Dilemmas in Regression Modeling

Frank E Harrell Jr
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine
Box 800717 Charlottesville VA 22908 USA
fharrell@virginia.edu
hesweb1.med.virginia.edu/biostat

STATISTICS / BIOSTATISTICS COLLOQUIUM
DEPARTMENT OF STATISTICS
UNIVERSITY OF VIRGINIA
5 NOVEMBER 1999

Outline

1. A modeling strategy yielding “optimum” predictions
2. Problems with seeking parsimonious models using variable selection
3. Problems using “full” models
4. Model approximation to achieve parsimony
5. Simulation study to compare accuracy of various strategies
6. What should be our default strategy?

A Modeling Strategy

- Assume *structure* of model is pre–specified
- Assume no missing data
- Decide how many degrees of freedom you can or want to spend
  - Number of categories for categorical predictors
  - Number of nonlinear components for continuous ones
  - Interaction terms
  - Number of predictors
- Estimate that many parameters (plus necessary intercept(s))
- Use penalized estimation if information in data will not support reliable estimation of that many parameters

Variable Selection

- Drop “insignificant” predictors / nonlinear terms / interactions
- Results in parsimonious models
- **BUT** . . .
  - Low probability of finding the “right” variables
  - Problems caused by collinearity
  - Predictions are not optimal due to deletion of marginally insignificant but important predictors
  - Estimate of σ^2 biased low
  - Estimate of β biased high in absolute value
  - Standard errors biased low
  - Test statistics don’t have claimed distribution
  - P–values invalid
  - Confidence limits for effects & predictions falsely narrow
  - R^2 and adjusted R^2 biased high

Full Models

- Yield correct estimates with correct statistical properties
- Optimum prediction
- Collinearity does not cause problems in most cases
  - need to test related predictors as a cluster
- **BUT** . . .
  - Some predictors affect predictions very little
  - Predictive instruments are cluttered
  - Predictions are normally conditional on *all* predictors
  - Variance of conditional prediction $>$ unconditional prediction

Example

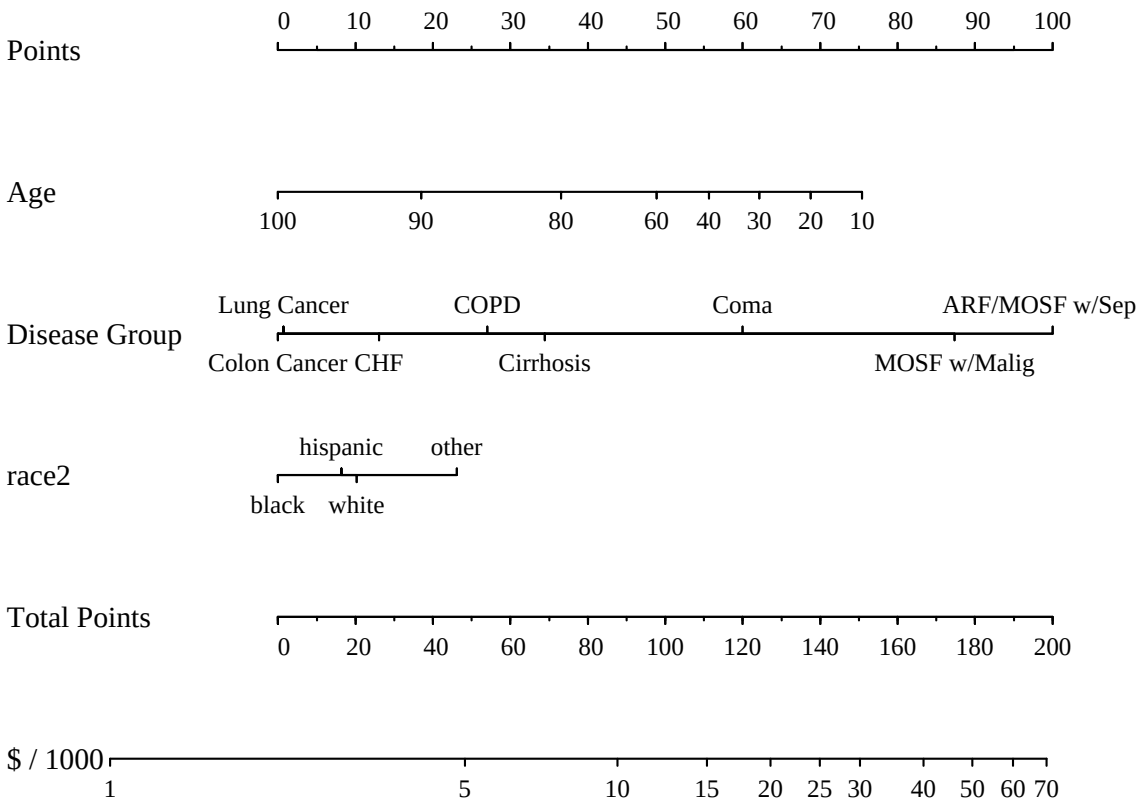
- Random sample of 1000 patients from SUPPORT study (biostat Web page)
- OLS used to predict log of hospital costs for those patients having nonzero charges
- S-PLUS command (restricted cubic spline in `age` with 4 knots; `dzgroup` has 8 levels; `race2` has 4)

```
f ← ols(log(totcst) ~ rcs(age,4) +
        dzgroup + race2)
anova(f)
```

