Regression Analysis of Survival in Randomized Clinical Trials

Frank E. Harrell Jr
Division of Biometry and The Heart Center
Duke University Medical Center
Box 3363 Durham NC 27710
feh@biostat.mc.duke.edu

Society for Clinical Trials

10 May 1992

Hypothetical Example
Trial of Coronary Bypass Surgery

Number of Diseased Vessels	n	Deaths	CABG:Medical Hazard Ratio	0.95 Confidence Limits	P
1	900	40	0.97	[0.57, 1.32]	0.521
2	600	70	0.80	[0.60, 1.01]	0.051
3	700	130	0.70	[0.55, 0.88]	0.012
Overall	2200	240	0.85	[0.59, 1.02]	0.055

Test	χ^2	$d.f.$	P
Standard test for treatment effect	3.7	1	0.055
Test for interaction	12.5	2	0.002
Test for treatment effect allowing interaction	14.2	3	0.003

Reasons for Survival Analysis in Randomized Clinical Trials

1. Test for and describe interactions with treatment
   (a) Model relative benefit (e.g., hazard ratio)
   (b) Test for and describe interactions with treatment (subgroup analyses can easily generate bogus results and they do not consider interacting factors in a dose–response manner)
   (c) Find out 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
   (a) Develop a model for survival probability
   (b) Compute differences in survival for treatments A and B
   (c) Differences will be due primarily to sickness (overall risk) of the patient and to

treatment interactions

4. Understand time course of treatment effect — period of maximum effect or period of any substantial effect
5. Gain power for testing treatment effect
6. Adjust for imbalances

Hazard and Survival Functions

$$
\begin{aligned}
T &= \text{ time until an event} \\
S(t) &= \text{Prob}\{T > t\} = 1 - F(t) \\
\lambda(t) &= \lim_{u \to 0} \frac{\text{Prob}\{t < T \leq t + u | T > t\}}{u} \\
&= \text{ instantaneous event rate}
\end{aligned}
$$

Assumptions	Model or Method
None	Kaplan–Meier Estimator of $S(t)$
$\lambda(t) =$ constant $S(t) = \exp(-\lambda t)$	Exponential Model (person–years)
$\lambda(t) = \alpha\gamma t^{\gamma-1}$ $S(t) = \exp(-\alpha t^{\gamma})$	Weibull Model

The Proportional Hazards Family

Predictor variables:

$$X = \{X_1, X_2, \ldots, X_p\}$$

X_i binary, ordinal (if linear), continuous. Regression effect:

$$X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

Assume the effect of predictors is to multiply the underlying hazard of the event:

$$\lambda(t|X) = \lambda(t)\exp(X\beta)$$

The effect on $S(t)$ is to raise it to a power:

$$S(t|X) = S(t)^{\exp(X\beta)}$$

Note absence of interaction between t and $X \Rightarrow$ effect of X is the same at all follow-up times $t \equiv$ *proportional hazards (PH)* assumption.

Regression coefficient for X_j, β_j, = increase in log hazard or log cumulative hazard at time t if X_j is increased by one unit and all other predictors are held constant and no predictors interact with X_j.

Example: $\beta_1 = 0.5$, raise X_1 from 0 to 1 increases log hazard by 0.5
Increases hazard by a factor of $\exp(0.5) = 1.65$ for all t
Worsens survival from $S(t)$ to $S(t)^{1.65}$
Example: $S(2\text{y}|X_1 = 0) = 0.8$,

$S(2\text{y}|X_1 = 1) = 0.69.$

Weibull PH Model:

$$\lambda(t|X) = \alpha\gamma t^{\gamma-1}\exp(X\beta)$$

Cox PH model [3] ($\lambda(t)$ unspecified):

$$\lambda(t|X) = \lambda(t)\exp(X\beta)$$

The Cox model contains the log–rank test as a special case (from the score test of the Cox regression coefficient for treatment). The log–rank test and Cox model test have the same assumptions.

