model, then convert to probability HbA$_{1c}$ exceeds any cutoff or estimate 0.9 quantile of HbA$_{1c}$

• Data: U.S. National Health and Nutrition Examination Survey (NHANES) from National Center for Health Statistics/CDC: http://www.cdc.gov/nchs/nhanes.htm

• age $\geq$ 80 coded as 80 by CDC

• Subset with age $\geq$ 21, neither diagnosed nor treated for DM

```r
require(rms)
options(prType='latex')  # for print, summary, anova
getHdata(nhgh)
w ← subset(nhgh, age $\geq$ 21 & dx==0 & tx==0, select=-c(dx,tx))
latex(describe(w), file='')
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>n</th>
<th>missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex</td>
<td>4629</td>
<td>0</td>
<td>2</td>
<td>56902</td>
<td>48.57</td>
<td>19.85</td>
<td>23.88</td>
<td>26.08</td>
<td>33.92</td>
<td>46.83</td>
<td>61.83</td>
<td>74.83</td>
<td>80.00</td>
</tr>
<tr>
<td>age [years]</td>
<td>4629</td>
<td>0</td>
<td>703</td>
<td>1</td>
<td>48.57</td>
<td>19.85</td>
<td>23.33</td>
<td>26.08</td>
<td>33.92</td>
<td>46.83</td>
<td>61.83</td>
<td>74.83</td>
<td>80.00</td>
</tr>
</tbody>
</table>

### Race/Ethnicity

<table>
<thead>
<tr>
<th>Race/Ethnicity</th>
<th>N</th>
<th>Missing</th>
<th>Distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mexican American</td>
<td>832</td>
<td>0.180</td>
<td></td>
</tr>
<tr>
<td>Other Hispanic</td>
<td>474</td>
<td>0.102</td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic White</td>
<td>2318</td>
<td>0.501</td>
<td></td>
</tr>
<tr>
<td>Non-Hispanic Black</td>
<td>756</td>
<td>0.163</td>
<td></td>
</tr>
<tr>
<td>Other Race Including Multi-Racial</td>
<td>249</td>
<td>0.054</td>
<td></td>
</tr>
</tbody>
</table>

### Income: Family Income

<table>
<thead>
<tr>
<th>Income Range</th>
<th>N</th>
<th>Missing</th>
<th>Distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0,5000)</td>
<td>162</td>
<td>0.037</td>
<td></td>
</tr>
<tr>
<td>[5000,10000)</td>
<td>216</td>
<td>0.049</td>
<td></td>
</tr>
<tr>
<td>[10000,15000)</td>
<td>371</td>
<td>0.085</td>
<td></td>
</tr>
<tr>
<td>[15000,20000)</td>
<td>535</td>
<td>0.122</td>
<td></td>
</tr>
<tr>
<td>[20000,25000)</td>
<td>421</td>
<td>0.091</td>
<td></td>
</tr>
<tr>
<td>[25000,35000)</td>
<td>346</td>
<td>0.079</td>
<td></td>
</tr>
<tr>
<td>[35000,45000)</td>
<td>237</td>
<td>0.059</td>
<td></td>
</tr>
<tr>
<td>[45000,55000)</td>
<td>287</td>
<td>0.063</td>
<td></td>
</tr>
<tr>
<td>[55000,65000)</td>
<td>188</td>
<td>0.043</td>
<td></td>
</tr>
<tr>
<td>[65000,75000)</td>
<td>149</td>
<td>0.034</td>
<td></td>
</tr>
<tr>
<td>&gt;750000</td>
<td>52</td>
<td>0.012</td>
<td></td>
</tr>
<tr>
<td>[75000,10000)</td>
<td>399</td>
<td>0.091</td>
<td></td>
</tr>
<tr>
<td>&gt;=100000</td>
<td>619</td>
<td>0.141</td>
<td></td>
</tr>
</tbody>
</table>

### wt: Weight [kg]

<table>
<thead>
<tr>
<th>wt Range</th>
<th>N</th>
<th>Missing</th>
<th>Distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>33.2</td>
<td>890</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>36.1</td>
<td>1675</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>37.9</td>
<td>1171</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>38.5</td>
<td>151.1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>38.7</td>
<td>154.4</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>38.9</td>
<td>160.1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>39.3</td>
<td>167.2</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>39.0</td>
<td>175.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>39.9</td>
<td>181.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>40.1</td>
<td>184.8</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

### ht: Standing Height [cm]

<table>
<thead>
<tr>
<th>ht Range</th>
<th>N</th>
<th>Missing</th>
<th>Distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>123.3</td>
<td>512</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>134.5</td>
<td>1675</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>137.5</td>
<td>1171</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>139.4</td>
<td>151.1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>139.8</td>
<td>154.4</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>140.2</td>
<td>160.1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>140.4</td>
<td>167.2</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>140.5</td>
<td>175.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>140.7</td>
<td>181.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>140.8</td>
<td>184.8</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

### bmi: Body Mass Index [kg/m^2]

<table>
<thead>
<tr>
<th>bmi Range</th>
<th>N</th>
<th>Missing</th>
<th>Distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>13.18</td>
<td>1994</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>14.59</td>
<td>28.59</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>15.02</td>
<td>6.965</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>15.49</td>
<td>20.02</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>15.90</td>
<td>21.35</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>16.49</td>
<td>24.12</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>16.90</td>
<td>27.60</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>17.50</td>
<td>31.88</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>18.48</td>
<td>36.75</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>19.05</td>
<td>40.68</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

### leg: Upper Leg Length [cm]

<table>
<thead>
<tr>
<th>leg Range</th>
<th>N</th>
<th>Missing</th>
<th>Distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.4</td>
<td>3839</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>24.9</td>
<td>4.301</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>25.0</td>
<td>32.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>25.1</td>
<td>33.5</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>25.2</td>
<td>36.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>25.3</td>
<td>38.4</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>25.4</td>
<td>41.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>25.5</td>
<td>43.3</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>25.6</td>
<td>44.6</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

### arml: Upper Arm Length [cm]

<table>
<thead>
<tr>
<th>arml Range</th>
<th>N</th>
<th>Missing</th>
<th>Distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>24.8</td>
<td>37.01</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>27.0</td>
<td>3.116</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>27.5</td>
<td>32.6</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>29.2</td>
<td>33.6</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>29.5</td>
<td>35.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>29.6</td>
<td>37.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>32.5</td>
<td>39.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>32.6</td>
<td>40.6</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>32.7</td>
<td>41.7</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

### armc: Arm Circumference [cm]

<table>
<thead>
<tr>
<th>armc Range</th>
<th>N</th>
<th>Missing</th>
<th>Distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>17.9</td>
<td>32.87</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>19.0</td>
<td>5.475</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>19.3</td>
<td>25.4</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>19.5</td>
<td>26.9</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>19.9</td>
<td>29.5</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>20.1</td>
<td>32.5</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>20.2</td>
<td>35.8</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>20.3</td>
<td>39.1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>20.4</td>
<td>41.4</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

### waist: Waist Circumference [cm]

<table>
<thead>
<tr>
<th>waist Range</th>
<th>N</th>
<th>Missing</th>
<th>Distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>59.7</td>
<td>97.62</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>60.0</td>
<td>17.18</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>61.5</td>
<td>74.8</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>62.1</td>
<td>78.6</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>62.4</td>
<td>86.9</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>62.8</td>
<td>96.3</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>63.2</td>
<td>107.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>63.5</td>
<td>117.8</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>63.8</td>
<td>125.0</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

### tri: Triceps Skinfold [mm]

<table>
<thead>
<tr>
<th>tri Range</th>
<th>N</th>
<th>Missing</th>
<th>Distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6</td>
<td>18.94</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3.1</td>
<td>9.463</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3.2</td>
<td>7.2</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3.3</td>
<td>8.8</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3.4</td>
<td>12.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3.5</td>
<td>18.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3.6</td>
<td>25.2</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3.7</td>
<td>31.0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>3.8</td>
<td>33.8</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

### sub: Subscapular Skinfold [mm]

<table>
<thead>
<tr>
<th>sub Range</th>
<th>N</th>
<th>Missing</th>
<th>Distinct</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.8</td>
<td>20.8</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>4.2</td>
<td>9.124</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>4.6</td>
<td>8.60</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>4.8</td>
<td>10.30</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>4.9</td>
<td>14.40</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5.0</td>
<td>20.30</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5.5</td>
<td>26.58</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>5.9</td>
<td>32.00</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>6.0</td>
<td>35.00</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
### gh: Glycohemoglobin [%]

<table>
<thead>
<tr>
<th>n</th>
<th>missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>4629</td>
<td>0</td>
<td>63</td>
<td>0.994</td>
<td>5.533</td>
<td>0.5411</td>
<td>4.8</td>
<td>5.0</td>
<td>5.2</td>
<td>5.5</td>
<td>5.8</td>
<td>6.0</td>
<td>6.3</td>
</tr>
</tbody>
</table>

Lowest: 4.0 4.1 4.2 4.3 4.4, highest: 11.9 12.0 12.1 12.3 14.5

### albumin: Albumin [g/dL]

<table>
<thead>
<tr>
<th>n</th>
<th>missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>4576</td>
<td>53</td>
<td>26</td>
<td>0.99</td>
<td>4.261</td>
<td>0.3528</td>
<td>3.7</td>
<td>3.9</td>
<td>4.1</td>
<td>4.3</td>
<td>4.5</td>
<td>4.7</td>
<td>4.8</td>
</tr>
</tbody>
</table>

Lowest: 2.6 2.7 3.0 3.1 3.2, highest: 4.9 5.0 5.1 5.2 5.3

### bun: Blood urea nitrogen [mg/dL]

<table>
<thead>
<tr>
<th>n</th>
<th>missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>4576</td>
<td>53</td>
<td>50</td>
<td>0.995</td>
<td>13.03</td>
<td>5.309</td>
<td>7</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>15</td>
<td>19</td>
<td>22</td>
</tr>
</tbody>
</table>

Lowest: 1 2 3 4 5, highest: 49 53 55 56 63

### Scr: Creatinine [mg/dL]

<table>
<thead>
<tr>
<th>n</th>
<th>missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>4576</td>
<td>53</td>
<td>167</td>
<td>1</td>
<td>0.8887</td>
<td>0.2697</td>
<td>0.58</td>
<td>0.62</td>
<td>0.72</td>
<td>0.84</td>
<td>0.99</td>
<td>1.14</td>
<td>1.25</td>
</tr>
</tbody>
</table>

Lowest: 0.34 0.38 0.39 0.40 0.41, highest: 5.98 6.34 9.13 10.98 15.66

\[ \text{dd ← datadist(w); options(datadist='dd')} \]
15.2 The Linear Model

The most popular multivariable model for analyzing a univariate continuous $Y$ is the linear model

$$E(Y|X) = X\beta,$$

where $\beta$ is estimated using ordinary least squares, that is, by solving for $\hat{\beta}$ to minimize $\sum(Y_i - X\hat{\beta})^2$.

To compute $P$-values and confidence limits using parametric methods (and for least squares estimates to coincide with maximum likelihood estimates) we would have to assume that $Y|X$ is normal with mean $X\beta$ and constant variance $\sigma^2$.

15.2.1 Checking Assumptions of OLS and Other Models

- First see if $gh$ would make a Gaussian residuals model fit
- Use ordinary regression on 4 key variables to collapse into one variable (predicted mean from OLS model)
- Stratify predicted mean into 6 quantile groups
- Apply the normal inverse ECDF of $gh$ to these strata and

---

$^a$The latter assumption may be dispensed with if we use a robust Huber–White or bootstrap covariance matrix estimate. Normality may sometimes be dispensed with by using bootstrap confidence intervals, but this would not fix inefficiency problems with OLS when residuals are non-normal.
check for normality and constant $\sigma^2$

- ECDF is for $\text{Prob}[Y \leq y|X]$ but for ordinal modeling we want to state models in terms of $\text{Prob}[Y \geq y|X]$ so take $1 - \text{ECDF}$ before inverse transforming

```r
f ← ols(gh ~ rcs(age,5) + sex + re + rcs(bmi, 3), data=w)
pgh ← fitted(f)

p ← function(fun, row, col) {
  f ← substitute(fun); g ← function(F) eval(f)
  z ← Ecdf(~ gh, groups=cut2(pgh, g=6),
            fun=function(F) g(1 - F),
            ylab=as.expression(f), xlim=c(4.5, 7.75), data=w,
            label.curve=FALSE)
  print(z, split=c(col, row, 2, 2), more=row < 2 | col < 2)
}
p(log(F/(1-F)), 1, 1)
p(qnorm(F), 1, 2)
p(-log(-log(F)), 2, 1)
p(log(-log(1-F)), 2, 2)

# Get slopes of pgh for some cutoffs of Y
# Use glm complementary log-log link on Prob(Y < cutoff) to
# get log-log link on Prob(Y ≥ cutoff)
r ← NULL
for(link in c('logit','probit','cloglog'))
  for(k in c(5, 5.5, 6)) {
    co ← coef(glm(gh < k ~ pgh, data=w, family=binomial(link)))
    r ← rbind(r, data.frame(link=link, cutoff=k,
                          slope=round(co[2],2))
  }
print(r, row.names=FALSE)
```

<table>
<thead>
<tr>
<th>link</th>
<th>cutoff</th>
<th>slope</th>
</tr>
</thead>
<tbody>
<tr>
<td>logit</td>
<td>5.0</td>
<td>-3.39</td>
</tr>
<tr>
<td>logit</td>
<td>5.5</td>
<td>-4.33</td>
</tr>
<tr>
<td>logit</td>
<td>6.0</td>
<td>-5.62</td>
</tr>
<tr>
<td>probit</td>
<td>5.0</td>
<td>-1.69</td>
</tr>
<tr>
<td>probit</td>
<td>5.5</td>
<td>-2.61</td>
</tr>
<tr>
<td>probit</td>
<td>6.0</td>
<td>-3.07</td>
</tr>
<tr>
<td>cloglog</td>
<td>5.0</td>
<td>-3.18</td>
</tr>
<tr>
<td>cloglog</td>
<td>5.5</td>
<td>-2.97</td>
</tr>
<tr>
<td>cloglog</td>
<td>6.0</td>
<td>-2.51</td>
</tr>
</tbody>
</table>

- Upper right curves are not linear, implying that a normal conditional distribution cannot work for gh^b

^bThey are not parallel either.
Figure 15.1: Examination of normality and constant variance assumption, and assumptions for various ordinal models
• There is non-parallelism for the logit model

• Other graphs will be used to guide selection of an ordinal model below
Quantile Regression

- Ruled out OLS and semiparametric proportional odds model
- Quantile regression \([120, 119]\) is a different approach to modeling \(Y\)
- No distributional assumptions other than continuity of \(Y\)
- All the usual right hand side assumptions
- When there is a single predictor that is categorical, quantile regression coincides with ordinary sample quantiles stratified by that predictor
- Is transformation invariant - pre-transforming \(Y\) not important

Let \(\rho_\tau(y) = y(\tau - [y < 0])\). The \(\tau^{th}\) sample quantile is the minimizer \(q\) of \(\sum_{i=1}^{n} \rho_\tau(y_i - q)\). For a conditional \(\tau^{th}\) quantile of \(Y|X\) the corresponding quantile regression estimator \(\hat{\beta}_\tau\) minimizes \(\sum_{i=1}^{n} \rho_\tau(Y_i - X\beta)\).

Quantile regression is not as efficient at estimating quantiles as is ordinary least squares at estimating the mean, if the latter’s assumptions hold.

Koenker’s quantreg package in R \([121]\) implements quantile re-
gression, and the rms package’s Rq function provides a front-end that gives rise to various graphics and inference tools.

If we model the median $g_{\theta}$ as a function of covariates, only the $X\beta$ structure need be correct. Other quantiles (e.g., 90\text{th} percentile) can be directly modeled but standard errors will be much larger as it is more difficult to precisely estimate outer quantiles.
15.4 Ordinal Regression Models for Continuous $Y$

- Advantages of semiparametric models (e.g., quantile regression and cumulative probability ordinal models)

- For ordinal cumulative probability models, there is no distributional assumption for $Y$ given a setting of $X$

- Assume only a connection between distributions of $Y$ for different $X$

- Applying an increasing 1–1 transformation to $Y$ results in no change to regression coefficient estimates\(^{c}\)

- Regression coefficient estimates are completely robust to extreme $Y$ values\(^{d}\)

- Estimates of quantiles of $Y$ are exactly transformation-preserving, e.g., estimate of median of $\log Y$ is exactly the log of the estimate of median $Y$

- Manuguerra [143] developed an ordinal model for continuous $Y$ which they incorrectly labeled semi-parametric and is actually a lower-dimensional flexible parametric model that instead of having intercepts has a spline function of $y$.

\(^{c}\)For symmetric distributions applying a decreasing transformation will negate the coefficients. For asymmetric distributions (e.g., Gumbel), reversing the order of $Y$ will do more than change signs.

\(^{d}\)Only an estimate of mean $Y$ from these $\beta$s is non-robust.
For a general continuous distribution function $F(y)$, an ordinal regression model based on cumulative probabilities may be stated as follows$^e$. Let the ordered unique values of $Y$ be denoted by $y_1, y_2, \ldots, y_k$ and let the intercepts associated with $y_1, \ldots, y_k$ be $\alpha_1, \alpha_2, \ldots, \alpha_k$, where $\alpha_1 = \infty$ because $\text{Prob}[Y \geq y_1] = 1$. Let $\alpha_y = \alpha_i, i : y_i = y$. Then

$$\text{Prob}[Y \geq y_i|X] = F(\alpha_i + X\beta) = F(\alpha_y + X\beta)$$

For the OLS fully parametric case, the model may be restated

$$\text{Prob}[Y \geq y|X] = \text{Prob}[\frac{Y - X\beta}{\sigma} \geq \frac{y - X\beta}{\sigma}] = 1 - \Phi(\frac{y - X\beta}{\sigma}) = \Phi(\frac{-y}{\sigma} + \frac{X\beta}{\sigma})$$

so that to within an additive constant$^f$ $\alpha_y = -\frac{y}{\sigma}$ (intercepts $\alpha$ are linear in $y$ whereas they are arbitrarily descending in the ordinal model), and $\sigma$ is absorbed in $\beta$ to put the OLS model into the new notation.

The general ordinal regression model assumes that for fixed $X_1, X_2$,

$$F^{-1}(\text{Prob}[Y \geq y|X_2]) - F^{-1}(\text{Prob}[Y \geq y|X_1]) = (X_2 - X_1)\beta$$

independent of the $\alpha$s (parallelism assumption). If $F = [1 + \exp(-y)]^{-1}$, this is the proportional odds assumption.

$^e$It is more traditional to state the model in terms of $\text{Prob}[Y \leq y|X]$ but we use $\text{Prob}[Y \geq y|X]$ so that higher predicted values are associated with higher $Y$.

$^f\alpha_y$ are unchanged if a constant is added to all $y$. 
Table 15.1: Distribution families used in ordinal cumulative probability models. Φ denotes the Gaussian cumulative distribution function. For the Connection column, $P_1 = \text{Prob}[Y \geq y|X_1], P_2 = \text{Prob}[Y \geq y|X_2], \Delta = (X_2 - X_1)\beta$. The connection specifies the only distributional assumption if the model is fitted semiparametrically, i.e., contains an intercept for every unique Y value less one. For parametric models, $P_1$ must be specified absolutely instead of just requiring a relationship between $P_1$ and $P_2$. For example, the traditional Gaussian parametric model specifies that $\text{Prob}[Y \geq y|X] = 1 - \Phi(\frac{y - X\beta}{\sigma}) = \Phi(-\frac{y + X\beta}{\sigma})$.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$F$</th>
<th>Inverse Link Name (Link Function)</th>
<th>Connection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>$[1 + \exp(-y)]^{-1}$</td>
<td>$\log(\frac{y}{1-y})$</td>
<td>$P_2 = \frac{P_1}{P_1 + \exp(\Delta)}$</td>
</tr>
<tr>
<td>Gaussian</td>
<td>$\Phi(y)$</td>
<td>$\Phi^{-1}(y)$</td>
<td>$P_2 = \Phi(\Phi^{-1}(P_1) + \Delta)$</td>
</tr>
<tr>
<td>Gumbel maximum value</td>
<td>$\exp(-\exp(-y))$</td>
<td>$\log(-\log(y))$</td>
<td>$P_2 = P_1^{\exp(\Delta)}$</td>
</tr>
<tr>
<td>Gumbel minimum value</td>
<td>$1 - \exp(-\exp(y))$</td>
<td>$\log(-\log(1 - y))$</td>
<td>$1 - P_2 = (1 - P_1)^{\exp(\Delta)}$</td>
</tr>
<tr>
<td>Cauchy</td>
<td>$\frac{1}{2}\tan^{-1}(y) + \frac{1}{2}$</td>
<td>$\tan[\pi(y - \frac{1}{2})]$</td>
<td>$\text{cauchit}$</td>
</tr>
</tbody>
</table>

Common choices of $F$, implemented in the \texttt{orm} function, are shown in Table 15.1. The Gumbel maximum value distribution is also called the extreme value type I distribution. This distribution ($\log - \log$ link) also represents a continuous time proportional hazards model. The hazard ratio when $X$ changes from $X_1$ to $X_2$ is $\exp(-(X_2 - X_1)\beta)$.

The mean of $Y|X$ is easily estimated by computing

$$\sum_{i=1}^{k} y_i \hat{\text{Prob}}[Y = y_i|X]$$

and the $q^{th}$ quantile of $Y|X$ is $y$ such that

$$F^{-1}(1 - q) - X\hat{\beta} = \hat{\alpha}_y.$$ 

The \texttt{orm} function in the \texttt{rms} package takes advantage of the information matrix being of a sparse tri-band diagonal form for the intercept parameters. This makes the computations

\footnote{The intercepts have to be shifted to the left one position in solving this equation because the quantile is such that $\hat{\text{Prob}}[Y \leq y] = q$ whereas the model is stated in terms of $\text{Prob}[Y \geq y]$.}
efficient even for hundreds of intercepts (i.e., unique values of \( Y \)). orm is made to handle continuous \( Y \).

Ordinal regression has nice properties in addition to those listed above, allowing for

- estimation of quantiles as efficiently as quantile regression if the parallel slopes assumptions hold
- efficient estimation of mean \( Y \)
- direct estimation of \( \text{Prob}[Y \geq y|X] \)
- arbitrary clumping of values of \( Y \), while still estimating \( \beta \) and mean \( Y \) efficiently\(^h\)
- solutions for \( \hat{\beta} \) using ordinary Newton-Raphson or other popular optimization techniques
- being based on a standard likelihood function, penalized estimation can be straightforward
- Wald, score, and likelihood ratio \( \chi^2 \) tests that are more powerful than tests from quantile regression

To summarize how assumptions of parametric models compare to assumptions of semiparametric models, consider the ordinary linear model or its special case the equal variance two-sample

\(^h\)But it is not sensible to estimate quantiles of \( Y \) when there are heavy ties in \( Y \) in the area containing the quantile.
t-test, vs. the probit or logit (proportional odds) ordinal model or their special cases the Van der Waerden (normal-scores) two-sample test or the Wilcoxon test. All the assumptions of the linear model other than independence of residuals are captured in the following (written in traditional $Y \leq y$ form):

$$F(y|X) = \Pr(Y \leq y|X) = \Phi\left(\frac{y - X\beta}{\sigma}\right)$$

$$\Phi^{-1}(F(y|X)) = \frac{y - X\beta}{\sigma}$$

On the other hand, ordinal models assume the following:

$$\Pr[Y \leq y|X] = F(g(y) - X\beta),$$

where $g$ is unknown and may be discontinuous.

From this point we revert back to $Y \geq y$ notation so that $Y$ increases as $X\beta$ increases.
Global Modeling Implications

- Ordinal regression invariant to choice of transformation of $Y$
- $Y$ needs to be ordinal
- Difference in two ordinal variables is not necessarily ordinal
- $\rightarrow$ Never analyze differences in regression
- $Y =$ final value, adjust for baseline values as covariates
15.5

Ordinal Regression Applied to HbA$_{1c}$

- In Figure 15.1, logit inverse curves are not parallel so proportional odds assumption does not hold

- log-log link yields highest degree of parallelism and most constant regression coefficients across cutoffs of gh so use this link in an ordinal regression model (linearity of the curves is not required)

15.5.1

Checking Fit for Various Models Using Age

Another way to examine model fit is to flexibly fit the single most important predictor (age) using a variety of methods, and comparing predictions to sample quantiles and means based on overlapping subsets on age, each subset being subjects having age $< 5$ years away from the point being predicted by the models. Here we predict the 0.5, 0.75, and 0.9 quantiles and the mean. For quantiles we can compare to quantile regression (discussed below) and for means we compare to OLS.

```r
ag ← 25:75
lag ← length(ag)
q2 ← q3 ← p90 ← means ← numeric(lag)
for(i in 1:lag) {
  s ← which(abs(w$age - ag[i]) < 5)
  y ← w$gh[s]
  a ← quantile(y, probs=c(.5, .75, .9))
  q2[i] ← a[1]
  q3[i] ← a[2]
  p90[i] ← a[3]
  means[i] ← mean(y)
}
It can be seen in Figure 15.3 that models dedicated to a specific task (quantile regression for quantiles and OLS for means) were best for those tasks. Although the log-log ordinal cumulative
Figure 15.3: Three estimated quantiles and estimated mean using 6 methods, compared against caliper-matched sample quantiles/means (circles). Numbers are mean absolute differences between predicted and sample quantities using overlapping intervals of age and caliper matching. QR: quantile regression.
probability model did not estimate the median as accurately as some other methods, it does well for the 0.75 and 0.9 quantiles and is the best compromise overall because of its ability to also directly predict the mean as well as quantiles such as \( \text{Prob}[\text{HbA}_{1c} > 7 | X] \).

For here on we focus on the log-log ordinal model.

