Regression Modeling Strategies using the R Package rms

Frank E Harrell Jr
Department of Biostatistics
Vanderbilt University School of Medicine
Nashville TN 37232
f.harrell@vanderbilt.edu
biostat.mc.vanderbilt.edu/rms

useR!
University of Warwick Coventry UK

The R User Conference
15 August 2011

Contents

1 Introduction **3**

1.1 Hypothesis Testing, Estimation, and Prediction . 3

1.2 Examples of Uses of Predictive Multivariable Modeling 5

1.3 Misunderstandings about Prediction vs. Classification 6

1.4 Planning for Modeling 10

1.5 Choice of the Model 14

1.6 Model uncertainty / Data-driven Model Specification 15

2 General Aspects of Fitting Regression Models **16**

2.1 Notation for Multivariable Regression Models . . 16

2.2 Model Formulations 17

2.3 Interpreting Model Parameters 18
2.3.1 Nominal Predictors 19
2.3.2 Interactions 20
2.3.3 Example: Inference for a Simple Model . 21
2.4 Review of Composite (Chunk) Tests 25
2.5 Relaxing Linearity Assumption for Continuous Predictors 25
2.5.1 Avoiding Categorization 25
2.5.2 Simple Nonlinear Terms 30
2.5.3 Splines for Estimating Shape of Regression Function and Determining Predictor Transformations 31
2.5.4 Cubic Spline Functions 34
2.5.5 Restricted Cubic Splines 34
2.5.6 Choosing Number and Position of Knots 38
2.5.7 Nonparametric Regression 40
2.5.8 Advantages of Regression Splines over Other Methods 42

2.6 Recursive Partitioning: Tree-Based Models . . . 43
2.7 New Directions in Predictive Modeling 45
2.8 Multiple Degree of Freedom Tests of Association 48
2.9 Assessment of Model Fit 50
2.9.1 Regression Assumptions 50
2.9.2 Modeling and Testing Complex Interactions 55
2.9.3 Fitting Ordinal Predictors 58
2.9.4 Distributional Assumptions 58

3 Multivariable Modeling Strategies 60

3.1 Prespecification of Predictor Complexity Without Later Simplification 61
3.1.1 Learning From a Saturated Model . . . 62
3.1.2 Using Marginal Generalized Rank Correlations 63
3.2 Checking Assumptions of Multiple Predictors Simultaneously 65
3.3 Variable Selection 65

3.4 Overfitting and Limits on Number of Predictors 69
3.5 Shrinkage 70
3.6 Collinearity 72
3.7 Data Reduction 74
3.7.1 Redundancy Analysis 75
3.7.2 Variable Clustering 76
3.7.3 Transformation and Scaling Variables Without Using Y 77
3.7.4 Simultaneous Transformation and Imputation 79
3.7.5 Simple Scoring of Variable Clusters . . . 84
3.7.6 Simplifying Cluster Scores 85
3.7.7 How Much Data Reduction Is Necessary? 85
3.8 Overly Influential Observations 88
3.9 Comparing Two Models 90
3.10 Summary: Possible Modeling Strategies 92
3.10.1 Developing Predictive Models 94
3.10.2 Developing Models for Effect Estimation 97

3.10.3 Developing Models for Hypothesis Testing 98

4 Describing, Resampling, Validating, and Simplifying the Model 99

4.1 Describing the Fitted Model 99
4.1.1 Interpreting Effects 99
4.1.2 Indexes of Model Performance 100
4.2 The Bootstrap 103
4.3 Model Validation 108
4.3.1 Introduction 108
4.3.2 Which Quantities Should Be Used in Validation? 109
4.3.3 Data-Splitting 110
4.3.4 Improvements on Data-Splitting: Resampling 112
4.3.5 Validation Using the Bootstrap 113
4.4 Simplifying the Final Model by Approximating It 118
4.4.1 Difficulties Using Full Models 118

4.4.2 Approximating the Full Model 119
4.5 How Do We Break Bad Habits? 120

5 S Software 122

5.1 The S Modeling Language 123
5.2 User-Contributed Functions 124
5.3 The `rms` Package 126
5.4 Other Functions 131

6 Logistic Model Case Study: Survival of Titanic Passengers 132

6.1 Descriptive Statistics 132
6.2 Exploring Trends with Nonparametric Regression 135
6.3 Binary Logistic Model with Casewise Deletion of Missing Values 136
6.4 Examining Missing Data Patterns 142
6.5 Single Conditional Mean Imputation 146
6.6 Multiple Imputation 150
6.7 Summarizing the Fitted Model 153

7 Case Study in Parametric Survival Modeling and Model Approximation 157

7.1 Descriptive Statistics 158
7.2 Checking Adequacy of Log-Normal Accelerated Failure Time Model 163
7.3 Summarizing the Fitted Model 173
7.4 Internal Validation of the Fitted Model Using the Bootstrap 174
7.5 Approximating the Full Model 178

Bibliography 191

Course Philosophy

- Satisfaction of model assumptions improves precision and increases statistical power
- 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
- Graphical methods should be married to formal inference
- 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
- 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.

- The bootstrap is a breakthrough for statistical modeling and model validation.
- 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 [104].

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 $\rightarrow$ 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
- 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 *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
- 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
- Far better is to use the full information in the data to develope a probability model, then develop classification rules on the basis of estimated probabilities
  - ↑ power, ↑ precision
- 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 [46]
- 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 negative) 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 [15, 19, 43, 49, 50, 113].

Accuracy score used to drive model building should be a continuous score that utilizes all of the information in the data.

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 your 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?

1.4 Planning for Modeling

- Chance that predictive model will be used[94]
- 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[68] 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 [125]: "generalized degrees of freedom" (GDF) for any "data mining" or model selection procedure based on least squares
  - Example: 20 candidate predictors, $n = 22$, forward stepwise, best 5-variable model: GDF=14.1
  - Example: CART, 10 candidate predictors, $n = 100$, 19 nodes: GDF=76
- See [79] for an approach involving adding noise to Y to improve variable selection

Chapter 2

General Aspects of Fitting Regression Models

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

Y	response (dependent) variable
X	$X_1, X_2, \ldots, X_p$ – list of predictors
β	$\beta_0, \beta_1, \ldots, \beta_p$ – regression coefficients
β_0	intercept parameter(optional)
$\beta_1, \ldots, \beta_p$	weights or regression coefficients
$X\beta$	$\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p, X_0 = 1$

Model: connection between X and Y
$C(Y|X)$: property of distribution of Y given X, e.g.
$C(Y|X) = \mathrm{E}(Y|X)$ or $\mathrm{Prob}\{Y = 1|X\}$.

2.2 Model Formulations

General regression model

$$C(Y|X) = g(X).$$

General linear regression model

$$C(Y|X) = g(X\beta).$$

Examples

$C(Y|X) =$ $E(Y|X) =$ $X\beta$,

$$Y|X \sim N(X\beta, \sigma^2)$$
$$C(Y|X) = \mathrm{Prob}\{Y = 1|X\} = (1 + \exp(-X\beta))^{-1}$$

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

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

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. $T = J, K, L, M$:

$$\begin{aligned} C(Y|T=J) &= \beta_0 \\ C(Y|T=K) &= \beta_0 + \beta_1 \\ C(Y|T=L) &= \beta_0 + \beta_2 \\ C(Y|T=M) &= \beta_0 + \beta_3. \end{aligned}$$

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

where

$$\begin{aligned} X_1 = 1 \text{ if } T = K,\ 0 \text{ otherwise} \\ X_2 = 1 \text{ if } T = L,\ 0 \text{ otherwise} \\ X_3 = 1 \text{ if } T = M,\ 0 \text{ otherwise.} \end{aligned}$$

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 $X_3 = X_1 X_2$ to model:

$$C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

$$\begin{aligned} C(Y|X_1+1, X_2) - C(Y|X_1, X_2) &= \beta_0 + \beta_1(X_1+1) + \beta_2 X_2 \\ &+ \beta_3(X_1+1)X_2 \\ &- [\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2] \\ &= \beta_1 + \beta_3 X_2. \end{aligned}$$

One-unit increase in X_2 on $C(Y|X)$: $\beta_2 + \beta_3 X_1$.
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).

2.3.3 Example: Inference for a Simple Model

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

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:

Parameter	Meaning
β_0	$C(Y\|age = 0, sex = m)$
β_1	$C(Y\|age = x + 1, sex = m) - C(Y\|age = x, sex = m)$
β_2	$C(Y\|age = 0, sex = f) - C(Y\|age = 0, sex = m)$
β_3	$C(Y\|age = x + 1, sex = f) - C(Y\|age = x, sex = f) -$ $[C(Y\|age = x + 1, sex = m) - C(Y\|age = x, sex = m)]$

β_3 is the difference in slopes (female – male).

[a] You can also think of the last part of the model as being $\beta_3 X_3$, where $X_3 = age \times I[sex = f]$.

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 age=0 or sex is male.

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 : age is associated with $C(Y)$, powered to detect a *linear* relationship"

Most Useful Tests for Linear age × sex Model

Null or Alternative Hypothesis	Mathematical Statement
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	$H_0 : \beta_3 = 0$
age interacts with sex age modifies effect of sex sex modifies effect of age sex and age are non-additive (synergistic)	$H_a : \beta_3 \neq 0$
age is not associated with Y age is associated with Y age is associated with Y for either females or males	$H_0 : \beta_1 = \beta_3 = 0$ $H_a : \beta_1 \neq 0$ or $\beta_3 \neq 0$
sex is not associated with Y sex is associated with Y sex is associated with Y for some value of age	$H_0 : \beta_2 = \beta_3 = 0$ $H_a : \beta_2 \neq 0$ or $\beta_3 \neq 0$
Neither age nor sex is associated with Y Either age or sex is associated with Y	$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$ or $\beta_3 \neq 0$

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.

2.4 Review of Composite (Chunk) Tests

- In the model

```
y ~ age + sex + weight + waist + 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)

2.5 Relaxing Linearity Assumption for Continuous Predictors

2.5.1 Avoiding Categorization

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

- Categorizing continuous predictors into intervals is a disaster[1, 2, 4, 8, 21, 44, 46, 64, 66, 73, 81, 84, 93, 96, 99, 108, 11
- 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 ≤ 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[b].
9. Categorization not blinded to $Y \rightarrow$ biased effect estimates[4, 99]
10. "Optimal" cutpoints do not replicate over studies. Hollander *et al.*[66] 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 lit-

[b] 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[66].

erature; 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 in 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."

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 ≤ 30, another for comparing BMI > 28 with BMI ≤ 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 [115].
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

2.5.2 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.

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

Draftman'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 $\beta_1, X \leq a$, $\beta_1 + \beta_2, X > a$.
- β_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

$$(u)_+ = \begin{cases} u, & u > 0, \\ 0, & u \leq 0. \end{cases}$$

$$f(X) = \begin{cases} \beta_0 + \beta_1 X, & X \leq a \\ \beta_0 + \beta_1 X + \beta_2(X-a) & a < X \leq b \\ \beta_0 + \beta_1 X + \beta_2(X-a) + \beta_3(X-b) & b < X \leq c \\ \beta_0 + \beta_1 X + \beta_2(X-a) & \\ \quad +\beta_3(X-b) + \beta_4(X-c) & c < X. \end{cases}$$

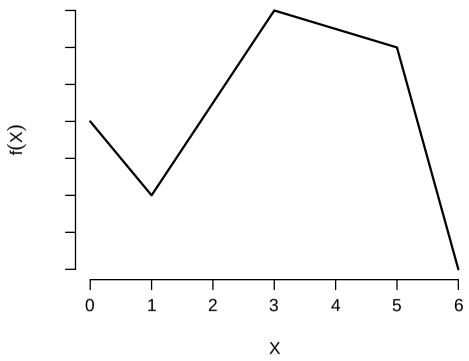


Figure 2.1: A linear spline function with knots at $a = 1, b = 3, c = 5$.

$$C(Y|X) = f(X) = X\beta,$$

where $X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$, and

$$X_1 = X \;\; X_2 = (X-a)_+$$
$$X_3 = (X-b)_+ \;\; X_4 = (X-c)_+.$$

Overall linearity in X can be tested by testing $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$.

2.5.4 Cubic Spline Functions

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

$$\begin{aligned} f(X) &= \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \\ &+ \beta_4 (X-a)_+^3 + \beta_5 (X-b)_+^3 + \beta_6 (X-c)_+^3 \\ &= X\beta \end{aligned}$$

$$X_1 = X \qquad X_2 = X^2$$
$$X_3 = X^3 \;\; X_4 = (X-a)_+^3$$
$$X_5 = (X-b)_+^3 \;\; X_6 = (X-c)_+^3.$$

k knots $\rightarrow k+3$ coefficients excluding intercept.

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

2.5.5 Restricted Cubic Splines

Stone and Koo [107]: cubic splines poorly behaved in tails. Constrain function to be linear in tails.

$k + 3 \rightarrow k - 1$ parameters[37].

To force linearity when $X < a$: X^2 and X^3 terms must be omitted
To force linearity when $X >$ last knot: last two βs are redundant, i.e., are just combinations of the other βs.

The restricted spline function with k knots $t_1, \ldots, t_k$ is given by[37]

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

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

$$\begin{aligned} X_{j+1} &= (X - t_j)_+^3 - (X - t_{k-1})_+^3 (t_k - t_j)/(t_k - t_{k-1}) \\ &+ (X - t_k)_+^3 (t_{k-1} - t_j)/(t_k - t_{k-1}). \end{aligned}$$

X_j is linear in X for $X \geq t_k$.

```
require(Hmisc)
x ← rcspline.eval(seq(0,1,.01),
                  knots=seq(.05,.95,length=5), inclx=TRUE)
xm ← x
xm[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)
```

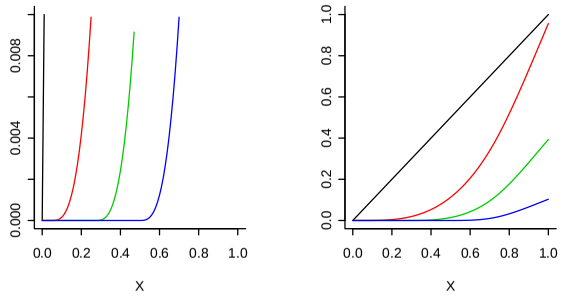



Figure 2.2: Restricted cubic spline component variables for $k = 5$ and knots at $X = .05, .275, .5, .725$, and $.95$. The left panel is a y–magnification of the right panel. Fitted functions such as those in Figure 2.3 will be linear combinations of these basis functions as long as knots are at the same locations used here.