Examining Assumptions of PH Model

1. Shape of $\lambda(t)$ if parametric, e.g. Weibull:
   Within "homogeneous" patient subsets plot $\log[-\log(S_{\mathrm{KM}}(t)$ vs. $\log t$ — should be a straight line.

2. PH Assumption:

   (a) Check parallelism of $\log[-\log(S_{\mathrm{KM}}(t))]$ estimates over t for different X.

   (b) Hazard ratio plots
   Use Cox model to estimate effects of X in an interval of time, e.g. effect of unstable angina vs. stable angina in months 6–12 after diagnosis.
   Can be used to estimate effect of treatment on instantaneous hazard rate for varying t.

   (c) Plots of smoothed Schoenfeld residuals [18] vs. t.

3. Effect of X on log hazard at a specific t.

$$\log \lambda(t|X) = \log \lambda(t) + X\beta$$

   Verify linearity in X or extend to allow non–linearity by generalizing the model:

$$\log \lambda(t|X) = \log \lambda(t) + f(X)$$

   Some choices of $f(\cdot)$:

   (a) polynomial $(\beta_1 X + \beta_2 X^2)$

   (b) piecewise cubic polynomials (splines) [11, 5, 19].
   Spline functions are more flexible and robust

and like polynomials contain the linear model as a special case ⇒ formal test of linearity.

(c) Nonparametric: smoothed martingale residual plot [21].

4. Joint Effects of Several Predictors

$$\log \lambda(t|X) = \log \lambda(t) + f(X_1, X_2)$$

$f(\cdot)$ may often be modeled with main effects and cross–products of variables making up each factor. Test of cross–product terms ⇒ test of additivity (no interaction).

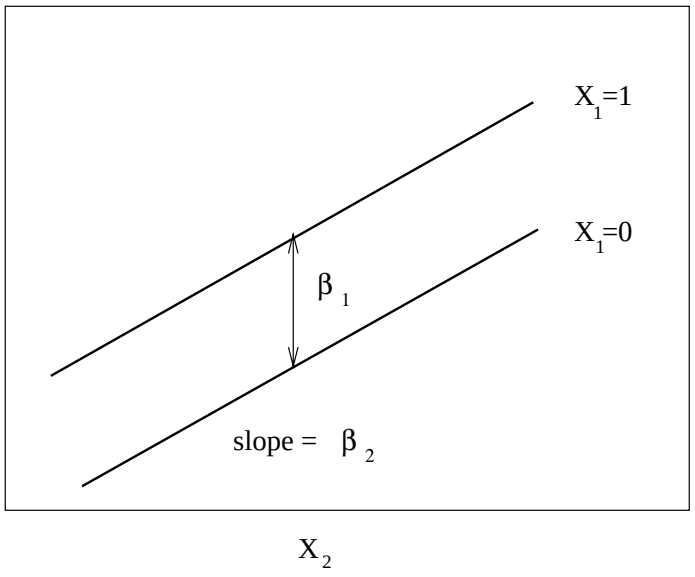

Figure 1: Regression assumptions, linear additive PH model with two predictors. Y–axis is $\log \lambda(t)$ or $\log \Lambda(t)$ for a fixed t.

Assumptions of the Proportional Hazards Model

$$\lambda(t|X) = \lambda(t)e^{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}$$