Going back to the bottom left of figure 15.1, let’s look at quantile groups of predicted HbA\(_{1c}\) by OLS and plot predicted distributions of actual HbA\(_{1c}\) against empirical distributions.

```r
w$pghg <- cut2(pgh, g=6)
f <- orm(gh ~ pghg, family=loglog, data=w)
lp <- predict(f, newdata=data.frame(pghg=levels(w$pghg)))
ep <- ExProb(f) # Exceedance prob. functn. generator in rms
z <- ep(lp)
j <- order(w$pghg) # puts in order of lp (levels of pghg)
plot(z, xlim=c(4, 7.5), data=w[j,c('pghg', 'gh')]) # Fig. 15.4
```

![Figure 15.4: Observed (dashed lines, open circles) and predicted (solid lines, closed circles) exceedance probability distributions from a model using 6-tiles of OLS-predicted HbA\(_{1c}\). Key shows quantile group intervals of predicted mean HbA\(_{1c}\).](image-url)

For here on we focus on the log-log ordinal model.
Agreement between predicted and observed exceedance probability distributions is excellent in Figure 15.4.

To return to the initial look at a linear model with assumed Gaussian residuals, fit a probit ordinal model and compare the estimated intercepts to the linear ordinal relationship with gh that is assumed by the normal distribution.

```r
f ← orm(gh ~ rcs(age,6), family=probit, data=w)
g ← ols(gh ~ rcs(age,6), data=w)
s ← g$stats['Sigma']
yu ← f$yunique[-1]
r ← quantile(w$gh, c(.005, .995))
alphas ← coef(f)[1: num.intercepts(f)]
plot(-yu / s, alphas, type='l', xlim = rev(- r / s), # Fig. 15.5
     xlab=expression(-y/ hat(sigma)), ylab=expression(alpha[y]))
```

Figure 15.5: Estimated intercepts from probit model

Figure 15.5 depicts a significant departure from that implied by Gaussian residuals.
15.5.2 Examination of BMI

Using the log-log model, we first check the adequacy of BMI as a summary of height and weight for estimating median \( gh \).

- Adjust for age (without assuming linearity) in every case
- Look at ratio of coefficients of log height and log weight
- Use AIC to judge whether BMI is an adequate summary of height and weight

```r
f ← orm(gh ~ rcs(age, 5) + log(ht) + log(wt),
        family=loglog, data=w)
```

**-log-log Ordinal Regression Model**

```r
orm(formula = gh ~ rcs(age, 5) + log(ht) + log(wt), data = w,
    family = loglog)
```

<table>
<thead>
<tr>
<th>Model Likelihood Ratio Test</th>
<th>Discrimination Indexes</th>
<th>Rank Discrim. Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs 4629</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distinct Y 63</td>
<td>LR ( \chi^2 ) 1126.94</td>
<td>( R^2 )  0.217</td>
</tr>
<tr>
<td>( Y_{0.5} ) 5.5</td>
<td>d.f. 6</td>
<td>( g ) 0.627</td>
</tr>
<tr>
<td>( \max</td>
<td>\frac{\partial \log L}{\partial \beta}</td>
<td>) \times 10^{-6}</td>
</tr>
<tr>
<td>Score ( \chi^2 ) 1262.81</td>
<td>Pr(( \chi^2 )) &lt;0.0001</td>
<td>(</td>
</tr>
</tbody>
</table>

| \( \hat{\beta} \) | S.E. | Wald Z | Pr(>|Z|) |
|-------------------|------|--------|----------|
| age               |  0.0398 |  0.0055 |  7.29    | <0.0001 |
| age'              | -0.0158 |  0.0275 | -0.57    |  0.5657 |
| age''             | -0.0072 |  0.0866 | -0.08    |  0.9333 |
| age'''            |  0.0309 |  0.1135 |  0.27    |  0.7853 |
| ht                | -3.0680 |  0.2789 | -11.00   | <0.0001 |
| wt                |  1.2748 |  0.0704 |  18.10   | <0.0001 |

**aic ← NULL**
The ratio of the coefficient of log height to the coefficient of log weight is -2.4, which is between what BMI uses and the more dimensionally reasonable weight / height$^3$. By AIC, a spline interaction surface between height and weight does slightly better than BMI in predicting HbA$_{1c}$, but a nonlinear function of BMI is barely worse. It will require other body size measures to displace BMI as a predictor.

As an aside, compare this model fit to that from the Cox proportional hazards model. The Cox model uses a conditioning argument to obtain a partial likelihood free of the intercepts $\alpha$ (and requires a second step to estimate these log discrete hazard components) whereas we are using a full marginal likelihood of the ranks of $Y$ [112].

<table>
<thead>
<tr>
<th>Obs</th>
<th>4629</th>
<th>4629</th>
<th>8.3792</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Tests</td>
<td>Discrimination Indexes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR $\chi^2$</td>
<td>1120.20</td>
<td>$R^2$</td>
<td>0.215</td>
</tr>
<tr>
<td>d.f.</td>
<td>6</td>
<td>$D_{xy}$</td>
<td>0.359</td>
</tr>
<tr>
<td>Pr($&gt; \chi^2$)</td>
<td>0.0000</td>
<td>$g$</td>
<td>0.622</td>
</tr>
<tr>
<td>Score $\chi^2$</td>
<td>1258.07</td>
<td>$g_r$</td>
<td>1.863</td>
</tr>
<tr>
<td>Pr($&gt; \chi^2$)</td>
<td>0.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| $\hat{\beta}$ | S.E. | Wald Z | Pr($>|Z|)$ |
|---------------|------|--------|-----------|
| age           | -0.0392 | 0.0054 | -7.24     | <0.0001   |
| age$^e$       | 0.0148  | 0.0274 | 0.54      | 0.5888    |
Back up and look at all body size measures, and examine their redundancies.

\[
\begin{align*}
\hat{\beta} & \quad \text{S.E.} & \quad \text{Wald} Z & \quad \text{Pr}(>|Z|) \\
age' & 0.0093 & 0.0862 & 0.11 & 0.9144 \\
age'' & -0.0321 & 0.1131 & -0.28 & 0.7767 \\
ht & 3.0477 & 0.2779 & 10.97 & <0.0001 \\
wt & -1.2653 & 0.0701 & -18.04 & <0.0001 \\
\end{align*}
\]

Back up and look at all body size measures, and examine their redundancies.

\[
\begin{align*}
\text{v} & \leftarrow \text{varclus} (\sim \text{wt} + \text{ht} + \text{bmi} + \text{leg} + \text{arml} + \text{armc} + \text{waist} + \\
& \quad \text{tri} + \text{sub} + \text{age} + \text{sex} + \text{re}, \text{data=}w) \\
\text{plot(v)} & \quad \# \text{Figure 15.6} \\
& \quad \# \text{Omit wt so it won’t be removed before bmi} \\
\text{redun} (\sim \text{ht} + \text{bmi} + \text{leg} + \text{arml} + \text{armc} + \text{waist} + \text{tri} + \text{sub}, \\
& \quad \text{data=}w, \text{r2}=.75) \\
\end{align*}
\]

Redundancy Analysis

\[
\text{redun} (\text{formula} = \sim \text{ht} + \text{bmi} + \text{leg} + \text{arml} + \text{armc} + \text{waist} + \text{tri} + \text{sub}, \\
& \quad \text{data=}w, \text{r2} = 0.75) \\
n: 3853 & \quad p: 8 \quad nk: 3 \\
\text{Number of NAs:} & 776 \\
\text{Frequencies of Missing Values Due to Each Variable} \\
\begin{align*}
\text{ht} & \quad \text{bmi} & \quad \text{leg} & \quad \text{arml} & \quad \text{armc} & \quad \text{waist} & \quad \text{tri} & \quad \text{sub} \\
0 & 0 & 155 & 127 & 130 & 164 & 334 & 655 \\
\end{align*}
\]

Transformation of target variables forced to be linear

\[
\text{R}^2 \text{ cutoff: 0.75} \quad \text{Type: ordinary} \\
\text{R}^2 \text{ with which each variable can be predicted from all other variables:} \\
\begin{align*}
\text{ht} & \quad \text{bmi} & \quad \text{leg} & \quad \text{arml} & \quad \text{armc} & \quad \text{waist} & \quad \text{tri} & \quad \text{sub} \\
0.829 & 0.924 & 0.682 & 0.748 & 0.843 & 0.864 & 0.531 & 0.594 \\
\end{align*}
\]

Redundant variables:

\[
\text{bmi} \quad \text{ht} \\
\]

Predicted from variables:

\[
\text{leg} \quad \text{arml} \quad \text{armc} \quad \text{waist} \quad \text{tri} \quad \text{sub} \\
\]

\begin{tabular}{lll}
\text{Variable Deleted} & \text{R}^2 & \text{R}^2 \text{ after later deletions} \\
1 & bmi & 0.924 & 0.909 \\
2 & ht & 0.792 & \\
\end{tabular}

Six size measures adequately capture the entire set. Height and
BMI are removed.

An advantage of removing height is that it is age-dependent in the elderly:

```r
f ← orm(ht ~ rcs(age,4)*sex, data=w) # Prop. odds model
qu ← Quantile(f); med ← function(x) qu(.5, x)
ggplot(Predict(f, age, sex, fun=med, conf.int=FALSE),
       ylab='Predicted Median Height, cm')
```

---

Figure 15.6: Variable clustering for all potential predictors

Figure 15.7: Estimated median height as a smooth function of age, allowing age to interact with sex, from a proportional odds model
**But** also see a change in leg length:

```r
f ← orm(leg ~ rcs(age,4)*sex, data=w)
qu ← Quantile(f); med ← function(x) qu(.5, x)
ggplot(Predict(f, age, sex, fun=med, conf.int=FALSE),
       ylab='Predicted Median Upper Leg Length, cm')
```

![Graph showing predicted median upper leg length as a function of age, allowing age to interact with sex.](image)

**Figure 15.8:** Estimated median upper leg length as a smooth function of age, allowing age to interact with sex, from a proportional odds model.

Next allocate d.f. according to generalized Spearman $\rho^2_i$.

```r
s ← spearman2(gh ~ age + sex + re + wt + leg + arml + armc +
               waist + tri + sub, data=w, p=2)
plot(s)
```

Parameters will be allocated in descending order of $\rho^2$. But note that subscapular skinfold has a large number of NAs and other predictors also have NAs. Suboptimal casewise deletion will be used until the final model is fitted.

Because there are many competing body measures, we use backwards stepdown to arrive at a set of predictors. The bootstrap will be used to penalize predictive ability for variable selection. First the full model is fit using casewise deletion, then

---

1Competition between collinear size measures hurts interpretation of partial tests of association in a saturated additive model.
we do a composite test to assess whether any of the frequently-missing predictors is important.

\[
f \leftarrow \text{orm}(\text{gh} \sim \text{rcs}(\text{age}, 5) + \text{sex} + \text{re} + \text{rcs}(\text{wt}, 3) + \text{rcs}(\text{leg}, 3) + \text{arml} + \text{rcs}(\text{armc}, 3) + \text{rcs}(\text{waist}, 4) + \text{tri} + \text{rcs}(\text{sub}, 3), \right.
\]
\[
\text{family=loglog, data=w, x=TRUE, y=TRUE)}
\]

print(f, coefs=FALSE)

-log-log Ordinal Regression Model

\[
\text{orm(formula = gh} \sim \text{rcs(age}, 5) + \text{sex} + \text{re} + \text{rcs(wt}, 3) + \text{rcs(leg, 3) + arml +}
\]
\[
\text{rcs(armc, 3) + rcs(waist, 4) + tri + rcs(sub, 3), data = w, x = TRUE, y = TRUE, family = loglog})
\]
## Model Likelihood

<table>
<thead>
<tr>
<th>Obs</th>
<th>3853</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distinct Y</td>
<td>60</td>
</tr>
<tr>
<td>$Y_{0.5}$</td>
<td>5.5</td>
</tr>
<tr>
<td>$\max</td>
<td>\frac{\partial \log L}{\partial \beta}</td>
</tr>
</tbody>
</table>

### Discrimination

<table>
<thead>
<tr>
<th>LR $\chi^2$</th>
<th>1180.13</th>
</tr>
</thead>
<tbody>
<tr>
<td>d.f.</td>
<td>22</td>
</tr>
<tr>
<td>$\Pr(\chi^2 &lt; 0.0001)$</td>
<td></td>
</tr>
</tbody>
</table>

### Score $\chi^2$

| 1298.88 |
| $\Pr(\chi^2 < 0.0001)$ |

#### Wald Statistics for $gh$

<table>
<thead>
<tr>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>leg</td>
<td>8.30</td>
<td>2</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>3.32</td>
<td>1</td>
</tr>
<tr>
<td>arml</td>
<td>0.16</td>
<td>1</td>
</tr>
<tr>
<td>armc</td>
<td>6.66</td>
<td>2</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>3.29</td>
<td>1</td>
</tr>
<tr>
<td>waist</td>
<td>29.40</td>
<td>3</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>4.29</td>
<td>2</td>
</tr>
<tr>
<td>tri</td>
<td>16.62</td>
<td>1</td>
</tr>
<tr>
<td>sub</td>
<td>40.75</td>
<td>2</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>4.50</td>
<td>1</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>14.95</td>
<td>5</td>
</tr>
<tr>
<td>TOTAL</td>
<td>128.29</td>
<td>11</td>
</tr>
</tbody>
</table>

The model yields Spearman $\rho = 0.52$, the rank correlation between predicted and observed HbA$_{1c}$.

Show predicted mean and median HbA$_{1c}$ as a function of age, adjusting other variables to median/mode. Compare the estimate of the median with that from quantile regression (discussed below).

\[
M \leftarrow \text{Mean}(f)
\]
\[
qu \leftarrow \text{Quantile}(f)
\]
\[
\text{med} \leftarrow \text{function}(x) \ qu(.5, x)
\]
\[
p90 \leftarrow \text{function}(x) \ qu(.9, x)
\]
\[
fq \leftarrow \text{Rq(formula}(f), \ data=w)
\]
\[
fq90 \leftarrow \text{Rq(formula}(f), \ data=w, \ tau=.9)
\]
\[
pmean \leftarrow \text{Predict}(f, \ age, \ fun=M, \ conf.int=FALSE)
\]
\[
pmed \leftarrow \text{Predict}(f, \ age, \ fun=med, \ conf.int=FALSE)
\]
\[
p90 \leftarrow \text{Predict}(f, \ age, \ fun=p90, \ conf.int=FALSE)
\]
pmedqr ← Predict(fq, age, conf.int=FALSE)
p90qr ← Predict(fq90, age, conf.int=FALSE)
z ← rbind('orm mean'=pmean, 'orm median'=pmed, 'orm P90'=p90,
       'QR median'=pmedqr, 'QR P90'=p90qr)
ggplot(z, groups='.set.',
      adj.subtitle=FALSE, legend.label=FALSE)

Figure 15.10: Estimated mean and 0.5 and 0.9 quantiles from the log-log ordinal model using casewise deletion, along with predictions of 0.5 and 0.9 quantiles from quantile regression (QR). Age is varied and other predictors are held constant to medians/modes.

Next do fast backward step-down in an attempt to get a model without so much competition among variables. The stepwise selection will be penalized for in the model validation.

print(fastbw(f, rule='p'), estimates=FALSE)

<table>
<thead>
<tr>
<th>Deleted</th>
<th>Chi-Sq</th>
<th>d.f.</th>
<th>P</th>
<th>Residual d.f.</th>
<th>P</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>arml</td>
<td>0.16</td>
<td>1</td>
<td>0.6924</td>
<td>0.16</td>
<td>1</td>
<td>0.6924</td>
</tr>
<tr>
<td>sex</td>
<td>0.45</td>
<td>1</td>
<td>0.5019</td>
<td>0.61</td>
<td>2</td>
<td>0.7381</td>
</tr>
<tr>
<td>wt</td>
<td>5.72</td>
<td>2</td>
<td>0.0572</td>
<td>6.33</td>
<td>4</td>
<td>0.1759</td>
</tr>
<tr>
<td>armc</td>
<td>3.32</td>
<td>2</td>
<td>0.1897</td>
<td>9.65</td>
<td>6</td>
<td>0.1400</td>
</tr>
</tbody>
</table>

Factors in Final Model

[1] age re leg waist tri sub

Validate the model, properly penalizing for variable selection
```r
set.seed(13) # so can reproduce results
v <- validate(f, B=100, bw=TRUE, estimates=FALSE, rule='p')
```

### Backwards Step-down - Original Model

<table>
<thead>
<tr>
<th>Deleted</th>
<th>Chi-Sq</th>
<th>d.f.</th>
<th>P</th>
<th>Residual d.f.</th>
<th>P</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>arml</td>
<td>0.16</td>
<td>1</td>
<td>0.6924</td>
<td>0.16</td>
<td>1</td>
<td>0.6924</td>
</tr>
<tr>
<td>sex</td>
<td>0.45</td>
<td>1</td>
<td>0.5019</td>
<td>0.61</td>
<td>2</td>
<td>0.7381</td>
</tr>
<tr>
<td>wt</td>
<td>5.72</td>
<td>2</td>
<td>0.0572</td>
<td>6.33</td>
<td>4</td>
<td>0.1759</td>
</tr>
<tr>
<td>armc</td>
<td>3.32</td>
<td>2</td>
<td>0.1897</td>
<td>9.65</td>
<td>6</td>
<td>0.1400</td>
</tr>
</tbody>
</table>

Factors in Final Model

[1] age re leg waist tri sub

---

# Show number of variables selected in first 30 boots
latex(v, B=30, file='', size='small')

<table>
<thead>
<tr>
<th>Index</th>
<th>Original Sample</th>
<th>Training Sample</th>
<th>Test Sample</th>
<th>Optimism</th>
<th>Corrected Index</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \rho )</td>
<td>0.5225</td>
<td>0.5279</td>
<td>0.5204</td>
<td>0.0076</td>
<td>0.5149</td>
<td>100</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.2712</td>
<td>0.2778</td>
<td>0.2689</td>
<td>0.0089</td>
<td>0.2623</td>
<td>100</td>
</tr>
<tr>
<td>Slope</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9790</td>
<td>0.0210</td>
<td>0.9790</td>
<td>100</td>
</tr>
<tr>
<td>( g )</td>
<td>1.2276</td>
<td>1.2483</td>
<td>1.2196</td>
<td>0.0287</td>
<td>1.1989</td>
<td>100</td>
</tr>
<tr>
<td>(</td>
<td>\Pr(Y \geq Y_{0.5}) - \frac{1}{2}</td>
<td></td>
<td>0.2007</td>
<td>0.2058</td>
<td>0.1988</td>
<td>0.0070</td>
</tr>
</tbody>
</table>

Factors Retained in Backwards Elimination

First 30 Resamples
Next fit the reduced model. Use multiple imputation to impute missing predictors.

Do an ANOVA for the reduced model, taking imputation into account.

```r
a ← aregImpute(~ gh + wt + ht + bmi + leg + arml + armc + waist + tri + sub + age + re, data=w, n.impute=5, pr=FALSE)
g ← fit.mult.impute(gh ~ rcs(age,5) + re + rcs(leg,3) + rcs(waist,4) + tri + rcs(sub,4), orm, a, family=loglog, data=w, pr=FALSE)
print(g, needspace='1.5in')
```
**-log-log Ordinal Regression Model**

```r
fit.mult.impute(formula = gh ~ rcs(age, 5) + re + rcs(leg, 3) +
                 rcs(waist, 4) + tri + rcs(sub, 4), fitter = orm, xtrans = a,
data = w, pr = FALSE, family = loglog)
```

<table>
<thead>
<tr>
<th></th>
<th>Model Likelihood Ratio Test</th>
<th>Discrimination Indexes</th>
<th>Rank Discrim. Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>4629</td>
<td>LR $\chi^2$ 1445.23</td>
<td>$R^2$ 0.269</td>
</tr>
<tr>
<td>Distinct $Y$</td>
<td>63</td>
<td>d.f. 17</td>
<td>$g$ 0.742</td>
</tr>
<tr>
<td>$Y_{0.5}$</td>
<td>5.5</td>
<td>Pr($&gt; \chi^2$) &lt;0.0001</td>
<td>$g_r$ 2.101</td>
</tr>
<tr>
<td>( \max</td>
<td>\frac{\partial \log L}{\partial \beta}</td>
<td>1 \times 10^{-5} )</td>
<td>Score $\chi^2$ 1566.55</td>
</tr>
</tbody>
</table>

|                  | $\hat{\beta}$ | S.E. | Wald $Z$ | Pr($>|Z|)$ |
|------------------|---------------|------|---------|-----------|
| age              | 0.0405        | 0.0055 | 7.36     | <0.0001   |
| age'             | -0.0228       | 0.0277 | -0.82    | 0.4094    |
| age'"            | 0.0123        | 0.0871 | 0.14     | 0.8880    |
| age'"'           | 0.0428        | 0.1143 | 0.37     | 0.7082    |
| re=Other Hispanic| -0.0795       | 0.0592 | -1.34    | 0.1794    |
| re=Non-Hispanic White | -0.4119 | 0.0451 | -9.14    | <0.0001   |
| re=Non-Hispanic Black | 0.0662 | 0.0563 | 1.18     | 0.2396    |
| re=Other Race Including Multi-Racial | -0.0509 | 0.0749 | -0.68    | 0.4964    |
| leg              | -0.0344       | 0.0092 | -3.75    | 0.0002    |
| leg'             | 0.0160        | 0.0106 | 1.51     | 0.1298    |
| waist            | 0.0071        | 0.0051 | 1.40     | 0.1618    |
| waist'           | 0.0318        | 0.0160 | 1.99     | 0.0469    |
| waist"           | -0.0950       | 0.0512 | -1.86    | 0.0634    |
| tri              | -0.0160       | 0.0027 | -5.86    | <0.0001   |
| sub              | -0.0023       | 0.0103 | -0.22    | 0.8220    |
| sub'             | 0.0655        | 0.0314 | 2.08     | 0.0372    |
| sub"            | -0.1838       | 0.1038 | -1.77    | 0.0766    |

\(\text{an} \leftarrow \text{anova}(g)\)

\(\text{print(an, caption='ANOVA for reduced model after multiple imputation, with addition of a combined effect for four size variables')}\)
ANOVA for reduced model after multiple imputation, with addition of a combined effect for four size variables

<table>
<thead>
<tr>
<th></th>
<th>( \chi^2 )</th>
<th>d.f.</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>698.03</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>( Nonlinear )</td>
<td>29.54</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>re</td>
<td>163.54</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>leg</td>
<td>24.19</td>
<td>2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>( Nonlinear )</td>
<td>2.29</td>
<td>1</td>
<td>0.1298</td>
</tr>
<tr>
<td>waist</td>
<td>128.33</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>( Nonlinear )</td>
<td>4.23</td>
<td>2</td>
<td>0.1208</td>
</tr>
<tr>
<td>tri</td>
<td>34.29</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>sub</td>
<td>41.27</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>( Nonlinear )</td>
<td>6.37</td>
<td>2</td>
<td>0.0414</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>46.91</td>
<td>8</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>1457.15</td>
<td>17</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\text{b} & \leftarrow \text{anova}(g, \text{leg, waist, tri, sub}) \\
\text{# Add new lines to the plot with combined effect of 4 size var.}
\text{s} & \leftarrow \text{rbind(an, size=b[‘TOTAL’, ])}
\text{class(s)} & \leftarrow \text{‘anova.rms’}
\text{plot(s)}
\end{align*}
\]

\[
\begin{align*}
\text{ggplot} \left( \text{Predict}(g), \text{abbrev=TRUE, ylab=NULL} \right) & \quad \# \text{Figure 15.11}
\text{M} & \leftarrow \text{Mean}(g)
\text{ggplot} \left( \text{Predict}(g, \text{fun=M), abbrev=TRUE, ylab=NULL} \right) & \quad \# \text{Figure 15.12}
\end{align*}
\]
Figure 15.11: Partial effects (log hazard or log-log cumulative probability scale) of all predictors in reduced model, after multiple imputation.
Figure 15.12: Partial effects (mean scale) of all predictors in reduced model, after multiple imputation
Compare the estimated age partial effects and confidence intervals with those from a model using casewise deletion, and with bootstrap nonparametric confidence intervals (also with casewise deletion).

```
gc ← orm(gh ~ rcs(age,5) + re + rcs(leg,3) +
         rcs(waist,4) + tri + rcs(sub,4),
         family=loglog, data=w, x=TRUE, y=TRUE)
gb ← bootcov(gc, B=300)
```

```
bootclb ← Predict(gb, age, boot.type='basic')
bootclp ← Predict(gb, age, boot.type='percentile')
multimp ← Predict(g, age)
plot(Predict(gc, age), addpanel=function(...) {
  with(bootclb, {llines(age, lower, col='blue')
                  llines(age, upper, col='blue')})
  with(bootclp, {llines(age, lower, col='blue', lty=2)
                  llines(age, upper, col='blue', lty=2)})
  with(multimp, {llines(age, lower, col='red')
                 llines(age, upper, col='red')
                 llines(age, yhat, col='red')})
  col.fill=gray(.9), adj.subtitle=FALSE)  # Figure 15.13
```

![Figure 15.13: Partial effect for age from multiple imputation (center red line) and casewise deletion (center blue line) with symmetric Wald 0.95 confidence bands using casewise deletion (gray shaded area), basic bootstrap confidence bands using casewise deletion (blue lines), percentile bootstrap confidence bands using casewise deletion (dashed blue lines), and symmetric Wald confidence bands accounting for multiple imputation (red lines).](image-url)
In OLS the mean equals the median and both are linearly related to any other quantiles. Semiparametric models are not this restrictive:

```r
M ← Mean(g)
qu ← Quantile(g)
med ← function(lp) qu(.5, lp)
q90 ← function(lp) qu(.9, lp)
lp ← predict(g)
lpr ← quantile(predict(g), c(.002, .998), na.rm=TRUE)
lps ← seq(lpr[1], lpr[2], length=200)
pmn ← M(lps)
pme ← med(lps)
p90 ← q90(lps)
plot(pmn, pme, # Figure 15.14
    xlab=expression(paste('Predicted Mean ', HbA["1c"])),
ylab='Median and 0.9 Quantile', type='l',
    xlim=c(4.75, 8.0), ylim=c(4.75, 8.0), bty='n')
box(col=gray(.8))
lines(pmn, p90, col='blue ')
abline(a=0, b=1, col=gray(.8))
text(6.5, 5.5, 'Median ')
text(5.5, 6.3, '0.9 ', col='blue ')
nint ← 350
scat1d(M(lp), nint=nint)
scat1d(med(lp), side=2, nint=nint)
scat1d(q90(lp), side=4, col='blue ', nint=nint)
```

Figure 15.14: Predicted mean HbA$_{1c}$ vs. predicted median and 0.9 quantile along with their marginal distributions
Draw a nomogram to compute 7 different predicted values for each subject.

```r
g ← Newlevels(g, list(re=abbreviate(levels(w$re))))
exprob ← ExProb(g)
nom ← nomogram(g, fun=list(Mean=M,
    'Median Glycohemoglobin' = med,
    '0.9 Quantile' = q90,
    'Prob(HbA1c ≥ 6.5)'
        = function(x) exprob(x, y=6.5),
    'Prob(HbA1c ≥ 7.0)'
        = function(x) exprob(x, y=7),
    'Prob(HbA1c ≥ 7.5)'
        = function(x) exprob(x, y=7.5)),
    fun.at=list(seq(5, 8, by=.5),
        c(5,5.25,5.5,5.75,6,6.25),
        c(5.5,6,6.5,7,8,10,12,14),
        c(.01,.05,.1,.2,.3,.4),
        c(.01,.05,.1,.2,.3,.4),
        c(.01,.05,.1,.2,.3,.4)))
plot(nom, lmgp=.28)  # Figure 15.15
```
Figure 15.15: Nomogram for predicting median, mean, and 0.9 quantile of glycohemoglobin, along with the estimated probability that $\text{HbA}_{1c} \geq 6.5$, 7, or 7.5, all from the log-log ordinal model.
Chapter 16

Models Using Nonparametric Transformations of $X$ and $Y$

See new Chapter 16 in the book.
Chapter 17

Introduction to Survival Analysis

17.1 Background

- Use when time to occurrence of event is important
- Don’t just count events; event at 6m worse than event at 9y
- Response called failure time, survival time, event time
- Ex: time until CV death, light bulb failure, pregnancy, ECG abnormality during exercise
- Allow for censoring
- Ex: 5y f/u study; subject still alive at 5y has failure time 5+
- Length of f/u can vary
• Even in a well-designed randomized clinical trial, survival modeling can allow one to

1. Test for and describe interactions with treatment. Sub-group analyses can easily generate spurious results and they do not consider interacting factors in a dose-response manner. Once interactions are modeled, relative treatment benefits can be estimated (e.g., hazard ratios), and analyses can be done to determine if some patients are too sick or too well to have even a relative benefit.