```
    knots ← seq(.05, .95, length=nk)
    xx ← rcspline.eval(x, knots=knots, inclx=TRUE)
    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)
            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)
      }
  }
```

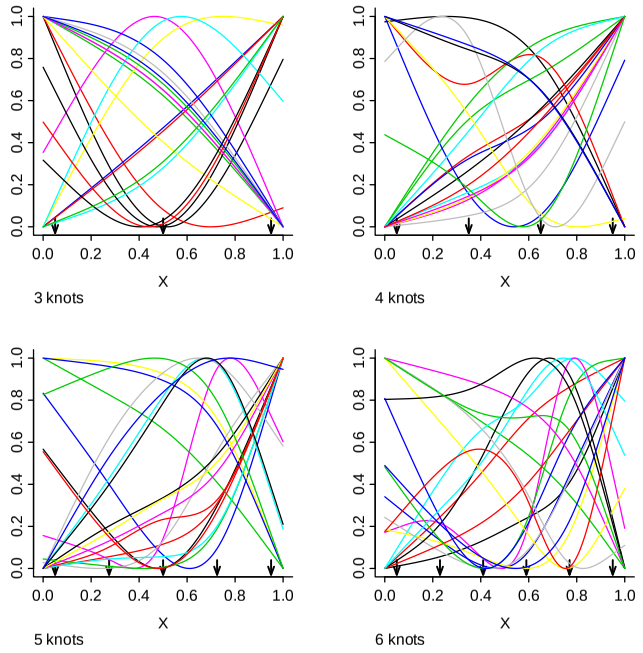


Figure 2.3: Some typical restricted cubic spline functions for $k = 3, 4, 5, 6$. The y–axis is $X\beta$. Arrows indicate knots. These curves were derived by randomly choosing values of β subject to standard deviations of fitted functions being normalized.

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 computing

$$\begin{aligned} \beta_k &= [\beta_2 (t_1 - t_k) + \beta_3 (t_2 - t_k) + \ldots \\ &\quad + \beta_{k-1} (t_{k-2} - t_k)]/(t_k - t_{k-1}) \\ \beta_{k+1} &= [\beta_2 (t_1 - t_{k-1}) + \beta_3 (t_2 - t_{k-1}) + \ldots \\ &\quad + \beta_{k-1} (t_{k-2} - t_{k-1})]/(t_{k-1} - t_k). \end{aligned}$$

A test of linearity in X can be obtained by testing

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

2.5.6 Choosing Number and Position of Knots

- Knots are specified in advance in regression splines
- Locations not important in most situations [39, 106]
- Place knots where data exist — fixed quantiles of predictor's marginal distribution
- Fit depends more on choice of k

k	Quantiles						
3			.10	.5	.90		
4			.05	.35	.65	.95	
5		.05	.275	.5	.725	.95	
6	.05	.23	.41	.59	.77	.95	
7	.025	.1833	.3417	.5	.6583	.8167	.975

$n < 100$ – replace outer quantiles with 5th smallest and 5th largest X [107].

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) [5, 111] to choose k
- This chooses k to maximize model likelihood ratio $\chi^2 - 2k$.

See [51] for a comparison of restricted cubic splines, fractional polynomials, and penalized splines.

2.5.7 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

X:	1	2	3	5	8
Y:	2.1	3.8	5.7	11.1	17.2

$$\hat{E}(Y|X = 2) = \frac{2.1 + 3.8 + 5.7}{3}$$

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

  - 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 [27] 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[c]
  - 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 great for binary Y — just turn off outlier detection

[c] Weight here means something different than regression coefficient. It means how much a point is emphasized in developing the regression coefficients.

- Other popular smoother: Friedman's "super smoother"
- For loess or supsmu amount of smoothing can be controlled by analyst
- Another alternative: smoothing splines[d]
- Smoothers are very useful for estimating trends in residual plots

2.5.8 Advantages of Regression Splines over Other Methods

Regression splines have several advantages [60]:

- 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

[d] These place knots at all the observed data points but penalize coefficient estimates towards smoothness.

$C(Y|X)$.

- 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. Multi-dimensional nonparametric estimators often require burdensome computations.

2.6 Recursive Partitioning: Tree-Based Models

Breiman, Friedman, Olshen, and Stone [18]: 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[28]; missing value imputation)

See[7].

2.7 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)[105, 110]
- elastic net (combination of L1 and L2 norms that handles the $p > n$ case better than the lasso)[129]
- adaptive lasso[116, 127]
- more flexible lasso to differentially penalize for variable selection and for regression coefficient estimation[92]
- group lasso to force selection of all or none of a group of related variables (e.g., dummy variables representing a polytomous predictor)
- group lasso-like procedures that also allow for variables within a group to be removed[117]
- adaptive group lasso (Wang & Leng)

- Breiman's nonnegative garrote[124]
- "preconditioning", i.e., model simplification after developing a "black box" predictive model[87]
- sparse principal components analysis to achieve parsimony in data reduction[77, 78, 121, 128]
- bagging, boosting, and random forests[62]

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 models[63, 122] should also be considered.

2.8 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 $\rightarrow$ 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[52] elegantly described the hazards of pretesting

- Studied quadratic regression
- Showed 2 d.f. test of association is nearly opti-

mal even when regression is linear if nonlinearity **entertained**

- Considered ordinary regression model $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$
- Two ways to test association between X and Y
- Fit quadratic model and test for linearity ($H_0 : \beta_2 = 0$)
- F-test for linearity significant at $\alpha = 0.05$ level $\rightarrow$ report as the final test of association the 2 d.f. F test of $H_0 : \beta_1 = \beta_2 = 0$
- 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

βs

- SSR from quadratic model $>$ SSR from linear model

2.9 Assessment of Model Fit

2.9.1 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.

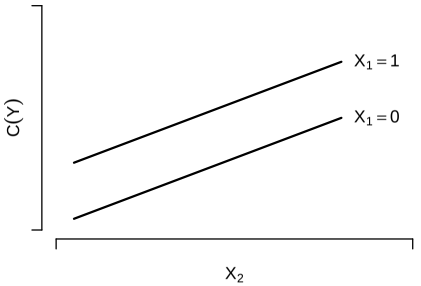


Figure 2.4: Regression assumptions for one binary and one continuous predictor.

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.

$$\begin{aligned}\hat{C}(Y|X) &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_2' + \hat{\beta}_4 X_2'' \\ &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{f}(X_2),\end{aligned}$$

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 $\rightarrow$ can fit separate functions by X_1
- Test of linearity: $H_0 : \beta_3 = \beta_4 = 0$
- Nonlinear $\rightarrow$ 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

$$\begin{aligned} C(Y|X) = & \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2' + \beta_4 X_2'' \\ & + \beta_5 X_3 + \beta_6 X_3' + \beta_7 X_3'', \end{aligned}$$

Overall test of linearity $H_0 : \beta_3 = \beta_4 = \beta_6 = \beta_7 = 0$, with 4 d.f.

2.9.2 Modeling and Testing Complex Interactions

X_1 binary or linear, X_2 continuous:

$$\begin{aligned} C(Y|X) = & \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2' + \beta_4 X_2'' \\ & + \beta_5 X_1 X_2 + \beta_6 X_1 X_2' + \beta_7 X_1 X_2'' \end{aligned}$$

Simultaneous test of linearity and additivity: $H_0 : \beta_3 = \ldots = \beta_7 = 0$.

- 2 continuous variables: could transform separately and form simple product
- Transformations depend on whether interaction terms adjusted for
- Fit interactions of the form $X_1 f(X_2)$ and $X_2 g(X_1)$:

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

- 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 $aX_1 + bX_2 + cX_1 X_2$ in corners
- Uses all $(k-1)^2$ cross-products of restricted cubic spline terms
- See Gray [53, 54, Section 3.2] for penalized splines

allowing control of effective degrees of freedom. See Berhane *et al.*[12] for a good discussion of tensor splines.

Other issues:

- Y non-censored (especially continuous) $\rightarrow$ multidimensional scatterplot smoother[22]
- Interactions of order > 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

2.9.3 Fitting Ordinal Predictors

- Small no. categories (3-4) $\rightarrow$ polytomous factor, dummy variables
- Design matrix for easy test of adequacy of initial codes $\rightarrow$ k original codes + $k - 2$ dummies
- More categories $\rightarrow$ score using data-driven trend. Later tests use $k - 1$ d.f. instead of 1 d.f.
- E.g., compute logit(mortality) vs. category

2.9.4 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[31]):
  $C(Y|X) = C(Y = y|X) = d(y) + X\beta$
  C = log hazard
- Check form of $d(y)$
- Show $d(y)$ does not interact with X

Chapter 3

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

3.1 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

3.1.1 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 χ^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 χ^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 χ^2 (the expected value of χ^2_p 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 `rms` package can be used to plot only what is needed. Here is an example for a logistic model.

```
f ← lrm(y ~ sex + race + rcs(age,5) + rcs(weight,5) +
          rcs(height,5) + rcs(blood.pressure,5))
plot(anova(f))
```

3.1.2 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 ρ—R^2 based on $rank(X)$ and $rank(X)^2$ vs. $rank(Y)$
- ρ^2 can detect U-shaped relationships
- For categorical X, ρ^2 is R^2 from dummy variables regressed against $rank(Y)$; this is tightly related to the Wilcoxon–Mann–Whitney–Kruskal–Wallis rank test for group differences[a]
- Sort variables by descending order of ρ^2
- Specify number of knots for continuous X, combine infrequent categories of categorical X based on ρ^2

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

- We already decided to keep variable in model no matter what ρ^2 or χ^2 values are seen
- ρ^2 and χ^2 do not reveal degree of nonlinearity; high value may be due solely to strong linear effect
- low ρ^2 or χ^2 for a categorical variable might lead

[a] This test statistic does not inform the analyst of *which* groups are different from one another.

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

3.2 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 $\rightarrow$ assess regression assumptions separately for each predictor

3.3 Variable Selection

- Series of potential predictors with no prior knowledge
- $\uparrow$ exploration $\rightarrow \uparrow$ shrinkage (overfitting)
- Summary of problem: $E(\hat{\beta}|\hat{\beta}$ "significant" $) \neq \beta$ [24]

- Biased R^2, $\hat{\beta}$, standard errors, P-values too small
- F and χ^2 statistics do not have the claimed distribution [52]
- Will result in residual confounding if use variable selection to find confounders [56]
- Derksen and Keselman [36] 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.
  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**

Variable selection methods[57]:

- Forward selection, backward elimination
- Stopping rule: "residual χ^2" with d.f. = no. candidates remaining at current step
- Test for significance or use Akaike's information criterion (AIC[5]), here $\chi^2 - 2 \times d.f.$
- Better to use subject matter knowledge!
- No currently available stopping rule was developed for stepwise, only for comparing **2** pre-specified models [16, Section 1.3]
- Roecker[95] 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 $> \frac{1}{2}$ candidate variables redundant or unnecessary
- Step-down is usually better than forward[80] and can be used efficiently with maximum likelihood estimation[74]
- Fruitless to try different stepwise methods to look for agreement[120]
- 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[3,16,67]
- Data reduction (grouping variables) can help
- Using the bootstrap to select important variables for inclusion in the final model[98] is problematic[6]
- It is not logical that a population regression coefficient would be exactly zero just because its estimate was "insignificant"

3.4 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 $< \frac{m}{15}$ [58, 59, 88, 89, 101, 114]
- p = number of parameters in full model or number of *candidate* parameters in a stepwise analysis

Table 3.1: Limiting Sample Sizes for Various Response Variables

Type of Response Variable	Limiting Sample Size m
Continuous	n (total sample size)
Binary	$\min(n_1, n_2)$ [b]
Ordinal (k categories)	$n - \frac{1}{n^2}\sum_{i=1}^{k} n_i^3$ [c]
Failure (survival) time	number of failures [d]

- Narrowly distributed predictor $\rightarrow$ even higher n
- p includes *all* variables screened for association with response, including interactions

[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 $\frac{n_1}{n}$ is near 0 or 1 [119, Eq. 10, 15]. Here n_1 and n_2 are the marginal frequencies of the two response levels [89].

[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) [119, Eq, 3]. If all cell sizes are equal, the relative efficiency of having k response categories compared to a continuous response is $1 - \frac{1}{k^2}$ [119, 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 [11].

- Univariable screening (graphs, crosstabs, etc.) **in no way** reduces multiple comparison problems of model building [109]

3.5 Shrinkage

- Slope of calibration plot; regression to the mean
- Statistical estimation procedure — "pre-shrunk" 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. $X\hat{\beta}$

```
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))
```

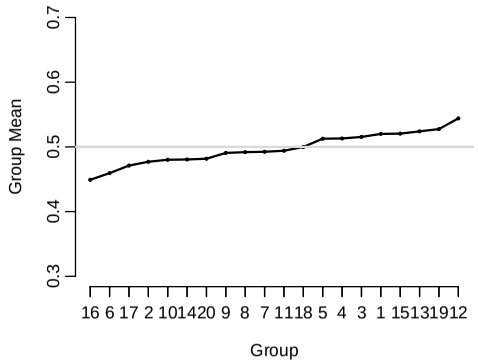


Figure 3.1: Sorted means from 20 samples of size 50 from a uniform $[0,1]$ distribution. The reference line at 0.5 depicts the true population value of all of the means.

- Prevent shrinkage by using pre-shrinkage
- Spiegelhalter [103]: 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[29]

- Ridge regression [75, 111]
- Penalized MLE [53, 61, 112]
- Heuristic shrinkage parameter of van Houwelingen and le Cessie [111, Eq. 77]

$$\hat{\gamma} = \frac{\text{model } \chi^2 - p}{\text{model } \chi^2},$$

- OLS: $\hat{\gamma} = \frac{n-p-1}{n-1} R^2_{\text{adj}}/R^2$
  $R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$
- p close to no. candidate variables
- Copas [29, Eq. 8.5] adds 2 to numerator

3.6 Collinearity

- When at least 1 predictor can be predicted well from others
- Can be a blessing (data reduction, transformations)
- ↑ s.e. of $\hat{\beta}$, ↓ power

- This is appropriate → asking too much of the data [25, p. 173]
- 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 [85, 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 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=diagor

of inverse[35, 118]

- See Belsley [9, pp. 28-30] for problems with VIF
- Easy approach: SAS `VARCLUS` procedure[97], S `varclus` function, other clustering techniques: group highly correlated variables
- Can score each group (e.g., first principal component, PC_1[34]); summary scores not collinear

3.7 Data Reduction

- Unless $n >> p$, 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.

3.7.1 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/S function `redun` in `Hmisc` package
- Related to *principal variables*[82] but faster

3.7.2 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[34, 97] (oblique-rotation PC analysis) $\rightarrow$ 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 [65].

3.7.3 Transformation and Scaling Variables Without Using Y

- Reduce p by estimating transformations using associations with other predictors
- Purely categorical predictors – correspondence analysis [26, 33, 55, 76, 83]
- Mixture of qualitative and continuous variables: qualitative principal components
- Maximum total variance (MTV) of Young, Takane, de Leeuw [83, 126]
  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 Xs

  4. Repeat 3-4 times until variation explained by PC_1 plateaus and transformations stabilize
- Maximum generalized variance (MGV) method of Sarle [72, 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 $aX_1 + bX_1^2$ 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` [72]
  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

3.7.4 Simultaneous Transformation and Imputation

S `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 Xs
- The imputed values can optionally be shrunk to avoid overfitting 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 S function `Function` will create a series of S 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 Figure 3.2

```
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)
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
```

```
plot(heart.rate, blood.pressure)
t ← w$transformed
plot(t[,'heart.rate'], t[,'blood.pressure'],
     xlab='Transformed hr', ylab='Transformed bp')
spe ← round(c(spearman(heart.rate, blood.pressure),
              spearman(t[,'heart.rate'], t[,'blood.pressure'])), 2)
```

ACE (Alternating Conditional Expectation) of Breiman and Friedman [17]

1. Uses nonparametric "super smoother"[48]
2. Allows monotonicity constraints, categorical vars.
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)

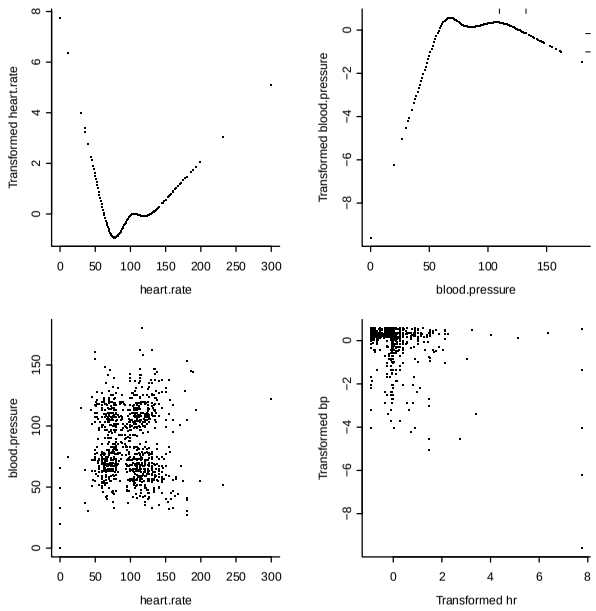


Figure 3.2: Transformations fitted using `transcan`. Tick marks indicate the two imputed values for blood pressure. The lower left plot contains raw data (Spearman $\rho = -0.02$); the lower right is a scatterplot of the corresponding transformed values ($\rho = -0.13$). Data courtesy of the SUPPORT study[70].

3. Expand original variable using spline, test additional predictive information over original transformation

3.7.5 Simple Scoring of Variable Clusters

- 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 positive
- Construct X_2 = number of positives

- Test whether original variables add to X_1 or X_2

3.7.6 Simplifying Cluster Scores

3.7.7 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 χ^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 $[\mathrm{LR} - (p-q) - q]/[\mathrm{LR} - (p-q)]$
- Solve for $q \rightarrow q \leq (\mathrm{LR} - p)/9$
- Under these assumptions, no hope unless original LR $> p + 9$

- No χ^2 lost by dimension reduction $\rightarrow q \leq \mathrm{LR}/10$

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 $\rightarrow$ use full model, adjust for multiple comparisons

Summary of Some Data Reduction Methods

Goals	Reasons	Methods
Group predictors so that each group represents a single dimension that can be summarized with a single score	• ↓ d.f. arising from multiple predictors • Make PC_1 more reasonable summary	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

3.8 Overly Influential Observations

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

Reasons for influence:

- Too few observations for complexity of model (see Sections 3.7, 3.3)
- 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[10]
- $h_{ii} > 2(p + 1)/n$ may signal a high leverage point[10]
- 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 $|\mathrm{DFFITS}| > 2\sqrt{(p+1)/(n-p-1)}$[10]
- Others examine entire distribution for "outliers"
- No substitute for careful examination of data[23, 102]
- Maximum likelihood estimation requires 1-step approximations

3.9 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 [86], model χ^2, Somers' D_{xy}, Spearman's ρ, 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 $\rightarrow$ model χ^2, R^2, examine extremes of distribution of $\hat{Y}$
- Examine differences in predicted values from the two models

- See [90, 91] for discussions and examples of low power for testing differences in ROC areas.

3.10 Summary: Possible Modeling Strategies

Greenland [56] 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)"

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

3.10.1 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[32]. 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 3.1)

6. Do data reduction if needed (pre-transformations, combinations), or use penalized estimation[61]
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[103]. 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[45] to penalize for the randomness inherent in the multiple steps.
20. Develop simplifications to the final model as needed.

3.10.2 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

3.10.3 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 4

Describing, Resampling, Validating, and Simplifying the Model

4.1 Describing the Fitted Model

4.1.1 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 interacting factors
- Subtract estimates, anti-log, e.g., to get inter-quartile-range odds or hazards ratios. Base C.L. on s.e. of difference.
- Plot effect of each predictor on $X\beta$ or some transformation of $X\beta$. See also [69].
- Nomogram
- Use regression tree to approximate the full model

4.1.2 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 ρ, Kendall τ, 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)
  - calibration slope and intercept
  - maximum absolute calibration error
  - mean absolute calibration error
  - 0.9 quantile of calibration error

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; $g = 11$mmHg is typical difference in $\hat{Y}$
- Independent of censoring etc.
- For models in which anti-log of difference in $\hat{Y}$ represent 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):
  $g_p =$ Gini's mean difference of $\hat{P}$
- These g–indexes represent e.g. "typical" odds ratios, "typical" risk differences
- Can define partial g

4.2 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.

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I(Y_i \leq y),$$

  where $I(w)$ is 1 if w is true, 0 otherwise.
- Example: sample of size $n = 30$ from a normal distribution with mean 100 and SD 10

```
set.seed(6)
x <- rnorm(30, 100, 20)
xs <- seq(50, 150, length=150)
cdf <- pnorm(xs, 100, 20)
plot(xs, cdf, type='l', ylim=c(0,1),
     xlab=expression(x),
     ylab=expression(paste("Prob[", X <= x, "]")))
lines(ecdf(x), cex=.5)
```

- 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 $Y_1, \ldots, Y_n$

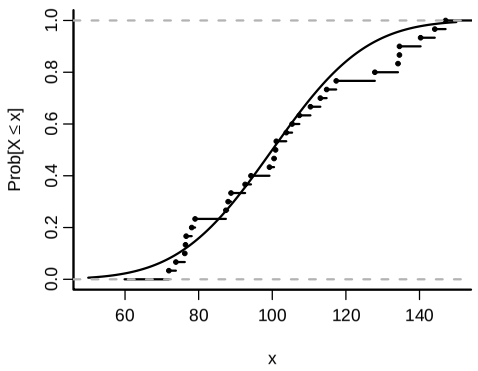


Figure 4.1: Empirical and population cumulative distribution functions

- 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* $\rightarrow$ 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).