Table 1: *Analysis of Variance for* `log(totcst)`

	$d.f.$	$PartialSS$	MS	F	P
age	3	23.62559	7.875198	9.24	< 0.0001
Nonlinear	2	6.11631	3.058155	3.59	0.0281
dzgroup	7	443.3492	63.3356	74.30	< 0.0001
race2	3	4.547531	1.515844	1.78	0.1497
TOTAL	13	513.0274	39.46365	46.30	< 0.0001
ERROR	875	745.865	0.8524171		

```
nomogram(f, age=c(10,20,30,40,60,80,90,100),
         lp=F,
         fun=list('$ / 1000'=
          function(y)exp(y)/1000),
         fun.lp.at=log(1000*c(1,5,10,15,20,
                              25,30,40,50,60,70)))
```


Points
0 10 20 30 40 50 60 70 80 90 100
Age
100 90 80 60 40 30 20 10
Disease Group
Lung Cancer
COPD
Coma
ARF/MOSF w/Sep
Colon Cancer
CHF
Cirrhosis
MOSF w/Malig
race2
hispanic
other
black
white
Total Points
0 20 40 60 80 100 120 140 160 180 200
$ / 1000
1 5 10 15 20 25 30 40 50 60 70

What to do about Race?

- Frequencies: black 129, hispanic 31, white 710, other 19
- Predicted log cost for 50 year old in a coma:

```
p ← predict(f, expand.grid(age=50,
                           dzgroup='Coma',
                           race2=levels(race2)),
                se.fit=T)
contrast(f, list(age=50, dzgroup='Coma',
                         race2=levels(race2)),
         type='average')
contrast(f, list(age=50, dzgroup='Coma',
                         race2=levels(race2)),
         type='average',
         weights=table(race2))
g ← ols(log(totcst) ~ rcs(age,4) +
         dzgroup)
predict(g,
        data.frame(age=50, dzgroup='Coma'),
        se.fit=T)
```

Race	$\hat{Y}$	S.E.
Other	10.01	0.25
White	9.79	0.14
Black	9.61	0.15
Hispanic	9.75	0.22
Unweighted avg.	9.79	0.15
Weighted avg.	9.77	0.14
Omit	9.76	0.14

- Predict $\hat{Y}$ from all predictors [2]
- $R^2 = 1$
- Simplify (omit predictors, simplify complex components)
- Different degrees of simplicity for different consumers
- Delete `race2` $\rightarrow R^2 = 0.99$, 0.95 quantile of $|$approximation error$| = 0.16$
- If also delete nonlinear age terms, $R^2 = 0.98$, 0.95 quantile of error = 0.22
- Use the model