2. Understand prognostic factors (strength and shape).

3. Model absolute clinical benefit. First, a model for the probability of surviving past time $t$ is developed. Then differences in survival probabilities for patients on treatments A and B can be estimated. The differences will be due primarily to sickness (overall risk) of the patient and to treatment interactions.

4. Understand time course of treatment effect. The period of maximum effect or period of any substantial effect can be estimated from a plot of relative effects of treatment over time.

5. Gain power for testing treatment effects.

6. Adjust for imbalances in treatment allocation.
17.2 Censoring, Delayed Entry, and Truncation

- Left-censoring
- Interval censoring
- Left-truncation (unknown subset of subjects who failed before qualifying for the study)
- Delayed entry (exposure after varying periods of survival)
- Choice of time zero important
- Take into account waiting time bias
- Usually have random type I censoring (on duration, not events)
- Must usually have non-informative censoring: censoring independent of impending failure
- Intention-to-treat is a preventative measure
17.3 Notation, Survival, and Hazard Functions

- $T =$ response variable

\[ S(t) = \text{Prob}\{T > t\} = 1 - F(t), \]

```
# Note: bb, dd not stated in usual Weibull form
aa <- .5
bb <- -.5
cc <- 10
dd <- 4

cumhaz <- (aa/(bb+1))*tt^(bb + 1) + (cc/(dd+1))*tt^(dd + 1)
survival <- exp(-cumhaz)
hazard <- ifelse(tt>.001, aa*tt^bb + cc*tt^dd, NA)

plot(tt, survival, type="l", xlab="t", ylab="Survival Function")
```

![Survival Function](image)

Figure 17.1: Survival function

```
plot(tt, cumhaz, type="l", xlab="t", ylab="Cumulative Hazard Function")
```

- Hazard function (force of mortality; instantaneous event rate)
\begin{itemize}
  \item $\approx \text{Prob\{event will occur in small interval around } t \text{ given has not occurred before } t\}$
  
  \item Very useful for learning about mechanisms and forces of risk over time
\end{itemize}

```r
plot(tt, hazard, type="l", xlab="t", ylab="Hazard Function")
```

Figure 17.2: Cumulative hazard function

Figure 17.3: Hazard function
\[ \lambda(t) = \lim_{u \to 0} \frac{\text{Prob}\{t < T \leq t + u | T > t\}}{u}, \]

which using the law of conditional probability becomes

\[ \lambda(t) = \lim_{u \to 0} \frac{\text{Prob}\{t < T \leq t + u\}/\text{Prob}\{T > t\}}{u} \]

\[ = \lim_{u \to 0} \frac{\left[ F(t + u) - F(t) \right]/u}{S(t)} \]

\[ = \frac{\partial F(t)}{\partial t}/S(t) \]

\[ = \frac{f(t)}{S(t)}, \]

\[ \frac{\partial \log S(t)}{\partial t} = \frac{\partial S(t)}{S(t)}/\frac{S(t)}{S(t)} = -\frac{f(t)}{S(t)}, \]

\[ \lambda(t) = -\frac{\partial \log S(t)}{\partial t}, \]

\[ \int_0^t \lambda(v)dv = -\log S(t). \]

\[ \Lambda(t) = -\log S(t), \]

\[ S(t) = \exp[-\Lambda(t)]. \]

- **Expected value of** \( \Lambda(T) = 1 \)

\[ T_q = S^{-1}(1 - q). \]

\[ T_{0.5} = S^{-1}(0.5). \]

\[ T_q = \Lambda^{-1}[-\log(1 - q)] \] and as a special case,
\[ T_{.5} = \Lambda^{-1}(\log 2). \]
\[ \mu = \int_0^\infty S(v)dv. \]

- Event time for subject \( i \): \( T_i \)
- Censoring time: \( C_i \)
- Event indicator:
  \[ e_i = 1 \text{ if the event was observed } (T_i \leq C_i), \]
  \[ = 0 \text{ if the response was censored } (T_i > C_i). \]

- The observed response is
  \[ Y_i = \min(T_i, C_i), \]

<table>
<thead>
<tr>
<th>( T_i )</th>
<th>( C_i )</th>
<th>( Y_i )</th>
<th>( e_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>81</td>
<td>75</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>76</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>68+</td>
<td>68</td>
<td>68</td>
<td>0</td>
</tr>
<tr>
<td>52+</td>
<td>52</td>
<td>52</td>
<td>0</td>
</tr>
<tr>
<td>20</td>
<td>56</td>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>

Termination of Study

Figure 17.4: Some censored data. Circles denote events.
17.4 Homogeneous Failure Time Distributions

- Exponential distribution: constant hazard
  \[
  \Lambda(t) = \lambda t \quad \text{and} \quad S(t) = \exp(-\Lambda(t)) = \exp(-\lambda t).
  \]
  \[
  T_{0.5} = \frac{\log(2)}{\lambda}.
  \]

- Weibull distribution
  \[
  \lambda(t) = \alpha \gamma t^{\gamma-1} \quad \Lambda(t) = \alpha t^\gamma \quad S(t) = \exp(-\alpha t^\gamma).
  \]

\[
\text{tt} \leftarrow \text{seq}(1e-5, 1.2, \text{length}=100) \\
\text{plot(tt, rep(0,100), ylim=c(0,7), type="n", xlab="t", ylab="")} \\
a \leftarrow 1 \\
i \leftarrow 0 \\
\text{for(b in c(.5,1,2,4))} \\
\{ \\
  i \leftarrow i + 1 \\
  \text{lines(tt, a*b*tt^\((b-1)\), lty=i)} \\
\} \\
\text{legend(.4, 6, c(".5","1","2","4"), lty=1:4, cex=.9, bty="n")}
\]
  \[
  T_{0.5} = \left[\frac{(\log 2)}{\alpha}\right]^{1/\gamma}.
  \]

- The restricted cubic spline hazard model with \(k\) knots is
  \[
  \lambda_k(t) = a + bt + \sum_{j=1}^{k-2} \gamma_j w_j(t),
  \]
Figure 17.5: Some Weibull hazard functions with $\alpha = 1$ and various values of $\gamma$. 
17.5

Nonparametric Estimation of $S$ and $\Lambda$

17.5.1

Kaplan–Meier Estimator

- No censoring $\rightarrow$

$$S_n(t) = \left[\text{number of } T_i > t\right]/n.$$  

- Kaplan–Meier (product-limit) estimator

<table>
<thead>
<tr>
<th>Day</th>
<th>No. Subjects</th>
<th>Deaths</th>
<th>Censored</th>
<th>Cumulative Survival</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>100</td>
<td>1</td>
<td>0</td>
<td>99/100 = .99</td>
</tr>
<tr>
<td>30</td>
<td>99</td>
<td>2</td>
<td>1</td>
<td>97/99 × 99/100 = .97</td>
</tr>
<tr>
<td>60</td>
<td>96</td>
<td>0</td>
<td>3</td>
<td>96/96 × .97 = .97</td>
</tr>
<tr>
<td>72</td>
<td>93</td>
<td>3</td>
<td>0</td>
<td>90/93 × .97 = .94</td>
</tr>
</tbody>
</table>

$$S_{KM}(t) = \prod_{i:t_i < t} (1 - d_i/n_i).$$

- The Kaplan–Meier estimator of $\Lambda(t)$ is $\Lambda_{KM}(t) = -\log S_{KM}(t)$.

- Simple example

$$1 \ 3 \ 3 \ 6^+ \ 8^+ \ 9 \ 10^+.$$

\[
\begin{array}{cccc}
i & t_i & n_i & d_i & (n_i - d_i)/n_i \\
1 & 1 & 7 & 1 & 6/7 \\
2 & 3 & 6 & 2 & 4/6 \\
3 & 9 & 2 & 1 & 1/2 \\
\end{array}
\]

\[
S_{KM}(t) = 1, \quad 0 \leq t < 1 \\
= 6/7 = .85, \quad 1 \leq t < 3 \\
= (6/7)(4/6) = .57, \quad 3 \leq t < 9 \\
= (6/7)(4/6)(1/2) = .29, \quad 9 \leq t < 10.
\]

```r
	tt <- c(1,3,3,6,8,9,10)
	stat <- c(1,1,1,0,0,1,0)
	S <- Surv(tt, stat)
	survplot(npsurv(S ~ 1), conf="bands", n.risk=TRUE, xlab="t")
	survplot(npsurv(S ~ 1, type="fleming-harrington", conf.int=FALSE),
	add=TRUE, lty=3)
```

Figure 17.6: Kaplan–Meier product-limit estimator with 0.95 confidence bands. The Altschuler–Nelson–Aalen–Fleming–Harrington estimator is depicted with the dashed lines.

\[
\text{Var}\{\log \Lambda_{KM}(t)\} = \frac{\sum_{i:t_i<t} d_i/[n_i(n_i - d_i)]}{\left\{\sum_{i:t_i<t} \log[(n_i - d_i)/n_i]\right\}^2}
\]
\[ S_{\text{KM}}(t)^{\exp(\pm zs)}. \]

### 17.5.2 Altschuler–Nelson Estimator

\[
\hat{\Lambda}(t) = \sum_{i: t_i < t} \frac{d_i}{n_i}
\]

\[ S_{\Lambda}(t) = \exp(-\hat{\Lambda}(t)) \]
17.6 Analysis of Multiple Endpoints

- Cancer trial: recurrence of disease or death
- CV trial: nonfatal MI or death
- Analyze usual way but watch out for differing risk factors

Analyzing multiple causes of terminating event → Cause-specific hazards, censor on cause not currently analyzed
Not assume mechanism for cause removal or correlations of causes
Problem if apply to a setting where causes are removed differently

- More complex if explicitly handle mixture of nonfatal outcomes with fatal outcome

17.6.1 Competing Risks

- Events independent → analyze separately, censoring others
- → unbiased estimate of cause-specific $\lambda(t)$ or $S(t)$ since censoring non-informative
• $1 - S_{KM}(t)$ estimates $\text{Prob}\{\text{failing from the event in absence of other events}\}$

• Method can work with dependent causes but interpretation difficult

• See Larson and Dinse for joint model of time until any failure and the type of failure

• See Duc and Wolbers [62] for an interesting smooth semi-nonparametric cumulative incidence estimator under competing risks, incorporating a joint distribution of event time and type

17.6.2 Competing Dependent Risks

• Ordinary K-M estimator biased

• Suppose cause $m$ is of interest

• Cause-specific hazard function:

$$\lambda_m(t) = \lim_{u \to 0} \frac{\Pr\{\text{fail from cause } m \text{ in } [t, t + u] | \text{alive at } t\}}{u}.$$ 

• The cumulative incidence function or probability of failure from cause $m$ by time $t$ is given by

$$F_m(t) = \int_0^t \lambda_m(u) S(u) du,$$
where $S(u)$ is the probability of surviving (ignoring cause of death), which equals
\[ \exp[-\int_0^u (\sum \lambda_m(x)) dx] \]

- $1 - F_m(t) = \exp[-\int_0^t \lambda_m(u) du]$ only if failures due to other causes are eliminated and if the cause-specific hazard of interest remains unchanged in doing so.

- **Nonparametric estimator of $F_m(t)$:**
  \[
  \hat{F}_m(t) = \sum_{i: t_i \leq t} \frac{d_{mi}}{n_i} S_{KM}(t_{i-1}),
  \]
  where $d_{mi}$ is the number of failures of type $m$ at time $t_i$ and $n_i$ is the number of subjects at risk of failure at $t_i$.

- Pepe *et al.* show how to use combo of K-M estimators to estimate $\text{Prob}\{\text{free of event 1 by time } t \text{ given event 2 not occur by } t\}$

- Suppose event 1 not a terminating event (e.g., death); even after event 1 subjects followed to find occurrences of event 2

- **$\text{Prob}\{T_1 > t \text{ given } T_2 > t\}$:**
  \[
  \text{Prob}\{T_1 > t | T_2 > t\} = \frac{\text{Prob}\{T_1 > t \text{ and } T_2 > t\}}{\text{Prob}\{T_2 > t\}} = \frac{S_{12}(t)}{S_2(t)}.
  \]
where $S_{12}(t)$ is the survival function for $\min(T_1, T_2)$

- Est. of $\text{Prob}\{\text{subject still alive at } t \text{ will be free of MI at } t\}$: $S_{KM12}/S_{KM2}$

- Can also easily compute $\text{Prob}\{\text{event 1 occurs by time } t \text{ and that event 2 has not occurred by } t\} \rightarrow S_2(t) - S_{12}(t) = [1 - S_{12}(t)] - [1 - S_2(t)]$

- **Crude survival functions** come from *marginal distributions*, i.e. $\text{Prob}\{T_1 > t \text{ whether or not event 2 occurs}\}$

  $$S_c(t) = 1 - F_1(t)$$
  $$F_1(t) = \text{Prob}\{T_1 \leq t\},$$

- $F_1(t)$ is **crude incidence function**

- $T_1 < t \rightarrow$ occurrence of event 1 is part of the prob. being computed

- Event 2 terminating \rightarrow some subjects can never suffer event 1 \rightarrow crude survival fctn. for $T_1$ never drops to zero

- Crude survival fctn interpreted as surv. dist. of $W$ where $W = T_1$ if $T_1 < T_2$ and $W = \infty$ otherwise
17.6.3 State Transitions and Multiple Types of Nonfatal Events

Multiple live states, one absorbing state (all-cause mortality)

- Strauss & Shavelle extended Kaplan–Meier estimator [200]
- Estimate $\pi_{ij}(t_1, t_2)$ that subject in state $i$ at time $t_1$ is in state $j$ $t_2$ time units later
- Define $S^i_{KM}(t|t_1) = \text{Kaplan–Meier estimate of } \text{Prob}\{\text{surviving } t \text{ additional years for cohort of subjects beginning follow-up at time } t_1 \text{ in state } i\}$

Then

$$\pi_{ij}(t_1, t_2) = \frac{n_{ij}(t_1, t_2)}{n_i(t_1, t_2)} S^i_{KM}(t_2|t_1),$$

where $n_i(t_1, t_2)$ is the number of subjects in live state $i$ at time $t_1$ who are alive and uncensored $t_2$ time units later, and $n_{ij}(t_1, t_2)$ is the number of such subjects in state $j$ $t_2$ time units beyond $t_1$.

17.6.4 Joint Analysis of Time and Severity of an Event

- Can give more weight to an event that occurs earlier or is more severe
• Berridge and Whitehead

• Ordinal scale for severity

• Severity measured at time of (single) event

• Ex: time until first headache / severity of headache

• “joint hazard function”, ordinal category $j$

$$\lambda_j(t) = \lambda(t)\pi_j(t),$$

• Allows shift in dist. of response severity as $t \uparrow$

---

**Analysis of Multiple Events**

• Ex: MI, ulcer, pregnancy, infection

• Analysis of time to first event may lose info

• Specialized multivariate failure time methods exist

• Simpler to model *marginal distributions*

• One record per event per subject

• Can have # previous events as covariable

• Correct variances for intra-subject correlation using clustered
sandwich estimator

- Method can handle multiple events of differing types
S Functions

- `event.chart` in Hmisc draws a variety of charts for displaying raw survival time data, for both single and multiple events per subject (see also `event.history`)

- Analyses in this chapter can be done as special cases of the Cox model

- Particular functions for this chapter (no covariables) from Therneau:

  - `Surv` function: Combines time to event variable and event/-censoring indicator into a single survival time matrix object

  - Right censoring: `Surv(y, event)`; event is event/censoring indicator, usually coded `TRUE/FALSE`, 0=censored 1=event or 1=censored 2=event. If the event status variable has other coding, e.g., 3 means death, use `Surv(y, s==3)`.

  - `survfit`: Kaplan–Meier and other nonparametric survival curves using the `survival` package

  - `npsurv`: rms package wrapper for `survfit`

```r
units(y) <- "Month" # Default is "Day" - used for axis labels, etc.
survfit(Surv(y, event) ~ svar1 + svar2 + ..., data, subset, na.action=na.delete, type=c("kaplan-meier", "fleming-harrington", "fh2"), error=c("greenwood", "tsiatis"), se.fit=TRUE, conf.int=.95,
```

conf.type=c("log", "plain", "log-log"))

If there are no stratification variables (svar1, ...), omit them. To print a table of estimates, use

```r
f <- survfit(...)
print(f)  # print brief summary of f
summary(f, times, censored=FALSE)
```

For failure times stored in days, use

```r
f <- survfit(Surv(futime, event) ~ sex)
summary(f, seq(30,180, by=30))
```

to print monthly estimates.

To plot the object returned by `survfit`, use

```r
plot(f, conf.int=TRUE, mark.time=TRUE, mark=3, col=1, lty=1, lwd=1, cex=1, log=FALSE, xscale=1, yscale=1, xlab="", ylab="", xaxs="S", ...)```

This invokes `plot.survfit`. More options: use `npsurv` and `survplot` in `rms` for other options that include automatic curve labeling and showing the number of subjects at risk at selected times. See code for Figure 17.6 above.

Stratified estimates, with four treatments distinguished by line type and curve labels, could be drawn by

```r
units(y) <- "Year"
f <- npsurv(Surv(y, stat) ~ treatment)
survplot(f, ylab="Fraction Pain-Free")
```

- `bootkm` function in `Hmisc` bootstraps Kaplan–Meier survival estimates or Kaplan–Meier estimates of quantiles of the survival time distribution. It is easy to use `bootkm` to compute for example a nonparametric confidence interval for the ratio of median survival times for two groups.
Chapter 18

Parametric Survival Models

18.1 Homogeneous Models (No Predictors)

Why use a parametric model?

1. easily compute selected quantiles of the survival distribution
2. estimate (usually by extrapolation) Regarding power, I think the paper by Moore explains it better than I can: Schoenfeld suggested that preliminary hypothesis testing for efficacy could be conducted with a high type I error rate (false positive rate up to 0.25) and that subsequent testing would provide a control over the chances of a false discovery in spite of the initial elevated error rate. the expected failure time
3. derive a concise equation and smooth function for estimating $S(t)$, $\Lambda(t)$, and $\lambda(t)$
4. estimate $S(t)$ more precisely than $S_{KM}(t)$ or $S_{A}(t)$ (Altschuler-Nelson-Fleming-Harrington estimator) if the parametric form is cor-
Note: Fitting more than two smooth survival curves and choosing the one that best reproduces the KM estimator will result in a true precision no better than KM due to model uncertainty.

18.1.1 Specific Models

- Seen exponential and Weibull already
- Many others obtained by assuming $\log(T)$ has a certain dist.
- Log-normal: $S(t) = 1 - \Phi\left(\frac{\log(t) - \mu}{\sigma}\right)$
- Log-logistic: $S(t) = \left[1 + \exp\left(-\frac{\log(t) - \mu}{\sigma}\right)\right]^{-1}$
- Log-extreme value: $S(t) = \exp\left[-\exp\left(\frac{\log(t) - \mu}{\sigma}\right)\right]$ another way of expressing Weibull

18.1.2 Estimation

- Log-likelihood for exponential distribution

$$\log L = \sum_{i:Y_i \text{ uncensored}} \log \lambda - \sum_{i=1}^{n} \lambda Y_i.$$ 

$$\hat{\lambda} = n_u/w$$
\[
\text{var}(\hat{\lambda}) = n_u/w^2 \\
\text{var}(\log \hat{\lambda}) = 1/n_u \\
\hat{\mu} = w/n_u \\
\hat{S}(t) = \exp(-\hat{\lambda}t)
\]

Consider these failure time data:

\[1\ 3\ 3\ 6^+\ 8^+\ 9\ 10^+\]

```
require(rms)
S <- Surv(c(1, 3, 3, 6, 8, 9, 10), c(1,1,1,0,0,1,0))
fe <- psm(S ~ 1, dist='exponential')
f2 <- psm(S ~ 1, dist='weibull')
```

\[
n_u = 4 \\
w = 40 \\
\hat{\mu} = 10 \pm 5 \\
T_{0.5} = 10 \log(2)
\]

- Weibull fit

\[
\hat{\alpha} = 0.0728 \\
\hat{\gamma} = 1.164 \\
\hat{S}(t) = \exp(-0.0728t^{1.164}) \\
\hat{S}^{-1}(0.5) = \left[(\log 2)/\hat{\alpha}\right]^{1/\hat{\gamma}} = 6.935
\]

(estimated median).
Assessment of Model Fit

- Example: Weibull
  \[ \log[-\log S(t)] = \log \Lambda(t) = \log \alpha + \gamma(\log t). \]

- Plot \( \log \hat{\Lambda}(t) \) versus \( \log t \)

- For assumed dist. \( S(t) \) plot \( S^{-1}[S_{\Lambda}(t)] \) or \( S^{-1}[S_{KM}(t)] \) against \( t \), check for linearity

- Log-distributions: plot vs. \( \log t \)

- Check log-normal: plot \( \Phi^{-1}[S_{\Lambda}(t)] \) vs. \( \log t \)

- Check log-logistic: plot \( \text{logit}[S_{\Lambda}(t)] \) vs. \( \log t \)

- Alternative: plot fitted \( \hat{S}(t) \) and \( S_{\Lambda}(t) \) vs. \( t \) on the same graph
18.2 Parametric Proportional Hazards Models

18.2.1 Model

\[ \lambda(t|X) = \lambda(t) \exp(X\beta). \]
\[ \Lambda(t|X) = \Lambda(t) \exp(X\beta) \]
\[ S(t|X) = \exp[-\Lambda(t) \exp(X\beta)] = \exp[-\Lambda(t)]^{\exp(X\beta)}. \]
\[ S(t|X) = S(t)^{\exp(X\beta)}, \]

18.2.2 Model Assumptions and Interpretation of Parameters

\[ \log \lambda(t|X) = \log \lambda(t) + X\beta \]
\[ \log \Lambda(t|X) = \log \Lambda(t) + X\beta. \]

Assumptions:

- Underlying functions \((\lambda, \Lambda, S)\)
- Linear effect of predictors on \(\log \lambda, \log \Lambda\)
- No interaction between \(X\) and \(t \rightarrow \) impact same over time
  \[ \beta_j = \log \lambda(t|X_1, X_2, \ldots, X_j + 1, X_{j+1}, \ldots, X_k) \]
\[
- \log \lambda(t|X_1, \ldots, X_j, \ldots, X_k),
\]

- Effect of increasing \( X_j \) by \( d \) is to increase \( \lambda \) by factor of \( \exp(\beta_j d) \)

- One binary predictor:

\[
\lambda(t|X_1 = 0) = \lambda(t)
\]

\[
\lambda(t|X_1 = 1) = \lambda(t) \exp(\beta_1).
\]

Here \( \exp(\beta_1) \) is the \( X_1 = 1 : X_1 = 0 \) hazard ratio.

- One continuous predictor:

\[
\lambda(t|X_1) = \lambda(t) \exp(\beta_1 X_1). \]

---

### 18.2.3 Hazard Ratio, Risk Ratio, and Risk Difference

\[
S_T = S_C^{0.5}
\]

<table>
<thead>
<tr>
<th>Subject</th>
<th>5-Year Survival Difference</th>
<th>Mortality Ratio (T/C)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>A</td>
<td>0.98</td>
<td>0.99</td>
</tr>
<tr>
<td>B</td>
<td>0.80</td>
<td>0.89</td>
</tr>
<tr>
<td>C</td>
<td>0.25</td>
<td>0.50</td>
</tr>
</tbody>
</table>

```r
plot(0, 0, type="n", xlab="Survival for Control Subject",
     ylab="Improvement in Survival",
     xlim=c(0,1), ylim=c(0,.7))
i <- 0
hr <- seq(.1, .9, by=.1)
for(h in hr) {
i <- i + 1
    p <- seq(.0001, .9999, length=200)
p2 <- p^h
d <- p2 - p
    lines(p, d, lty=i)
}```
Exponential:

\[
\begin{align*}
\lambda(t|X) &= \lambda \exp(X \beta) \\
S(t|X) &= \exp[-\lambda t \exp(X \beta)] = \exp(-\lambda t) \exp(X \beta).
\end{align*}
\]

\[
E\{T|X\} = \frac{1}{\lambda \exp(X \beta)}
\]

\[
T_{0.5}|X = \frac{(\log 2)}{\lambda \exp(X \beta)}.
\]
• Weibull:

\[
\begin{align*}
\lambda(t|X) &= \alpha \gamma t^{\gamma-1} \exp(X\beta) \\
\Lambda(t|X) &= \alpha t^\gamma \exp(X\beta) \\
S(t|X) &= \exp[-\alpha t^\gamma \exp(X\beta)] \\
&= \exp(-\alpha t^\gamma) \exp(X\beta).
\end{align*}
\]

\[
T_{0.5}|X = \left\{\log 2/\left[\alpha \exp(X\beta)\right]\right\}^{1/\gamma}.
\]

For numerical reasons, re-write:

\[
S(t|X) = \exp(-\Lambda(t|X)), \quad \text{where}
\]

\[
\Lambda(t|X) = \exp(\gamma \log t + X\beta).
\]

See also spline hazard models [103, 102, 122] and the generalized gamma distribution [51].