```
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
        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
  6  10   7   8   2  23  43  75  59  66  47  42  11   1
```

```
hist(M, nclass=length(unique(M)), xlab="", main="")
```

First 20 samples:

Bootstrap Sample	Sample Median
1 6 6 7 8 9	6.5
1 5 5 5 6 8	5.0
5 7 8 9 9 9	8.5
7 7 7 8 8 9	7.5
1 5 7 7 9 9	7.0
1 5 6 6 7 8	6.0
7 8 8 8 8 8	8.0
5 5 5 7 9 9	6.0
1 5 5 7 7 9	6.0
1 5 5 7 7 8	6.0
1 1 5 5 7 7	5.0
1 1 5 5 7 8	5.0
1 5 5 7 7 8	6.0
1 5 6 7 8 8	6.5
1 5 6 7 9 9	6.5
6 6 7 7 8 9	7.0
1 5 7 8 8 9	7.5
6 6 8 9 9 9	8.5
1 1 5 5 6 9	5.0
1 6 8 9 9 9	8.5

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

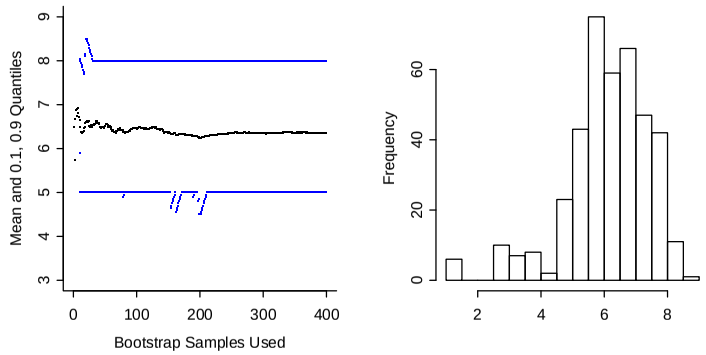


Figure 4.2: Estimating properties of sample median using the bootstrap

- Need high B for quantiles, low for variance (but see [14])

4.3 Model Validation

4.3.1 Introduction

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

4.3.2 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

4.3.3 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 $R^2 = 1 - SSE/SST$ 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 ($n = 100$), $R^2 = 1 - \frac{\Sigma(Y_i - X_i\hat{\beta})^2}{\Sigma(Y_i - Y)^2}$.
- Disadvantages of data splitting:
  1. Costly in $\downarrow n$ [16, 95]
  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

4.3.4 Improvements on Data-Splitting: Resampling

- No sacrifice in sample size
- Work when modeling process automated
- Bootstrap excellent for studying arbitrariness of variable selection [98]
- Cross-validation solves many problems of data splitting [40, 100, 111, 123]
- 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

4.3.5 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_{app}) =0.4
  - how inflated is R^2_{app}?
  - get resamples of size 1000 with replacement from original 1000
  - for each resample compute R^2_{boot} = apparent R^2 in bootstrap sample
  - freeze these coefficients (call them $\hat{\beta}_{boot}$), apply to original (whole) sample (X_{orig}, Y_{orig}) to get $R^2_{orig} = R^2(X_{orig}\hat{\beta}_{boot}, Y_{orig})$
  - optimism = $R^2_{boot} - R^2_{orig}$
  - average over $B = 100$ optimisms to get $\overline{optimism}$
  - $R^2_{overfitting\ corrected} = R^2_{app} - \overline{optimism}$

- Is estimating unconditional (not conditional on X) distribution of R^2, etc. [45, 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 [40], [41, p. 253], [42]

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 γ_0, γ_1
- $\hat{\gamma}_1 < 1$ if overfit, $\hat{\gamma}_0 > 0$ to compensate
- $\hat{\gamma}_1$ quantifies overfitting and useful for improving calibration [103]
- Use Efron's method to estimate optimism in $(0, 1)$, estimate (γ_0, γ_1) by subtracting optimism from $(0, 1)$

- See also Copas [30] and van Houwelingen and le Cessie [111, p. 1318]

See [47] for warnings about the bootstrap, and [40] 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 [103]
- Stepwise modeling often reduces optimism but this is not offset by loss of information from deleting marginal var.

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

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 [111] for many interesting ideas
- Faraway[45] shows how bootstrap is used to penalize for choosing transformations for Y, outlier and influence checking, variable selection, etc. simultaneously

- Brownstone [20, 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[13] and Chatfield[24] for more interesting examples of problems resulting from data-driven analyses.

4.4 Simplifying the Final Model by Approximating It

4.4.1 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.

4.4.2 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 σ 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'

4.5 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

Chapter 5

S Software

S allows interaction spline functions, wide variety of predictor parameterizations, wide variety of models, unifying model formula language, model validation by resampling.

S is comprehensive:

- Easy to write S functions for new models → wide variety of modern regression models implemented (trees, nonparametric, ACE, AVAS, survival models for multiple events)
- Designs can be generated for any model → all handle "class" var, interactions, nonlinear expan-

sions

- Single S objects (e.g., fit object) can be self-documenting → automatic hypothesis tests, predictions for new data
- Superior graphics
- Classes and generic functions

5.1 The S Modeling Language

S statistical modeling language:

```
response ~ terms
y ~ age + sex              # age + sex main effects
y ~ age + sex + age:sex    # add second-order interaction
y ~ age*sex                # second-order interaction +
                           # all main effects
y ~ (age + sex + pressure)^2
                           # age+sex+pressure+age:sex+age:pressure...
y ~ (age + sex + pressure)^2 - sex:pressure
                           # all main effects and all 2nd order
                           # interactions except sex:pressure
y ~ (age + race)*sex       # age+race+sex+age:sex+race:sex
y ~ treatment*(age*race + age*sex) # no interact. with race,sex
sqrt(y) ~ sex*sqrt(age) + race
# functions, with dummy variables generated if
# race is an S factor (classification) variable
y ~ sex + poly(age,2)      # poly generates orthogonal polynomials
race.sex ← interaction(race,sex)
y ~ age + race.sex         # for when you want dummy variables for
                           # all combinations of the factors
```


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.
```

5.2 User-Contributed Functions

- S is high-level object-oriented language.
- S-PLUS (UNIX, Linux, Microsoft Windows)
- R (UNIX, Linux, Mac, Windows)
- Multitude of user-contributed functions freely available
- International community of users

[a] `lrm` and `rcs` are in the `rms` package.

Some S 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,S-PLUS))
- 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`)

5.3 The `rms` Package

- `datadist` function to compute predictor distribution summaries

```
y ~ sex + lsp(age,c(20,30,40,50,60)) +
    sex %ia% lsp(age,c(20,30,40,50,60))
```

E.g. restrict age $\times$ cholesterol interaction to be of form $AF(B) + BG(A)$:

```
y ~ lsp(age,30) + rcs(cholesterol,4) +
    lsp(age,30) %ia% rcs(cholesterol,4)
```

Special fitting functions by Harrell to simplify procedures described in these notes:

Table 5.1: `rms` Fitting Functions

Function	Purpose	Related S Functions
`ols`	Ordinary least squares linear model	`lm`
`lrm`	Binary and ordinal logistic regression model Has options for penalized MLE	`glm`
`psm`	Accelerated failure time parametric survival models	`survreg`
`cph`	Cox proportional hazards regression	`coxph`
`bj`	Buckley-James censored least squares model	`survreg,lm`
`Glm`	`rms` version of `glm`	`glm`
`Gls`	`rms` version of `gls`	`gls` (`nlme` package)
`Rq`	`rms` version of `rq`	`rq` (`quantreg` package)

Table 5.2: `rms` Transformation Functions

Function	Purpose	Related S Functions
`asis`	No post-transformation (seldom used explicitly)	`I`
`rcs`	Restricted cubic splines	`ns`
`pol`	Polynomial using standard notation	`poly`
`lsp`	Linear spline	
`catg`	Categorical predictor (seldom)	`factor`
`scored`	Ordinal categorical variables	`ordered`
`matrx`	Keep variables as group for `anova` and `fastbw`	`matrix`
`strat`	Non-modeled stratification factors (used for `cph` only)	`strata`

Function	Purpose	Related Functions
`print`	Print parameters and statistics of fit	
`coef`	Fitted regression coefficients	
`formula`	Formula used in the fit	
`specs`	Detailed specifications of fit	
`vcov`	Fetch covariance matrix	
`logLik`	Fetch maximized log-likelihood	
`AIC`	Fetch AIC with option to put on chi-square basis	
`lrtest`	Likelihood ratio test for two nested models	
`univarLR`	Compute all univariable LR χ^2	
`robcov`	Robust covariance matrix estimates	
`bootcov`	Bootstrap covariance matrix estimates and bootstrap distributions of estimates	
`pentrace`	Find optimum penalty factors by tracing effective AIC for a grid of penalties	
`effective.df`	Print effective d.f. for each type of variable in model, for penalized fit or `pentrace` result	
`summary`	Summary of effects of predictors	
`plot.summary`	Plot continuously shaded confidence bars for results of `summary`	
`anova`	Wald tests of most meaningful hypotheses	
`plot.anova`	Graphical depiction of anova	
`contrast`	General contrasts, C.L., tests	
`gendata`	Easily generate predictor combinations	
`predict`	Obtain predicted values or design matrix	
`Predict`	Obtain predicted values and confidence limits easily varying a subset of predictors and others set at default values	
`plot.Predict`	Plot effects of predictors	
`fastbw`	Fast backward step-down variable selection	`step`
`residuals`	(or `resid`) Residuals, influence stats from fit	
`sensuc`	Sensitivity analysis for unmeasured confounder	
`which.influence`	Which observations are overly influential	`residuals`
`latex`	LaTeX representation of fitted model	`Function`
`Function`	S function analytic representation of $X\hat{\beta}$ from a fitted regression model	`latex`

Function	Purpose	Related Functions
`Hazard`	S function analytic representation of a fitted hazard function (for `psm`)	
`Survival`	S function analytic representation of fitted survival function (for `psm`, `cph`)	
`Quantile`	S function analytic representation of fitted function for quantiles of survival time (for `psm`, `cph`)	
`Mean`	S function analytic representation of fitted function for mean survival time or for ordinal logistic	
`nomogram`	Draws a nomogram for the fitted model	`latex`, `plot`
`survest`	Estimate survival probabilities (`psm`, `cph`)	`survfit`
`survplot`	Plot survival curves (`psm`, `cph`)	`plot.survfit`
`validate`	Validate indexes of model fit using resampling	
`val.prob`	External validation of a probability model	`lrm`
`val.surv`	External validation of a survival model	`calibrate`
`calibrate`	Estimate calibration curve using resampling	`val.prob`
`vif`	Variance inflation factors for fitted model	
`naresid`	Bring elements corresponding to missing data back into predictions and residuals	
`naprint`	Print summary of missing values	
`impute`	Impute missing values	`aregImpute`

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(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

```
require(rms)                        # make new functions available
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))
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)
anova(fit, treat, cholesterol)      # Test these 2 by themselves
plot(anova(fit))                    # Summarize anova graphically
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.

p ← Predict(fit, age=seq(20,80,length=100), treat, conf.int=FALSE)
plot(p)                             # Plot relationship between age and log
                                    # odds, separate curve for each treat,
                                    # no C.I.
plot(p, ~ age | treat)              # 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
# 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
# Could also use
#    logit ← Predict(f, age=ag, ...)$yhat,
# which allows evaluation of the shape for any level of interacting
#  factors.  When age does not interact with anything, the result from
# predict(f, ..., type="terms") would equal the result from
# Predict if all other terms were ignored

# Could also specify
#    logit ← predict(fit, gendata(fit, age=ag, cholesterol=...))
# Un-mentioned variables set to reference values

plot(ag^.5, logit)                  # try square root vs. spline transform.
plot(ag^1.5, logit)                 # try 1.5 power
latex(fit)                          # invokes latex.lrm, creates fit.tex
# Draw a nomogram for the model fit
plot(nomogram(fit))
# Compose S function to evaluate linear predictors analytically
g ← Function(fit)
g(treat='b', cholesterol=260, age=50)
# Letting num.diseases default to reference value
```

To examine interactions in a simpler way, you may want to group age into tertiles:

```
age.tertile ← cut2(age, g=3)
# For automatic ranges later, add age.tertile to datadist input
fit ← lrm(y ~ age.tertile * rcs(cholesterol))
```

5.4 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

```
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 6

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 R, S-PLUS, and Excel formats from `biostat.mc.vanderbilt.edu/DataSets` under the name `titanic3`.

6.1 Descriptive Statistics

```
require(rms)
getHdata(titanic3)      # get dataset from web site
units(titanic3$age) ← 'years'
# List of names of variables to analyze
v ← c('pclass','survived','age','sex','sibsp','parch')
```

```
latex(describe(titanic3[,v]), file='')
```

titanic3[, v]
6 Variables 1309 Observations

pclass

```
    n missing unique
 1309       0      3
```

```
1st (323, 25%), 2nd (277, 21%), 3rd (709, 54%)
```

survived : Survived

```
    n missing unique Sum Mean
 1309       0      2 500 0.382
```

age : Age [years]

```
    n missing unique  Mean   .05   .10   .25   .50   .75   .90   .95
 1046     263     98 29.88     5    14    21    28    39    50    57
```

```
lowest :  0.1667  0.3333  0.4167  0.6667  0.7500
highest: 70.5000 71.0000 74.0000 76.0000 80.0000
```

sex

```
    n missing unique
 1309       0      2
```

```
female (466, 36%), male (843, 64%)
```

sibsp : Number of Siblings/Spouses Aboard

```
    n missing unique   Mean
 1309       0      7 0.4989
```

```
            0   1  2  3  4 5 8
Frequency 891 319 42 20 22 6 9
%          68  24  3  2  2 0 1
```

parch : Number of Parents/Children Aboard

```
    n missing unique   Mean
 1309       0      8  0.385
```

```
            0   1   2 3 4 5 6 9
Frequency 1002 170 113 8 6 6 2 2
%           77  13   9 1 0 0 0 0
```

```
dd <- datadist(titanic3[,v])
# describe distributions of variables to rms
options(datadist='dd')
attach(titanic3[,v])
options(digits=2)
s <- summary(survived ~ age + sex + pclass +
             cut2(sibsp,0:3) + cut2(parch,0:3))
latex(s, file='', label='titanic-summary.table')  # create LaTeX code for Table
```


Table 6.1: Survived N=1309

	N	survived
Age years		
[0.167,22.0)	290	0.43
[22.000,28.5)	246	0.39
[28.500,40.0)	265	0.42
[40.000,80.0]	245	0.39
Missing	263	0.28
sex		
female	466	0.73
male	843	0.19
pclass		
1st	323	0.62
2nd	277	0.43
3rd	709	0.26
Number of Siblings/Spouses Aboard		
0	891	0.35
1	319	0.51
2	42	0.45
[3,8]	57	0.16
Number of Parents/Children Aboard		
0	1002	0.34
1	170	0.59
2	113	0.50
[3,9]	24	0.29
Overall		
	1309	0.38

```
plot(s, main='', subtitles=FALSE) # convert table to dot plot (Figure 6.1)
```

Show 4-way relationships after collapsing levels. Suppress estimates based on < 25 passengers.

```
agec <- ifelse(age<21,'child','adult')
sibsp.parch <-
  paste(ifelse(sibsp==0,'no sib/spouse','sib/spouse'),
  ifelse(parch==0,'no parent/child','parent/child'),
  sep=' / ')
g <- function(y) if(length(y) < 25) NA else mean(y)
s <- summarize(survived,
               llist(agec, sex, pclass, sibsp.parch),g)
# llist, summarize, Dotplot in Hmisc package
require(lattice)    # trellis for S-PLUS
## To remove color background from strip labels do the following:
## ltheme <- canonical.theme(color = FALSE)
## ltheme$strip.background$col <- "transparent"
```

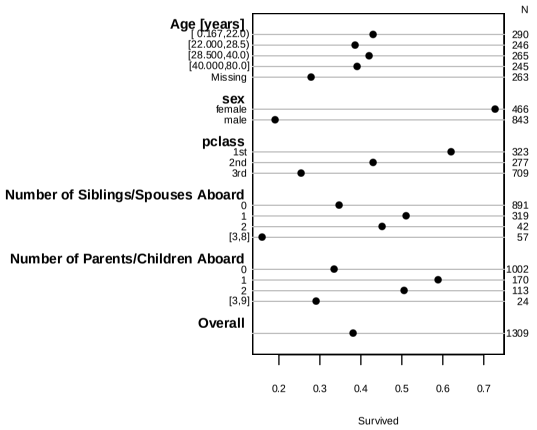


Figure 6.1: Univariable summaries of Titanic survival

```
## lattice.options(default.theme = ltheme)      ## set as default

i ← s$agec != 'NA'
print(Dotplot(pclass ~ survived | sibsp.parch*agec,
              groups=sex[i], data=s, subset=i, pch=c(1,4), col=c(1,1),
              xlab='Proportion Surviving',
              par.strip.text=list(cex=.6)))  # Figure 6.2
Key(.07)
```

6.2 Exploring Trends with Nonparametric Regression

```
# Figure 6.3
plsmo(age, survived, datadensity=TRUE)
plsmo(age, survived, group=sex, datadensity=TRUE)
plsmo(age, survived, group=pclass, datadensity=TRUE)
plsmo(age, survived, group=interaction(pclass,sex),
      datadensity=TRUE, lty=c(1,1,1,2,2,2))
```

```
# Figure 6.4
plsmo(age, survived, group=cut2(sibsp,0:2), datadensity=TRUE)
plsmo(age, survived, group=cut2(parch,0:2), datadensity=TRUE)
```

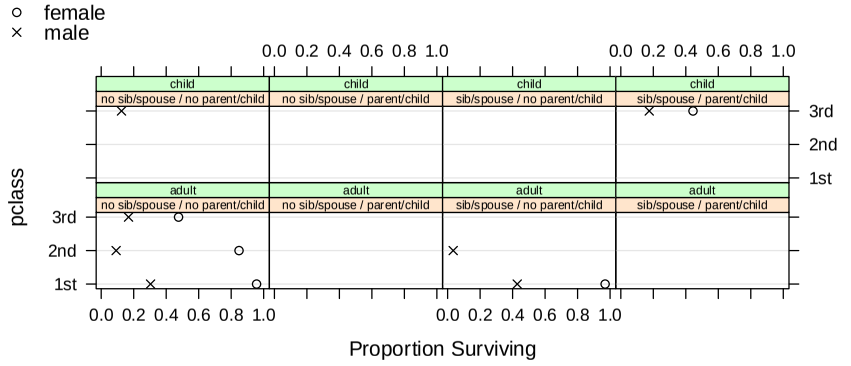



Figure 6.2: Multi-way summary of Titanic survival

6.3 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.