Variables	Assumptions	Verification
Response Variable T Time Until Event	Shape of $\lambda(t\|X)$ for fixed X as $t \uparrow$ Cox: none Weibull: t^θ	Shape of $S_{\mathrm{KM}}(t)$
Interaction between X and T	Proportional hazards — effect of X does not depend on T. E.g. treatment effect is constant over time.	Binary X: check parallelism of stratified $\log[-\log S_{\mathrm{KM}}(t)]$ plots as $t \uparrow$ Continuous X: estimate β as a function of t — hazard ratio plots (time interval–specific hazard ratios); smoothed Schoenfeld residual plots
Individual Predictors X	Shape of $\lambda(t\|X)$ for fixed t as $X \uparrow$ Linear: $\log\lambda(t\|X) = \log\lambda(t) + \beta X$ Nonlinear: $\log\lambda(t\|X) = \log\lambda(t) + f(X)$	k–level ordinal X : linear term + $k-2$ dummy variables Continuous X: Polynomials, spline functions, smoothed martingale residual plots
Interaction between X_1 and X_2	Additive effects: effect of X_1 on $\log\lambda$ is independent of X_2 and vice–versa	Test non–additive terms, e.g. products

Overfitting and Data Reduction

Model will fail to validate (predictions will be inaccurate) on a new sample and will give undue influence to quirks in the data if number of *candidate* predictor degrees of freedom (p) $> d/15$ where d is the number of deaths (events) [10].

Univariable screening of candidates **in no way** gets around this problem.

Variable Selection

May yield stable (replicable) model if $p < d/15$.
Clinical intuition is usually better.
Variable selection has these drawbacks:

1. The list of variables selected is highly affected by variances of predictors and by strong correlations among predictors and by quirks in the data.
2. Variable selection results in over–optimistic model fits and inflated estimates of β (regression to the mean).
3. Traditional confidence intervals derived from the reduced model are invalid [1].
4. Variable selection often trades predictive accuracy for parsimony [20].
5. Stopping rules are sometimes arbitrary.
6. Selection assumes that some variables have $\beta = 0$, even clinically relevant ones.

Data Reduction

Can solve many of the problems of variable selection and overfitting [10]. Possible methods:

1. Clinical intuition — derivation of clinical summary scores.
2. Principal components analysis [12].
3. Qualitative principal components [22].
4. Correspondence analysis [4].
5. Variable clustering [17].

Model Validation

Verify that the model discriminates outcomes as well as you think it does.
If the model is used to predict probabilities, need to validate that the model is calibrated over the entire risk range.
Model will almost always *seem* to fit when tested on same data used to fit it.

Validation Methods:

1. Data–splitting [15]
2. Jackknifing (leave out 1)
3. Cross–validation (leave out m)
4. Bootstrap [6, 8, 7, 9]

For any of the methods you must be sure to replicate all steps in model building (i.e., preliminary hypothesis tests, variable selection).

Bootstrapping, which uses 50–200 samples of size n with replacement from the original sample, has the following advantages:

1. It does not require holding back data, so $n \uparrow$.
2. It evaluates in a nearly unbiased way the accuracy of the model developed on the entire sample of size n.
3. It is more precise, i.e. requires fewer re–samples to estimate an index of model fit with a given accuracy.
4. It validates and graphically depicts variability in the variable selection process.

Example of bootstrapping to estimate over–optimism:

Method	Apparent Rank Correlation Of Predicted vs. Observed	Over–Optimism	Bias–Corrected Correlation
Full Model	0.50	0.06	0.44
Stepwise Model	0.47	0.05	0.42

Estimating Absolute Clinical Benefit

1. Estimate $S(t|X)$ [13].
2. Let X_1 be the (binary) treatment variable and $A = \{X_2, \ldots, X_p\}$ be adjustment variables.
3. Over the distribution of A, plot $S(t|X_1 = 1, A) - S(t|X_1 = 0, A)$ vs. the survival in the untreated patients, $S(t|X_1 = 0, A)$ and also vs. any factors interacting with X_1.