---

18.2.5

**Assessment of Model Fit**

If \( \lambda(t) \) is Weibull, the two curves will be linear if \( \log t \) is plotted instead of \( t \) on the \( x \)-axis.

• Weibull: Stratify on \( X \), plot \( \log \Lambda_{KM}(t|X \text{ stratum}) \) vs. \( \log t \).

• Assesses PH in addition to shape assumptions—all curves should be parallel as well as straight.
Figure 18.2: PH Model with one binary predictor. Y-axis is $\log \lambda(t)$ or $\log \Lambda(t)$. For $\log \Lambda(t)$, the curves must be non-decreasing. For $\log \lambda(t)$, they may be any shape.

Figure 18.3: PH model with one continuous predictor. Y-axis is $\log \lambda(t)$ or $\log \Lambda(t)$. For $\log \Lambda(t)$, drawn for $t_2 > t_1$. The slope of each line is $\beta_1$. 
Figure 18.4: PH model with one continuous predictor. Y-axis is $\log \lambda(t)$ or $\log \Lambda(t)$. For $\log \lambda$, the functions need not be monotonic.

Figure 18.5: Regression assumptions, linear additive PH or AFT model with two predictors. For PH, Y-axis is $\log \lambda(t)$ or $\log \Lambda(t)$ for a fixed $t$. For AFT, Y-axis is $\log(T)$.
18.3

Accelerated Failure Time Models

18.3.1

Model

- Specifies that predictors act multiplicatively on failure time
- Alters rate subject proceeds along time axis

\[ S(t|X) = \psi \left( \frac{\log(t) - X\beta}{\sigma} \right), \]

\[ \frac{\log(T) - X\beta}{\sigma} \sim \psi \]

\[ \log(T) = X\beta + \sigma \epsilon \]

\[ \epsilon \sim \psi \]

- Weibull (and exponential) members of PH and AFT

18.3.2

Model Assumptions and Interpretation of Parameters

\[ \psi^{-1}[S(t|X)] = \frac{\log(t) - X\beta}{\sigma}. \]

Letting \( \epsilon \sim \psi \)

\[ \log(T) = X\beta + \sigma \epsilon. \]

Check that residuals \( \log(T) - X\hat{\beta} \sim \psi \) (within scale factor).
The assumptions of the AFT model are thus

1. The true form of \( \psi \) (the distributional family) is correctly specified.

2. In the absence of nonlinear and interaction terms, each \( X_j \) affects \( \log(T) \) or \( \psi^{-1}[S(t|X)] \) linearly.

3. Implicit in these assumptions is that \( \sigma \) is a constant independent of \( X \).

1-unit change in \( X_j = \beta_j \) change in \( \log T \), or increase \( T \) by factor of \( \exp(\beta_j) \).

Median survival time:

\[
T_{0.5}|X = \exp[X\beta + \sigma \psi^{-1}(0.5)]
\]

### Specific Models

- **Extreme value**: \( \psi(u) = \exp(-\exp(u)) \)

- **Logistic**: \( \psi(u) = [1 + \exp(u)]^{-1} \)

- **Normal**: \( \psi(u) = 1 - \Phi(u) \)

- **Log-normal**: 

\[
S(t|X) = 1 - \Phi\left(\frac{\log(t) - X\beta}{\sigma}\right),
\]
• Log-logistic:

\[
S(t|X) = \left[1 + \exp\left(\frac{\log(t) - X\beta}{\sigma}\right)\right]^{-1}.
\]

**Estimation**

Works better if \(\sigma\) parameterized as \(\exp(\delta)\).

\[
\hat{S}(t|X) = \psi\left(\frac{\log(t) - X\hat{\beta}}{\hat{\sigma}}\right)
\]

\[
\hat{T}_{0.5}|X = \exp[X\hat{\beta} + \hat{\sigma}\psi^{-1}(0.5)].
\]

Normal and logistic: \(\hat{T}_{0.5}|X = \exp(X\hat{\beta})\).

\[
\psi\left(\frac{\log(t) - X\hat{\beta}}{\hat{\sigma}}\right) \pm z_{1-\alpha/2} \times s.
\]

**Residuals**

For an AFT model, standardized residuals are simply

\[
r = (\log(T) - X\hat{\beta})/\sigma.
\]

When \(T\) is right-censored, \(r\) is right-censored.

**Assessment of Model Fit**
Figure 18.6: AFT model with one predictor. Y-axis is $\psi^{-1}[S(t|X)] = \frac{\log(t) - X\beta}{\sigma}$. Drawn for $d > c$. The slope of the lines is $\sigma^{-1}$.

Figure 18.7: AFT model with one continuous predictor. Y-axis is $\psi^{-1}[S(t|X)] = \frac{\log(t) - X\beta}{\sigma}$. Drawn for $t_2 > t_1$. The slope of each line is $\beta_1/\sigma$ and the difference between the lines is $\frac{1}{\sigma} \log(t_2/t_1)$. 
CHAPTER 18. PARAMETRIC SURVIVAL MODELS

<table>
<thead>
<tr>
<th>Group 1</th>
<th>143</th>
<th>164</th>
<th>188</th>
<th>188</th>
<th>190</th>
<th>192</th>
<th>206</th>
<th>209</th>
<th>213</th>
<th>216</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>220</td>
<td>227</td>
<td>230</td>
<td>234</td>
<td>246</td>
<td>265</td>
<td>304</td>
<td>216+</td>
<td>244+</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group 2</th>
<th>142</th>
<th>156</th>
<th>163</th>
<th>198</th>
<th>205</th>
<th>232</th>
<th>232</th>
<th>233</th>
<th>233</th>
<th>233</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>233</td>
<td>239</td>
<td>240</td>
<td>261</td>
<td>280</td>
<td>280</td>
<td>296</td>
<td>296</td>
<td>323</td>
<td>204+</td>
</tr>
<tr>
<td></td>
<td>344+</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

getHdata(kprats)
kprats$group ← factor(kprats$group, 0:1, c('Group 1', 'Group 2'))
dd ← datadist(kprats); options(datadist="dd")

S ← with(kprats, Surv(t, death))
f ← npsurv(S ~ group, type="fleming", data=kprats)
survplot(f, n.risk=TRUE, type="fleming", data=kprats)

# Figure 18.8
label.curves=list(keys='lines'), levels.only=TRUE)
title(sub="Nonparametric estimates", adj=0, cex=.7)

# Check fits of Weibull, log-logistic, log-normal
xl ← c(4.8, 5.9)
survplot(f, loglog=TRUE, logt=TRUE, conf="none", xlim=xl,
        label.curves=list(keys='lines'), levels.only=TRUE)
title(sub="Weibull (extreme value)", adj=0, cex=.7)
survplot(f, fun=function(y) log(y/(1-y)), ylab="logit S(t)",
         logt=TRUE, conf="none", xlim=xl,
         label.curves=list(keys='lines'), levels.only=TRUE)
title(sub="Log-logistic", adj=0, cex=.7)
survplot(f, fun=qnorm, ylab="Inverse Normal S(t)",
         logt=TRUE, conf="none",
         xlim=xl, cex.label=.7,
         label.curves=list(keys='lines'), levels.only=TRUE)
title(sub="Log-normal", adj=0, cex=.7)

Fit Weibull (in aft form), log-logistic, and log-normal models.

fw ← psm(S ~ group, data=kprats, dist='weibull')
fl ← psm(S ~ group, data=kprats, dist='loglogistic',
        y=TRUE)
fn ← psm(S ~ group, data=kprats, dist='lognormal')
cat('\\textbf{Weibull default form}: \\

\textbf{Weibull default form:}

latex(fw)

\[
\text{Prob}\{ T \geq t \} = \exp[- \exp(\frac{\log(t) - X\beta}{0.1832976})] \quad \text{where}
\]

\[
X\hat{\beta} = 5.450859
\]
Figure 18.8: Altschuler-Nelson-Fleming-Harrington nonparametric survival estimates for rats treated with DMBA [167], along with various transformations of the estimates for checking distributional assumptions of 3 parametric survival models.
and \([c] = 1\) if subject is in group \(c\), 0 otherwise

**Weibull PH form:**

\[
\text{Prob}\{T \geq t\} = \exp\{-t^{5.455608} \exp(X\hat{\beta})\} \quad \text{where}
\]

\[
X\hat{\beta} = \\
-29.73775 \\
-0.7200475\text{[Group 2]}
\]

and \([c] = 1\) if subject is in group \(c\), 0 otherwise

**Log-logistic:**

\[
\text{Prob}\{T \geq t\} = \left[1 + \exp\left(\frac{\log(t) - X\beta}{0.1159753}\right)\right]^{-1} \quad \text{where}
\]

\[
X\hat{\beta} = \\
5.375675 \\
+0.1051005\text{[Group 2]}
\]

and \([c] = 1\) if subject is in group \(c\), 0 otherwise

**Log-normal**

\[
\text{Prob}\{T \geq t\} = \exp\{-t^{5.455608} \exp(X\hat{\beta})\} \quad \text{where}
\]

\[
X\hat{\beta} = \\
-29.73775 \\
-0.7200475\text{[Group 2]}
\]
\[ \text{Prob}\{T \geq t\} = 1 - \Phi\left( \frac{\log(t) - X\beta}{0.2100184} \right) \]

where

\[ X\hat{\beta} = 5.375328 + 0.0930606 [\text{Group 2}] \]

and \([c] = 1\) if subject is in group \(c\), 0 otherwise

<table>
<thead>
<tr>
<th>Model</th>
<th>Group 2:1 Failure Time Ratio</th>
<th>Median Survival Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Group 1</td>
<td>Group 2</td>
</tr>
<tr>
<td>Extreme Value (Weibull)</td>
<td>1.14</td>
<td>217</td>
</tr>
<tr>
<td>Log-logistic</td>
<td>1.11</td>
<td>217</td>
</tr>
<tr>
<td>Log-normal</td>
<td>1.10</td>
<td>217</td>
</tr>
</tbody>
</table>

- More general approach to verifying distributional assumptions:

- Plot nonparametric estimate of survival distribution of \(r\)

- Superimpose theoretical standardized distribution

- Can get distribution of residuals separately by strata — should all have same standardized distribution (e.g., same \(\sigma\))
Figure 18.9: Agreement between fitted log-logistic model and nonparametric survival estimates for rat vaginal cancer data

Figure 18.10: Kaplan-Meier estimates of distribution of standardized, censored residuals from the log-logistic model, along with the assumed standard log-logistic distribution (blue). Red step function is the estimated distribution of all residuals; black step functions are the estimated distributions of residuals stratified by group, as indicated.
Derive R code for median, mean, hazard, survival functions

```r
med ← Quantile(fl)
l latex(med, fi='', type='Sinput')

med ← function (q = 0.5, lp = NULL, parms = -2.15437773933124)
{
  names(parms) ← NULL
  f ← function(lp, q, parms) lp + exp(parms) * logb(q/(1 - q))
  names(q) ← format(q)
  drop(exp(outer(lp, q, FUN = f, parms = parms)))
}

meant ← Mean(fl)
l latex(meant, fi='', type='Sinput')

meant ← function (lp = NULL, parms = -2.15437773933124)
{
  names(parms) ← NULL
  if (exp(parms) > 1)
    rep(Inf, length(lp))
  else exp(lp) * pi * exp(parms)/sin(pi * exp(parms))
}

haz ← Hazard(fl)
l latex(haz, fi='', type='Sinput')

haz ← function (times = NA, lp = NULL, parms = -2.15437773933124)
{
  t.trans ← logb(times)
  t.deriv ← 1/times
  scale ← exp(parms)
  names(t.trans) ← format(times)
  t.deriv/scale/(1 + exp(-(t.trans - lp)/scale))
}

surv ← Survival(fl)
l latex(surv, fi='', type='Sinput')

surv ← function (times = NULL, lp = NULL, parms = -2.15437773933124)
{
  1/(1 + exp((logb(times) - lp)/exp(parms)))
}

Show fitted hazard function from log-logistic, and add median survival time to graph

```r
# Plot estimated hazard functions and add median survival times to graph
survplot(fl, group, what='hazard')  # Figure 18.11
# Compute median survival time
m ← med(lp=predict(fl,
                 data.frame(group=levels(kprats$group))))
m
```
Figure 18.11: Estimated hazard functions for log-logistic fit to rat vaginal cancer data, along with median survival times

Validating the Fitted Model

- Check distributional shape
• Group predicted $t$-year survival and plot Kaplan-Meier estimate at $t$ vs. mean predicted $\hat{S}$

• Cox-Snell residuals — check against $U[0, 1]$

• loess smooth of $F(T|X) - 0.5F(C|X)$ against $X\hat{\beta}$ or $\frac{2F(T|X)}{F(C|X)}$ vs. $X\hat{\beta}$ if $C$ is known

See the val.surv function in the rms package.
18.4 R Functions

- **survival package**: `survreg` for Weibull, log-normal, log-logistic, etc.

- **rms package**: `psm` front-end for `survreg`

- **rstpm2 package**: [cran.r-project.org/web/packages/rstpm2](https://cran.r-project.org/web/packages/rstpm2) which has more general AFT models

- Many other R packages: [cran.r-project.org/web/views/Survival.html](https://cran.r-project.org/web/views/Survival.html)
Data source: Random sample of 1000 patients from Phases I & II of SUPPORT (Study to Understand Prognoses Preferences Outcomes and Risks of Treatment, funded by the Robert Wood Johnson Foundation). See [117]. The dataset is available from https://hbiostat.org/data.

- Analyze acute disease subset of SUPPORT (acute respiratory failure, multiple organ system failure, coma) — the shape of the survival curves is different between acute and chronic disease categories
- Patients had to survive until day 3 of the study to qualify
- Baseline physiologic variables measured during day 3
19.1 Descriptive Statistics

Create a variable `acute` to flag categories of interest; print univariate descriptive statistics.

```r
require(rms)
options(prType='latex')  # for print, summary, anova
getHdata(support)        # Get data frame from web site
acute ← support$dzclass %in% c('ARF/MOSF','Coma')
ltx ← latex(describe(support[acute,]), file='')
```

```latex
\begin{verbatim}
support[acute, ]
\end{verbatim}

\begin{verbatim}
35 Variables 537 Observations
\end{verbatim}

\begin{verbatim}
age : Age
   n missing distinct  Info  Mean  Gmd .05  .10  .25  .50  .75  .90  .95
537     0      529    0   1  60.7   19.98  32.49  47.93  63.67  74.49 81.54 85.56

lowest: 18.04 18.41 19.76 20.29 20.31, highest: 91.62 91.82 91.93 92.74 95.51

death : Death at any time up to NDI date:31DEC94
   n missing distinct  Info  Sum  Mean  Gmd
537     0      2    0.67 356   0.66 0.44

sex
   n missing distinct
537     0      2

Value  female  male
Frequency 251 286
Proportion 0.467 0.533

hospdead : Death in Hospital
   n missing distinct  Info  Sum  Mean  Gmd
537     0      2   0.70 201  0.37 0.47

slos : Days from Study Entry to Discharge
   n missing distinct  Info  Mean  Gmd .05  .10  .25  .50  .75  .90  .95
537     0      85    0 23.44  22.24  4.0  5.0  9.0 15.0 27.0 47.4 68.2

lowest: 3 4 5 6 7, highest: 145 164 202 236 241

d.time : Days of Follow-Up
   n missing distinct  Info  Mean  Gmd .05  .10  .25  .50  .75  .90  .95
537     0     340    0 446.1 566.1  4.6  6.0 16.0 18.2 72.4 142.1 174.2


dzgroup
   n missing distinct
537     0      3

Value  ARF/MOSF w/Sepsis  Coma  MOSF w/Malig
Frequency 391 60 88
Proportion 0.728 0.112 0.160

dzclass
   n missing distinct
537     0      2

Value  ARF/MOSF  Coma
Frequency 477 60
Proportion 0.888 0.112
\end{verbatim}
```
### Table 19.1: Descriptive Statistics for Numerical Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
</tr>
</thead>
<tbody>
<tr>
<td>num.co: Number of comorbidities</td>
<td>537</td>
<td>0</td>
<td>7</td>
<td>0.926</td>
<td>1.525</td>
</tr>
<tr>
<td>edu: Years of Education</td>
<td>411</td>
<td>126</td>
<td>22</td>
<td>0.957</td>
<td>12.03</td>
</tr>
<tr>
<td>income</td>
<td>335</td>
<td>202</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>scoma: SUPPORT Coma Score based on Glasgow D3</td>
<td>537</td>
<td>0</td>
<td>11</td>
<td>0.022</td>
<td>19.24</td>
</tr>
<tr>
<td>charges: Hospital Charges</td>
<td>517</td>
<td>20</td>
<td>516</td>
<td>1</td>
<td>86652</td>
</tr>
<tr>
<td>totcst: Total RCC cost</td>
<td>471</td>
<td>66</td>
<td>471</td>
<td>1</td>
<td>46195</td>
</tr>
<tr>
<td>avtisst: Average TISS, Days 3-25</td>
<td>536</td>
<td>1</td>
<td>205</td>
<td>1</td>
<td>29.83</td>
</tr>
<tr>
<td>race</td>
<td>535</td>
<td>2</td>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>meanbp: Mean Arterial Blood Pressure Day 3</td>
<td>537</td>
<td>0</td>
<td>109</td>
<td>1</td>
<td>83.28</td>
</tr>
</tbody>
</table>

### Table 19.2: Categorical Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
</tr>
</thead>
<tbody>
<tr>
<td>lowest</td>
<td>0 1 2 3 4</td>
<td>highest: 2 3 4 5 6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>edu</td>
<td>lowest: 0 1 2 3 4</td>
<td>highest: 17 18 19 20 22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>income</td>
<td>Value under $11k, $11-$25k, $25-$50k, &gt;$50k</td>
<td>Frequency</td>
<td>Proportion</td>
<td></td>
<td></td>
</tr>
<tr>
<td>race</td>
<td>lowest: white black asian other hispanic, highest: white black asian other hispanic</td>
<td>Proportion</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>meanbp</td>
<td>lowest: 0 20 27 30 32</td>
<td>highest: 155 158 161 162 180</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>wblc</td>
<td>lowest: 0.04999542 0.06999207 0.09999004 0.14999390 0.19998169</td>
<td>highest: 61.38943759 81.19551260 61.19551260 79.39062500 100.00000000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Chapter 19. Parametric Survival Modeling and Model Approximation

#### hrt: Heart Rate Day 3

<table>
<thead>
<tr>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>537</td>
<td>0</td>
<td>111</td>
<td>0.999</td>
<td>105</td>
<td>38.59</td>
<td>51</td>
<td>60</td>
<td>75</td>
<td>111</td>
<td>126</td>
<td>140</td>
</tr>
</tbody>
</table>

**lowest:** 0  11  30  36  40, **highest:** 189  193  199  232  300

#### resp: Respiration Rate Day 3

<table>
<thead>
<tr>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>537</td>
<td>0</td>
<td>45</td>
<td>0.997</td>
<td>23.72</td>
<td>12.65</td>
<td>8</td>
<td>10</td>
<td>12</td>
<td>24</td>
<td>32</td>
<td>39</td>
</tr>
</tbody>
</table>

**lowest:** 0  4  6  7  8, **highest:** 48  49  52  60  64

#### temp: Temperature (Celsius) Day 3

<table>
<thead>
<tr>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>537</td>
<td>0</td>
<td>61</td>
<td>0.999</td>
<td>37.52</td>
<td>1.505</td>
<td>35.50</td>
<td>35.80</td>
<td>36.40</td>
<td>37.80</td>
<td>38.50</td>
<td>39.09</td>
</tr>
</tbody>
</table>

**lowest:** 32.500000 34.000000 34.09375 34.89844 35.000000, **highest:** 40.19531 40.59375 40.89844 41.000000 41.19531

#### pafi: PaO2/(.01*FiO2) Day 3

<table>
<thead>
<tr>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>37</td>
<td>357</td>
<td>1</td>
<td>277.2</td>
<td>125</td>
<td>86.99</td>
<td>105.08</td>
<td>137.88</td>
<td>202.56</td>
<td>290.00</td>
<td>390.49</td>
</tr>
</tbody>
</table>

**lowest:** 45.000000 48.000000 53.32812 54.000000 55.000000, **highest:** 574.300000 595.125000 600.000000 605.375000

#### alb: Serum Albumin Day 3

<table>
<thead>
<tr>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>346</td>
<td>191</td>
<td>34</td>
<td>0.997</td>
<td>2.668</td>
<td>0.7219</td>
<td>1.700</td>
<td>1.900</td>
<td>2.225</td>
<td>2.600</td>
<td>3.100</td>
<td>3.400</td>
</tr>
</tbody>
</table>

**lowest:** 1.0999854 1.199991 1.299990 1.399990 1.500000, **highest:** 4.0999914 4.199991 4.500000 4.699991 4.799991

#### bili: Bilirubin Day 3

<table>
<thead>
<tr>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>386</td>
<td>151</td>
<td>88</td>
<td>0.997</td>
<td>2.678</td>
<td>3.507</td>
<td>0.3000</td>
<td>0.4000</td>
<td>0.6000</td>
<td>0.8999</td>
<td>2.0000</td>
<td>6.5996</td>
</tr>
</tbody>
</table>

**lowest:** 0.09999084 0.19998169 0.29998779 0.39996338 0.50000000, **highest:** 22.59765620 30.00000000 31.50000000 35.00000000 39.29687500

#### crea: Serum creatinine Day 3

<table>
<thead>
<tr>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>537</td>
<td>0</td>
<td>84</td>
<td>0.998</td>
<td>2.232</td>
<td>1.997</td>
<td>0.6000</td>
<td>0.7000</td>
<td>0.8999</td>
<td>1.3999</td>
<td>2.5996</td>
<td>5.2395</td>
</tr>
</tbody>
</table>

**lowest:** 0.2999878 0.3999634 0.5000000 0.5999756 0.6999612, **highest:** 10.3984375 10.5960941 11.1992188 11.5960941 11.7988281

#### sod: Serum sodium Day 3

<table>
<thead>
<tr>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>537</td>
<td>0</td>
<td>38</td>
<td>0.997</td>
<td>138.1</td>
<td>7.471</td>
<td>129</td>
<td>131</td>
<td>134</td>
<td>137</td>
<td>142</td>
<td>147</td>
</tr>
</tbody>
</table>

**lowest:** 118  120  121  126  127, **highest:** 156  157  158  168  175

#### ph: Serum pH (arterial) Day 3

<table>
<thead>
<tr>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>37</td>
<td>49</td>
<td>0.996</td>
<td>7.418</td>
<td>0.06775</td>
<td>7.170</td>
<td>7.319</td>
<td>7.380</td>
<td>7.426</td>
<td>7.470</td>
<td>7.510</td>
</tr>
</tbody>
</table>

**lowest:** 6.959961 6.989258 7.069336 7.119141 7.129883, **highest:** 7.559570 7.569336 7.589844 7.599609 7.659180

#### glucose: Glucose Day 3

<table>
<thead>
<tr>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>297</td>
<td>240</td>
<td>179</td>
<td>1</td>
<td>167.7</td>
<td>92.13</td>
<td>76.0</td>
<td>89.0</td>
<td>106.0</td>
<td>141.0</td>
<td>200.0</td>
<td>292.4</td>
</tr>
</tbody>
</table>

**lowest:** 30  42  52  68, **highest:** 446  468  492  576  598

#### bun: BUN Day 3

<table>
<thead>
<tr>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>304</td>
<td>233</td>
<td>100</td>
<td>1</td>
<td>38.91</td>
<td>31.12</td>
<td>8.00</td>
<td>11.00</td>
<td>16.75</td>
<td>30.00</td>
<td>56.00</td>
<td>79.70</td>
</tr>
</tbody>
</table>

**lowest:** 1  3  4  5  6, **highest:** 123  124  126  128  146

#### urine: Urine Output Day 3

<table>
<thead>
<tr>
<th>n missing</th>
<th>distinct</th>
<th>Info</th>
<th>Mean</th>
<th>Gmd</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>303</td>
<td>234</td>
<td>262</td>
<td>1</td>
<td>2095</td>
<td>20.3</td>
<td>364.0</td>
<td>1156.5</td>
<td>1870.0</td>
<td>2790.5</td>
<td>4008.6</td>
<td>4817.5</td>
</tr>
</tbody>
</table>

**lowest:** 0  5  8  15  20, **highest:** 6885  6920  7360  7560  7750
show patterns of missing data
plot(nacclus(support[acute,]))  # Figure 19.1

Show associations between predictors using a general non-monotonic measure of dependence (Hoeffding $D$).