```
f1 ← lrm(survived ~ sex*pclass*rcs(age,5) +
         rcs(age,5)*(sibsp + parch))
latex(anova(f1), file='', label='titanic-anova3')   # Table 6.2
```

3-way interactions, `parch` clearly insignificant, so drop

```
f ← lrm(survived ~ (sex + pclass + rcs(age,5))^2 +
        rcs(age,5)*sibsp)
print(f, latex=TRUE)
```

Logistic Regression Model

```
lrm(formula = survived ~ (sex + pclass + rcs(age, 5))^2 + rcs(age,
    5) * sibsp)
```

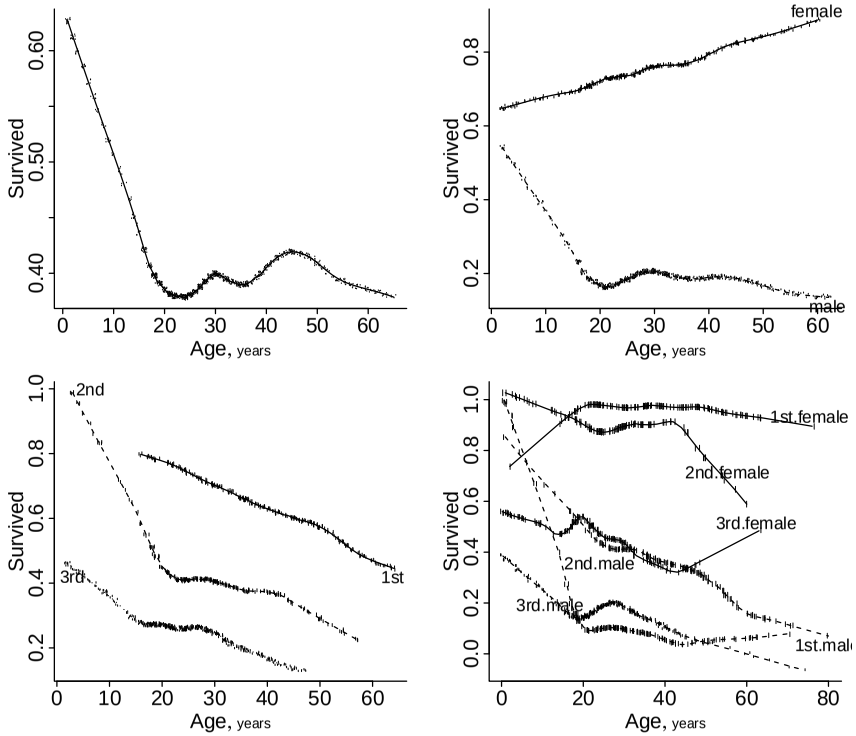


Figure 6.3: Nonparametric regression (loess) estimates of the relationship between age and the probability of surviving the Titanic. The top left panel shows unstratified estimates. The top right panel depicts relationships stratified by sex. The bottom left and right panels show respectively estimates stratified by class and by the cross-classification of sex and class of the passenger. Tick marks are drawn at actual age values for each strata.

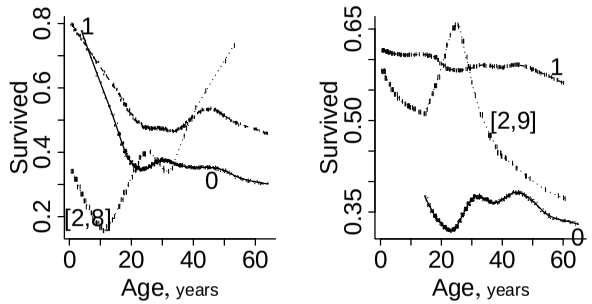


Figure 6.4: Relationship between age and survival stratified by the number of siblings or spouses on board (left panel) or by the number of parents or children of the passenger on board (right panel)

Table 6.2: Wald Statistics for `survived`

	χ^2	*d.f.*	*P*
sex (Factor+Higher Order Factors)	187.15	15	< 0.0001
All Interactions	59.74	14	< 0.0001
pclass (Factor+Higher Order Factors)	100.10	20	< 0.0001
All Interactions	46.51	18	0.0003
age (Factor+Higher Order Factors)	56.20	32	0.0052
All Interactions	34.57	28	0.1826
Nonlinear (Factor+Higher Order Factors)	28.66	24	0.2331
sibsp (Factor+Higher Order Factors)	19.67	5	0.0014
All Interactions	12.13	4	0.0164
parch (Factor+Higher Order Factors)	3.51	5	0.6217
All Interactions	3.51	4	0.4761
sex × pclass (Factor+Higher Order Factors)	42.43	10	< 0.0001
sex × age (Factor+Higher Order Factors)	15.89	12	0.1962
Nonlinear (Factor+Higher Order Factors)	14.47	9	0.1066
Nonlinear Interaction : f(A,B) vs. AB	4.17	3	0.2441
pclass × age (Factor+Higher Order Factors)	13.47	16	0.6385
Nonlinear (Factor+Higher Order Factors)	12.92	12	0.3749
Nonlinear Interaction : f(A,B) vs. AB	6.88	6	0.3324
age × sibsp (Factor+Higher Order Factors)	12.13	4	0.0164
Nonlinear	1.76	3	0.6235
Nonlinear Interaction : f(A,B) vs. AB	1.76	3	0.6235
age × parch (Factor+Higher Order Factors)	3.51	4	0.4761
Nonlinear	1.80	3	0.6147
Nonlinear Interaction : f(A,B) vs. AB	1.80	3	0.6147
sex × pclass × age (Factor+Higher Order Factors)	8.34	8	0.4006
Nonlinear	7.74	6	0.2581
TOTAL NONLINEAR	28.66	24	0.2331
TOTAL INTERACTION	75.61	30	< 0.0001
TOTAL NONLINEAR + INTERACTION	79.49	33	< 0.0001
TOTAL	241.93	39	< 0.0001

```
Frequencies of Missing Values Due to Each Variable
survived      sex   pclass      age    sibsp
       0        0        0      263        0
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	1046	LR χ^2	553.87	R^2	0.555	C	0.878
0	619	d.f.	26	g	2.427	D_{xy}	0.756
1	427	Pr($> \chi^2$)	< 0.0001	g_r	11.325	γ	0.758
max |deriv|	6×10^{-6}			g_p	0.365	τ_a	0.366
				Brier	0.130		

	Coef	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 * pclass=2nd	-0.7682	0.7083	-1.08	0.2781
sex=male * pclass=3rd	2.1520	0.6214	3.46	0.0005
sex=male * age	-0.2191	0.0722	-3.04	0.0024
sex=male * age'	1.0842	0.3886	2.79	0.0053
sex=male * age''	-6.5578	2.6511	-2.47	0.0134
sex=male * age'''	8.3716	3.8532	2.17	0.0298
pclass=2nd * age	-0.5446	0.2653	-2.05	0.0401
pclass=3rd * age	-0.1634	0.1308	-1.25	0.2118
pclass=2nd * age'	1.9156	1.0189	1.88	0.0601
pclass=3rd * age'	0.8205	0.6091	1.35	0.1780
pclass=2nd * age''	-8.9545	5.5027	-1.63	0.1037
pclass=3rd * age''	-5.4276	3.6475	-1.49	0.1367
pclass=2nd * age'''	9.3926	6.9559	1.35	0.1769
pclass=3rd * age'''	7.5403	4.8519	1.55	0.1202
age * sibsp	0.0357	0.0340	1.05	0.2933
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

```
latex(anova(f), file='', label='titanic-anova2')   # Table 6.3
```


Table 6.3: Wald Statistics for `survived`

	χ^2	*d.f.*	*P*
sex (Factor+Higher Order Factors)	199.42	7	< 0.0001
All Interactions	56.14	6	< 0.0001
pclass (Factor+Higher Order Factors)	108.73	12	< 0.0001
All Interactions	42.83	10	< 0.0001
age (Factor+Higher Order Factors)	47.04	20	0.0006
All Interactions	24.51	16	0.0789
Nonlinear (Factor+Higher Order Factors)	22.72	15	0.0902
sibsp (Factor+Higher Order Factors)	19.95	5	0.0013
All Interactions	10.99	4	0.0267
sex × pclass (Factor+Higher Order Factors)	35.40	2	< 0.0001
sex × age (Factor+Higher Order Factors)	10.08	4	0.0391
Nonlinear	8.17	3	0.0426
Nonlinear Interaction : f(A,B) vs. AB	8.17	3	0.0426
pclass × age (Factor+Higher Order Factors)	6.86	8	0.5516
Nonlinear	6.11	6	0.4113
Nonlinear Interaction : f(A,B) vs. AB	6.11	6	0.4113
age × sibsp (Factor+Higher Order Factors)	10.99	4	0.0267
Nonlinear	1.81	3	0.6134
Nonlinear Interaction : f(A,B) vs. AB	1.81	3	0.6134
TOTAL NONLINEAR	22.72	15	0.0902
TOTAL INTERACTION	67.58	18	< 0.0001
TOTAL NONLINEAR + INTERACTION	70.68	21	< 0.0001
TOTAL	253.18	26	< 0.0001

Show the many effects of predictors.

```
p ← Predict(f, age, pclass, sex, fun=plogis)
plot(p, adj.subtitle=FALSE)        # Fig. 6.5
# To take control of panel vs groups assignment use:
# plot(p, ~ age | sex, groups='pclass', adj.subtitle=FALSE)
```

```
plot(Predict(f, sibsp, age=c(10,15,20,50), conf.int=FALSE)) # Fig. 6.6
```

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

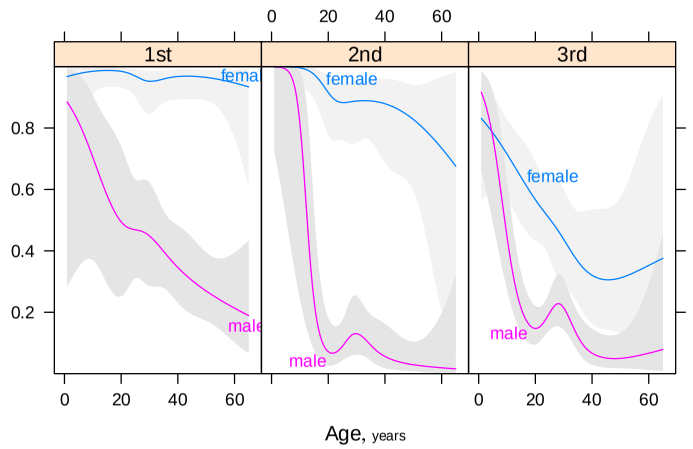


Figure 6.5: Effects of predictors on probability of survival of Titanic passengers, estimated for zero siblings or spouses. Lines for females are black; males are drawn using gray scale.

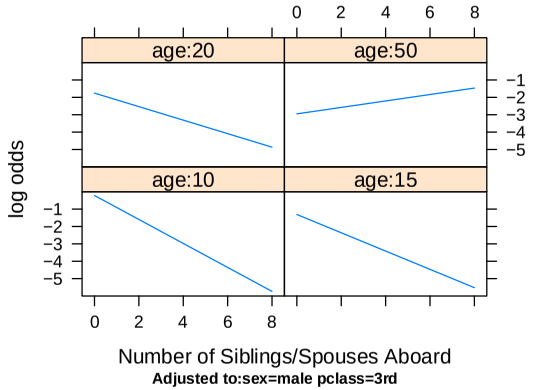


Figure 6.6: Effect of number of siblings and spouses on the log odds of surviving, for third class males. Numbers next to lines are ages in years.

tests.

```
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=80), digits=2, size='Ssize')
```

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.76	0.77	0.74	0.03	0.72	80
R^2	0.55	0.58	0.53	0.05	0.50	80
Intercept	0.00	0.00	−0.09	0.09	−0.09	80
Slope	1.00	1.00	0.86	0.14	0.86	80
$E_{\max}$	0.00	0.00	0.05	0.05	0.05	80
D	0.53	0.56	0.49	0.07	0.46	80
U	0.00	0.00	0.01	−0.01	0.01	80
Q	0.53	0.56	0.49	0.08	0.45	80
B	0.13	0.12	0.13	−0.01	0.14	80
g	2.43	2.79	2.38	0.40	2.02	80
g_p	0.37	0.37	0.35	0.02	0.35	80

```
cal ← calibrate(f, B=80)       # Figure 6.7
plot(cal)
```

```
n=1046   Mean absolute error=0.012   Mean squared error=0.00018
0.9 Quantile of absolute error=0.018
```

But moderate problem with missing data

6.4 Examining Missing Data Patterns

```
na.patterns ← naclus(titanic3)
require(rpart)       # Recursive partitioning package
who.na ← rpart(is.na(age) ~ sex + pclass + survived +
               sibsp + parch, minbucket=15)
naplot(na.patterns, 'na per var')
plot(na.patterns)
options(digits=5)
plot(who.na, margin=.1); text(who.na) # Figure 6.8
```

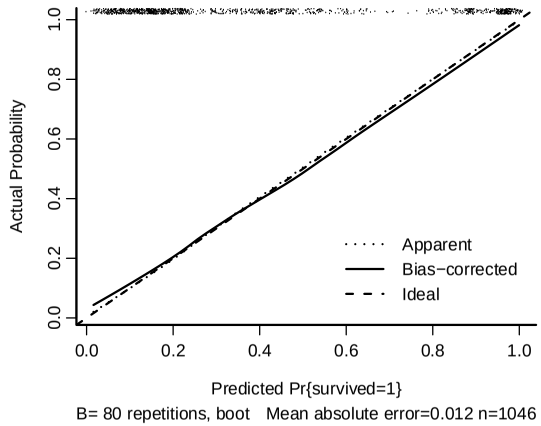


Figure 6.7: Bootstrap overfitting-corrected loess nonparametric calibration curve for casewise deletion model

```
plot(summary(is.na(age) ~ sex + pclass + survived +
             sibsp + parch))  # Figure 6.9
```

```
m ← lrm(is.na(age) ~ sex * pclass + survived + sibsp + parch)
print(m, latex=TRUE)
```

Logistic Regression Model

```
lrm(formula = is.na(age) ~ sex * pclass + survived + sibsp +
    parch)
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	1309	LR χ^2	114.99	R^2	0.133	C	0.703
FALSE	1046	d.f.	8	g	1.015	D_{xy}	0.406
TRUE	263	Pr($> \chi^2$)	< 0.0001	g_r	2.759	γ	0.452
max |deriv|	5×10^{-6}			g_p	0.126	τ_a	0.131
				Brier	0.148		

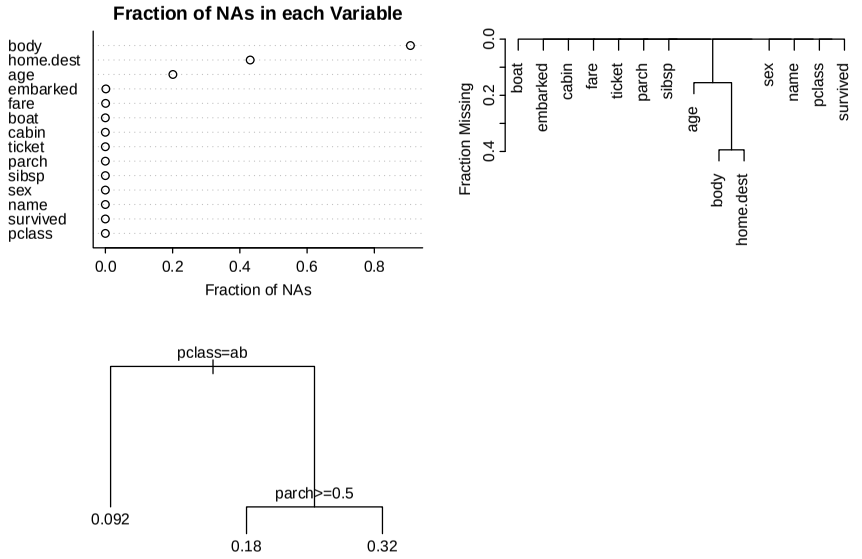


Figure 6.8: Patterns of missing data. Upper left panel shows the fraction of observations missing on each predictor. Upper right panel depicts a hierarchical cluster analysis of missingness combinations. The similarity measure shown on the Y-axis is the fraction of observations for which both variables are missing. Lower left panel shows the result of recursive partitioning for predicting `is.na(age)`. The `rpart` function found only strong patterns according to passenger class.

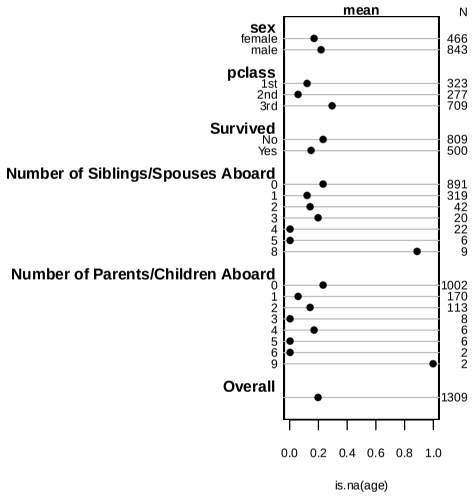


Figure 6.9: Univariable descriptions of proportion of passengers with missing age

	Coef	S.E.	Wald Z	Pr($>\|Z\|$)
Intercept	-2.2030	0.3641	-6.05	<0.0001
sex=male	0.6440	0.3953	1.63	0.1033
pclass=2nd	-1.0079	0.6658	-1.51	0.1300
pclass=3rd	1.6124	0.3596	4.48	<0.0001
survived	-0.1806	0.1828	-0.99	0.3232
sibsp	0.0435	0.0737	0.59	0.5548
parch	-0.3526	0.1253	-2.81	0.0049
sex=male * pclass=2nd	0.1347	0.7545	0.18	0.8583
sex=male * pclass=3rd	-0.8563	0.4214	-2.03	0.0422