Examples

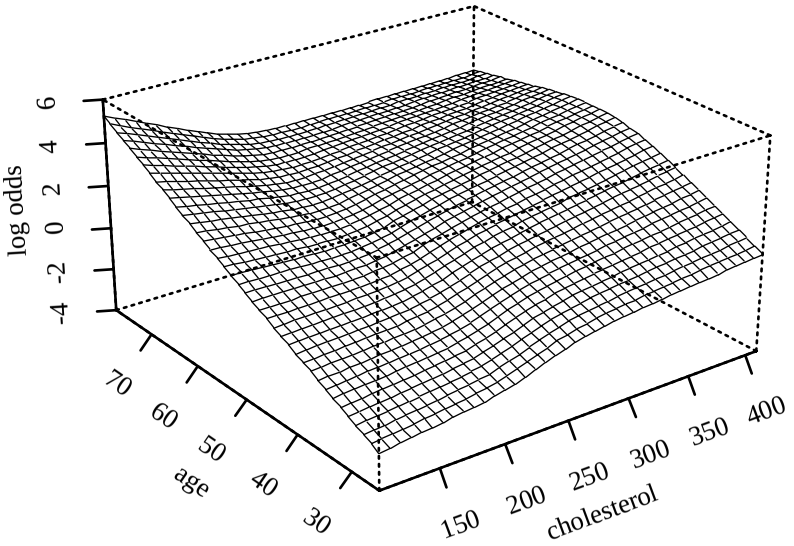

Figure 2: Restricted cubic spline fit with age × spline(cholesterol) and cholesterol × spline(age)

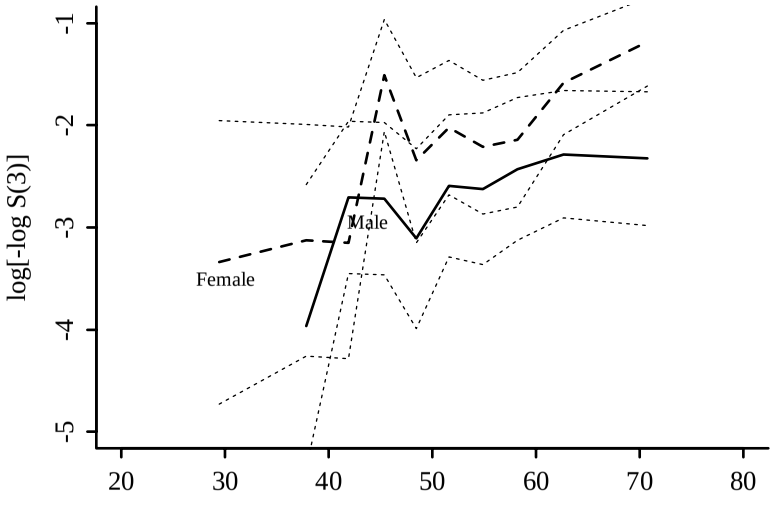

Figure 3: Kaplan–Meier log cumulative hazard estimates by sex and deciles of age

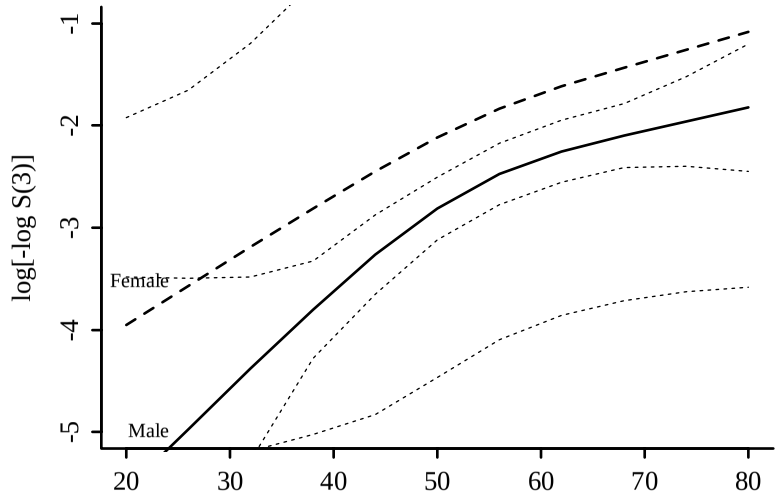

Figure 4: Cox PH model stratified on sex, with interaction between age spline and sex.