```r
ac <- support[acute,]
ac$dzgroup <- ac$dzgroup[drop=TRUE]  # Remove unused levels
attach(ac)
vc <- varclus(~ age+sex+dzgroup+num.co+edu+income+scoma+race+
    meanbp+wblc+hrt+resp+temp+pafi+alb+bili+crea+sod+
    ph+glucose+bu+n+urine+adlsc, sim='hoeffding')
plot(vc)  # Figure 19.2
```
Figure 19.1: Cluster analysis showing which predictors tend to be missing on the same patients.

Figure 19.2: Hierarchical clustering of potential predictors using Hoeffding $D$ as a similarity measure. Categorical predictors are automatically expanded into dummy variables.
19.2

Checking Adequacy of Log-Normal Accelerated Failure Time Model

```r
dd <- datadist(ac)
# describe distributions of variables to rms
options(datadist='dd')

# Generate right-censored survival time variable
years <- d.time/365.25
units(years) <- 'Year'
S <- Surv(years, death)

# Show normal inverse Kaplan-Meier estimates
# stratified by dzgroup
survplot(npsurv(S ~ dzgroup), conf='none',
         fun=qnorm, logt=TRUE)  # Figure 19.3
```

![Figure 19.3: $\Phi^{-1}(S_{KM}(t))$ stratified by dzgroup. Linearity and semi-parallelism indicate a reasonable fit to the log-normal accelerated failure time model with respect to one predictor.](image)

More stringent assessment of log-normal assumptions: check distribution of residuals from an adjusted model:

```r
f <- psm(S ~ dzgroup + rcs(age,5) + rcs(meanbp,5),
          dist='lognormal', y=TRUE)  # dist='gaussian' for S+
r <- resid(f)
survplot(r, dzgroup, label.curve=FALSE)
survplot(r, age, label.curve=FALSE)
survplot(r, meanbp, label.curve=FALSE)
random.number <- runif(length(age))
```
Figure 19.4: Kaplan-Meier estimates of distributions of normalized, right-censored residuals from the fitted log-normal survival model. Residuals are stratified by important variables in the model (by quartiles of continuous variables), plus a random variable to depict the natural variability (in the lower right plot). Theoretical standard Gaussian distributions of residuals are shown with a thick solid line. The upper left plot is with respect to disease group.

The fit for `dzgroup` is not great but overall fit is good.

Remove from consideration predictors that are missing in $> 0.2$ of the patients. Many of these were only collected for the second phase of SUPPORT.

Of those variables to be included in the model, find which ones have enough potential predictive power to justify allowing for nonlinear relationships or multiple categories, which spend more d.f. For each variable compute Spearman $\rho^2$ based on multiple
linear regression of \( \text{rank}(x) \), \( \text{rank}(x)^2 \) and the survival time, truncating survival time at the shortest follow-up for survivors (356 days). This rids the data of censoring but creates many ties at 356 days.

\[
\text{shortest.follow.up} \leftarrow \min(\text{d.time}[\text{death==0}], \text{na.rm=TRUE}) \\
\text{d.timet} \leftarrow \text{pmin}(\text{d.time}, \text{shortest.follow.up}) \\
\]

\[
w \leftarrow \text{spearman2(\text{d.timet} \sim \text{age} + \text{num.co} + \text{scoma} + \text{meanbp} + \text{hrt} + \text{resp} + \text{temp} + \text{crea} + \text{sod} + \text{adlsc} + \text{wblc} + \text{pafi} + \text{ph} + \text{dzgroup} + \text{race}, \text{p}=2)} \\
\]

\[
\text{plot}(w, \text{main}='') \quad \# \text{Figure 19.5}
\]

\[
\begin{array}{ll}
\text{scoma} & 537 2 \\
\text{meanbp} & 537 2 \\
\text{dzgroup} & 537 2 \\
\text{crea} & 537 2 \\
\text{pafi} & 537 2 \\
\text{ph} & 537 2 \\
\text{sod} & 537 2 \\
\text{hrt} & 537 2 \\
\text{adlsc} & 537 2 \\
\text{temp} & 537 2 \\
\text{wblc} & 537 2 \\
\text{num.co} & 537 2 \\
\text{age} & 537 2 \\
\text{resp} & 537 2 \\
\text{race} & 537 2 \\
\end{array}
\]

Figure 19.5: Generalized Spearman \( \rho^2 \) rank correlation between predictors and truncated survival time

A better approach is to use the complete information in the failure and censoring times by computing Somers’ \( D_{xy} \) rank correlation allowing for censoring.

\[
w \leftarrow \text{rcorrcens(S} \sim \text{age} + \text{num.co} + \text{scoma} + \text{meanbp} + \text{hrt} + \text{resp} + \text{temp} + \text{crea} + \text{sod} + \text{adlsc} + \text{wblc} + \text{pafi} + \text{ph} + \text{dzgroup} + \text{race}) \\
\]

\[
\text{plot}(w, \text{main}='') \quad \# \text{Figure 19.6}
\]

\[
\text{# Compute number of missing values per variable} \\
\text{sapply(\text{llist}(\text{age}, \text{num.co}, \text{scoma}, \text{meanbp}, \text{hrt}, \text{resp}, \text{temp}, \text{crea}, \text{sod}, \text{adlsc}, \text{wblc}, \text{pafi}, \text{ph}), \text{function(x) sum(is.na(x)));})}
\]

| age num.co scoma meanbp hrt resp temp crea sod adlsc wblc |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 5 37 37 |
Do a formal redundancy analysis using more than pairwise associations, and allow for non-monotonic transformations in predicting each predictor from all other predictors. This analysis requires missing values to be imputed so as to not greatly reduce the sample size.

```r
redun(~ crea + age + sex + dzgroup + num.co + scoma + adlsc + race2 + meanbp + hrt + resp + temp + sod + wblc.i + pafi.i + ph.i, nk=4)
```

Redundancy Analysis

```r
redun(formula = ~crea + age + sex + dzgroup + num.co + scoma + adlsc + race2 + meanbp + hrt + resp + temp + sod + wblc.i + pafi.i + ph.i, nk = 4)
```
n: 537  p: 16  nk: 4
Number of NAs: 0
Transformation of target variables forced to be linear
\( R^2 \) cutoff: 0.9  Type: ordinary
\( R^2 \) with which each variable can be predicted from all other variables:

<table>
<thead>
<tr>
<th>Variable</th>
<th>( R^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>crea</td>
<td>0.133</td>
</tr>
<tr>
<td>age</td>
<td>0.246</td>
</tr>
<tr>
<td>sex</td>
<td>0.132</td>
</tr>
<tr>
<td>dzgroup</td>
<td>0.451</td>
</tr>
<tr>
<td>num.co</td>
<td>0.147</td>
</tr>
<tr>
<td>scoma</td>
<td>0.418</td>
</tr>
<tr>
<td>adlsc</td>
<td>0.153</td>
</tr>
<tr>
<td>race2</td>
<td>0.151</td>
</tr>
<tr>
<td>meanbp</td>
<td>0.178</td>
</tr>
<tr>
<td>hrt</td>
<td>0.258</td>
</tr>
<tr>
<td>resp</td>
<td>0.131</td>
</tr>
<tr>
<td>temp</td>
<td>0.197</td>
</tr>
<tr>
<td>sod</td>
<td>0.135</td>
</tr>
<tr>
<td>wblc.i</td>
<td>0.093</td>
</tr>
<tr>
<td>pafi.i</td>
<td>0.143</td>
</tr>
<tr>
<td>ph.i</td>
<td>0.171</td>
</tr>
</tbody>
</table>

No redundant variables

Better approach to gauging predictive potential and allocating d.f.:

- Allow all continuous variables to have a the maximum number of knots entertained, in a log-normal survival model
- Must use imputation to avoid losing data
- Fit a “saturated” main effects model
- Makes full use of censored data
- Had to limit to 4 knots, force \texttt{scoma} to be linear, and omit \texttt{ph.i} to avoid singularity

```r
k ← 4
f ← psm(S ~ rcs(age,k)+sex+dzgroup+pol(num.co,2)+scoma+
       pol(adlsc,2)+race+rcs(meanbp,k)+rcs(hrt,k)+rcs(resp,k)+
       rcs(temp,k)+rcs(crea,3)+rcs(sod,k)+rcs(wblc.i,k)+
       rcs(pafi.i,k), dist='lognormal')
plot(anova(f)) # Figure 19.7
```

- Figure 19.7 properly blinds the analyst to the form of effects
Fit a log-normal survival model with number of parameters corresponding to nonlinear effects determined from Figure 19.7. For the most promising predictors, five knots can be allocated, as there are fewer singularity problems once less promising predictors are simplified.

Note: Since the audio was recorded, a bug in `psm` was fixed on 2017-03-12. Discrimination indexes shown in the table below are correct but the audio is incorrect for $g$ and $g_r$.

```r
f <- psm(S ~ rcs(age, 5) + sex + dzgroup + num.co +
         scoma + pol(adlsc, 2) + race2 + rcs(meanbp, 5) +
         rcs(hrt, 3) + rcs(resp, 3) + temp +
         rcs(crea, 4) + sod + rcs(wblc.i, 3) + rcs(pafi.i, 4),
         dist = 'lognormal')  # 'gaussian' for S
print(f)
```

Parametric Survival Model: Log Normal Distribution

```r
psm(formula = S ~ rcs(age, 5) + sex + dzgroup + num.co + scoma +
     pol(adlsc, 2) + race2 + rcs(meanbp, 5) + rcs(hrt, 3) + rcs(resp, 3) +
     temp + rcs(crea, 4) + sod + rcs(wblc.i, 3) + rcs(pafi.i, 4),
     dist = "lognormal")
```
<table>
<thead>
<tr>
<th></th>
<th>Model Likelihood Ratio Test</th>
<th>Discrimination Indexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>537</td>
<td></td>
</tr>
<tr>
<td>Events</td>
<td>356</td>
<td></td>
</tr>
<tr>
<td>$\sigma$</td>
<td>2.230782</td>
<td></td>
</tr>
<tr>
<td>LR $\chi^2$</td>
<td>236.83</td>
<td>$R^2$ 0.594</td>
</tr>
<tr>
<td>d.f.</td>
<td>30</td>
<td>$D_{xy}$ 0.485</td>
</tr>
<tr>
<td>$\Pr(\chi^2) &lt;$0.0001</td>
<td></td>
<td>$g$ 1.959</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$g_r$ 7.095</td>
</tr>
</tbody>
</table>

|        | $\hat{\beta}$ | S.E. | Wald $Z$ | $Pr(> |Z|)$ |
|--------|----------------|------|----------|-----------|
| (Intercept) | -5.6883       | 3.7851 | -1.50    | 0.1329    |
| age     | -0.0148        | 0.0309 | -0.48    | 0.6322    |
| age'    | -0.0412        | 0.1078 | -0.38    | 0.7024    |
| age''   | 0.1670         | 0.5594 | 0.30     | 0.7653    |
| age'''' | -0.2099        | 1.3707 | -0.15    | 0.8783    |
| sex=male                          | -0.0737        | 0.2181 | -0.34    | 0.7354    |
| dzgroup=Coma                       | -2.0676        | 0.4062 | -5.09    | $<0.0001$ |
| dzgroup=MOE w/Malig                | -1.4664        | 0.3112 | -4.71    | $<0.0001$ |
| num.co                              | -0.1917        | 0.0858 | -2.23    | 0.0255    |
| scoma                                | -0.0142        | 0.0044 | -3.25    | 0.0011    |
| adlsc                                | -0.3735        | 0.1520 | -2.46    | 0.0140    |
| adlsc$^2$                          | 0.0442         | 0.0243 | 1.82     | 0.0691    |
| race2=other                         | 0.2979         | 0.2658 | 1.12     | 0.2624    |
| meanbp                              | 0.0702         | 0.0210 | 3.34     | 0.0008    |
| meanbp'                             | -0.3080        | 0.2261 | -1.36    | 0.1732    |
| meanbp''                            | 0.8438         | 0.8556 | 0.99     | 0.3241    |
| meanbp''''                          | -0.5715        | 0.7707 | -0.74    | 0.4584    |
| hrt                                 | -0.0171        | 0.0069 | -2.46    | 0.0140    |
| hrt'                                | 0.0064         | 0.0063 | 1.02     | 0.3090    |
| resp                                | 0.0454         | 0.0230 | 1.97     | 0.0483    |
| resp'                               | -0.0851        | 0.0291 | -2.93    | 0.0034    |
| temp                                | 0.0523         | 0.0834 | 0.63     | 0.5308    |
| crea                                | -0.4585        | 0.6727 | -0.68    | 0.4955    |
| crea'                               | -11.5176       | 19.0027 | -0.61    | 0.5444    |
| crea''                              | 21.9840        | 31.0113 | 0.71    | 0.4784    |
| sod                                 | 0.0044         | 0.0157 | 0.28     | 0.7792    |
| wblc.i                              | 0.0746         | 0.0331 | 2.25     | 0.0242    |
| wblc.i'                             | -0.0880        | 0.0377 | -2.34    | 0.0195    |
| pafi.i                              | 0.0169         | 0.0055 | 3.07     | 0.0021    |
| pafi.i'                             | -0.0569        | 0.0239 | -2.38    | 0.0173    |
| pafi.i''                            | 0.1088         | 0.0482 | 2.26     | 0.0239    |
| Log(scale)                           | 0.8024         | 0.0401 | 19.99    | $<0.0001$ |

a ← anova(f)
19.3

**Summarizing the Fitted Model**

- Plot the shape of the effect of each predictor on log survival time.

- All effects centered: can be placed on common scale

- Wald $\chi^2$ statistics, penalized for d.f., plotted in descending order

```r
ggplot(Predict(f, ref.zero=TRUE), vnames='names', 
       sepdiscrete='vertical', anova=a)  # Figure 19.8

print(a, size='tsz')
```

Wald Statistics for $S$
Figure 19.8: Effect of each predictor on log survival time. Predicted values have been centered so that predictions at predictor reference values are zero. Pointwise 0.95 confidence bands are also shown. As all Y-axes have the same scale, it is easy to see which predictors are strongest.
<table>
<thead>
<tr>
<th>Variable</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>15.99</td>
<td>4</td>
<td>0.0030</td>
</tr>
<tr>
<td>$Nonlinear$</td>
<td>0.23</td>
<td>3</td>
<td>0.9722</td>
</tr>
<tr>
<td>sex</td>
<td>0.11</td>
<td>1</td>
<td>0.7354</td>
</tr>
<tr>
<td>dzgroup</td>
<td>45.69</td>
<td>2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>num.co</td>
<td>4.99</td>
<td>1</td>
<td>0.0255</td>
</tr>
<tr>
<td>scoma</td>
<td>10.58</td>
<td>1</td>
<td>0.0011</td>
</tr>
<tr>
<td>adlsc</td>
<td>8.28</td>
<td>2</td>
<td>0.0159</td>
</tr>
<tr>
<td>$Nonlinear$</td>
<td>3.31</td>
<td>1</td>
<td>0.0691</td>
</tr>
<tr>
<td>race2</td>
<td>1.26</td>
<td>1</td>
<td>0.2624</td>
</tr>
<tr>
<td>meanbp</td>
<td>27.62</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>$Nonlinear$</td>
<td>10.51</td>
<td>3</td>
<td>0.0147</td>
</tr>
<tr>
<td>hrt</td>
<td>11.83</td>
<td>2</td>
<td>0.0027</td>
</tr>
<tr>
<td>$Nonlinear$</td>
<td>1.04</td>
<td>1</td>
<td>0.3090</td>
</tr>
<tr>
<td>resp</td>
<td>11.10</td>
<td>2</td>
<td>0.0039</td>
</tr>
<tr>
<td>$Nonlinear$</td>
<td>8.56</td>
<td>1</td>
<td>0.0034</td>
</tr>
<tr>
<td>temp</td>
<td>0.39</td>
<td>1</td>
<td>0.5308</td>
</tr>
<tr>
<td>crea</td>
<td>33.63</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>$Nonlinear$</td>
<td>21.27</td>
<td>2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>sod</td>
<td>0.08</td>
<td>1</td>
<td>0.7792</td>
</tr>
<tr>
<td>wblc.i</td>
<td>5.47</td>
<td>2</td>
<td>0.0649</td>
</tr>
<tr>
<td>$Nonlinear$</td>
<td>5.46</td>
<td>1</td>
<td>0.0195</td>
</tr>
<tr>
<td>pafi.i</td>
<td>15.37</td>
<td>3</td>
<td>0.0015</td>
</tr>
<tr>
<td>$Nonlinear$</td>
<td>6.97</td>
<td>2</td>
<td>0.0307</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>60.48</td>
<td>14</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>261.47</td>
<td>30</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

`plot(a)` # Figure 19.9

`options(digits=3)`

`plot(summary(f), log=TRUE, main='')` # Figure 19.10
Figure 19.9: Contribution of variables in predicting survival time in log-normal model

Figure 19.10: Estimated survival time ratios for default settings of predictors. For example, when age changes from its lower quartile to the upper quartile (47.9y to 74.5y), median survival time decreases by more than half. Different shaded areas of bars indicate different confidence levels (0.9, 0.95, 0.99).
19.4

Internal Validation of the Fitted Model Using the Bootstrap

Validate indexes describing the fitted model.

```r
# First add data to model fit so bootstrap can re-sample
# from the data
g <- update(f, x=TRUE, y=TRUE)
set.seed(717)
latex(validate(g, B=120, dxy=TRUE), digits=2, size='Ssize')
```

<table>
<thead>
<tr>
<th>Index</th>
<th>Original Sample</th>
<th>Training Sample</th>
<th>Test Sample</th>
<th>Optimism</th>
<th>Corrected Index</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{xy}$</td>
<td>0.49</td>
<td>0.51</td>
<td>0.46</td>
<td>0.05</td>
<td>0.44</td>
<td>120</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.59</td>
<td>0.59</td>
<td>-0.03</td>
<td>0.03</td>
<td>-0.03</td>
<td>120</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.03</td>
<td>0.03</td>
<td>-0.03</td>
<td>120</td>
</tr>
<tr>
<td>Slope</td>
<td>1.00</td>
<td>1.00</td>
<td>0.90</td>
<td>0.10</td>
<td>0.90</td>
<td>120</td>
</tr>
<tr>
<td>$D$</td>
<td>0.48</td>
<td>0.48</td>
<td>0.90</td>
<td>0.10</td>
<td>0.90</td>
<td>120</td>
</tr>
<tr>
<td>$U$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>120</td>
</tr>
<tr>
<td>$Q$</td>
<td>0.48</td>
<td>2.04</td>
<td>1.86</td>
<td>0.18</td>
<td>1.78</td>
<td>120</td>
</tr>
</tbody>
</table>

- From $D_{xy}$ and $R^2$ there is a moderate amount of overfitting.
- Slope shrinkage factor (0.90) is not troublesome
- Almost unbiased estimate of future predictive discrimination on similar patients is the corrected $D_{xy}$ of 0.43.

Validate predicted 1-year survival probabilities. Use a smooth approach that does not require binning [122] and use less precise Kaplan-Meier estimates obtained by stratifying patients by the predicted probability, with at least 60 patients per group.

```r
set.seed(717)
cal <- calibrate(g, u=1, B=120)
```
plot(cal, subtitles=FALSE)
cal ← calibrate(g, cmethod='KM', u=1, m=60, B=120, pr=FALSE)

plot(cal, add=TRUE)  # Figure 19.11

Figure 19.11: Bootstrap validation of calibration curve. Dots represent apparent calibration accuracy; × are bootstrap estimates corrected for overfitting, based on binning predicted survival probabilities and computing Kaplan-Meier estimates. Black curve is the estimated observed relationship using \texttt{hare} and the blue curve is the overfitting-corrected \texttt{hare} estimate. The gray-scale line depicts the ideal relationship.
19.5

Approximating the Full Model

The fitted log-normal model is perhaps too complex for routine use and for routine data collection. Let us develop a simplified model that can predict the predicted values of the full model with high accuracy ($R^2 = 0.96$). The simplification is done using a fast backward stepdown against the full model predicted values.

\[
Z \leftarrow \text{predict(f)} \quad \# X \ast \beta \hat{\text{h}}
\]

\[
a \leftarrow \text{ols(Z} \sim \text{rcs(age,5)+sex+dzgroup+num.co+}
\]

\[
\text{scoma+pol(adlsc,2)+race2+}
\]

\[
\text{rcs(meanbp,5)+rcs(hrt,3)+rcs(resp,3)+}
\]

\[
\text{temp+rcs(crea,4)+sod+rcs(wblc.i,3)+}
\]

\[
\text{rcs(pafi.i,4), sigma=1)
\]

\[
\# \text{sigma=1 is used to prevent } \hat{\sigma} \text{ from being zero when}
\]

\[
\# R^2=1.0 \text{ since we start out by approximating } Z \text{ with all}
\]

\[
\# \text{component variables}
\]

\[
\text{fastbw(a, aics=10000)} \quad \# \text{fast backward stepdown}
\]

<table>
<thead>
<tr>
<th>Deleted</th>
<th>Chi-Sq</th>
<th>d.f.</th>
<th>P</th>
<th>Residual</th>
<th>d.f.</th>
<th>P</th>
<th>AIC</th>
<th>R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>sod</td>
<td>0.43</td>
<td>1</td>
<td>0.512</td>
<td>0.43</td>
<td>1</td>
<td>0.5117</td>
<td>-1.57</td>
<td>1.000</td>
</tr>
<tr>
<td>sex</td>
<td>0.57</td>
<td>1</td>
<td>0.451</td>
<td>1.00</td>
<td>2</td>
<td>0.6073</td>
<td>-3.00</td>
<td>0.999</td>
</tr>
<tr>
<td>temp</td>
<td>2.20</td>
<td>1</td>
<td>0.138</td>
<td>3.20</td>
<td>3</td>
<td>0.3621</td>
<td>-2.80</td>
<td>0.998</td>
</tr>
<tr>
<td>race2</td>
<td>6.81</td>
<td>1</td>
<td>0.009</td>
<td>10.01</td>
<td>4</td>
<td>0.0402</td>
<td>2.01</td>
<td>0.994</td>
</tr>
<tr>
<td>wblc.i</td>
<td>29.52</td>
<td>2</td>
<td>0.000</td>
<td>39.53</td>
<td>6</td>
<td>0.0000</td>
<td>27.53</td>
<td>0.976</td>
</tr>
<tr>
<td>num.co</td>
<td>30.84</td>
<td>1</td>
<td>0.009</td>
<td>70.36</td>
<td>7</td>
<td>0.0000</td>
<td>56.36</td>
<td>0.957</td>
</tr>
<tr>
<td>resp</td>
<td>54.18</td>
<td>2</td>
<td>0.000</td>
<td>124.55</td>
<td>9</td>
<td>0.0000</td>
<td>106.55</td>
<td>0.924</td>
</tr>
<tr>
<td>adlsc</td>
<td>52.46</td>
<td>2</td>
<td>0.000</td>
<td>177.00</td>
<td>11</td>
<td>0.0000</td>
<td>155.00</td>
<td>0.892</td>
</tr>
<tr>
<td>pafi.i</td>
<td>66.78</td>
<td>3</td>
<td>0.000</td>
<td>243.79</td>
<td>14</td>
<td>0.0000</td>
<td>215.79</td>
<td>0.851</td>
</tr>
<tr>
<td>scoma</td>
<td>78.07</td>
<td>1</td>
<td>0.000</td>
<td>321.86</td>
<td>15</td>
<td>0.0000</td>
<td>291.86</td>
<td>0.803</td>
</tr>
<tr>
<td>hrt</td>
<td>83.17</td>
<td>2</td>
<td>0.000</td>
<td>405.02</td>
<td>17</td>
<td>0.0000</td>
<td>371.02</td>
<td>0.752</td>
</tr>
<tr>
<td>age</td>
<td>68.08</td>
<td>4</td>
<td>0.000</td>
<td>473.10</td>
<td>21</td>
<td>0.0000</td>
<td>431.10</td>
<td>0.710</td>
</tr>
<tr>
<td>crea</td>
<td>314.47</td>
<td>3</td>
<td>0.000</td>
<td>787.57</td>
<td>24</td>
<td>0.0000</td>
<td>739.57</td>
<td>0.517</td>
</tr>
<tr>
<td>meanbp</td>
<td>403.04</td>
<td>2</td>
<td>0.000</td>
<td>1190.61</td>
<td>28</td>
<td>0.0000</td>
<td>1134.61</td>
<td>0.270</td>
</tr>
<tr>
<td>dzgroup</td>
<td>441.28</td>
<td>2</td>
<td>0.000</td>
<td>1631.89</td>
<td>30</td>
<td>0.0000</td>
<td>1571.89</td>
<td>0.000</td>
</tr>
</tbody>
</table>

**Approximate Estimates after Deleting Factors**

Coef    S.E.  Wald  Z   P
[1,]  -0.5928 0.04315 -13.74 0

**Factors in Final Model**

None
f.approx <- ols(Z ~ dzgroup + rcs(meanbp,5) + rcs(crea,4) + rcs(age,5) +
    rcs(hrt,3) + scoma + rcs(pafi.i,4) + pol(adlsc,2) +
    rcs(resp,3), x=TRUE)
f.approx$stats

<table>
<thead>
<tr>
<th>n</th>
<th>Model L.R.</th>
<th>d.f.</th>
<th>R2</th>
<th>g</th>
</tr>
</thead>
<tbody>
<tr>
<td>537.000</td>
<td>1688.225</td>
<td>23.000</td>
<td>0.957</td>
<td>1.915</td>
</tr>
</tbody>
</table>

- Estimate variance–covariance matrix of the coefficients of reduced model

- This covariance matrix does not include the scale parameter

\[ V \leftarrow \text{vcov}(f, \text{regcoef.only = TRUE}) \]  # \text{var(full model)}
\[ X \leftarrow \text{cbind(Intercept=1, g$x)} \]  # \text{full model design}
\[ x \leftarrow \text{cbind(Intercept=1, f.approx$x)} \]  # \text{approx. model design}
\[ w \leftarrow \text{solve(t(x) \times x, t(x)) \times X} \]  # contrast matrix
\[ v \leftarrow w \times V \times t(w) \]

Compare variance estimates (diagonals of \( v \)) with variance estimates from a reduced model that is fitted against the actual outcomes.

\[ f.sub \leftarrow \text{psm(S ~ dzgroup + rcs(meanbp,5) + rcs(crea,4) + rcs(age,5) +}
    \hspace{1em} \text{rcs(hrt,3) + scoma + rcs(pafi.i,4) + pol(adlsc,2) +}
    \hspace{1em} \text{rcs(resp,3), dist='lognormal') \hspace{1em} # 'gaussian' for S+} \]
\[ r \leftarrow \text{diag(v)/diag(vcov(f.sub, regcoef.only = TRUE))} \]
\[ r[\text{which.min(r), which.max(r))}] \]

- hrt  age
  0.976  0.982

\[ f.approx$var \leftarrow v \]
\[ \text{print(anova(f.approx, test='Chisq', ss=FALSE), size='tsz')} \]
Wald Statistics for \( Z \)

<table>
<thead>
<tr>
<th></th>
<th>( \chi^2 )</th>
<th>d.f.</th>
<th>( P )</th>
</tr>
</thead>
<tbody>
<tr>
<td>dzgroup</td>
<td>55.94</td>
<td>2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>meanbp</td>
<td>29.87</td>
<td>4</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>9.84</td>
<td>3</td>
<td>0.0200</td>
</tr>
<tr>
<td>crea</td>
<td>39.04</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>24.37</td>
<td>2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>age</td>
<td>18.12</td>
<td>4</td>
<td>0.0012</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>0.34</td>
<td>3</td>
<td>0.9517</td>
</tr>
<tr>
<td>hrt</td>
<td>9.87</td>
<td>2</td>
<td>0.0072</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>0.40</td>
<td>1</td>
<td>0.5289</td>
</tr>
<tr>
<td>scoma</td>
<td>9.85</td>
<td>1</td>
<td>0.0017</td>
</tr>
<tr>
<td>pafi.i</td>
<td>14.01</td>
<td>3</td>
<td>0.0029</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>6.66</td>
<td>2</td>
<td>0.0357</td>
</tr>
<tr>
<td>adlsc</td>
<td>9.71</td>
<td>2</td>
<td>0.0078</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>2.87</td>
<td>1</td>
<td>0.0904</td>
</tr>
<tr>
<td>resp</td>
<td>9.65</td>
<td>2</td>
<td>0.0080</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>7.13</td>
<td>1</td>
<td>0.0076</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>58.08</td>
<td>13</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>252.32</td>
<td>23</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

Equation for simplified model:

\[
\text{E}(Z) = X\beta, \quad \text{where}
\]

\[
X\beta = \\
-2.51 \\
-1.94 [\text{Coma}] - 1.75 [\text{MOSF w/Malig}] \\
+0.068 \text{meanbp} - 3.08 \times 10^{-5} (\text{meanbp} - 41.8)_+^3 + 7.9 \times 10^{-5} (\text{meanbp} - 61)_+^3 \\
-4.91 \times 10^{-5} (\text{meanbp} - 73)_+^3 + 2.61 \times 10^{-6} (\text{meanbp} - 109)_+^3 - 1.7 \times 10^{-6} (\text{meanbp} - 135)_+^3 \\
-0.553 \text{crea} - 0.229 (\text{crea} - 0.6)_+^3 + 0.45 (\text{crea} - 1.1)_+^3 - 0.233 (\text{crea} - 1.94)_+^3 \\
+0.0131 (\text{crea} - 7.32)_+^3 \\
-0.0165 \text{age} - 1.13 \times 10^{-5} (\text{age} - 28.5)_+^3 + 4.05 \times 10^{-5} (\text{age} - 49.5)_+^3 \\
-2.15 \times 10^{-5} (\text{age} - 63.7)_+^3 - 2.68 \times 10^{-5} (\text{age} - 72.7)_+^3 + 1.9 \times 10^{-5} (\text{age} - 85.6)_+^3 \\
-0.0136 \text{hrt} + 6.09 \times 10^{-7} (\text{hrt} - 60)_+^3 - 1.68 \times 10^{-6} (\text{hrt} - 111)_+^3 + 1.07 \times 10^{-6} (\text{hrt} - 140)_+^3 \\
-0.0135 \text{scoma}
\]
+0.0161pafi.i - 4.77 \times 10^{-7} (pafi.i - 88)^3 + 9.11 \times 10^{-7} (pafi.i - 167)^3 \\
-5.02 \times 10^{-7} (pafi.i - 276)^3 + 6.76 \times 10^{-8} (pafi.i - 426)^3 - 0.369 \text{adlsc} + 0.0409 \text{adlsc}^2 \\
+0.0394 \text{resp} - 9.11 \times 10^{-5} (\text{resp} - 10)^3 + 0.000176 (\text{resp} - 24)^3 - 8.5 \times 10^{-5} (\text{resp} - 39)^3

and \([c] = 1\) if subject is in group \(c\), 0 otherwise; \((x)_+ = x\) if \(x > 0\), 0 otherwise

### Nomogram for predicting median and mean survival time, based on approximate model:

```r
# Derive S functions that express mean and quantiles # of survival time for specific linear predictors # analytically
expected.surv ← Mean(f)
quantile.surv ← Quantile(f)
lxet(expected.surv, file='', type='Sinput')

expected.surv ← function (lp = NULL, parms = 0.802352037606488) 
{ 
    names(parms) ← NULL
    exp(lp + exp(2 * parms)/2)
}
lxet(quantile.surv, file='', type='Sinput')

quantile.surv ← function (q = 0.5, lp = NULL, parms = 0.802352037606488) 
{ 
    names(parms) ← NULL
    f ← function(lp, q, parms) lp + exp(parms) * qnorm(q)
    names(q) ← format(q)
    drop(exp(outer(lp, q, FUN = f, parms = parms)))
}

median.surv ← function(x) quantile.surv(lp=x)

# Improve variable labels for the nomogram
f.approx ← Newlabels(f.approx, c('Disease Group', 'Mean Arterial BP', 'Creatinine', 'Age', 'Heart Rate', 'SUPPORT Coma Score', 'PaO2/(.01 * FiO2)', 'ADL', 'Resp. Rate'))
nom ← nomogram(f.approx, 
pafi.i=c(0, 50, 100, 200, 300, 500, 600, 700, 800, 900), 
fun=list('Median Survival Time'=median.surv, 
'Mean Survival Time' =expected.surv), 
fun.at=c(.1,.25,.5,1,2,5,10,20,40))
plot(nom, cex.var=1, cex.axis=.75, lmgp=.25)
# Figure 19.12
```
## Figure 19.12: Nomogram for predicting median and mean survival time, based on approximation of full model
S Packages and Functions Used

<table>
<thead>
<tr>
<th>Packages</th>
<th>Purpose</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hmisc</td>
<td>Miscellaneous functions</td>
<td>describe, ecdf, naclus, varclus, llist, spearman2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>describe, impute, latex</td>
</tr>
<tr>
<td>rms</td>
<td>Modeling</td>
<td>datadist, psm, rcs, ols, fastbw</td>
</tr>
<tr>
<td></td>
<td>Model presentation</td>
<td>survplot, Newlabels, Function, Mean, Quantile, nomogram</td>
</tr>
<tr>
<td></td>
<td>Model validation</td>
<td>validate, calibrate</td>
</tr>
</tbody>
</table>

Note: All packages are available from CRAN
Chapter 20

Cox Proportional Hazards Regression Model

20.1

Model

20.1.1

Preliminaries

- Most popular survival model

- Semi-parametric (nonparametric hazard; parametric regression)

- Usually more interest in effects of \( X \) than on shape of \( \lambda(t) \)

- Uses only rank ordering of failures/censoring times \( \rightarrow \) more robust, easier to write protocol

- Even if parametric PH assumptions true, Cox still fully efficient for \( \beta \)
• Model diagnostics are advanced

• Log-rank test is a special case with one binary $X$

**Model Definition**

$$\lambda(t|X) = \lambda(t) \exp(X\beta).$$

• No intercept parameter

• Shape of $\lambda$ not important

• When a predictor say $X_1$ is
  – binary
  – doesn’t interact with other predictors
  – has coefficient $\beta_1$
  – satisfies the proportional hazards (PH) assumption so that $X_1$ does not interact with time

→ hazard ratio (HR) $\exp(\beta_1)$ is the ratio of hazard functions for $X_1 = 1$ vs. $X_1 = 0$

– $\lambda(t)$ cancels out

– by the PH assumption, the HR does not depend on $t$; $X_1$
has a constant effect on $\lambda$ over time

- under PH and absence of covariate interactions, HR is a good overall effect estimate for binary $X_1$

HR is the ratio of two instantaneous event rates

## Estimation of $\beta$

- The objective function to optimize is the Cox’s partial likelihood function

- Partial likelihood only covers the $\beta$ part of the model, not the $\lambda$ or underlying survival curve part
  - these are estimated in a separate step once $\hat{\beta}$ is obtained

- Obtain maximum likelihood estimates of $\beta$ (formally, maximum partial likelihood estimates)

- See text for details

## Model Assumptions and Interpretation of Parameters

- Similar to other models; interpretation is on the log relative hazard scale

- Equivalent to using $\log(-\log(S(t)))$ scale
• HR of 2 is equivalent to raising the entire survival curve for a control subject to the second power to get the survival curve for an exposed subject

− Example: if a control subject has 5y survival probability of 0.7 and the exposed:control HR is 2, the exposed subject has a 5y survival probability of 0.49

− If the HR is \( \frac{1}{2} \), the exposed subject has a survival curve that is the square root of the control, so \( S(5) \) would be \( \sqrt{0.7} = 0.837 \)

---

### Example

```r
require(rms)
options(prType='latex')
group ← c(rep('Group 1',19),rep('Group 2',21))
group ← factor(group)
dd ← datadist(group); options(datadist='dd')
days ← c(143,164,188,188,190,192,206,209,213,216,220,227,230,
   234,246,265,304,216,244,142,156,163,198,205,232,232,
   233,233,233,239,240,261,280,280,296,296,323,204,344)
death ← rep(1,40)
death[c(18,19,39,40)] ← 0
units(days) ← 'Day'
df ← data.frame(days, death, group)
S ← Surv(days, death)

f ← npsurv(S ~ group, type='fleming')
for(meth in c('exact', 'breslow', 'efron')) {
  g ← cph(S ~ group, method=meth, surv=TRUE, x=TRUE, y=TRUE)
  # print(g) to see results
}
f.exp ← psm(S ~ group, dist='exponential')
fw ← psm(S ~ group, dist='weibull')
phform ← pphsm(fw)

co ← gray(c(0, .8))
survplot(f, lty=c(1, 1), lwd=c(1, 3), col=co,
   label.curves=FALSE, conf='none')
survplot(g, lty=c(3, 3), lwd=c(1, 3), col=co, # Efron approx.
   add=TRUE, label.curves=FALSE, conf.type='none')
```
CHAPTER 20. COX PROPORTIONAL HAZARDS REGRESSION MODEL

```
legend(c(2, 160), c(.38 , .54 ),
      c('Nonparametric Estimates ', 'Cox-Breslow Estimates '),
      lty=c(1 , 3) , cex=.8 , bty='n')
legend(c(2, 160) , c(.18 , .34 ), cex=.8 ,
      c('Group 1', 'Group 2'), lwd=c (1 ,3) , col=co , bty='n')
```

![Graph showing survival probability over follow-up time for Group 1 and Group 2 with nonparametric and Cox-Breslow estimates.](image)

**Figure 20.1:** Altschuler–Nelson–Fleming–Harrington nonparametric survival estimates and Cox-Breslow estimates for rat data [167]

<table>
<thead>
<tr>
<th>Model</th>
<th>Group Regression Coefficient</th>
<th>S.E.</th>
<th>Wald p Value</th>
<th>Group 2:1 Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cox (Exact)</td>
<td>-0.629</td>
<td>0.361</td>
<td>0.08</td>
<td>0.533</td>
</tr>
<tr>
<td>Cox (Efron)</td>
<td>-0.569</td>
<td>0.347</td>
<td>0.10</td>
<td>0.566</td>
</tr>
<tr>
<td>Cox (Breslow)</td>
<td>-0.596</td>
<td>0.348</td>
<td>0.09</td>
<td>0.551</td>
</tr>
<tr>
<td>Exponential</td>
<td>-0.093</td>
<td>0.334</td>
<td>0.78</td>
<td>0.911</td>
</tr>
<tr>
<td>Weibull (AFT)</td>
<td>0.132</td>
<td>0.061</td>
<td>0.03</td>
<td>0.911</td>
</tr>
<tr>
<td>Weibull (PH)</td>
<td>-0.721</td>
<td></td>
<td></td>
<td>0.486</td>
</tr>
</tbody>
</table>

---

**20.1.6 Design Formulations**

- $k - 1$ dummies for $k$ treatments, one treatment $\rightarrow \lambda(t)$
- Only provides relative effects
Extending the Model by Stratification

- Is a unique feature of the Cox model
- Adjust for non-modeled factors
- Factors too difficult to model or fail PH assumption
- Commonly used in RCTs to adjust for site variation
- Allow form of $\lambda$ to vary across strata
- Rank failure times within strata
- $b$ strata, stratum ID is $C$

$$\lambda(t|X, C = j) = \lambda_j(t) \exp(X\beta), \quad \text{or} \quad S(t|X, C = j) = S_j(t) \exp(X\beta).$$

- Not assume connection between shapes of $\lambda_j$
- By default, assume common $\beta$
- Ex: model age, stratify on sex
  Estimates common age slope pooling F and M
  No assumption about effect of sex except no age interact.
- Can stratify on multiple factors (cross-classify)
• Loss of efficiency not bad unless number of events in strata very small

• Stratum with no events is ignored

• Estimate $\beta$ by getting separate log-likelihood for each stratum and adding up (independence)

• No inference about strat. factors

• Useful for checking PH and linearity assumptions: Model, then stratify on an $X$

• Can extend to strata $\times$ covariable interaction

\[
\begin{align*}
\lambda(t|X_1, C = 1) &= \lambda_1(t) \exp(\beta_1 X_1) \\
\lambda(t|X_1, C = 2) &= \lambda_2(t) \exp(\beta_1 X_1 + \beta_2 X_1). \\
\lambda(t|X_1, C = j) &= \lambda_j(t) \exp(\beta_1 X_1 + \beta_2 X_2)
\end{align*}
\]

• $X_2$ is product interaction term (0 for F, $X_1$ for M)

• Testing interaction with sex without modeling main effect!
20.2 Estimation of Survival Probability and Secondary Parameters

- Kalbfleisch-Prentice discrete hazard model method $\rightarrow$ K-M if $\hat{\beta} = 0$

- Breslow method $\rightarrow$ Nelson et al. if $\hat{\beta} = 0$

$$\hat{S}(t|X) = \hat{S}(t)\exp(X\hat{\beta}).$$

- Stratified model $\rightarrow$ estimate underlying hazard parameters separately within strata

- “Adjusted K-M estimates”

Figure 20.2: Unadjusted (Kaplan–Meier) and adjusted (Cox–Kalbfleisch–Prentice) estimates of survival. Left, Kaplan–Meier estimates for patients treated medically and surgically at Duke University Medical Center from November 1969 through December 1984. These survival curves are not adjusted for baseline prognostic factors. Right, survival curves for patients treated medically or surgically after adjusting for all known important baseline prognostic characteristics. [14]
\[ \hat{\Lambda}(t) = \sum_{i: t_i < t} \frac{d_i}{\sum_{Y_i \geq t_i} \exp(X_i \hat{\beta})}. \]

For any \( X \), the estimates of \( \Lambda \) and \( S \) are

\[
\hat{\Lambda}(t|X) = \hat{\Lambda}(t) \exp(X\hat{\beta}) \\
\hat{S}(t|X) = \exp[-\hat{\Lambda}(t) \exp(X\hat{\beta})].
\]
20.3 Sample Size Considerations

- Consider case with no covariates and want to estimate $S(t)$; results to Kaplan-Meier

- As detailed in the text, one may need 184 subjects with an event, or censored late, to estimate $S(t)$ to within a margin of error of 0.1 everywhere, at the 0.95 confidence level

- Instead consider the case where the model has a single binary covariate and we want to estimate the hazard ratio to within a specified multiplicative margin of error (MMOE) with confidence $1 - \alpha$

- Assume equal sample size for $X = 0$ and $X = 1$ and let $e_0$ and $e_1$ denote the number of events in the two $X$ groups

- Variance of log HR is approximately $v = \frac{1}{e_0} + \frac{1}{e_1}$.

- Let $z$ denote the $1 - \alpha/2$ standard normal critical value

- MMOE with confidence $1 - \alpha$ is $\exp(z\sqrt{v})$

- To achieve a MMOE of 1.2 in estimating $e^{\hat{\beta}}$ with equal numbers of events in the two groups and $\alpha = 0.05$ requires a total of 462 events:

```r
z <- qnorm(1 - .05/2)
# v = (log(mmoe) / z)^2
```
• Sample size for external validation: at least 200 events [45]
20.4 Test Statistics

- Score test $= \text{log-rank } \chi^2$ test statistic

- Score test in a stratified PH model is the stratified log-rank statistic
## 20.5 Residuals

<table>
<thead>
<tr>
<th>Residual</th>
<th>Purposes</th>
</tr>
</thead>
<tbody>
<tr>
<td>martingale</td>
<td>Assessing adequacy of a hypothesized predictor transformation; Graphing an estimate of a predictor transformation (Section 20.6.1)</td>
</tr>
<tr>
<td>score</td>
<td>Detecting overly influential observations</td>
</tr>
<tr>
<td>Schoenfeld</td>
<td>Testing PH assumption (Section 20.6.2)</td>
</tr>
<tr>
<td></td>
<td>Graphing estimate of hazard ratio function (Section 20.6.2)</td>
</tr>
</tbody>
</table>
20.6 Assessment of Model Fit

20.6.1 Regression Assumptions

- Stratified KM estimates have problems
- 2000 simulated subject, $d = 368, 1196\ M, 804\ F$
- Exponential with known log hazard, linear in age, additive in sex

$$\lambda(t|X_1, X_2) = .02\ exp[.8X_1 + .04(X_2 - 50)]$$

```r
n <- 2000
set.seed(3)
age <- 50 + 12 * rnorm(n)
label(age) <- 'Age'
sex <- factor(1 + (runif(n) <= .4), 1:2, c('Male', 'Female'))
cens <- 15 * runif(n)
h <- .02 * exp(.04 * (age - 50) + .8 * (sex == 'Female'))
ft <- -log(runif(n)) / h
e <- ifelse(ft <= cens, 1, 0)
print(table(e))
```

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>e</td>
<td>1611</td>
<td>389</td>
</tr>
</tbody>
</table>

```r
ft <- pmin(ft, cens)
units(ft) <- 'Year'
Srv <- Surv(ft, e)
age.dec <- cut2(age, g=10, levels.mean=TRUE)
label(age.dec) <- 'Age'
dd <- datadist(age, sex, age.dec);
options(datadist='dd')
f.np <- cph(Srv ~ strat(age.dec) + strat(sex), surv=TRUE)
# surv=TRUE speeds up computations, and confidence limits when
# there are no covariates are still accurate.
p <- Predict(f.np, age.dec, sex, time=3, loglog=TRUE)
```

```r
# Treat age.dec as a numeric variable (means within deciles)
p$age.dec <- as.numeric(as.character(p$age.dec))
```
Better: A 4-knot spline Cox PH model in two variables $(X_1, X_2)$ which assumes linearity in $X_1$ and no $X_1 \times X_2$ interaction

$$\lambda(t|X) = \lambda(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X'_2 + \beta_4 X''_2),$$

$$= \lambda(t) \exp(\beta_1 X_1 + f(X_2)),$$

$$f(X_2) = \beta_2 X_2 + \beta_3 X'_2 + \beta_4 X''_2.$$

$$\log \lambda(t|X) = \log \lambda(t) + \beta_1 X_1 + f(X_2).$$

To not assume PH in $X_1$, stratify on it:

$$\log \lambda(t|X_2, C = j) = \log \lambda_j(t) + \beta_1 X_2 + \beta_2 X'_2 + \beta_3 X''_2$$

$$= \log \lambda_j(t) + f(X_2).$$
print(anova(f.noia), size='normalsize')

Wald Statistics for Srv

<table>
<thead>
<tr>
<th></th>
<th>(\chi^2)</th>
<th>d.f.</th>
<th>(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>72.33</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>0.69</td>
<td>2</td>
<td>0.7067</td>
</tr>
<tr>
<td>TOTAL</td>
<td>72.33</td>
<td>3</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

\(p \leftarrow \text{Predict(f.noia, age, sex, time=3, loglog=TRUE)}\)

ggplot(p, ylim=c(-5, -.5))

![Graph showing Cox PH model stratified on sex, using spline function for age, no interaction. 0.95 confidence limits also shown. Solid line is for males, dashed line is for females.](image)

Figure 20.4: Cox PH model stratified on sex, using spline function for age, no interaction. 0.95 confidence limits also shown. Solid line is for males, dashed line is for females.

Formal test of linearity: \(H_0 : \beta_2 = \beta_3 = 0, \chi^2 = 4.84, 2\) d.f., \(P = 0.09\).

- Model allowing interaction with sex strata:

\[
\log \lambda(t|X_2, C = j) = \log \lambda_j(t) + \beta_1 X_2 + \beta_2 X_2' + \beta_3 X_2'' + \beta_4 X_1 X_2 + \beta_5 X_1 X_2' + \beta_6 X_1 X_2''.
\]

Test for interaction: \(P = 0.33\).

f.ia \(\leftarrow\) cph(Srv \sim \text{rcs(age,4)} * \text{strat(sex)}, x=TRUE, y=TRUE, surv=TRUE)

w \(\leftarrow\) latex(f.ia, file='f.ia.tex', inline=TRUE, digits=3)

print(anova(f.ia), size='normalsize')
### Chapter 20. Cox Proportional Hazards Regression Model

#### Wald Statistics for Srv

<table>
<thead>
<tr>
<th>Term</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>age (Factor+Higher Order Factors)</td>
<td>72.82</td>
<td>6</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>1.05</td>
<td>3</td>
<td>0.7886</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>1.80</td>
<td>4</td>
<td>0.7728</td>
</tr>
<tr>
<td>age $\times$ sex (Factor+Higher Order Factors)</td>
<td>1.05</td>
<td>3</td>
<td>0.7886</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>1.05</td>
<td>2</td>
<td>0.5911</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>1.05</td>
<td>2</td>
<td>0.5911</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>1.80</td>
<td>4</td>
<td>0.7728</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>1.80</td>
<td>5</td>
<td>0.8763</td>
</tr>
<tr>
<td>TOTAL</td>
<td>72.82</td>
<td>6</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

```r
p ← Predict(f.ia, age, sex, time=3, loglog=TRUE)
ggplot(p, ylim=c(-5, -.5))
```

![Graph showing Cox PH model stratified on sex, with interaction between age spline and sex.](image)

**Figure 20.5:** Cox PH model stratified on sex, with interaction between age spline and sex. 0.95 confidence limits are also shown. Solid line is for males, dashed line for females.

- Example of modeling a single continuous variable (left ventricular ejection fraction), outcome = time to cardiovascular death

$$LVEF' = \begin{cases} 
  LVEF & \text{if } LVEF \leq 0.5, \\
  0.5 & \text{if } LVEF > 0.5,
\end{cases}$$

The AICs for 3, 4, 5, and 6-knots spline fits were respectively 126, 124, 122, and 120.

Smoothed residual plot: Martingale residuals, loess smoother
Figure 20.6: Restricted cubic spline estimate of relationship between LVEF and relative log hazard from a sample of 979 patients and 198 cardiovascular deaths. Data from the Duke Cardiovascular Disease Databank.

Figure 20.7: Three smoothed estimates relating martingale residuals \([200]\) to LVEF.

- One vector of residuals no matter how many covariables
- Unadjusted estimates of regression shape obtained by fixing \(\hat{\beta} = 0\) for all \(X\)s
<table>
<thead>
<tr>
<th>Purpose</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate transformation for a single variable</td>
<td>Force $\hat{\beta}_1 = 0$ and compute residuals off of the null regression</td>
</tr>
<tr>
<td>Check linearity assumption for a single variable</td>
<td>Compute $\hat{\beta}_1$ and compute residuals off of the linear regression</td>
</tr>
<tr>
<td>Estimate marginal transformations for $p$ variables</td>
<td>Force $\hat{\beta}_1, \ldots, \hat{\beta}_p = 0$ and compute residuals off the global null model</td>
</tr>
<tr>
<td>Estimate transformation for variable $i$ adjusted for other $p - 1$ variables</td>
<td>Estimate $p - 1$ $\hat{\beta}$s, forcing $\hat{\beta}_i = 0$ and compute residuals off of mixed global/null model</td>
</tr>
</tbody>
</table>

20.6.2

### Proportional Hazards Assumption

- Parallelism of $\log \Lambda$ plots
- Comparison of stratified and modeled estimates of $S(t)$
- Plot actual ratio of estimated $\Lambda$, or get differences in $\log \Lambda$
- Plot $\hat{\Lambda}$ vs. cumulative number of events as $t \uparrow$
- Stratify time, get interval-specific Cox regression coefficients: In an interval, exclude all subjects with event/censoring time before start of interval.

Censor all events at end of interval.