```
latex(anova(m), file='', label='titanic-anova.na')    # Table 6.4
```

`pclass` and `parch` are the important predictors of missing age.

Table 6.4: Wald Statistics for `is.na(age)`

	χ^2	$d.f.$	P
sex (Factor+Higher Order Factors)	5.61	3	0.1324
All Interactions	5.58	2	0.0614
pclass (Factor+Higher Order Factors)	68.43	4	<0.0001
All Interactions	5.58	2	0.0614
survived	0.98	1	0.3232
sibsp	0.35	1	0.5548
parch	7.92	1	0.0049
sex × pclass (Factor+Higher Order Factors)	5.58	2	0.0614
TOTAL	82.90	8	<0.0001

6.5 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=titanic3)
summary(xtrans)
```

```
transcan(x = ~I(age) + sex + pclass + sibsp + parch, imputed = TRUE,
    pr = FALSE, pl = FALSE, data = titanic3)

Iterations: 5

R^2 achieved in predicting each variable:

   age    sex pclass  sibsp  parch
 0.258  0.078  0.244  0.241  0.288

Adjusted R^2:

   age    sex pclass  sibsp  parch
 0.254  0.074  0.240  0.238  0.285

Coefficients of canonical variates for predicting each (row) variable
```

```
       age   sex   pclass sibsp parch
age           0.89 -6.13  -1.81 -2.77
sex     0.02        0.56  -0.10 -0.71
pclass -0.08  0.26        -0.07 -0.25
sibsp  -0.02 -0.04 -0.07         0.87
parch  -0.03 -0.29 -0.22   0.75

Summary of imputed values

age
      n missing  unique    Mean     .05     .10     .25
    263       0      24   28.41   16.76   21.66   26.17
    .50     .75     .90     .95
  28.04   28.04   42.92   42.92

lowest :  7.563  9.425 14.617 16.479 16.687
highest: 33.219 34.749 38.588 41.058 42.920

Starting estimates for imputed values:

   age    sex pclass  sibsp  parch
    28      2      3      0      0
```

```
  # Look at mean imputed values by sex,pclass and observed means
  # age.i is age, filled in with conditional mean estimates
  age.i ← impute(xtrans, age, data=titanic3)
  i ← is.imputed(age.i)
  tapply(age.i[i], list(sex[i],pclass[i]), mean)
```

```
         1st    2nd    3rd
female 39.137 31.357 22.926
male   42.920 33.219 26.715
```

```
  tapply(age, list(sex,pclass), mean, na.rm=TRUE)
```

```
         1st    2nd    3rd
female 37.038 27.499 22.185
male   41.029 30.815 25.962
```

```
  dd   ← datadist(dd, age.i)
  f.si ← lrm(survived ~ (sex + pclass + rcs(age.i,5))^2 +
             rcs(age.i,5)*sibsp)
  print(f.si, coefs=FALSE, latex=TRUE)
```

Logistic Regression Model

```
lrm(formula = survived ~ (sex + pclass + rcs(age.i, 5))^2 + rcs(age.i,
```


Table 6.5: Wald Statistics for survived

	χ^2	*d.f.*	*P*
sex (Factor+Higher Order Factors)	245.53	7	< 0.0001
All Interactions	52.80	6	< 0.0001
pclass (Factor+Higher Order Factors)	112.02	12	< 0.0001
All Interactions	36.77	10	0.0001
age.i (Factor+Higher Order Factors)	49.25	20	0.0003
All Interactions	25.53	16	0.0610
Nonlinear (Factor+Higher Order Factors)	19.86	15	0.1772
sibsp (Factor+Higher Order Factors)	21.74	5	0.0006
All Interactions	12.25	4	0.0156
sex × pclass (Factor+Higher Order Factors)	30.25	2	< 0.0001
sex × age.i (Factor+Higher Order Factors)	8.95	4	0.0622
Nonlinear	5.63	3	0.1308
Nonlinear Interaction : f(A,B) vs. AB	5.63	3	0.1308
pclass × age.i (Factor+Higher Order Factors)	6.04	8	0.6427
Nonlinear	5.44	6	0.4882
Nonlinear Interaction : f(A,B) vs. AB	5.44	6	0.4882
age.i × sibsp (Factor+Higher Order Factors)	12.25	4	0.0156
Nonlinear	2.04	3	0.5639
Nonlinear Interaction : f(A,B) vs. AB	2.04	3	0.5639
TOTAL NONLINEAR	19.86	15	0.1772
TOTAL INTERACTION	66.83	18	< 0.0001
TOTAL NONLINEAR + INTERACTION	69.48	21	< 0.0001
TOTAL	305.58	26	< 0.0001

```
5) * sibsp)
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	1309	LR χ^2	641.01	R^2	0.526	C	0.861
0	809	d.f.	26	g	2.227	D_{xy}	0.722
1	500	$\Pr(> \chi^2) < 0.0001$		g_r	9.272	γ	0.728
max |deriv|	4×10^{-4}			g_p	0.346	τ_a	0.341
				Brier	0.133		

```
p1 ← Predict(f,    age,   pclass, sex, fun=plogis)
p2 ← Predict(f.si, age.i, pclass, sex, fun=plogis)
p  ← rbind('Casewise Deletion'=p1, 'Single Imputation'=p2,
           rename=c(age.i='age'))   # creates .set. variable
plot(p, ~ age | pclass*.set., groups='sex',
     ylab='Probability of Surviving', adj.subtitle=FALSE)
# Figure 6.10
```

```
latex(anova(f.si), file='', label='titanic-anova.si')   # Table 6.5
```

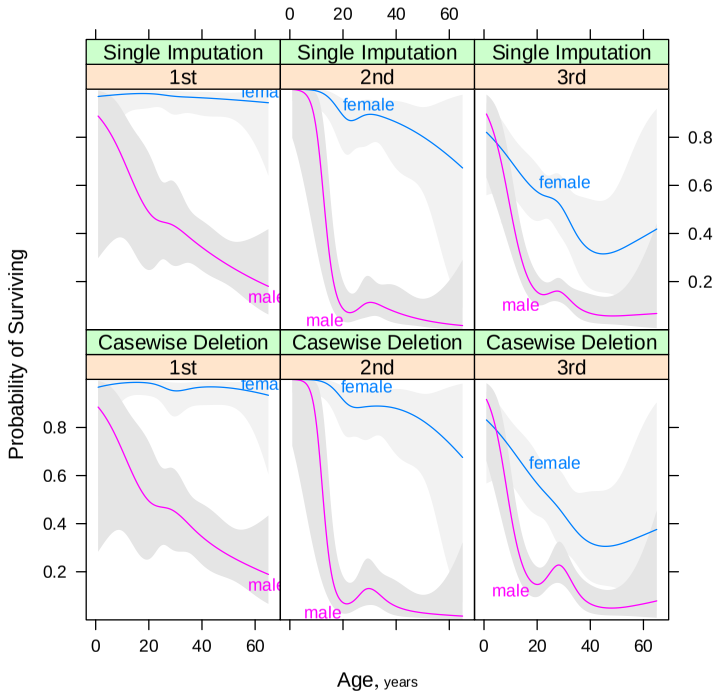


Figure 6.10: Predicted probability of survival for males from fit using casewise deletion (left panel) and single conditional mean imputation (right panel). `sibsp` is set to zero for these predicted values.

6.6 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).

```
set.seed(17)          # so can reproduce random aspects
mi ← aregImpute(~ age + sex + pclass +
                  sibsp + parch + survived,
                  n.impute=5, nk=4, pr=FALSE)
mi
```

```
Multiple Imputation using Bootstrap and PMM

aregImpute(formula = ~age + sex + pclass + sibsp + parch + survived,
    n.impute = 5, nk = 4, pr = FALSE)

n: 1309         p: 6    Imputations: 5          nk: 4

Number of NAs:
     age      sex   pclass    sibsp    parch survived
     263        0        0        0        0        0

          type d.f.
age          s    1
sex          c    1
pclass       c    2
sibsp        s    2
parch        s    2
survived     l    1

Transformation of Target Variables Forced to be Linear

R-squares for Predicting Non-Missing Values for Each Variable
Using Last Imputations of Predictors
  age
```

```
0.344
```

```
# Print the 5 imputations for the first 10 passengers
#  having missing age
mi$imputed$age[1:10,]
```

```
    [,1] [,2] [,3] [,4] [,5]
16  28.5 60.0 32.5   46   71
38  26.0 26.0 29.0   49   51
41  47.0 62.0 47.0   55   42
47  45.0 47.0 17.0   46   39
60  39.0 27.0 42.0   39   18
70  39.0 39.0 23.0   30   41
71  29.0 42.0 47.0   47   61
75  46.0 28.5 32.5   17   36
81  47.0 48.0 30.0   55   40
107 62.0 50.0 23.0   33   17
```

Show the distribution of imputed (black) and actual ages (gray).

```
plot(mi)
Ecdf(age, add=TRUE, col='gray', lwd=2, subtitles=FALSE) # Figure 6.11
```

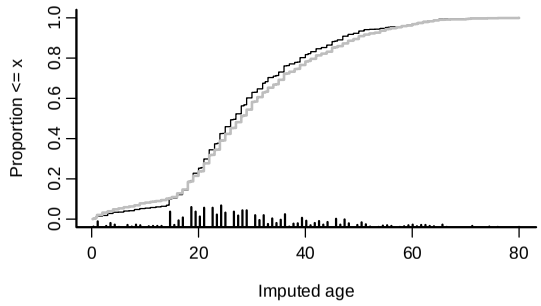


Figure 6.11: Distributions of imputed and actual ages for the Titanic dataset

Fit logistic models for 5 completed datasets and print the ratio of imputation-corrected variances to aver-

Table 6.6: Wald Statistics for `survived`

	χ^2	*d.f.*	*P*
sex (Factor+Higher Order Factors)	236.24	7	< 0.0001
All Interactions	52.20	6	< 0.0001
pclass (Factor+Higher Order Factors)	109.82	12	< 0.0001
All Interactions	37.09	10	0.0001
age (Factor+Higher Order Factors)	49.09	20	0.0003
All Interactions	22.73	16	0.1211
Nonlinear (Factor+Higher Order Factors)	21.38	15	0.1251
sibsp (Factor+Higher Order Factors)	23.68	5	0.0003
All Interactions	11.00	4	0.0266
sex × pclass (Factor+Higher Order Factors)	33.48	2	< 0.0001
sex × age (Factor+Higher Order Factors)	9.22	4	0.0559
Nonlinear	7.18	3	0.0663
Nonlinear Interaction : f(A,B) vs. AB	7.18	3	0.0663
pclass × age (Factor+Higher Order Factors)	3.66	8	0.8861
Nonlinear	3.27	6	0.7739
Nonlinear Interaction : f(A,B) vs. AB	3.27	6	0.7739
age × sibsp (Factor+Higher Order Factors)	11.00	4	0.0266
Nonlinear	1.90	3	0.5925
Nonlinear Interaction : f(A,B) vs. AB	1.90	3	0.5925
TOTAL NONLINEAR	21.38	15	0.1251
TOTAL INTERACTION	65.11	18	< 0.0001
TOTAL NONLINEAR + INTERACTION	68.89	21	< 0.0001
TOTAL	302.90	26	< 0.0001

age ordinary variances

```
f.mi ← fit.mult.impute(survived ~ (sex + pclass + rcs(age,5))^2 +
                       rcs(age,5)*sibsp,
                       lrm, mi, data=titanic3, pr=FALSE)
latex(anova(f.mi), file='', label='titanic-anova.mi')    # Table 6.6
```

The Wald χ^2 for age is reduced by accounting for imputation but is increased by using patterns of association with survival status to impute missing age.

Show estimated effects of age by classes.

```
p1 ← Predict(f.si,  age.i, pclass, sex, fun=plogis)
p2 ← Predict(f.mi,  age,   pclass, sex, fun=plogis)
p  ← rbind('Single Imputation'=p1, 'Multiple Imputation'=p2,
           rename=c(age.i='age'))
plot(p, ~ age | pclass*.set., groups='sex',
```

```
          ylab='Probability of Surviving', adj.subtitle=FALSE)
# Figure 6.12
```

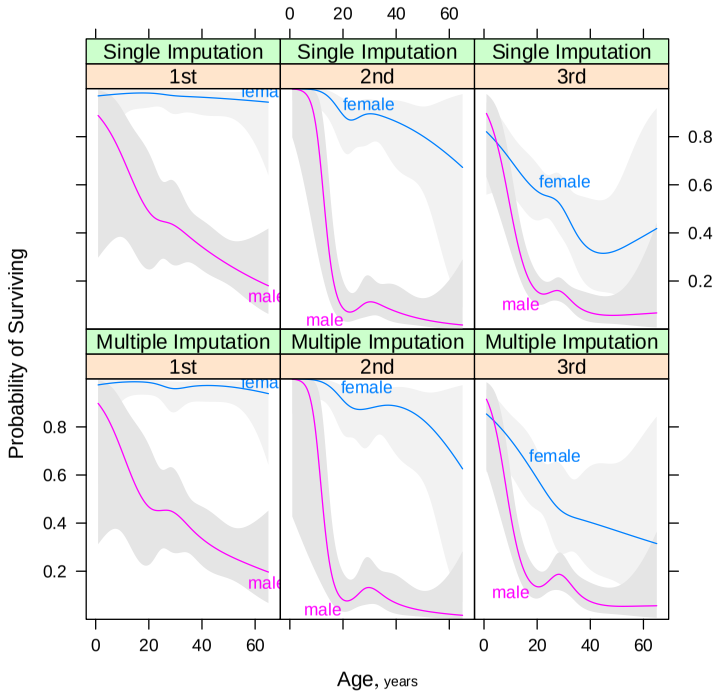


Figure 6.12: Predicted probability of survival for males from fit using single conditional mean imputation again (left panel) and multiple random draw imputation (right panel). Both sets of predictions are for `sibsp=0`.

6.7 Summarizing the Fitted Model

Show odds ratios for changes in predictor values


```
s ← summary(f.mi, age=c(1,30), sibsp=0:1)
# override default ranges for 3 variables
plot(s, log=TRUE, main='')              # Figure 6.13
```

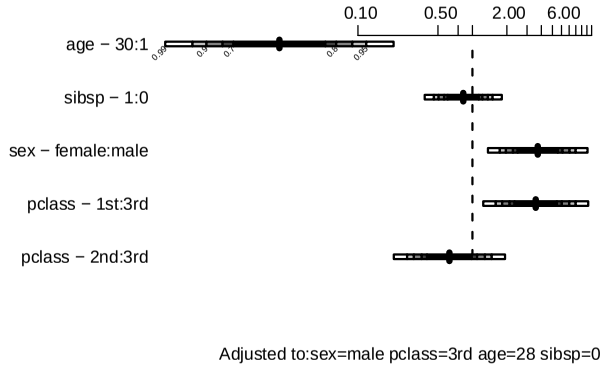


Figure 6.13: Odds ratios for some predictor settings

Get predicted values for certain types of passengers

```
phat ← predict(f.mi,
               combos ←
         expand.grid(age=c(2,21,50),sex=levels(sex),
                     pclass=levels(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)
```

	age	sex	pclass	sibsp	phat
1	2	female	1st	0	0.98
2	21	female	1st	0	0.98
3	50	female	1st	0	0.97
4	2	male	1st	0	0.88
5	21	male	1st	0	0.46
6	50	male	1st	0	0.27
7	2	female	2nd	0	1.00
8	21	female	2nd	0	0.90
9	50	female	2nd	0	0.83
10	2	male	2nd	0	1.00
11	21	male	2nd	0	0.08

```
12   50    male    2nd     0 0.04
13    2  female    3rd     0 0.84
14   21  female    3rd     0 0.57
15   50  female    3rd     0 0.37
16    2    male    3rd     0 0.89
17   21    male    3rd     0 0.14
18   50    male    3rd     0 0.05
```

```
options(digits=5)
```

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')
```

```
pred.logit ← function (sex = "male", pclass = "3rd", age = 28, sibsp = 0)
{
    3.5810728 - 1.2694669 * (sex == "male") + 5.227106 * (pclass == "2nd") -
        1.7471648 * (pclass == "3rd") + 0.072213655 * age - 0.00021294639 *
        pmax(age - 4, 0)^3 + 0.0015984839 * pmax(age - 21, 0)^3 - 0.0023265999 *
        pmax(age - 28, 0)^3 + 0.0010212127 * pmax(age - 36.15, 0)^3 - 8.0150336e-05 *
        pmax(age - 56, 0)^3 - 1.1339431 * sibsp + (sex == "male") * (-0.46284486 *
        (pclass == "2nd") + 2.0884806 * (pclass == "3rd")) + (sex == "male") *
        (-0.22398928 * age + 0.0003578076 * pmax(age - 4, 0)^3 - 0.002354863 *
            pmax(age - 21, 0)^3 + 0.0032067241 * pmax(age - 28, 0)^3 -
            0.0013085171 * pmax(age - 36.15, 0)^3 + 9.8848428e-05 * pmax(age -
            56, 0)^3) + (pclass == "2nd") * (-0.4600114 * age + 0.00052411339 *
        pmax(age - 4, 0)^3 - 0.0025239553 * pmax(age - 21, 0)^3 + 0.0026577424 *
        pmax(age - 28, 0)^3 - 0.00067164981 * pmax(age - 36.15, 0)^3 +
        1.3749304e-05 * pmax(age - 56, 0)^3) + (pclass == "3rd") * (-0.14784979 *
        age + 0.00021831279 * pmax(age - 4, 0)^3 - 0.001437761 * pmax(age -
        21, 0)^3 + 0.0020012161 * pmax(age - 28, 0)^3 - 0.00085968161 *
        pmax(age - 36.15, 0)^3 + 7.7913743e-05 * pmax(age - 56, 0)^3) +
        sibsp * (0.045169115 * age - 2.90579e-05 * pmax(age - 4, 0)^3 +
            0.00025289589 * pmax(age - 21, 0)^3 - 0.00048983359 * pmax(age -
            28, 0)^3 + 0.00032115845 * pmax(age - 36.15, 0)^3 - 5.5162848e-05 *
            pmax(age - 56, 0)^3)
}
```

```
# Run the newly created function
plogis(pred.logit(age=c(2,21,50), sex='male', pclass='3rd'))
```

```
[1] 0.886318 0.135294 0.054266
```


A nomogram could be used to obtain predicted values manually, but this is not feasible when so many interaction terms are present.