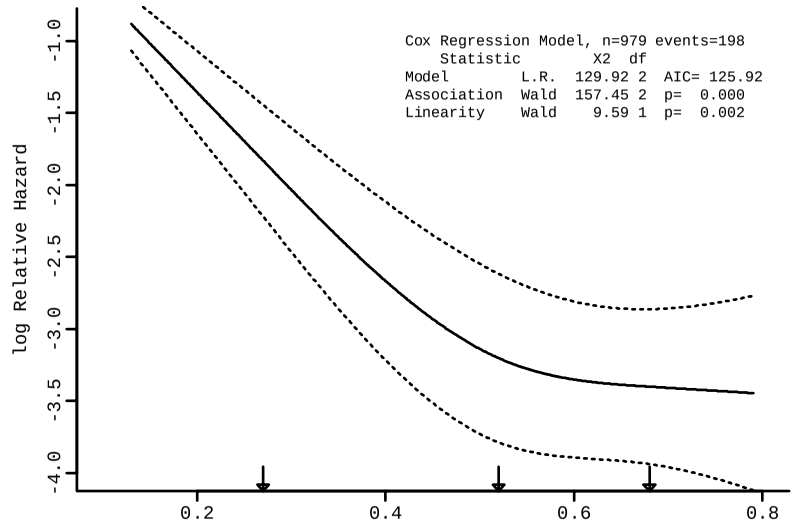

Figure 5: Restricted cubic spline estimate of relationship between LVEF and relative log hazard from a sample of 979 patients and 198 cardiovascular deaths. Data from the Duke Cardiovascular Disease Databank.

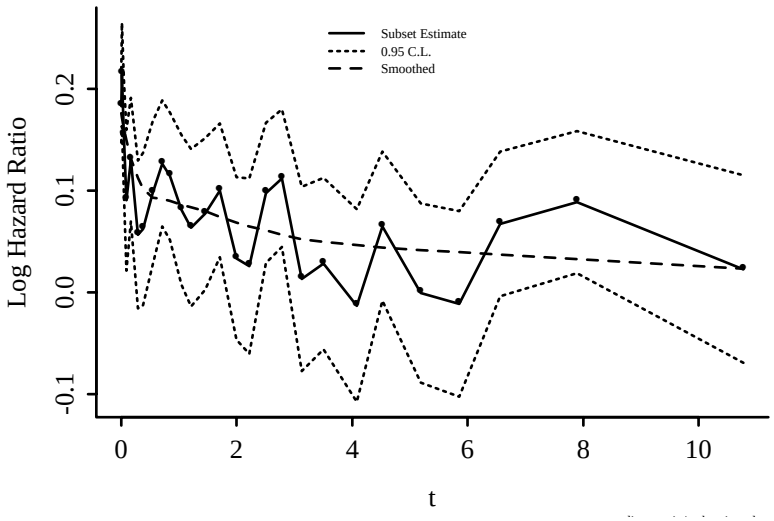

Figure 6: Stratified hazard ratios for pain/ischemia index over time. Data from the Duke Cardiovascular Disease Databank.