  $\sim$ `rcs(age,4) + dzgroup`

Problems and Advantages

- Problems
  - What to use for standard errors, CLs for $\hat{\beta}_j$?
  - Parameter estimates of this approximate model = those from re–fit against actual Y in the case of OLS

    If T is a subset of columns of X,
    $(T'T)^{-1}T'X(X'X)^{-1}X'Y =$
    $(T'T)^{-1}T'Y$
    $T'X(X'X)^{-1}X'Y = T'Y$

- Advantages
  - Will select different variables than stepwise regression against response variable
  - Easy to use OLS approximations to $X\hat{\beta}$ from non–OLS models
  - If original model used shrinkage, approximate model will inherit the shrinkage

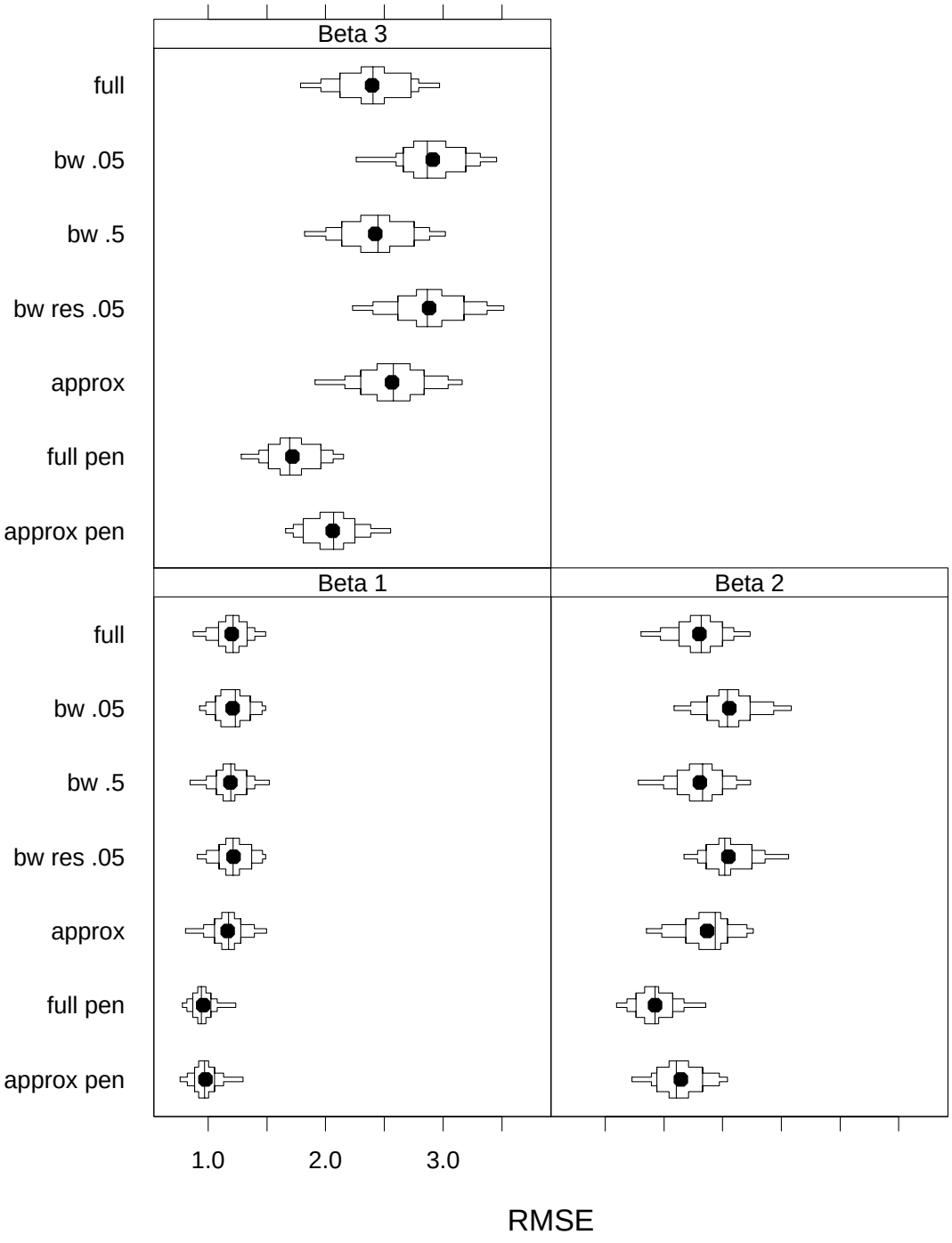
- Population model is additive, linear, i.i.d. Gaussian residuals
- $n = 500$, 20 predictors $N(0, 1)$ all with correlation $\rho = 0.2$
- Accuracy measure: $[\frac{1}{500} \sum_{i=1}^{500} (X_i\beta - X_i\hat{\beta})]^{\frac{1}{2}}$
- Three X matrices, 25 simulations of Y for each one
- Three sets of population β and σ
  1. $10\beta^1 = [1 - 10, 5 \times 10, 5 \times 0]$
  2. $10\beta^2 = [1 - 13, 7 \times 1.5]$
  3. $10\beta^3 = [1 - 13, 7 \times 3.5]$