R/ S-PLUS Software Used

Package	Purpose	Functions
Hmisc	Miscellaneous functions	summary,plsmo,naclus,llist,latex
		summarize,Dotplot,describe,dataRep
Hmisc	Imputation	transcan,impute,fit.mult.impute,aregImpute
rms	Modeling	datadist,lrm,rcs
	Model presentation	plot,summary,nomogram,Function
	Model validation	validate,calibrate
rpart[a]	Recursive partitioning	rpart

[a]Written by Atkinson & Therneau

Chapter 7

Case Study in Parametric Survival Modeling and Model Approximation

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 [70]. The dataset is available from `http://biostat.mc.vanderbilt.edu/DataSets`.

- 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

7.1 Descriptive Statistics

Create a variable `acute` to flag categories of interest; print univariable descriptive statistics.

```
require(rms)
getHdata(support)      # Get data frame from web site
acute ← support$dzclass %in% c('ARF/MOSF','Coma')
latex(describe(support[acute,]), file='')
```

support[acute,]

35 Variables 537 Observations

age : Age

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	529	60.7	28.49	35.22	47.93	63.67	74.49	81.54	85.56

lowest : 18.04 18.41 19.76 20.30 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	unique	Sum	Mean
537	0	2	356	0.6629

sex

n	missing	unique
537	0	2

female (251, 47%), male (286, 53%)

hospdead : Death in Hospital

n	missing	unique	Sum	Mean
537	0	2	201	0.3743

slos : Days from Study Entry to Discharge

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	85	23.44	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	unique	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	340	446.1	4	6	16	182	724	1421	1742

lowest : 3 4 5 6 7, highest: 1977 1979 1982 2011 2022

dzgroup

n	missing	unique
537	0	3

ARF/MOSF w/Sepsis (391, 73%), Coma (60, 11%), MOSF w/Malig (86, 16%)

dzclass

n	missing	unique
537	0	2

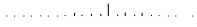
ARF/MOSF (477, 89%), Coma (60, 11%)

num.co : number of comorbidities

n	missing	unique	Mean
537	0	7	1.525

	0	1	2	3	4	5	6
Frequency	111	196	133	51	31	10	5
%	21	36	25	9	6	2	1

edu : Years of Education



n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
411	126	22	12.03	7	8	10	12	14	16	17

lowest : 0 1 2 3 4, highest: 17 18 19 20 22

income

n	missing	unique
335	202	4

under $11k (158, 47%), $11-$25k (79, 24%), $25-$50k (63, 19%)
>$50k (35, 10%)

scoma : SUPPORT Coma Score based on Glasgow D3

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	11	19.24	0	0	0	0	37	55	100

	0	9	26	37	41	44	55	61	89	94	100
Frequency	301	50	44	19	17	43	11	6	8	6	32
%	56	9	8	4	3	8	2	1	1	1	6

charges : Hospital Charges

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
517	20	516	86652	11075	15180	27389	51079	100904	205562	283411

lowest : 3448 4432 4574 5555 5849
highest: 504660 538323 543761 706577 740010

totcst : Total RCC cost

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
471	66	471	46360	6359	8449	15412	29308	57028	108927	141569

lowest : 0 2071 2522 3191 3325
highest: 269057 269131 338955 357919 390460

totmcst : Total micro-cost

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
331	206	328	39022	6131	8283	14415	26323	54102	87495	111920

lowest : 0 1562 2478 2626 3421
highest: 144234 154709 198047 234876 271467

avtisst : Average TISS, Days 3-25

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
536	1	205	29.83	12.46	14.50	19.62	28.00	39.00	47.17	50.37

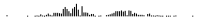
lowest : 4.000 5.667 8.000 9.000 9.500
highest: 58.500 59.000 60.000 61.000 64.000

race

n	missing	unique
535	2	5

	white	black	asian	other	hispanic
Frequency	417	84	4	8	22
%	78	16	1	1	4

meanbp : Mean Arterial Blood Pressure Day 3



n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	109	83.28	41.8	49.0	59.0	73.0	111.0	124.4	135.0

lowest : 0 20 27 30 32, highest: 155 158 161 162 180

wblc : White Blood Cell Count Day 3

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
532	5	241	14.1	0.8999	4.5000	7.9749	12.3984	18.1992	25.1891	30.1873

```
lowest :   0.05000   0.06999   0.09999   0.14999   0.19998
highest:  51.39844  58.19531  61.19531  79.39062 100.00000
```

hrt : Heart Rate Day 3

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	111	105	51	60	75	111	126	140	155

```
lowest :   0  11  30  36  40, highest: 189 193 199 232 300
```

resp : Respiration Rate Day 3

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	45	23.72	8	10	12	24	32	39	40

```
lowest :  0  4  6  7  8, highest: 48 49 52 60 64
```

temp : Temperature (celcius) Day 3

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	61	37.52	35.50	35.80	36.40	37.80	38.50	39.09	39.50

```
lowest : 32.50 34.00 34.09 34.90 35.00
highest: 40.20 40.59 40.90 41.00 41.20
```

pafi : PaO2/(.01*FiO2) Day 3

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
500	37	357	227.2	86.99	105.08	137.88	202.56	290.00	390.49	433.31

```
lowest :  45.00  48.00  53.33  54.00  55.00
highest: 574.00 595.12 640.00 680.00 869.38
```

alb : Serum Albumin Day 3

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
346	191	34	2.668	1.700	1.900	2.225	2.600	3.100	3.400	3.800

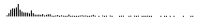
```
lowest : 1.100 1.200 1.300 1.400 1.500
highest: 4.100 4.199 4.500 4.699 4.800
```

bili : Bilirubin Day 3

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
386	151	88	2.678	0.3000	0.4000	0.6000	0.8999	2.0000	6.5996	13.1743

```
lowest :  0.09999  0.19998  0.29999  0.39996  0.50000
highest: 22.59766 30.00000 31.50000 35.00000 39.29688
```

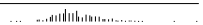

crea : Serum creatinine Day 3



n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	84	2.232	0.6000	0.7000	0.8999	1.3999	2.5996	5.2395	7.3197

```
lowest :  0.3  0.4  0.5  0.6  0.7, highest: 10.4 10.6 11.2 11.6 11.8
```

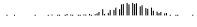
sod : Serum sodium Day 3



n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	38	138.1	129	131	134	137	142	147	150

```
lowest : 118 120 121 126 127, highest: 156 157 158 168 175
```

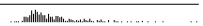
ph : Serum pH (arterial) Day 3



n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
500	37	49	7.416	7.270	7.319	7.380	7.420	7.470	7.510	7.529

```
lowest : 6.960 6.989 7.069 7.119 7.130
highest: 7.560 7.569 7.590 7.600 7.659
```

glucose : Glucose Day 3



n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
297	240	179	167.7	76.0	89.0	106.0	141.0	200.0	292.4	347.2

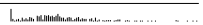
```
lowest :  30  42  52  55  68, highest: 446 468 492 576 598
```

bun : BUN Day 3

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
304	233	100	38.91	8.00	11.00	16.75	30.00	56.00	79.70	100.70

```
lowest :   1   3   4   5   6, highest: 123 124 125 128 146
```

urine : Urine Output Day 3



n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
303	234	262	2095	20.3	364.0	1156.5	1870.0	2795.0	4008.6	4817.5

```
lowest :    0    5    8   15   20, highest: 6865 6920 7360 7560 7750
```

adlp : ADL Patient Day 3

n	missing	unique	Mean
104	433	8	1.577

	0	1	2	3	4	5	6	7
Frequency	51	19	7	6	4	7	8	2
%	49	18	7	6	4	7	8	2

adls : ADL Surrogate Day 3

```
   n   missing   unique    Mean
 392       145        8    1.86


             0  1  2  3  4  5  6  7
Frequency  185 68 22 18 17 20 39 23
%           47 17  6  5  4  5 10  6
```

sfdm2

```
   n   missing   unique
 468        69        5


no(M2 and SIP pres) (134, 29%), adl>=4 (>=5 if sur) (78, 17%)
SIP>=30 (30, 6%), Coma or Intub (5, 1%), <2 mo. follow-up (221, 47%)
```

adlsc : Imputed ADL Calibrated to Surrogate

```
   n   missing   unique    Mean     .05     .10     .25     .50     .75     .90     .95
 537         0      144   2.119   0.000   0.000   0.000   1.839   3.375   6.000   6.000


lowest : 0.0000 0.4948 0.4948 1.0000 1.1667
highest: 5.7832 6.0000 6.3398 6.4658 7.0000
```

```
# Show patterns of missing data
plot(naclus(support[acute,]))       # Figure 7.1
```

Show associations between predictors using a general non-monotonic measure of dependence (Hoeffding D).

```
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+bun+urine+adlsc, sim='hoeffding')
plot(vc)                            # Figure 7.2
```

7.2 Checking Adequacy of Log-Normal Accelerated Failure Time Model

```
dd ← datadist(ac)
# describe distributions of variables to rms
options(datadist='dd')
```

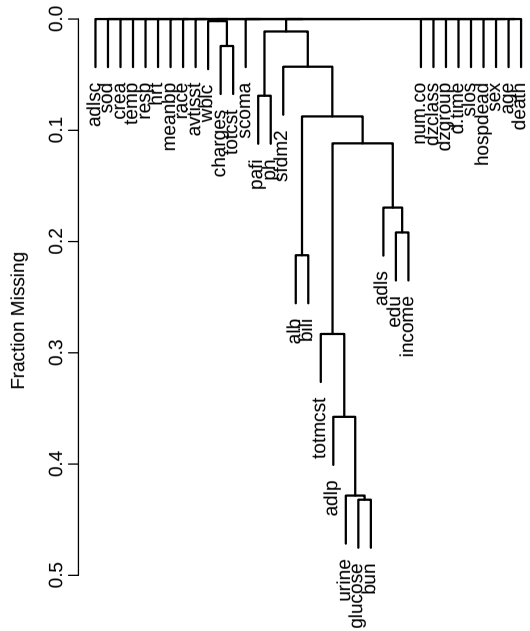



Figure 7.1: Cluster analysis showing which predictors tend to be missing on the same patients

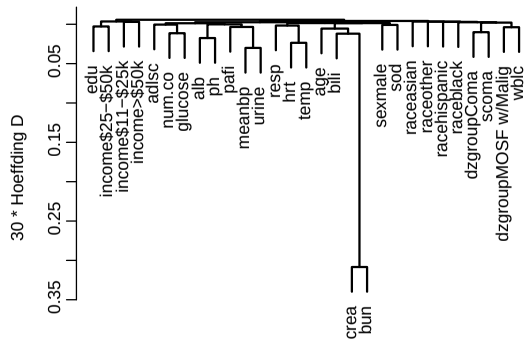


Figure 7.2: Hierarchical clustering of potential predictors using Hoeffding D as a similarity measure. Categorical predictors are automatically expanded into dummy variables.

```
# 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(survfit(S ~ dzgroup), conf='none',
         fun=qnorm, logt=TRUE)    # Figure 7.3
```

More stringent assessment of log-normal assumptions: check distribution of residuals from an adjusted model:

```
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))
survplot(r, random.number, label.curve=FALSE)    # Figure 7.4
```

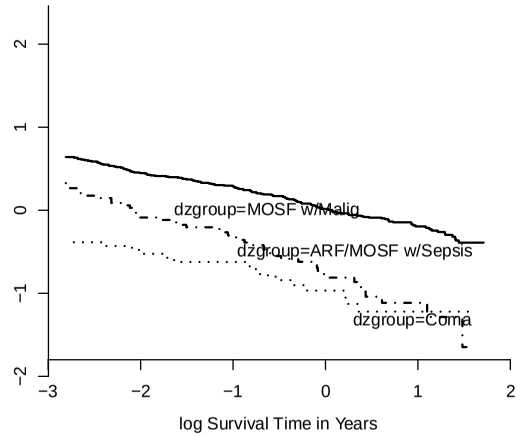



Figure 7.3: $\Phi^{-1}(S_{KM}(t))$ stratified by `dzgroup`. Linearity and semi-parallelism indicate a reasonable fit to the log-normal accelerated failure time model with respect to one predictor.

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 ρ^2 based on multiple linear regression of rank(x), rank(x)^2 and the survival time,

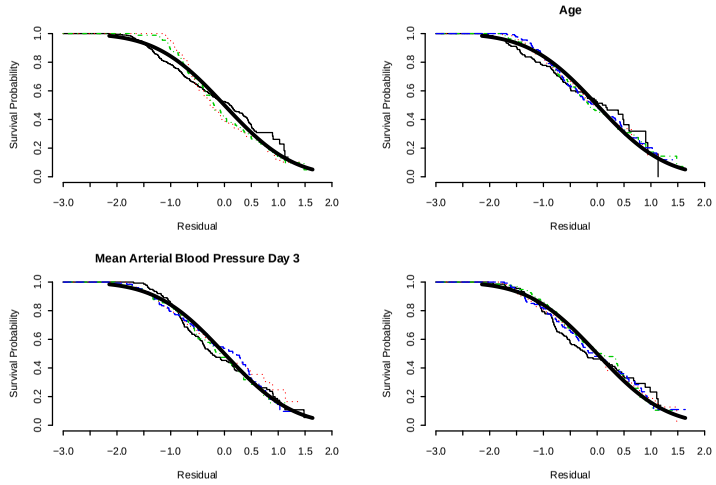


Figure 7.4: Kaplan-Meier estimates of distributions of normalized, right-censored residuals from the fitted log-normal survival model. Residuals are stratified by important variables in the model (by quartiles of continuous variables), plus a random variable to depict the natural variability (in the 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.

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.

```
shortest.follow.up ← min(d.time[death==0], na.rm=TRUE)
d.timet ← pmin(d.time, shortest.follow.up)
w ← spearman2(d.timet ~ age + num.co + scoma + meanbp +
              hrt + resp + temp + crea + sod + adlsc +
              wblc + pafi + ph + dzgroup + race, p=2)
plot(w, main='')          # Figure 7.5
```

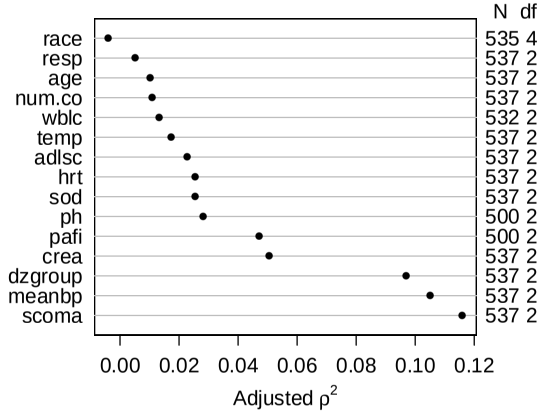


Figure 7.5: Generalized Spearman ρ^2 rank correlation between predictors and truncated survival time

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 ← rcorrcens(S ~ age + num.co + scoma + meanbp + hrt + resp +
                  temp + crea + sod + adlsc + wblc + pafi + ph +
                  dzgroup + race)
```

```
plot(w, main='')          # Figure 7.6
```

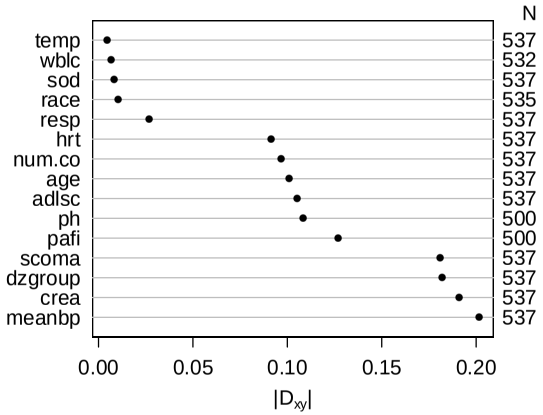


Figure 7.6: Somers' D_{xy} rank correlation between predictors and original survival time. For `dzgroup` or `race`, the correlation coefficient is the maximum correlation from using a dummy variable to represent the most frequent or one to represent the second most frequent category.

```
# Compute number of missing values per variable
sapply(llist(age,num.co,scoma,meanbp,hrt,resp,temp,crea,sod,adlsc,
             wblc,pafi,ph), function(x) sum(is.na(x)))
```

```
  age num.co  scoma meanbp    hrt   resp   temp   crea    sod  adlsc
    0      0      0      0      0      0      0      0      0      0
 wblc   pafi     ph
    5     37     37
```

```
# Can also do naplot(naclus(support[acute,]))
# Can also use the Hmisc naclus and naplot functions to do this
# Impute missing values with normal or modal values
wblc.i ← impute(wblc, 9)
pafi.i ← impute(pafi, 333.3)
ph.i   ← impute(ph,   7.4)
race2  ← race
levels(race2) ← list(white='white',other=levels(race)[-1])
race2[is.na(race2)] ← 'white'
dd ← datadist(dd, wblc.i, pafi.i, ph.i, race2)
```


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.