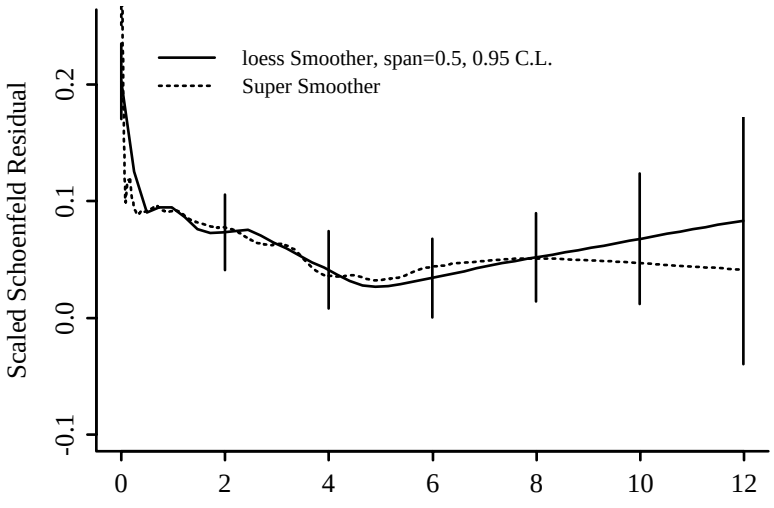

Figure 7: Smoothed Schoenfeld [18] residuals for the same data in Figure 6

Figure 8: Calibration of random predictions using Efron's bootstrap with B=40 resamples and 20 patients per interval. Dataset has n=200, 100 uncensored observations, 20 random predictors. •: apparent calibration; X: bias–corrected calibration.

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index
D_{yx}	-0.16	-0.33	-0.11	-0.22	0.05
Slope	1.00	1.00	0.24	0.76	0.24

Figure 9: A display of an interaction between treatment, extent of disease, and calendar year of start of treatment [2]

Figure 10: Cox–Kalbfleisch–Prentice survival estimates stratifying on treatment and adjusting for several predictors [16]

Figure 11: Cox model predictions with respect to a continuous variable [14]

References

[1] D. G. Altman and P. K. Andersen. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, 8:771–783, 1989.

[2] R. M. Califf, F. E. Harrell, K. L. Lee, J. S. Rankin, et al. The evolution of medical and surgical therapy for coronary artery disease. *Journal of the American Medical Association*, 261:2077–2086, 1989.

[3] D. R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, 34:187–220, 1972.

[4] N. J. Crichton and J. P. Hinde. Correspondence analysis as a screening method for indicants for clinical diagnosis. *Statistics in Medicine*, 8:1351–1362, 1989.

[5] S. Durrleman and R. Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8:551–561, 1989.

[6] B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331, 1983.

[7] B. Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81:461–470, 1986.

[8] B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician*, 37:36–48, 1983.

[9] F. E. Harrell. Comparison of strategies for validating binary logistic regression models. Unpublished manuscript, 1991.

[10] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati. Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine*, 3:143–152, 1984.

[11] F. E. Harrell, K. L. Lee, and B. G. Pollock. Regression models in clinical studies: Determining relationships between predictors and response. *Journal of the National Cancer Institute*, 80:1198–1202, 1988.

[12] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[13] J. D. Kalbfleisch and R. L. Prentice. *The Statistical Analysis of Failure Time Data*. Wiley, New York, 1980.

[14] D. B. Mark, M. A. Hlatky, F. E. Harrell, K. L. Lee, R. M. Califf, and D. B. Pryor. Exercise treadmill score for predicting prognosis in coronary artery disease. *Annals of Internal Medicine*, 106:53–55, 1987.

[15] R. R. Picard and K. N. Berk. Data splitting. *American Statistician*, 44:140–147, 1990.

[16] D. B. Pryor, F. E. Harrell, J. S. Rankin, et al. The changing survival benefits of coronary revascularization over time. *Circulation (Supplement V)*, 76:13–21, 1987.

[17] W. S. Sarle. The VARCLUS procedure. In *SAS/STAT User's Guide*, volume 2, chapter 43, pages 1641–1659. SAS Institute, Inc., Cary NC, Fourth edition, 1990.

[18] D. Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69:239–241, 1982.

[19] L. A. Sleeper and D. P. Harrington. Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, 85:941–949, 1990.

[20] D. J. Spiegelhalter. Probabilistic prediction in patient management. *Statistics in Medicine*, 5:421–433, 1986.

[21] T. M. Therneau, P. M. Grambsch, and T. R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77:216–218, 1990.

[22] F. W. Young, Y. Takane, and J. de Leeuw. The principal components of mixed measurement level multivariate data: an alternating least squares method with optimal scaling features. *Psychometrika*, 43:279–281, 1978.