```r
f <- cph(S ~ strat(group), surv=TRUE)
# For both strata, eval. S(t) at combined set of death times
times <- sort(unique(days[death == 1]))
est <- survest(f, data.frame(group=levels(group)),
              times=times, conf.type="none")$surv

cumhaz <- -log(est)
plot(times, cumhaz[2,] / cumhaz[1,], xlab="Days",
     ylab="Cumulative Hazard Ratio", type="s")
```
Figure 20.8: Estimate of $\Lambda_2/\Lambda_1$ based on $-\log$ of Altschuler–Nelson–Fleming–Harrington nonparametric survival estimates.

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Observations</th>
<th>Deaths</th>
<th>Log Hazard Ratio</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 209)</td>
<td>40</td>
<td>12</td>
<td>-0.47</td>
<td>0.59</td>
</tr>
<tr>
<td>[209, 234)</td>
<td>27</td>
<td>12</td>
<td>-0.72</td>
<td>0.58</td>
</tr>
<tr>
<td>234 +</td>
<td>14</td>
<td>12</td>
<td>-0.50</td>
<td>0.64</td>
</tr>
</tbody>
</table>

Overall Cox $\hat{\beta} = -0.57$.

- VA Lung Cancer dataset, squamous vs. (small, adeno)
Figure 20.9: Stratified hazard ratios for pain/ischemia index over time. Data from the Duke Cardiovascular Disease Databank.

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Observations</th>
<th>Deaths</th>
<th>Log Hazard Ratio</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 21)</td>
<td>110</td>
<td>26</td>
<td>-0.46</td>
<td>0.47</td>
</tr>
<tr>
<td>[21, 52)</td>
<td>84</td>
<td>26</td>
<td>-0.90</td>
<td>0.50</td>
</tr>
<tr>
<td>[52, 118)</td>
<td>59</td>
<td>26</td>
<td>-1.35</td>
<td>0.50</td>
</tr>
<tr>
<td>118 +</td>
<td>28</td>
<td>26</td>
<td>-1.04</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Estimates for Karnofsky performance status weight over time:

<table>
<thead>
<tr>
<th>Time Interval</th>
<th>Observations</th>
<th>Deaths</th>
<th>Log Hazard Ratio</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 19)</td>
<td>137</td>
<td>27</td>
<td>-0.053</td>
<td>0.010</td>
</tr>
<tr>
<td>[19, 49)</td>
<td>112</td>
<td>26</td>
<td>-0.047</td>
<td>0.009</td>
</tr>
<tr>
<td>[49, 99)</td>
<td>85</td>
<td>27</td>
<td>-0.036</td>
<td>0.012</td>
</tr>
<tr>
<td>99 +</td>
<td>28</td>
<td>26</td>
<td>-0.012</td>
<td>0.014</td>
</tr>
</tbody>
</table>

- Schoenfeld residuals computed at each unique failure time
- Partial derivative of $\log L$ with respect to each $X$ in turn
- Grambsch and Therneau scale to yield estimates of $\beta(t)$
Can form a powerful test of PH

\[ \hat{\beta} + dR\hat{V}, \]

Can test PH by testing \( t \times X \) interaction using time-dependent covariables

Separate parametric fits, e.g. Weibull with differing \( \gamma \); hazard ratio is

\[ \frac{\alpha \gamma t^{\gamma-1}}{\delta \theta t^{\theta-1}} = \frac{\alpha \gamma}{\delta \theta} t^{\gamma-\theta}. \]
<table>
<thead>
<tr>
<th>$t$</th>
<th>log Hazard Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>-0.36</td>
</tr>
<tr>
<td>36</td>
<td>-0.64</td>
</tr>
<tr>
<td>83.5</td>
<td>-0.83</td>
</tr>
<tr>
<td>200</td>
<td>-1.02</td>
</tr>
</tbody>
</table>

- Interaction between $X$ and spline function of $t$:

$$\log \lambda(t|X) = \log \lambda(t) + \beta_1 X + \beta_2 Xt + \beta_3 Xt' + \beta_4 Xt'',$$

The $X + 1 : X$ log hazard ratio function is estimated by

$$\hat{\beta}_1 + \hat{\beta}_2 t + \hat{\beta}_3 t' + \hat{\beta}_4 t''.$$
### Assumptions of the Proportional Hazards Model

\[ \lambda(t|X) = \lambda(t)e^{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p} \]

<table>
<thead>
<tr>
<th>Variables</th>
<th>Assumptions</th>
<th>Verification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Response Variable</strong> T Time Until Event</td>
<td>Shape of ( \lambda(t</td>
<td>X) ) for fixed ( X ) as ( t \uparrow )</td>
</tr>
<tr>
<td>Cox: none</td>
<td>Weibull: ( t^\theta )</td>
<td></td>
</tr>
<tr>
<td><strong>Interaction between ( X ) and ( T )</strong></td>
<td>Proportional hazards – effect of ( X ) does not depend on ( T ). E.g. treatment effect is constant over time.</td>
<td></td>
</tr>
<tr>
<td><strong>Individual Predictors</strong> ( X )</td>
<td>Shape of ( \lambda(t</td>
<td>X) ) for fixed ( t ) as ( X \uparrow )</td>
</tr>
<tr>
<td>Linear: [ \log \lambda(t</td>
<td>X) = \log \lambda(t) + \beta X ]</td>
<td></td>
</tr>
<tr>
<td>Nonlinear: [ \log \lambda(t</td>
<td>X) = \log \lambda(t) + f(X) ]</td>
<td></td>
</tr>
<tr>
<td><strong>Interaction between ( X_1 ) and ( X_2 )</strong></td>
<td>Additive effects: effect of ( X_1 ) on ( \log \lambda ) is independent of ( X_2 ) and vice-versa</td>
<td>Test non-additive terms, e.g. products</td>
</tr>
<tr>
<td>k-level ordinal ( X ): linear term + ( k - 2 ) dummy variables</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Continuous ( X ): Polynomials, spline functions, smoothed martingale residual plots</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Method

<table>
<thead>
<tr>
<th></th>
<th>Requires Grouping X</th>
<th>Requires Grouping t</th>
<th>Computational Efficiency</th>
<th>Yields Formal Test</th>
<th>Yields Estimate of $\lambda_2(t)/\lambda_1(t)$</th>
<th>Requires Fitting 2 Models</th>
<th>Must Choose Smoothing Parameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>log[-log], Muenz, Arjas plots</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Dabrowska log $\hat{\lambda}$ difference plots</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Stratified vs. Modeled Estimates</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Hazard ratio plot</td>
<td>x</td>
<td>?</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>?</td>
</tr>
<tr>
<td>Schoenfeld residual plot</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Schoenfeld residual correlation test</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Fit time-dependent covariables</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratio of parametric estimates of $\lambda(t)$</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>
20.7 What to Do When PH Fails

- Test of association not needed and the key variable is categorical $\rightarrow$ stratify

- Key results display: covariate-adjusted cumulative incidence curves by strata with confidence bands for the difference in the two curves
  - allows curves to cross

- $P$-value for testing variable may still be useful (conservative)

- Survival estimates wrong in certain time intervals

- Can model non-PH:
  
  \[
  \lambda(t|X) = \lambda_0(t) \exp(\beta_1 X + \beta_2 X \times \log(t + 1))
  \]

  For this model, Breslow, Edler, Berger [28] derived a simple 2 d.f. score test for whether one group has a different hazard rate than the other group at any time $t$

- Can also use time intervals:
  
  \[
  \lambda(t|X) = \lambda_0(t) \exp(\beta_1 X + \beta_2 X \times [t > c])
  \]

- Or fit one model for early follow-up, one for late
- Try another model, e.g. log-normal, log-logistic can have effects of $X$ changing constantly over time

- Differences in mean restricted life length can be useful in comparing therapies when PH fails [113], but see bit.ly/datamethods-rmst

See [172, 164, 151].
20.8 Collinearity

20.9 Overly Influential Observations

20.10 Quantifying Predictive Ability

\[ R_{LR}^2 = 1 - \exp\left(\frac{-LR}{n}\right) = 1 - \frac{\omega^2}{n}, \]

- \( \omega \) = null model likelihood divided by the fitted model likelihood

- Divide by max attainable value to get \( R_N^2 \)

\( c = \) concordance probability (predicted vs. observed)

- All possible pairs of subjects whose ordering of failure times can be determined

- Fraction of these for which \( X \) ordered same as \( Y \)

- Somers’ \( D_{xy} = 2(c - 0.5) \)

See fharrell.com/post/addvalue for more about the most sensi-
tive values for assessing predictive discrimination and comparing competing models.
Validating the Fitted Model

Separate bootstrap validations for calibration and for discrimination. For external validation, a sample containing at least 200 events is needed [45].

Validation of Model Calibration

- Calibration at fixed $t$
- Get $\hat{S}(t \mid X)$ for all subjects
- Divide into intervals each containing say 50 subjects
- Compare mean predicted survival with K-M
- Bootstrap this process to add back optimism in difference of these 2, due to overfitting
- Ex: 20 random predictors, $n = 200$

```r
n ← 200
p ← 20
set.seed(6)
xx ← matrix(rnorm(n * p), nrow = n, ncol = p)
y ← runif(n)
units(y) ← "Year"
e ← c(rep(0, n / 2), rep(1, n / 2))
f ← cph(Surv(y, e) ~ xx, x=TRUE, y=TRUE,
time.inc=.5, surv=TRUE)
cal ← calibrate(f, u=.5, B=200)
```
CHAPTER 20. COX PROPORTIONAL HAZARDS REGRESSION MODEL

Using Cox survival estimates at 0.5 Years

```r
plot(cal, ylim=c(.4, 1), subtitles=FALSE)
calkm ← calibrate(f, u=.5, m=40, cmethod='KM', B=200)
```

Using Cox survival estimates at 0.5 Years

```r
plot(calkm, add=TRUE) # Figure 20.11
```

![Figure 20.11: Calibration of random predictions using Efron's bootstrap with $B = 200$ resamples. Dataset has $n = 200$, 100 uncensored observations, 20 random predictors, model $\chi^2_{20} = 19$. The smooth black line is the apparent calibration estimated by adaptive linear spline hazard regression [122], and the blue line is the bootstrap bias– (overfitting–) corrected calibration curve estimated also by hazard regression. The gray scale line is the line of identity representing perfect calibration. Black dots represent apparent calibration accuracy obtained by stratifying into intervals of predicted 0.5y survival containing 40 events per interval and plotting the mean predicted value within the interval against the stratum's Kaplan-Meier estimate. The blue × represent bootstrap bias-corrected Kaplan-Meier estimates.]

---

### 20.11.2 Validation of Discrimination and Other Statistical Indexes

Validate slope calibration (estimate shrinkage from overfitting):

$$\lambda(t|X) = \lambda(t) \exp(\gamma Xb).$$

```latex
\text{latex(validate(f, B=200), digits=3, file='',
caption='Bootstrap validation of a Cox model with random predictors',
table.env=TRUE, label='tab:cox-val-random')}
```
Table 20.1: Bootstrap validation of a Cox model with random predictors

<table>
<thead>
<tr>
<th>Index</th>
<th>Original Sample</th>
<th>Training Sample</th>
<th>Test Sample</th>
<th>Optimism</th>
<th>Corrected Index</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{xy}$</td>
<td>0.213</td>
<td>0.332</td>
<td>0.144</td>
<td>0.188</td>
<td>0.025</td>
<td>200</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.092</td>
<td>0.192</td>
<td>0.041</td>
<td>0.151</td>
<td>$-0.060$</td>
<td>200</td>
</tr>
<tr>
<td>Slope</td>
<td>1.000</td>
<td>1.000</td>
<td>0.383</td>
<td>0.617</td>
<td>0.383</td>
<td>200</td>
</tr>
<tr>
<td>$D$</td>
<td>0.021</td>
<td>0.048</td>
<td>0.008</td>
<td>0.040</td>
<td>$-0.019$</td>
<td>200</td>
</tr>
<tr>
<td>$U$</td>
<td>$-0.002$</td>
<td>$-0.002$</td>
<td>0.030</td>
<td>$-0.032$</td>
<td>0.030</td>
<td>200</td>
</tr>
<tr>
<td>$Q$</td>
<td>0.023</td>
<td>0.050</td>
<td>$-0.022$</td>
<td>0.072</td>
<td>$-0.049$</td>
<td>200</td>
</tr>
<tr>
<td>$g$</td>
<td>0.516</td>
<td>0.872</td>
<td>0.332</td>
<td>0.539</td>
<td>$-0.023$</td>
<td>200</td>
</tr>
</tbody>
</table>
Figure 20.12: A display of an interaction between treatment and extent of disease, and between treatment and calendar year of start of treatment. Comparison of medical and surgical average hazard ratios for patients treated in 1970, 1977, and 1984 according to coronary artery disease severity. Circles represent point estimates; bars represent 0.95 confidence limits of hazard ratios. Hazard ratios < 1 indicate that surgery is more effective. [34]

### Describing the Fitted Model

- Can use coefficients if linear and additive
- Can use e.g. inter-quartile-range hazard ratios for various levels of interacting factors if linearity holds approximately

```r
p ← Predict(f.ia, age, sex, time=3)
ggplot(p)
```

- Nomogram to compute $X\hat{\beta}$
- Also $\hat{S}(t|X)$ for a few $t$
- Can have axis for median failure time if sample is high risk
Figure 20.13: Cox–Kalbfleisch–Prentice survival estimates stratifying on treatment and adjusting for several predictors, showing a secular trend in the efficacy of coronary artery bypass surgery. Estimates are for patients with left main disease and normal (LVEF=0.6) or impaired (LVEF=0.4) ventricular function. [170]

Figure 20.14: Cox model predictions with respect to a continuous variable. X-axis shows the range of the treadmill score seen in clinical practice and Y-axis shows the corresponding 5-year survival probability predicted by the Cox regression model for the 2842 study patients [144].
Figure 20.15: Survival estimates for model stratified on sex, with interaction.
**R Functions**

### 20.13.1 Power and Sample Size Calculations, Hmisc Package

- **cpower**: computes power for a two-sample Cox test with random patient entry over a fixed duration and a given length of minimum follow-up, using exponential distribution with handling of dropout and drop-in [124]

- **ciapower**: computes power of the Cox interaction test in a $2 \times 2$ setup using the method of Peterson and George [166]

- **spower**: simulates power for 2-sample tests (the log-rank test by default) allowing for very complex conditions such as continuously varying treatment effect and non-compliance probabilities.

### 20.13.2 Cox Model using rms Package

- **cph**: slight modification of Therneau’s survival package `coxph` function

- **print** method prints the Nagelkerke index $R_N^2$ (Section 20.10)

- **cph** works with generic functions such as `specs`, `predict`,
CHAPTER 20. COX PROPORTIONAL HAZARDS REGRESSION MODEL

summary, anova, fastbw, which.influence, latex, residuals, coef, nomogram, plot, ggplot, plotp,

• plot, ggplot, plotp have an additional argument time for plotting cph fits. It also has an argument loglog which if \( T \) causes instead log -log survival to be plotted on the \( y \)-axis.

• Survival.cph, Quantile.cph, Mean.cph create other R functions to evaluate survival probabilities, survival time quantiles, and mean and mean restricted lifetimes, based on a cph fit with surv=TRUE

• Quantile and Mean are especially useful with plot, ggplot, plotp and nomogram. Survival is useful with nomogram

\[
f \leftarrow \text{cph}(\ldots, \text{surv}=\text{T})
\]
\[
\text{med} \leftarrow \text{Quantile}(f)
\]
\[
\text{plot(nomogram(f, fun=function(lp) \text{med}(lp=x), funlabel=’Median Survival Time’))}
\]

\[
\# \text{fun transforms the linear predictors}
\]
\[
\text{srv} \leftarrow \text{Survival}(f)
\]
\[
\text{rmean} \leftarrow \text{Mean}(f, \text{tmax}=3, \text{method}=’approx’)
\]
\[
\text{plot(nomogram(f, fun=list(function(x) srv(3, x), rmean), funlabel=c(’3-Year Survival Prob.’, ’Restricted Mean’)))}
\]

\[
\# \text{med, srv, expected are more complicated if strata are present}
\]

The R program below demonstrates how several cph-related functions work well with the nomogram function to display this last fit. Here predicted 3-year survival probabilities and median survival time (when defined) are displayed against age and sex. The fact that a nonlinear effect interacts with a stratified factor is taken into account.

\[
surv \leftarrow \text{Survival}(f.ia)
\]
\[
surv.f \leftarrow \text{function(lp) surv(3, lp, stratum=’sex=Female’)}
\]
\[
surv.m \leftarrow \text{function(lp) surv(3, lp, stratum=’sex=Male’)}
\]
\[
\text{quant} \leftarrow \text{Quantile}(f.ia)
\]
\[
\text{Prob}\{T \geq t \mid \text{sex} = i\} = S_i(t)e^{X\hat{\beta}}, \quad \text{where}
\]

\[
X\hat{\beta} =
\]

\[-1.8
\]

\[+0.0493\text{age} - 2.15 \times 10^{-6}(\text{age} - 30.3)^3_+ - 2.82 \times 10^{-5}(\text{age} - 45.1)
\]

\[+5.18 \times 10^{-5}(\text{age} - 54.6)^3_+ - 2.15 \times 10^{-5}(\text{age} - 69.6)^3_+
\]

\[+[\text{Female}][-0.0366\text{age} + 4.29 \times 10^{-5}(\text{age} - 30.3)^3_+ - 0.00011(\text{age})
\]

\[+6.74 \times 10^{-5}(\text{age} - 54.6)^3_+ - 2.32 \times 10^{-7}(\text{age} - 69.6)^3_+]
\]

and \([c] = 1\) if subject is in group \(c\), 0 otherwise; \((x)_+ = x\) if \(x > 0\), 0 otherwise
<table>
<thead>
<tr>
<th>$t$</th>
<th>$S_{Male}(t)$</th>
<th>$S_{Female}(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>1</td>
<td>0.993</td>
<td>0.902</td>
</tr>
<tr>
<td>2</td>
<td>0.984</td>
<td>0.825</td>
</tr>
<tr>
<td>3</td>
<td>0.975</td>
<td>0.725</td>
</tr>
<tr>
<td>4</td>
<td>0.967</td>
<td>0.648</td>
</tr>
<tr>
<td>5</td>
<td>0.956</td>
<td>0.576</td>
</tr>
<tr>
<td>6</td>
<td>0.947</td>
<td>0.520</td>
</tr>
<tr>
<td>7</td>
<td>0.938</td>
<td>0.481</td>
</tr>
<tr>
<td>8</td>
<td>0.928</td>
<td>0.432</td>
</tr>
<tr>
<td>9</td>
<td>0.920</td>
<td>0.395</td>
</tr>
<tr>
<td>10</td>
<td>0.909</td>
<td>0.358</td>
</tr>
<tr>
<td>11</td>
<td>0.904</td>
<td>0.314</td>
</tr>
<tr>
<td>12</td>
<td>0.892</td>
<td>0.268</td>
</tr>
<tr>
<td>13</td>
<td>0.886</td>
<td>0.223</td>
</tr>
<tr>
<td>14</td>
<td>0.877</td>
<td>0.203</td>
</tr>
</tbody>
</table>
Figure 20.16: Nomogram from a fitted stratified Cox model that allowed for interaction between age and sex, and nonlinearity in age. The axis for median survival time is truncated on the left where the median is beyond the last follow-up time.
Chapter 21

Case Study in Cox Regression

21.1 Choosing the Number of Parameters and Fitting the Model

- Clinical trial of estrogen for prostate cancer
- Response is time to death, all causes
- Base analysis on Cox proportional hazards model [52]
- \( S(t|X) = \text{probability of surviving at least to time } t \text{ given set of predictor values } X \)
- \( S(t|X) = S_0(t)^{\exp(X\beta)} \)
- Censor time to death at time of last follow-up for patients still alive at end of study (treat survival time for pt. censored at 24m as 24m+)
• Use simple, partial approaches to data reduction

• Use transcan for single imputation

• Again combine last 2 categories for ekg,pf

• See if we can use a full additive model (4 knots for continuous $X$)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Name</th>
<th>d.f.</th>
<th>Original Levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dose of estrogen</td>
<td>rx</td>
<td>3</td>
<td>placebo, 0.2, 1.0, 5.0 mg estrogen</td>
</tr>
<tr>
<td>Age in years</td>
<td>age</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Weight index: wt(kg)-ht(cm)+200</td>
<td>wt</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Performance rating</td>
<td>pf</td>
<td>2</td>
<td>normal, in bed &lt; 50% of time, in bed &gt; 50%, in bed always</td>
</tr>
<tr>
<td>History of cardiovascular disease</td>
<td>hx</td>
<td>1</td>
<td>present/absent</td>
</tr>
<tr>
<td>Systolic blood pressure/10</td>
<td>sbp</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Diastolic blood pressure/10</td>
<td>dbp</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Electrocardiogram code</td>
<td>ekg</td>
<td>5</td>
<td>normal, benign, rhythm disturb., block, strain, old myocardial infarction, new MI</td>
</tr>
<tr>
<td>Serum hemoglobin (g/100ml)</td>
<td>hg</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Tumor size (cm$^2$)</td>
<td>sz</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Stage/histologic grade combination</td>
<td>sg</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Serum prostatic acid phosphatase</td>
<td>ap</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Bone metastasis</td>
<td>bm</td>
<td>1</td>
<td>present/absent</td>
</tr>
</tbody>
</table>

• Total of 36 candidate d.f.

• Impute missings and estimate shrinkage

```r
require(rms)
options(prType='latex') # for print, summary, anova
getHdata(prostate)
levels(prostate$ekg)[levels(prostate$ekg) %in% c('old MI','recent MI')] <- 'MI'
# combines last 2 levels and uses a new name, MI
```
PROSTATE$pf.coded ← as.integer(PROSTATE$pf)
# save original pf, re-code to 1-4
levels(prostate$pf) ← c(levels(prostate$pf)[1:3],
levels(prostate$pf)[3])
# combine last 2 levels

w ← transcan(~ sz + sg + ap + sbp + dbp + age +
wt + hg + ekg + pf + bm + hx,
imputed=TRUE, data=prostate, pl=FALSE, pr=FALSE)

attach(prostate)
sz ← impute(w, sz, data=prostate)
sg ← impute(w, sg, data=prostate)
age ← impute(w, age, data=prostate)
wt ← impute(w, wt, data=prostate)
ekg ← impute(w, ekg, data=prostate)

dd ← datadist(prostate)
options(datadist='dd')

units(dtime) ← 'Month'
S ← Surv(dtime, status!='alive')

f ← cph(S ~ rx + rcs(age, 4) + rcs(wt, 4) + pf + hx +
rcs(sbp, 4) + rcs(dbp, 4) + ekg + rcs(hg, 4) +
rcs(sg, 4) + rcs(sz, 4) + rcs(log(ap), 4) + bm)

print(f, coefs=FALSE)

Cox Proportional Hazards Model

cph(formula = S ~ rx + rcs(age, 4) + rcs(wt, 4) + pf + hx + rcs(sbp,
4) + rcs(dbp, 4) + ekg + rcs(hg, 4) + rcs(sg, 4) + rcs(sz,
4) + rcs(log(ap), 4) + bm)

<table>
<thead>
<tr>
<th></th>
<th>Model Tests</th>
<th>Discrimination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs 502</td>
<td>LR $\chi^2$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>Events 354</td>
<td>d.f. 36</td>
<td>$D_{xy}$</td>
</tr>
<tr>
<td>Center -2.9933</td>
<td>Pr($&gt;\chi^2$</td>
<td>$g$</td>
</tr>
<tr>
<td></td>
<td>0.0000</td>
<td>$g_r$</td>
</tr>
<tr>
<td>Score $\chi^2$</td>
<td>143.62</td>
<td>0.238</td>
</tr>
<tr>
<td>Pr($&gt;\chi^2$</td>
<td>0.0000</td>
<td>0.333</td>
</tr>
</tbody>
</table>

- Global LR $\chi^2$ is 135 and very significant $\rightarrow$ modeling warranted
• AIC on $\chi^2$ scale = $136.2 - 2 \times 36 = 64.2$

• Rough shrinkage: $0.74 \left( \frac{136.2 - 36}{136.2} \right)$

• Informal data reduction (increase for $ap$)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Reductions</th>
<th>d.f. Saved</th>
</tr>
</thead>
<tbody>
<tr>
<td>$wt$</td>
<td>Assume variable not important enough for 4 knots; use 3 knots</td>
<td>1</td>
</tr>
<tr>
<td>$pf$</td>
<td>Assume linearity</td>
<td>1</td>
</tr>
<tr>
<td>$hx, ekg$</td>
<td>Make new 0,1,2 variable and assume linearity: $2=hx$ and $ekg$ not normal or benign, $1=market$, $0=none$</td>
<td>5</td>
</tr>
<tr>
<td>$sbp, dbp$</td>
<td>Combine into mean arterial bp and use 3 knots: $map=\frac{2}{3} dbp + \frac{1}{3} sbp$</td>
<td>4</td>
</tr>
<tr>
<td>$sg$</td>
<td>Use 3 knots</td>
<td>1</td>
</tr>
<tr>
<td>$sz$</td>
<td>Use 3 knots</td>
<td>1</td>
</tr>
<tr>
<td>$ap$</td>
<td>Look at shape of effect of $ap$ in detail, and take log before expanding as spline to achieve numerical stability: add 1 knot</td>
<td>-1</td>
</tr>
</tbody>
</table>
heart <- hx + ekg %nin% c('normal', 'benign')
label(heart) <- 'Heart Disease Code'
map <- (2*dbp + sbp)/3
label(map) <- 'Mean Arterial Pressure/10'
dd <- datadist(dd, heart, map)

f <- cph(S ~ rx + rcs(age,4) + rcs(wt,3) + pf.coded +
    heart + rcs(map,3) + rcs(hg,4) +
    rcs(sg,3) + rcs(sz,3) + rcs(log(ap),5) + bm,
    x=TRUE, y=TRUE, surv=TRUE, time.inc=5*12)
print(f, coefs=FALSE)

Cox Proportional Hazards Model

cph(formula = S ~ rx + rcs(age, 4) + rcs(wt, 3) + pf.coded +
    heart + rcs(map, 3) + rcs(hg, 4) + rcs(sg, 3) + rcs(sz, 3) +
    rcs(log(ap), 5) + bm, x = TRUE, y = TRUE, surv = TRUE, time.inc = 5 * 12)

<table>
<thead>
<tr>
<th>Obs</th>
<th>502</th>
<th>Events</th>
<th>354</th>
<th>Center</th>
<th>-2.4307</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Tests</td>
<td>LR $\chi^2$</td>
<td>118.37</td>
<td>d.f.</td>
<td>24</td>
<td>$D_{xy}$</td>
</tr>
<tr>
<td></td>
<td>Pr($&gt;\chi^2$)</td>
<td>0.0000</td>
<td></td>
<td></td>
<td>$g$</td>
</tr>
<tr>
<td></td>
<td>Score $\chi^2$</td>
<td>125.58</td>
<td></td>
<td></td>
<td>$g_r$</td>
</tr>
<tr>
<td></td>
<td>Pr($&gt;\chi^2$)</td>
<td>0.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

# x, y for predict, validate, calibrate;
# surv, time.inc for calibrate
anova(f)

Wald Statistics for S
<table>
<thead>
<tr>
<th></th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>rx</td>
<td>8.01</td>
<td>3</td>
<td>0.0459</td>
</tr>
<tr>
<td>age</td>
<td>13.84</td>
<td>3</td>
<td>0.0031</td>
</tr>
<tr>
<td></td>
<td>Nonlinear</td>
<td>9.06</td>
<td>2 0.0108</td>
</tr>
<tr>
<td>wt</td>
<td>8.21</td>
<td>2</td>
<td>0.0165</td>
</tr>
<tr>
<td></td>
<td>Nonlinear</td>
<td>2.54</td>
<td>1 0.1110</td>
</tr>
<tr>
<td>pf.coded</td>
<td>3.79</td>
<td>1</td>
<td>0.0517</td>
</tr>
<tr>
<td>heart</td>
<td>23.51</td>
<td>1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>map</td>
<td>0.04</td>
<td>2</td>
<td>0.9779</td>
</tr>
<tr>
<td></td>
<td>Nonlinear</td>
<td>0.04</td>
<td>1 0.8345</td>
</tr>
<tr>
<td>hg</td>
<td>12.52</td>
<td>3</td>
<td>0.0058</td>
</tr>
<tr>
<td></td>
<td>Nonlinear</td>
<td>8.25</td>
<td>2 0.0162</td>
</tr>
<tr>
<td>sg</td>
<td>1.64</td>
<td>2</td>
<td>0.4406</td>
</tr>
<tr>
<td></td>
<td>Nonlinear</td>
<td>0.05</td>
<td>1 0.8304</td>
</tr>
<tr>
<td>sz</td>
<td>12.73</td>
<td>2</td>
<td>0.0017</td>
</tr>
<tr>
<td></td>
<td>Nonlinear</td>
<td>0.06</td>
<td>1 0.7990</td>
</tr>
<tr>
<td>ap</td>
<td>6.51</td>
<td>4</td>
<td>0.1639</td>
</tr>
<tr>
<td></td>
<td>Nonlinear</td>
<td>6.22</td>
<td>3 0.1012</td>
</tr>
<tr>
<td>bm</td>
<td>0.03</td>
<td>1</td>
<td>0.8670</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>23.81</td>
<td>11</td>
<td>0.0136</td>
</tr>
<tr>
<td>TOTAL</td>
<td>119.09</td>
<td>24</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>

- Savings of 12 d.f.
- AIC=70, shrinkage 0.80
21.2 Checking Proportional Hazards

- This is our tentative model
- Examine distributional assumptions using scaled Schoenfeld residuals
- Complication arising from predictors using multiple d.f.
- Transform to 1 d.f. empirically using \( X \hat{\beta} \)
- Following analysis approx. since internal coefficients estimated