```
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

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:

    crea      age      sex dzgroup   num.co    scoma    adlsc    race2
   0.133    0.246    0.132    0.451    0.147    0.418    0.153    0.151
  meanbp      hrt     resp     temp      sod   wblc.i   pafi.i    ph.i
   0.178    0.258    0.131    0.197    0.135    0.093    0.143    0.171

No redundant variables
```

Better approach to gauging predictive potential and allocating d.f.:

- Allow all continuous variables to have a the maxi-

mum 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 `scoma` to be linear, and omit `ph.i` to avoid singularity

```
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 7.7
```

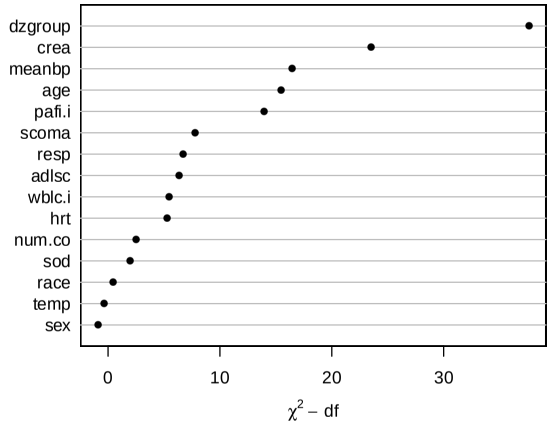


Figure 7.7: Partial χ^2 statistics for association of each predictor with response from saturated main effects model, penalized for d.f.

- Figure 7.7 properly blinds the analyst to the form of effects (tests of linearity).
- Fit a log-normal survival model with number of parameters corresponding to nonlinear effects determined from Figure 7.7. For the most promising predictors, five knots can be allocated, as there are fewer singularity problems once less promising predictors are simplified.

```
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, latex=TRUE)
```

Parametric Survival Model: Log Normal Distribution

```
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")
```

		Model Likelihood Ratio Test		Discrimination Indexes	
Obs	537	LR χ^2	236.83	R^2	0.594
Events	356	d.f.	30	g	1.959
σ	2.2308	$\Pr(>\chi^2) < 0.0001$		g_r	7.095

	Coef	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

	Coef	S.E.	Wald Z	$\Pr(>\|Z\|)$
dzgroup=Coma	-2.0676	0.4062	-5.09	< 0.0001
dzgroup=MOSF 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

7.3 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 χ^2 statistics, penalized for d.f., plotted in

Table 7.2: Wald Statistics for S

	χ^2	$d.f.$	P
age	15.99	4	0.0030
Nonlinear	0.23	3	0.9722
sex	0.11	1	0.7354
dzgroup	45.69	2	< 0.0001
num.co	4.99	1	0.0255
scoma	10.58	1	0.0011
adlsc	8.28	2	0.0159
Nonlinear	3.31	1	0.0691
race2	1.26	1	0.2624
meanbp	27.62	4	< 0.0001
Nonlinear	10.51	3	0.0147
hrt	11.83	2	0.0027
Nonlinear	1.04	1	0.3090
resp	11.10	2	0.0039
Nonlinear	8.56	1	0.0034
temp	0.39	1	0.5308
crea	33.63	3	< 0.0001
Nonlinear	21.27	2	< 0.0001
sod	0.08	1	0.7792
wblc.i	5.47	2	0.0649
Nonlinear	5.46	1	0.0195
pafi.i	15.37	3	0.0015
Nonlinear	6.97	2	0.0307
TOTAL NONLINEAR	60.48	14	< 0.0001
TOTAL	261.47	30	< 0.0001

descending order

```
plot(Predict(f, ref.zero=TRUE))
# Figure 7.8

latex(anova(f), file='', label='support-anovat')    # Table 7.2

plot(anova(f))                 # Figure 7.9

options(digits=3)
plot(summary(f), log=TRUE, main='')    # Figure 7.10
```

7.4 Internal Validation of the Fitted Model Using the Bootstrap

Validate indexes describing the fitted model.

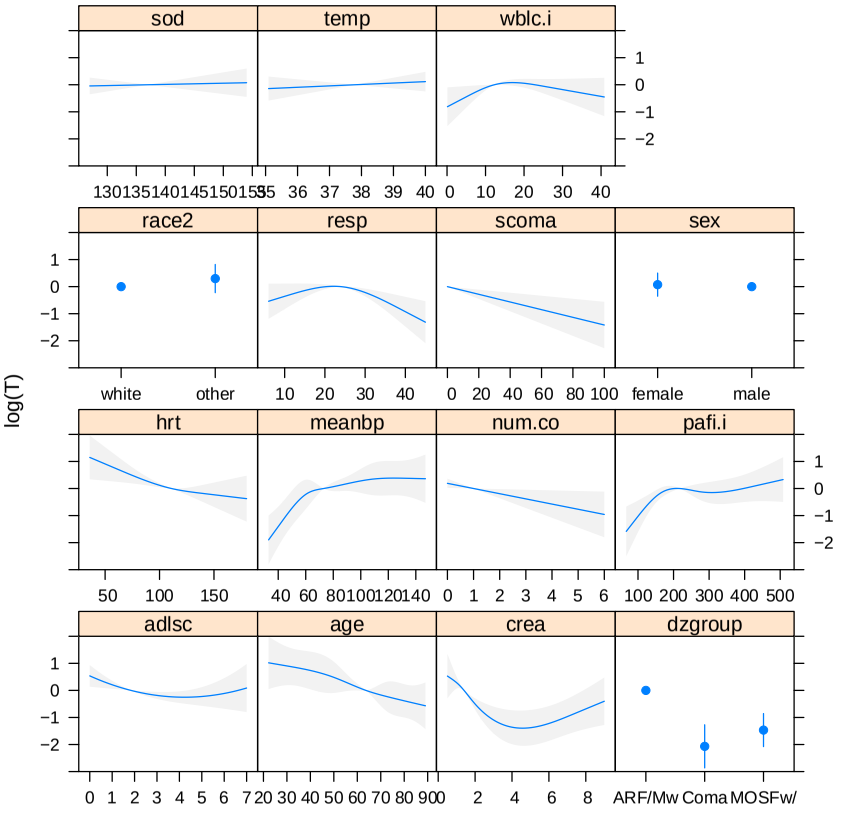


Figure 7.8: Effect of each predictor on log survival time. Predicted values have been centered so that predictions at predictor reference values are zero. Pointwise 0.95 confidence bands are also shown. As all Y-axes have the same scale, it is easy to see which predictors are strongest.

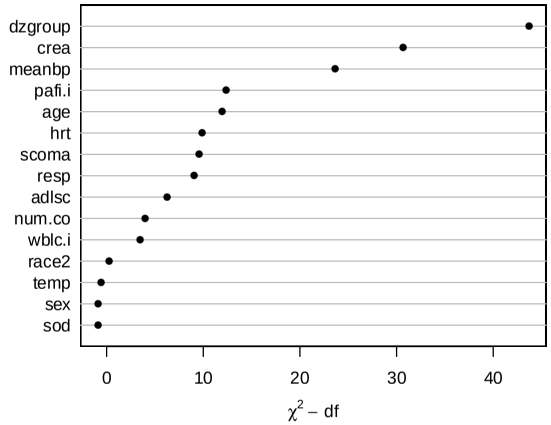


Figure 7.9: Contribution of variables in predicting survival time in log-normal model

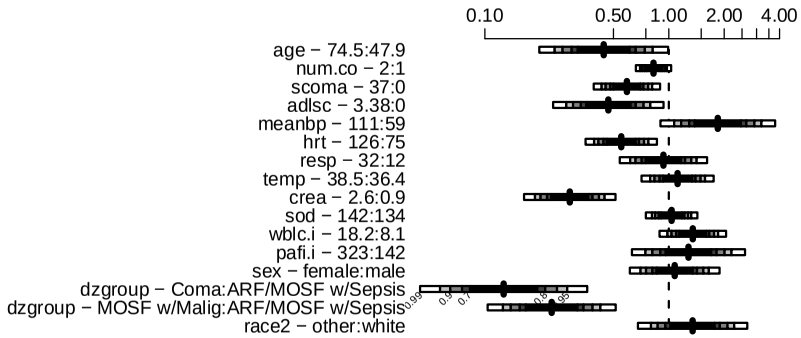


Figure 7.10: Estimated survival time ratios for default settings of predictors. For example, when age changes from its lower quartile to the upper quartile (47.9y to 74.5y), median survival time decreases by more than half. Different shaded areas of bars indicate different confidence levels, ranging from 0.7 to 0.99.

```
# 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')
```

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.49	0.51	0.46	0.05	0.43	120
R^2	0.59	0.66	0.54	0.12	0.47	120
Intercept	0.00	0.00	−0.06	0.06	−0.06	120
Slope	1.00	1.00	0.90	0.10	0.90	120
D	0.48	0.55	0.42	0.13	0.35	120
U	0.00	0.00	−0.01	0.01	−0.01	120
Q	0.48	0.55	0.43	0.12	0.36	120
g	1.96	2.06	1.86	0.19	1.76	120

- 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[71] and use less precise Kaplan-Meier estimates obtained by stratifying patients by the predicted probability, with at least 60 patients per group.

```
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 7.11
```

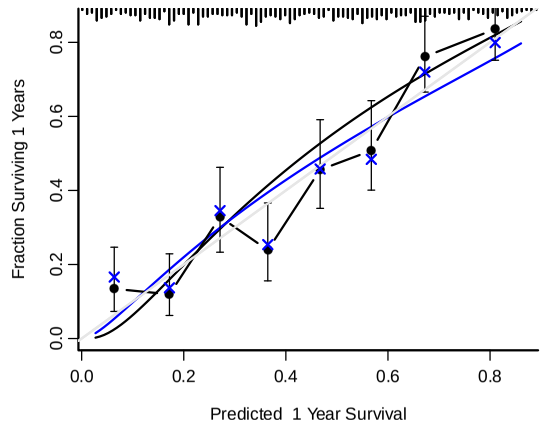


Figure 7.11: Bootstrap validation of calibration curve. Dots represent apparent calibration accuracy; × are bootstrap estimates corrected for overfitting, based on binning predicted survival probabilities and and computing Kaplan-Meier estimates. Black curve is the estimated observed relationship using `hare` and the blue curve is the overfitting-corrected `hare` estimate. The gray-scale line depicts the ideal relationship.

7.5 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 ← predict(f)     # X*beta hat
a ← ols(Z ~ 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), sigma=1)
# sigma=1 is used to prevent sigma hat from being zero when
# R2=1.0 since we start out by approximating Z with all
#  component variables
fastbw(a, aics=10000)     # fast backward stepdown
```

```
 Deleted Chi-Sq d.f. P      Residual d.f. P      AIC      R2
 sod        0.43 1    0.512    0.43   1   0.5117   -1.57 1.000
 sex        0.57 1    0.451    1.00   2   0.6073   -3.00 0.999
 temp       2.20 1    0.138    3.20   3   0.3621   -2.80 0.998
 race2      6.81 1    0.009   10.01   4   0.0402    2.01 0.994
 wblc.i    29.52 2    0.000   39.53   6   0.0000   27.53 0.976
 num.co    30.84 1    0.000   70.36   7   0.0000   56.36 0.957
 resp      54.18 2    0.000  124.55   9   0.0000  106.55 0.924
 adlsc     52.46 2    0.000  177.00  11   0.0000  155.00 0.892
 pafi.i    66.78 3    0.000  243.79  14   0.0000  215.79 0.851
 scoma     78.07 1    0.000  321.86  15   0.0000  291.86 0.803
 hrt       83.17 2    0.000  405.02  17   0.0000  371.02 0.752
 age       68.08 4    0.000  473.10  21   0.0000  431.10 0.710
 crea     314.47 3    0.000  787.57  24   0.0000  739.57 0.517
 meanbp   403.04 4    0.000 1190.61  28   0.0000 1134.61 0.270
 dzgroup  441.28 2    0.000 1631.89  30   0.0000 1571.89 0.000

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
```

```
        n Model L.R.       d.f.         R2          g      Sigma
```



```
  537.000    1688.225      23.000       0.957       1.915       0.370
```

- Estimate variance–covariance matrix of the coefficients of reduced model
- This covariance matrix does not include the scale parameter

```
V ← vcov(f,regcoef.only=TRUE)        # var(full model)
X ← g$x                              # full model design
x ← f.approx$x                       # approx. model design
w ← solve(t(x) %*% x, t(x)) %*% X    # contrast matrix
v ← w %*% V %*% t(w)
```

Compare variance estimates (diagonals of `v`) with variance estimates from a reduced model that is fitted against the actual outcomes.

```
f.sub ← psm(S ~ dzgroup + rcs(meanbp,5) + rcs(crea,4) + rcs(age,5) +
              rcs(hrt,3) + scoma + rcs(pafi.i,4) + pol(adlsc,2)+
              rcs(resp,3), dist='lognormal')  # 'gaussian' for S+
diag(v)/diag(vcov(f.sub,regcoef.only=TRUE))
```

```
          Intercept dzgroup=Coma dzgroup=MOSF w/Malig
              0.981          0.979                0.979
            meanbp         meanbp'              meanbp''
              0.977          0.979                0.979
          meanbp'''           crea                crea'
              0.979          0.979                0.979
           crea''             age                 age'
              0.979          0.982                0.981
            age''           age'''                 hrt
              0.981          0.980                0.978
             hrt'           scoma               pafi.i
              0.976          0.979                0.980
          pafi.i'         pafi.i''               adlsc
              0.980          0.980                0.981
          adlsc^2           resp                 resp'
```

Table 7.3: Wald Statistics for Z

	χ^2	$d.f.$	P
dzgroup	55.94	2	< 0.0001
meanbp	29.87	4	< 0.0001
Nonlinear	9.84	3	0.0200
crea	39.04	3	< 0.0001
Nonlinear	24.37	2	< 0.0001
age	18.12	4	0.0012
Nonlinear	0.34	3	0.9517
hrt	9.87	2	0.0072
Nonlinear	0.40	1	0.5289
scoma	9.85	1	0.0017
pafi.i	14.01	3	0.0029
Nonlinear	6.66	2	0.0357
adlsc	9.71	2	0.0078
Nonlinear	2.87	1	0.0904
resp	9.65	2	0.0080
Nonlinear	7.13	1	0.0076
TOTAL NONLINEAR	58.08	13	< 0.0001
TOTAL	252.32	23	< 0.0001

```
0.981                0.978                0.977
```

The ratios ranged from 0.978 to 0.982.

```
f.approx$var ← v
latex(anova(f.approx, test='Chisq', ss=FALSE), file='',
      label='suport.anovaa')
```

Equation for simplified model:

```
# Typeset mathematical form of approximate model
latex(f.approx, file='')
```

$$\mathrm{E(Z)} = X\beta, \quad \text{where}$$

$X\hat{\beta} =$

-2.51
$-1.94\{\text{Coma}\} - 1.75\{\text{MOSF w/Malig}\}$
$+0.068\text{meanbp} - 3.08\times10^{-5}(\text{meanbp}-41.8)^3_+ + 7.9\times10^{-5}(\text{meanbp}-61)^3_+$
$-4.91\times10^{-5}(\text{meanbp}-73)^3_+ + 2.61\times10^{-6}(\text{meanbp}-109)^3_+ - 1.7\times10^{-6}(\text{meanbp}-135)^3_+$
$-0.553\text{crea} - 0.229(\text{crea}-0.6)^3_+ + 0.45(\text{crea}-1.1)^3_+ - 0.233(\text{crea}-1.94)^3_+$

$+0.0131(\text{crea}-7.32)^3_+$
$-0.0165\text{age} - 1.13\times10^{-5}(\text{age}-28.5)^3_+ + 4.05\times10^{-5}(\text{age}-49.5)^3_+$
$-2.15\times10^{-5}(\text{age}-63.7)^3_+ - 2.68\times10^{-5}(\text{age}-72.7)^3_+ + 1.9\times10^{-5}(\text{age}-85.6)^3_+$
$-0.0136\text{hrt} + 6.09\times10^{-7}(\text{hrt}-60)^3_+ - 1.68\times10^{-6}(\text{hrt}-111)^3_+ + 1.07\times10^{-6}(\text{hrt}-140)^3_+$
$-0.0135\,\text{scoma}$
$+0.0161\text{pafi.i} - 4.77\times10^{-7}(\text{pafi.i}-88)^3_+ + 9.11\times10^{-7}(\text{pafi.i}-167)^3_+$
$-5.02\times10^{-7}(\text{pafi.i}-276)^3_+ + 6.76\times10^{-8}(\text{pafi.i}-426)^3_+$
$-0.3693\,\text{adlsc} + 0.0409\,\text{adlsc}^2$
$+0.0394\text{resp} - 9.11\times10^{-5}(\text{resp}-10)^3_+ + 0.000176(\text{resp}-24)^3_+ - 8.5\times10^{-5}(\text{resp}-39)^3_+$

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:

```
# Derive S functions that express mean and quantiles
# of survival time for specific linear predictors
# analytically
expected.surv ← Mean(f)
quantile.surv ← Quantile(f)
latex(expected.surv, file='', type='Sinput')
```

```
expected.surv ← function (lp = NULL, parms = 0.802352037606488)
{
    names(parms) ← NULL
    exp(lp + exp(2 * parms)/2)
}
```

```
latex(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 7.12
```

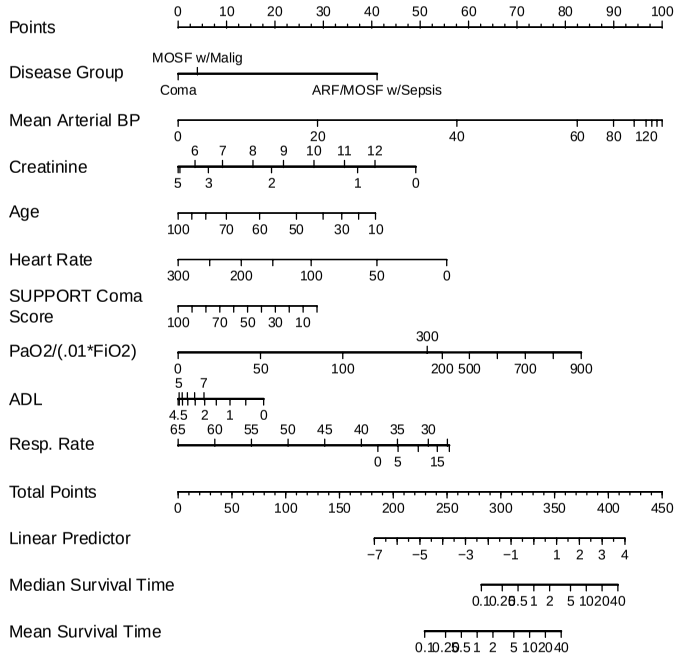


Figure 7.12: Nomogram for predicting median and mean survival time, based on approximation of full model