  $\sigma = 6, 9, 12$, respectively

1. Full model fit
2. Backward stepdown, $\alpha = 0.05$
3. Backward stepdown, $\alpha = 0.5$
4. Backward stepdown using global test, $\alpha = 0.05$
5. Full model approximation using $R^2 = 0.95$
6. Full penalized model, penalty chosen using effective AIC
7. Full penalized model approximation using $R^2 = 0.95$

Results (Average of 75 RMSEs)

Strategy	β^1	β^2	β^3
Full	1.20	1.80	2.39
BW 0.05	1.21	2.06	2.91
BW 0.5	1.19	1.80	2.42
BW res	1.22	2.05	2.88
Approx	1.17	1.87	2.57
Full pen.	0.96	1.42	1.72
Approx pen.	0.98	1.64	2.06


Beta 3
full
bw .05
bw .5
bw res .05
approx
full pen
approx pen
Beta 1
Beta 2
full
bw .05
bw .5
bw res .05
approx
full pen
approx pen
1.0
2.0
3.0
RMSE

What Default Strategy?

- Full model
- Backward step–down with large α
- Penalization is clearly advantageous
- Model approximation to achieve parsimony
- Tibshirani *lasso* [3]: MLE with penalty for large absolute values of β_j

  Variable selection with shrinkage
- Breiman nonnegative garrote [1]

  Solve for optimum penalties to coefficients estimated by ordinary MLE
- Many other choices, e.g. MARS, projection pursuit, neural net

References

[1] Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37:373–384, 1995.

[2] Frank E. Harrell, Peter A. Margolis, Sandy Gove, Karen E. Mason, E. Kim Mulholland, Deborah Lehmann, Lulu Muhe, Salvacion Gatchalian, and Heinz F. Eichenwald. Development of a clinical prediction model for an ordinal outcome: The World Health Organization ARI Multicentre Study of clinical signs and etiologic agents of pneumonia, sepsis, and meningitis in young infants. *Statistics in Medicine*, 17:909–944, 1998.

[3] Robert Tibshirani. The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16:385–395, 1997.

Abstract

Regression modeling comprises many steps including model selection, variable selection, estimation of variable transformations, diagnostics, and handling missing data. When the model and its variables are non pre-specified, searching for models and for 'important' variables lead to biased parameter estimates and invalid statistical tests. On the other hand, fitting a model containing all pre-specified predictors leads to complex models and difficulties in obtaining predicted values that are not conditional on all of the predictors. Model approximation, based on representing the predicted values more simply, can help, but its properties are not fully understood. This talk will discuss such current dilemmas in regression modeling.