```r
z ← predict(f, type='terms')
# required x=T above to store design matrix
f.short ← cph(S ~ z, x=TRUE, y=TRUE)
# store raw x, y so can get residuals

ph.test ← cox.zph(f.short, transform='identity')
ph.test
```

- Fit \( f_.short \) has same LR \( \chi^2 \) of 118 as the fit \( f \), but with falsely low d.f.
- All \( \beta = 1 \)

<table>
<thead>
<tr>
<th></th>
<th>rho</th>
<th>chisq</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>rx</td>
<td>0.10232</td>
<td>4.00823</td>
<td>0.0453</td>
</tr>
<tr>
<td>age</td>
<td>-0.05483</td>
<td>1.05850</td>
<td>0.3036</td>
</tr>
<tr>
<td>wt</td>
<td>0.01838</td>
<td>0.11632</td>
<td>0.7331</td>
</tr>
<tr>
<td>pf.coded</td>
<td>-0.03429</td>
<td>0.41884</td>
<td>0.5175</td>
</tr>
<tr>
<td>heart</td>
<td>0.02650</td>
<td>0.30052</td>
<td>0.5836</td>
</tr>
<tr>
<td>map</td>
<td>0.02055</td>
<td>0.14135</td>
<td>0.7069</td>
</tr>
<tr>
<td>hq</td>
<td>-0.00362</td>
<td>0.00511</td>
<td>0.9430</td>
</tr>
<tr>
<td>sg</td>
<td>-0.05137</td>
<td>0.94589</td>
<td>0.3308</td>
</tr>
<tr>
<td>sz</td>
<td>-0.01554</td>
<td>0.08330</td>
<td>0.7729</td>
</tr>
<tr>
<td>ap</td>
<td>0.01720</td>
<td>0.11858</td>
<td>0.7306</td>
</tr>
</tbody>
</table>
\textbf{Figure 21.1:} Raw and spline-smoothed scaled Schoenfeld residuals for dose of estrogen, nonlinearly coded from the Cox model fit, with $\pm 2$ standard errors.

- Only the drug effect significantly changes over time
- Global test of PH $P = 0.78$
## Testing Interactions

- Will ignore non-PH for dose even though it makes sense

- More accurate predictions could be obtained using stratification or time dep. cov.

- Test all interactions with dose
  
  Reduce to 1 d.f. as before

```
z.dose <- z[, "rx"]  # same as saying z[, 1] - get first column
z.other <- z[, -1]    # all but the first column of z
f.ia <- cph(S ~ z.dose * z.other)
print(anova(f.ia), size='tsz')
```

### Wald Statistics for $S$

<table>
<thead>
<tr>
<th></th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>z.dose (Factor+Higher Order Factors)</td>
<td>18.74</td>
<td>11</td>
<td>0.0660</td>
</tr>
<tr>
<td>All Interactions</td>
<td>12.17</td>
<td>10</td>
<td>0.2738</td>
</tr>
<tr>
<td>z.other (Factor+Higher Order Factors)</td>
<td>125.89</td>
<td>20</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>12.17</td>
<td>10</td>
<td>0.2738</td>
</tr>
<tr>
<td>z.dose $\times$ z.other (Factor+Higher Order Factors)</td>
<td>12.17</td>
<td>10</td>
<td>0.2738</td>
</tr>
<tr>
<td>TOTAL</td>
<td>129.10</td>
<td>21</td>
<td>&lt;0.0001</td>
</tr>
</tbody>
</table>
21.4 Describing Predictor Effects

- Plot relationship between each predictor and $\log \lambda$

```r
ggplot(Predict(f), sepdiscrete='vertical', nlevels=4, vnames='names') # Figure 21.2
```

Figure 21.2: Shape of each predictor on log hazard of death. $Y$-axis shows $X\hat{\beta}$, but the predictors not plotted are set to reference values. Note the highly non-monotonic relationship with $ap$, and the increased slope after age 70 which has been found in outcome models for various diseases.
Validating the Model

- Validate for $D_{xy}$ and slope shrinkage

```r
set.seed(1)  # so can reproduce results
v ← validate(f, B=300)
lutex(v, file=’’)
```

<table>
<thead>
<tr>
<th>Index</th>
<th>Original Sample</th>
<th>Training Sample</th>
<th>Test Sample</th>
<th>Optimism</th>
<th>Corrected Index</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{xy}$</td>
<td>0.3208</td>
<td>0.3467</td>
<td>0.2953</td>
<td>0.0514</td>
<td>0.2695</td>
<td>300</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.2101</td>
<td>0.2458</td>
<td>0.1751</td>
<td>0.0706</td>
<td>0.1395</td>
<td>300</td>
</tr>
<tr>
<td>Slope</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.7900</td>
<td>0.2100</td>
<td>0.7900</td>
<td>300</td>
</tr>
<tr>
<td>$D$</td>
<td>0.0292</td>
<td>0.0351</td>
<td>0.0238</td>
<td>0.0113</td>
<td>0.0179</td>
<td>300</td>
</tr>
<tr>
<td>$U$</td>
<td>−0.0005</td>
<td>−0.0005</td>
<td>0.0023</td>
<td>−0.0028</td>
<td>0.0023</td>
<td>300</td>
</tr>
<tr>
<td>$Q$</td>
<td>0.0297</td>
<td>0.0356</td>
<td>0.0214</td>
<td>0.0141</td>
<td>0.0155</td>
<td>300</td>
</tr>
<tr>
<td>$g$</td>
<td>0.7174</td>
<td>0.7950</td>
<td>0.6265</td>
<td>0.1685</td>
<td>0.5489</td>
<td>300</td>
</tr>
</tbody>
</table>

- Shrinkage surprisingly close to heuristic estimate of 0.79

- Now validate 5-year survival probability estimates

```r
cal ← calibrate(f, B=300, u=5*12, maxdim=4)
plot(cal)
```

Using Cox survival estimates at 60 Months
Figure 21.3: Bootstrap estimate of calibration accuracy for 5-year estimates from the final Cox model, using adaptive linear spline hazard regression. Line nearer the ideal line corresponds to apparent predictive accuracy. The blue curve corresponds to bootstrap-corrected estimates.
21.6 Presenting the Model

- Display hazard ratios, overriding default for \( ap \)

\[
\text{plot(summary(f, ap=c(1,20)), log=TRUE, main='')}\]

![Figure 21.4: Hazard ratios and multi-level confidence bars for effects of predictors in model, using default ranges except for \( ap \)](image)

- Draw nomogram, with predictions stated 4 ways

```r
surv ← Survival(f)
surv3 ← function(x) surv(3*12, lp=x)
surv5 ← function(x) surv(5*12, lp=x)
quan ← Quantile(f)
med ← function(x) quan(lp=x)/12
ss ← c(.05,.1,.2,.3,.4,.5,.6,.7,.8,.9,.95)

nom ← nomogram(f, ap=c(.1,.5,1,2,3,4,5,10,20,30,40),
    fun=list(surv3, surv5, med),
    funlabel=c('3-year Survival','5-year Survival','Median Survival Time (years)'),
    fun.at=list(ss, ss, c(.5,1:6)))
plot(nom, xfrac=.65, lmgp=.35)
```
Figure 21.5: Nomogram for predicting death in prostate cancer trial
Annotated Bibliography


[11] Peter C. Austin, Jack V. Tu, and Douglas S. Lee. "Logistic Regression Had Superior Performance Compared with Regression Trees for Predicting In-Hospital Mortality in Patients Hospitalized with Heart Failure". In: *J Clin Epi* 63 (2010). ROC areas for logistic models varied from 0.747 to 0.775 whereas they varied from 0.620-0.651 for recursive partitioning; repeated data simulation showed large variation in tree structure, pp. 1145–1155 (cit. on p. 2-35).

[12] Sunni A. Barnes, Stacy R. Lindborg, and John W. Seaman. "Multiple Imputation Techniques in Small Sample Clinical Trials". In: *Stat Med* 25 (2006). bad performance of LOCF including high bias and poor confidence interval coverage; simulation setup; longitudinal data; serial data; RCT; dropout; assumed missing at random (MAR); approximate Bayesian bootstrap; Bayesian least squares; missing data; nice background summary; new completion score method based on fitting a Poisson model for the number of completed clinic visits and using donors and approximate Bayesian bootstrap, pp. 233–245 (cit. on p. 3-15).
[13] Federica Barzi and Mark Woodward. “Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies”. In: *Am J Epi* 160 (2004), excellent review article for multiple imputation; list of variables to include in imputation model; imputation models should ideally include all covariates that are related to the missing data mechanism, have distributions that differ between the respondents and nonrespondents, are associated with cholesterol, and will be included in the analyses of the final complete data sets; detailed comparison of results (cholesterol effect and confidence limits) for various imputation methods, pp. 34–45 (cit. on pp. 3-8, 3-15).


[23] James G. Booth and Somnath Sarkar. “Monte Carlo Approximation of Bootstrap Variances”. In: *Am Statistician* 52 (1998). number of resamples required to estimate variances, quantiles; 800 resamples may be required to guarantee with 0.95 confidence that the relative error of a variance estimate is 0.1; Efron’s original suggestions for as low as 25 resamples were based on comparing stability of bootstrap estimates to sampling error, but small relative effects can significantly change P-values; number of bootstrap resamples, pp. 354–357 (cit. on p. 5-10).


... ROC curves lack or obscure several quantities that are necessary for evaluating the operational effectiveness of diagnostic tests. ... ROC curves were first used to check how radio receivers (like radar receivers) operated over a range of frequencies. ... This is not how most ROC curves are used now, particularly in medicine. The receiver of a diagnostic measurement ... wants to make a decision based on some $x_c$, and is not especially interested in how well he would have done had he used some different cutoff."

In the discussion David Hand states "when integrating to yield the overall AUC measure, it is necessary to decide what weight to give each value in the integration. The AUC implicitly does this using a weighting derived empirically from the data. This is nonsensical. The relative importance of misclassifying a case as a noncase, compared to the reverse, cannot come from the data itself. It must come externally, from considerations of the severity one attaches to the different kinds of misclassifications."; see Lin, Kvam, Lu Stat Med 28:798-813;2009, pp. 250–261 (cit. on p. 1-7).


C. Chatfield. “Model Uncertainty, Data Mining and Statistical Inference (with Discussion)”. In: J Roy Stat Soc A 158 (1995). bias by selecting model because it fits the data well; bias in standard errors;P. 420: ... need for a better balance in the literature and in statistical teaching between techniques and problem solving strategies. P. 421: It is ‘well known’ to be ‘logically unsound and practically misleading’ (Zhang, 1992) to make inferences as if a model is known to be true when it has, in fact, been selected from the same data to be used for estimation purposes. However, although statisticians may admit this privately (Breiman (1992) calls it a ‘quiet scandal’), they (we) continue to ignore the difficulties because it is not clear what else could or should be done. P. 421: Estimation errors for regression coefficients are usually smaller than errors from failing to take into account model specification. P. 422: Statisticians must stop pretending that model uncertainty does not exist and begin to find ways of coping with it. P. 426: It is indeed strange that we often admit model uncertainty by searching for a best model but then ignore this uncertainty by making inferences and predictions as if certain that the best fitting model is actually true. P. 427: The analyst needs to assess the model selection process and not just the best fitting model. P. 432: The use of subset selection methods is well known to introduce alarming biases. P. 433: ... the AIC can be highly biased in data-driven model selection situations. P. 434: Prediction intervals will generally be too narrow. In the discussion, Jamal R. M. Ameen states that a model should be (a) satisfactory in performance relative to the stated objective, (b) logically sound, (c) representative, (d) questionable and subject to on-line interrogation, (e) able to accommodate external or expert information and (f) able to convey information., pp. 419–466 (cit. on pp. 4-10, 5-22).


[48] Nancy R. Cook. “Use and Misues of the Receiver Operating Characteristic Curve in Risk Prediction”. In: *Circ* 115 (2007). example of large change in predicted risk in cardiovascular disease with tiny change in ROC area;possible limits to c index when calibration is perfect;importance of calibration accuracy and changes in predicted risk when new variables are added, pp. 928–935 (cit. on p. 10-38).


[73] Valerii Fedorov, Frank Mannino, and Rongmei Zhang. “Consequences of Dichotomization”. In: *Pharm Stat* 8 (2009), optimal cutpoint depends on unknown parameters; should only entertain dichotomization when “estimating a value of the cumulative distribution and when the assumed model is very different from the true model”, nice graphics, pp. 50–61. doi: 10.1002/pst.331. url: http://dx.doi.org/10.1002/pst.331 (cit. on pp. 1-6, 2-13).


Practical Longitudinal Data Analysis


[83] Usha S. Govindarajulu et al. “Comparing Smoothing Techniques in Cox Models for Exposure-Response Relationships”. In: *Stat Med* 26 (2007). authors wrote a SAS macro for restricted cubic splines even though such a macro as existed since 1984; would have gotten more useful results had simulation been used so would know the true regression shape; measure of agreement of two estimated curves by computing the area between them, standardized by average of areas under the two; penalized spline and rcs were closer to each other than to fractional polynomials, pp. 3735–3752 (cit. on p. 2-29).


Amendment and corrections in 82: 668 (1995)


[89] Sander Greenland. “When Should Epidemiologic Regressions Use Random Coefficients?” In: *Biometrics* 56 (2000). use of statistics in epidemiology is largely primitive; stepwise variable selection on confounders leaves important confounders uncontrolled; composition matrix; example with far too many significant predictors with many regression coefficients absurdly inflated when overfit; lack of evidence for dietary effects mediated through constituents; shrinkage instead of variable selection; larger effect on confidence interval width than on point estimates with variable selection; uncertainty about variance of random effects is just uncertainty about prior opinion; estimation of variance is pointless; instead the analysis should be repeated using different values; “if one feels compelled to estimate $\tau$^2$, I would recommend giving it a proper prior concentrated amount contextually reasonable values; claim about ordinary MLE being unbiased is misleading because it assumes the model is correct and is the only model entertained; shrinkage towards compositional model; ‘models need to be complex to capture uncertainty about the relations... an honest uncertainty assessment requires parameters for all effects that we know may be present. This advice is implicit in an antiparsimony principle often attributed to L. J. Savage ‘All models should be as big as an elephant (see Draper, 1995)’”. See also gus06per., pp. 915–921. doi: 10.1111/j.0006-341X.2000.00915.x. url: http://dx.doi.org/10.1111/j.0006-341X.2000.00915.x (cit. on pp. 4-10, 4-44).

[90] Jian Guo et al. “Principal Component Analysis with Sparse Fused Loadings”. In: *J Comp Graph Stat* 19.4 (2011). incorporates blocking structure in the variables; selects different variables for different components; encourages loadings of highly correlated variables to have same magnitude, which aids in interpretation, pp. 930–946 (cit. on p. 4-29).


[92] Ofer Harel and Xiao-Hua Zhou. “Multiple Imputation: Review of Theory, Implementation and Software”. In: *Stat Med* 26 (2007). failed to review aregImpute; excellent overview; ugly S code; nice description of different statistical tests including combining likelihood ratio tests (which appears to be complex, requiring an out-of-sample log likelihood computation); congeniality of imputation and analysis models; Bayesian approximation or approximate Bayesian bootstrap overview; “Although missing at random (MAR) is a non-testable assumption, it has been pointed out in the literature that we can get very close to MAR if we include enough variables in the imputation models... it would be preferred if the missing data modelling was done by the data constructors and not by the users... MI yields valid inferences not only in congenial settings, but also in certain un congenial ones as well—where the imputer’s model (1) is more general (i.e. makes fewer assumptions) than the complete-data estimation method, or...
when the imputer’s model makes additional assumptions that are well-founded.”, pp. 3057–3077 (cit. on pp. 3-1, 3-8, 3-12, 3-15).


ISBN 9780412343902.


[106] Norbert Holländer, Willi Sauerbrei, and Martin Schumacher. “Confidence Intervals for the Effect of a Prognostic Factor after Selection of an ‘optimal’ Cutpoint”. In: *Stat Med* 23 (2004), true type I error can be much greater than nominal level; one example where nominal is 0.05 and true is 0.5; minimum P-value method; CART; recursive partitioning; bootstrap method for correcting confidence interval; based on heuristic shrinkage coefficient; “It should be noted, however, that the optimal cutpoint approach has disadvantages. One of these is that in almost every study where this method is applied, another cutpoint will emerge. This makes comparisons across studies extremely difficult or even impossible. Altman et al. point out this problem for studies of the prognostic relevance of the S-phase fraction in breast cancer published in the literature. They identified 19 different cutpoints used in the literature; some of them were solely used because they emerged as the ‘optimal’ cutpoint in a specific data set. In a meta-analysis on the relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients, 12 studies were included with 12 different cutpoints ... Interestingly, neither cathepsin-D nor the S-phase fraction are recommended to be used as prognostic markers in breast cancer in the recent update of the American Society of Clinical Oncology.”; dichotomization; categorizing continuous variables; refs alt94dan, sch94out, alt98sub, pp. 1701–1713. doi: [10.1002/sim.1611](http://dx.doi.org/10.1002/sim.1611) url: [http://dx.doi.org/10.1002/sim.1611](http://dx.doi.org/10.1002/sim.1611) (cit. on pp. 2-13, 2-15).


R package version 4.38.


[130] Katherine J. Lee and John B. Carlin. “Recovery of Information from Multiple Imputation: A Simulation Study”. In: Emerg Themes Epi 9.1 (June 2012). Not sure that the authors satisfactorily dealt with nonlinear predictor effects in the absence of strong auxiliary information, there is little to gain from multiple imputation with missing data in the exposure-of-interest. In fact, the authors went further to say that multiple imputation can introduce bias not present in a complete case analysis if a poorly fitting imputation model is used [from Yong Hao Pua], pp. 3+. issn: 1742-7622. doi: 10.1186/1742-7622-9-3. pmid: 22695083. url: http://dx.doi.org/10.1186/1742-7622-9-3 (cit. on p. 3-4).


[134] Kung-Yee Liang and Scott L. Zeger. “Longitudinal Data Analysis of Continuous and Discrete Responses for Pre-Post Designs”. In: Sankhyâ 62 (2000). makes an error in assuming the baseline variable will have the same univariate distribution as the response except for a shift; baseline may have for example a truncated distribution based on a trial’s inclusion criteria; if correlation between baseline and response is zero, ANCOVA will be twice as efficient as simple analysis of change scores; if correlation is one they may be equally efficient, pp. 134–148 (cit. on p. 7-5).


[159] Michael J. Pencina, Ralph B. D’Agostino, and Ewout W. Steyerberg. "Extensions of Net Reclassification Improvement Calculations to Measure Usefulness of New Biomarkers". In: Stat Med 30 (2011). lack of need for NRI to be category-based; arbitrariness of categories; category-less or continuous NRI is the most objective and versatile measure of improvement in risk prediction; authors misunderstood the inadequacy of three categories if categories are used; comparison of NRI to change in C index; example of continuous plot of risk for old model vs. risk for new model, pp. 11–21. doi: 10.1002/sim.4085. url: http://dx.doi.org/10.1002/sim.4085 (cit. on p. 4-43).


[173] Peter Radchenko and Gareth M. James. “Variable Inclusion and Shrinkage Algorithms”. In: *J Am Stat Assoc* 103.483 (2008). solves problem caused by lasso using the same penalty parameter for variable selection and shrinkage which causes lasso to have to keep too many variables in the model to avoid overshrinking the remaining predictors; does not handle scaling issue well, pp. 1304–1315 (cit. on p. 2-36).


[187] Stephen Senn. “Change from Baseline and Analysis of Covariance Revisited”. In: Stat Med 25 (2006). shows that claims that in a 2-arm study it is not true that ANCOVA requires the population means at baseline to be identical;refutes some claims of lia00lon;problems with counterfactuals;temporal additivity ("amounts to supposing that despite the fact that groups are difference at baseline they would show the same evolution over time");causal additivity;is difficult to design trials for which simple analysis of change scores is unbiased; ANCOVA is biased, and a causal interpretation can be given;temporally and logically, a "baseline cannot be a <i>response</i> to treatment", so baseline and response cannot be modeled in an integrated framework as Laird and Ware’s model has been used;’one should focus clearly on ‘outcomes’ as being the only values that can be influenced by treatment and examine critically any schemes that assume that these are linked in some rigid and deterministic view to ‘baseline’ values. An alternative tradition sees a baseline as being merely one of a number of measurements capable of improving predictions of outcomes and models it in this way.";”You cannot establish necessary conditions for an estimator to be valid by nominating a model and seeing what the model implies unless the model is universally agreed to be impeccable. On the contrary it is appropriate to start with the estimator and see what assumptions are implied by valid conclusions.”;this is in distinction to lia00lon, pp. 4334–4344 (cit. on p. 7-5).


[218] Andrew J. Vickers. “Decision Analysis for the Evaluation of Diagnostic Tests, Prediction Models, and Molecular Markers”. In: Am Statistician 62.4 (2008). limitations of accuracy metrics;incorporating clinical consequences;nice example of calculation of expected outcome;drawbacks of conventional decision analysis, especially because of the difficulty of eliciting the expected harm of a missed diagnosis;use of a threshold on the probability of disease for taking some action;decision curve;has other good references to decision analysis, pp. 314–320 (cit. on p. 1-7).


[220] Eric Vittinghoff and Charles E. McCulloch. “Relaxing the Rule of Ten Events per Variable in Logistic and Cox Regression”. In: Am J Epi 165 (2006). the authors may have not been quite stringent enough in their assessment of adequacy of predictions;letter to the editor submitted, pp. 710–718 (cit. on p. 4-19).


[230] Ian R. White and Patrick Royston. “Imputing Missing Covariate Values for the Cox Model”. In: Stat Med 28 (2009), approach to using event time and censoring indicator as predictors in the imputation model for missing baseline covariates;recommended an approximation using the event indicator and the cumulative hazard transformation of time, without their interaction, pp. 1982–1998 (cit. on p. 3-3).
[231] Ian R. White, Patrick Royston, and Angela M. Wood. “Multiple Imputation Using Chained Equations: Issues and Guidance for Practice”. In: Stat Med 30.4 (2011). practical guidance for the use of multiple imputation using chained equations;MICE;imputation models for different types of target variables;PMM choosing at random from among a few closest matches;choosing number of multiple imputations by a reproducibility argument, suggesting 100f imputations when f is the fraction of cases that are incomplete, pp. 377–399 (cit. on pp. 3-1, 3-10, 3-15, 3-19).


See letter to editor SM 15:1065-6 for binary case; see errata in SM 13:871 1994; see kol95com, jul96sam .


[234] Daniela M. Witten and Robert Tibshirani. “Testing Significance of Features by Lassoed Principal Components”. In: Ann Appl Stat 2.3 (2008). reduction in false discovery rates over using a vector of t-statistics;borrowing strength across genes;one would not expect a single gene to be associated with the outcome, since, in practice, many genes work together to effect a particular phenotype. LPC effectively down-weights individual genes that are associated with the outcome but that do not share an expression pattern with a larger group of genes, and instead favors large groups of genes that appear to be differentially-expressed.”;regress principal components on outcome;sparse principal components, pp. 986–1012 (cit. on p. 2-37).


[237] Shifeng Xiong. “Some Notes on the Nonnegative Garroote”. In: Technometrics 52.3 (2010). “… to select tuning parameters, it may be unnecessary to optimize a model selectin criterion repeatedly”;natural selection of penalty function, pp. 349–361 (cit. on p. 2-37).


[240] Recai M. Yucel and Alan M. Zaslavsky. “Using Calibration to Improve Rounding in Imputation”. In: Am Statistician 62.2 (2008). using rounding to impute binary variables using techniques for continuous data;uses the method to solve for the cutpoint for a continuous estimate to be converted into a binary value;method should be useful in more general situations;idea is to duplicate the entire dataset and in the second half of the new datasets to set all non-missing values of the target variable to missing;multiply impute these now-missing values and compare them to the actual values, pp. 125–129 (cit. on p. 3-17).


R packages written by FE Harrell are freely available from CRAN, and are managed at github.com/harrelfe.

To obtain a book with detailed examples and case studies and notes on the theory and applications of survival analysis, logistic regression, ordinal regression, linear models, and longitudinal models, order Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd Edition by FE Harrell from Springer NY (2015). Steyerberg [195] and Dupont [63] are excellent texts for accompanying the book.

To obtain a glossary of statistical terms and other handouts related to diagnostic and prognostic modeling, see hbiostat.org/doc/glossary.pdf