S Packages and Functions Used

Packages	Purpose	Functions
`Hmisc`	Miscellaneous functions	`describe,ecdf,naclus,` `varclus,llist,spearman2` `describe,impute,latex`
`rms`	Modeling	`datadist,psm,rcs,ols,fastbw`
	Model presentation	`survplot,Newlabels,Function,` `Mean,Quantile,nomogram`
	Model validation	`validate,calibrate`

Note: All packages are available from CRAN

Bibliography

[1] D. G. Altman. Categorising continuous covariates (letter to the editor). *Brit J Cancer*, 64:975, 1991. [26]

[2] D. G. Altman. Suboptimal analysis using 'optimal' cutpoints. *Brit J Cancer*, 78:556–557, 1998. [26]

[3] D. G. Altman and P. K. Andersen. Bootstrap investigation of the stability of a Cox regression model. *Stat Med*, 8:771–783, 1989. [68]

[4] D. G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher. Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *J Nat Cancer Inst*, 86:829–835, 1994. [26, 28]

[5] A. C. Atkinson. A note on the generalized information criterion for choice of a model. *Biometrika*, 67:413–418, 1980. [39, 67]

[6] P. C. Austin. Bootstrap model selection had similar performance for selecting authentic and noise variables compared to backward variable elimination: a simulation study. *J Clin Epi*, 61:1009–1017, 2008. [68]

[7] P. C. Austin, J. V. Tu, and D. S. Lee. Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *J Clin Epi*, 63:1145–1155, 2010. [45]

[8] H. Belcher. The concept of residual confounding in regression models and some applications. *Stat Med*, 11:1747–1758, 1992. [26]

[9] D. A. Belsley. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. Wiley, New York, 1991. [74]

[10] D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York, 1980. [89, 90]

[11] J. K. Benedetti, P. Liu, H. N. Sather, J. Seinfeld, and M. A. Epton. Effective sample size for tests of censored survival data. *Biometrika*, 69:343–349, 1982. [69]

[12] K. Berhane, M. Hauptmann, and B. Langholz. Using tensor product splines in modeling exposure–time–response relationships: Application to the Colorado Plateau Uranium Miners cohort. *Stat Med*, 27:5484–5496, 2008. [57]

[13] M. Blettner and W. Sauerbrei. Influence of model-building strategies on the results of a case-control study. *Stat Med*, 12:1325–1338, 1993. [118]

[14] J. G. Booth and S. Sarkar. Monte Carlo approximation of bootstrap variances. *Am Statistician*, 52:354–357, 1998. [108]

[15] R. Bordley. Statistical decisionmaking without math. *Chance*, 20(3):39–44, 2007. [8]

[16] L. Breiman. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J Am Stat Assoc*, 87:738–754, 1992. [67, 68, 111]

[17] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation (with discussion). *J Am Stat Assoc*, 80:580–619, 1985. [82]

[18] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1984. [43]

[19] W. M. Briggs and R. Zaretzki. The skill plot: A graphical technique for evaluating continuous diagnostic tests (with discussion). *Biometrics*, 64:250–261, 2008. [8]

[20] D. Brownstone. Regression strategies. In *Proceedings of the 20th Symposium on the Interface between Computer Science and Statistics*, pages 74–79, Washington, DC, 1988. American Statistical Association. [118]

[21] P. Buettner, C. Garbe, and I. Guggenmoos-Holzmann. Problems in defining cutoff points of continuous prognostic factors: Example of tumor thickness in primary cutaneous melanoma. *J Clin Epi*, 50:1201–1210, 1997. [26]

[22] J. M. Chambers and T. J. Hastie, editors. *Statistical Models in S*. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1992. [57]

[23] C. Chatfield. Avoiding statistical pitfalls (with discussion). *Statistical Sci*, 6:240–268, 1991. [90]

[24] C. Chatfield. Model uncertainty, data mining and statistical inference (with discussion). *J Roy Stat Soc A*, 158:419–466, 1995. [65, 118]

[25] S. Chatterjee and B. Price. *Regression Analysis by Example*. Wiley, New York, second edition, 1991. [73]

[26] A. Ciampi, J. Thiffault, J.-P. Nakache, and B. Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition. *Comp Stat Data Analysis*, 1986:185–204, 1986. [77]

[27] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc*, 74:829–836, 1979. [41]

[28] E. F. Cook and L. Goldman. Asymmetric stratification: An outline for an efficient method for controlling confounding in cohort studies. *Am J Epi*, 127:626–639, 1988. [45]

[29] J. B. Copas. Regression, prediction and shrinkage (with discussion). *J Roy Stat Soc B*, 45:311–354, 1983. [71, 72]

[30] J. B. Copas. Cross-validation shrinkage of regression predictors. *J Roy Stat Soc B*, 49:175–183, 1987. [116]

[31] D. R. Cox. Regression models and life-tables (with discussion). *J Roy Stat Soc B*, 34:187–220, 1972. [59]

[32] S. L. Crawford, S. L. Tennstedt, and J. B. McKinlay. A comparison of analytic methods for non-random missingness of outcome data. *J Clin Epi*, 48:209–219, 1995. [94]

[33] N. J. Crichton and J. P. Hinde. Correspondence analysis as a screening method for indicants for clinical diagnosis. *Stat Med*, 8:1351–1362, 1989. [77]

[34] R. B. D'Agostino, A. J. Belanger, E. W. Markson, M. Kelly-Hayes, and P. A. Wolf. Development of health risk appraisal functions in the presence of multiple indicators: The Framingham Study nursing home institutionalization model. *Stat Med*, 14:1757–1770, 1995. [74, 76]

[35] C. E. Davis, J. E. Hyde, S. I. Bangdiwala, and J. J. Nelson. An example of dependencies among variables in a conditional logistic regression. In S. Moolgavkar and R. Prentice, editors, *Modern Statistical Methods in Chronic Disease Epidemiology*, pages 140–147. Wiley, New York, 1986. [74]

[36] S. Derksen and H. J. Keselman. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British J Math Stat Psych*, 45:265–282, 1992. [66]

[37] T. F. Devlin and B. J. Weeks. Spline functions for logistic regression modeling. In *Proceedings of the Eleventh Annual SAS Users Group International Conference*, pages 646–651, Cary, NC, 1986. SAS Institute, Inc. [35]

[38] W. D. Dupont. *Statistical Modeling for Biomedical Researchers*. Cambridge University Press, Cambridge, UK, second edition, 2008. [192]

[39] S. Durrleman and R. Simon. Flexible regression models with cubic splines. *Stat Med*, 8:551–561, 1989. [38]

[40] B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J Am Stat Assoc*, 78:316–331, 1983. [112, 115, 116]

[41] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993. [115]

[42] B. Efron and R. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *J Am Stat Assoc*, 92:548–560, 1997. [115]

[43] J. Fan and R. A. Levine. To amnio or not to amnio: That is the decision for Bayes. *Chance*, 20(3):26–32, 2007. [8]

[44] D. Faraggi and R. Simon. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Stat Med*, 15:2203–2213, 1996. [26]

[45] J. J. Faraway. The cost of data analysis. *J Comp Graph Stat*, 1:213–229, 1992. [97, 115, 117]

[46] V. Fedorov, F. Mannino, and R. Zhang. Consequences of dichotomization. *Pharm Stat*, 8:50–61, 2009. [7, 26]

[47] D. Freedman, W. Navidi, and S. Peters. *On the Impact of Variable Selection in Fitting Regression Equations*, pages 1–16. Lecture Notes in Economics and Mathematical Systems. Springer-Verlag, New York, 1988. [116]

[48] J. H. Friedman. A variable span smoother. Technical Report 5, Laboratory for Computational Statistics, Department of Statistics, Stanford University, 1984. [82]

[49] M. H. Gail and R. M. Pfeiffer. On criteria for evaluating models of absolute risk. *Biostatistics*, 6(2):227–239, 2005. [8]

[50] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc*, 102:359–378, 2007. [8]

[51] U. S. Govindarajulu, D. Spiegelman, S. W. Thurston, B. Ganguli, and E. A. Eisen. Comparing smoothing techniques in Cox models for exposure-response relationships. *Stat Med*, 26:3735–3752, 2007. [39]

[52] P. M. Grambsch and P. C. O'Brien. The effects of transformations and preliminary tests for non-linearity in regression. *Stat Med*, 10:697–709, 1991. [48, 66]

[53] R. J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J Am Stat Assoc*, 87:942–951, 1992. [56, 72]

[54] R. J. Gray. Spline-based tests in survival analysis. *Biometrics*, 50:640–652, 1994. [56]

[55] M. J. Greenacre. Correspondence analysis of multivariate categorical data by weighted least-squares. *Biometrika*, 75:457–467, 1988. [77]

[56] S. Greenland. When should epidemiologic regressions use random coefficients? *Biometrics*, 56:915–921, 2000. [66, 92]

[57] F. E. Harrell. The LOGIST Procedure. In *SUGI Supplemental Library Users Guide*, pages 269–293. SAS Institute, Inc., Cary, NC, Version 5 edition, 1986. [67]

[58] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati. Regression modeling strategies for improved prognostic prediction. *Stat Med*, 3:143–152, 1984. [69]

[59] F. E. Harrell, K. L. Lee, D. B. Matchar, and T. A. Reichert. Regression models for prognostic prediction: Advantages, problems, and suggested solutions. *Cancer Treatment Reports*, 69:1071–1077, 1985. [69]

[60] F. E. Harrell, K. L. Lee, and B. G. Pollock. Regression models in clinical studies: Determining relationships between predictors and response. *J Nat Cancer Inst*, 80:1198–1202, 1988. [42]

[61] F. E. Harrell, P. A. Margolis, S. Gove, K. E. Mason, E. K. Mulholland, D. Lehmann, L. Muhe, S. Gatchalian, and H. 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. *Stat Med*, 17:909–944, 1998. [72, 95]

[62] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, New York, second edition, 2008. ISBN-10: 0387848576; ISBN-13: 978-0387848570. [47]

[63] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC, Boca Raton, FL, 1990. ISBN 9780412343902. [47]

[64] S. G. Hilsenbeck and G. M. Clark. Practical p-value adjustment for optimally selected cutpoints. *Stat Med*, 15:103–112, 1996. [26]

[65] W. Hoeffding. A non-parametric test of independence. *Ann Math Stat*, 19:546–557, 1948. [77]

[66] N. Holländer, W. Sauerbrei, and M. Schumacher. Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint. *Stat Med*, 23:1701–1713, 2004. [26, 28]

[67] C. M. Hurvich and C. L. Tsai. The impact of model selection on inference in linear regression. *Am Statistician*, 44:214–217, 1990. [68]

[68] L. I. Iezzoni. Dimensions of risk. In L. I. Iezzoni, editor, *Risk Adjustment for Measuring Health Outcomes*, chapter 2, pages 29–118. Foundation of the American College of Healthcare Executives, Ann Arbor, MI, 1994. [13]

[69] J. Karvanen and F. E. Harrell. Visualizing covariates in proportional hazards model. *Stat Med*, 28:1957–1966, 2009. PMID 19378282. [100]

[70] W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, P. Layde, R. K. Oye, P. E. Bellamy, R. B. Hakim, and D. P. Wagner. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Ann Int Med*, 122:191–203, 1995. [83, 157]

[71] C. Kooperberg, C. J. Stone, and Y. K. Truong. Hazard regression. *J Am Stat Assoc*, 90:78–94, 1995. [177]

[72] W. F. Kuhfeld. The PRINQUAL procedure. In *SAS/STAT 9.2 User's Guide*. SAS Publishing, Cary NC, second edition, 2009. [78]

[73] B. Lausen and M. Schumacher. Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Comp Stat Data Analysis*, 1996. [26]

[74] J. F. Lawless and K. Singhal. Efficient screening of nonnormal regression models. *Biometrics*, 34:318–327, 1978. [68]

[75] S. le Cessie and J. C. van Houwelingen. Ridge estimators in logistic regression. *Appl Stat*, 41:191–201, 1992. [72]

[76] A. Leclerc, D. Luce, F. Lert, J. F. Chastang, and P. Logeay. Correspondance analysis and logistic modelling: Complementary use in the analysis of a health survey among nurses. *Stat Med*, 7:983–995, 1988. [77]

[77] S. Lee, J. Z. Huang, and J. Hu. Sparse logistic principal components analysis for binary data. *Ann Appl Stat*, 4(3):1579–1601, 2010. [47]

[78] C. Leng and H. Wang. On general adaptive sparse principal component analysis. *J Comp Graph Stat*, 18(1):201–215, 2009. [47]

[79] X. Luo, L. A. Stfanski, and D. D. Boos. Tuning variable selection procedures by adding noise. *Technometrics*, 48:165–175, 2006. [15]

[80] N. Mantel. Why stepdown procedures in variable selection. *Technometrics*, 12:621–625, 1970. [68]

[81] S. E. Maxwell and H. D. Delaney. Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, 113:181–190, 1993. [26]

[82] G. P. McCabe. Principal variables. *Technometrics*, 26:137–144, 1984. [76]

[83] G. Michailidis and J. de Leeuw. The Gifi system of descriptive multivariate analysis. *Statistical Sci*, 13:307–336, 1998. [77, 77]

[84] B. K. Moser and L. P. Coombs. Odds ratios for a continuous outcome variable without dichotomizing. *Stat Med*, 23:1843–1860, 2004. [26]

[85] R. H. Myers. *Classical and Modern Regression with Applications*. PWS-Kent, Boston, 1990. [73]

[86] N. J. D. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78:691–692, 1991. [91]

[87] D. Paul, E. Bair, T. Hastie, and R. Tibshirani. "preconditioning" for feature selection and regression in high-dimensional problems. *Ann Stat*, 36(4):1595–1619, 2008. [47]

[88] P. Peduzzi, J. Concato, A. R. Feinstein, and T. R. Holford. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epi*, 48:1503–1510, 1995. [69]

[89] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epi*, 49:1373–1379, 1996. [69, 69]

[90] N. Peek, D. G. T. Arts, R. J. Bosman, P. H. J. van der Voort, and N. F. de Keizer. External validation of prognostic models for critically ill patients required substantial sample sizes. *J Clin Epi*, 60:491–501, 2007. [92]

[91] M. J. Pencina, R. B. D'Agostino Sr, R. B. D'Agostino Jr, and R. S. Vasan. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med*, 27:157–172, 2008. [92]

[92] P. Radchenko and G. M. James. Variable inclusion and shrinkage algorithms. *J Am Stat Assoc*, 103(483):1304–1315, 2008. [46]

[93] D. R. Ragland. Dichotomizing continuous outcome variables: Dependence of the magnitude of association and statistical power on the cutpoint. *Epidemiology*, 3:434–440, 1992. [26]

[94] B. M. Reilly and A. T. Evans. Translating clinical research into clinical practice: Impact of using prediction rules to make decisions. *Ann Int Med*, 144:201–209, 2006. [10]

[95] E. B. Roecker. Prediction error and its estimation for subset-selected models. *Technometrics*, 33:459–468, 1991. [67, 111]

[96] P. Royston, D. G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*, 25:127–141, 2006. [26]

[97] 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. [74, 76]

[98] W. Sauerbrei and M. Schumacher. A bootstrap resampling procedure for model building: Application to the Cox regression model. *Stat Med*, 11:2093–2109, 1992. [68, 112]

[99] G. Schulgen, B. Lausen, J. Olsen, and M. Schumacher. Outcome-oriented cutpoints in quantitative exposure. *Am J Epi*, 120:172–184, 1994. [26, 28]

[100] J. Shao. Linear model selection by cross-validation. *J Am Stat Assoc*, 88:486–494, 1993. [112]

[101] L. R. Smith, F. E. Harrell, and L. H. Muhlbaier. Problems and potentials in modeling survival. In M. L. Grady and H. A. Schwartz, editors, *Medical Effectiveness Research Data Methods (Summary Report), AHCPR Pub. No. 92-0056*, pages 151–159. US Dept. of Health and Human Services, Agency for Health Care Policy and Research, Rockville, MD, 1992. Available from `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/FrankHarrell/smi92pro.pdf`. [69]

[102] I. Spence and R. F. Garrison. A remarkable scatterplot. *Am Statistician*, 47:12–19, 1993. [90]

[103] D. J. Spiegelhalter. Probabilistic prediction in patient management and clinical trials. *Stat Med*, 5:421–433, 1986. [71, 96, 115, 116]

[104] E. W. Steyerberg. *Clinical Prediction Models*. Springer, New York, 2009. [2, 192]

[105] E. W. Steyerberg, M. J. C. Eijkemans, F. E. Harrell, and J. D. F. Habbema. Prognostic modelling with logistic regression analysis: A comparison of selection and estimation methods in small data sets. *Stat Med*, 19:1059–1079, 2000. [46]

[106] C. J. Stone. Comment: Generalized additive models. *Statistical Sci*, 1:312–314, 1986. [38]

[107] C. J. Stone and C. Y. Koo. Additive splines in statistics. In *Proceedings of the Statistical Computing Section ASA*, pages 45–48, Washington, DC, 1985. [34, 39]

[108] S. Suissa and L. Blais. Binary regression with continuous outcomes. *Stat Med*, 14:247–255, 1995. [26]

[109] G. Sun, T. L. Shook, and G. L. Kay. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epi*, 49:907–916, 1996. [70]

[110] R. Tibshirani. Regression shrinkage and selection via the lasso. *J Roy Stat Soc B*, 58:267–288, 1996. [46]

[111] J. C. van Houwelingen and S. le Cessie. Predictive value of statistical models. *Stat Med*, 9:1303–1325, 1990. [39, 72, 72, 112, 116, 117]

[112] P. Verweij and H. C. van Houwelingen. Penalized likelihood in Cox regression. *Stat Med*, 13:2427–2436, 1994. [72]

[113] A. J. Vickers. Decision analysis for the evaluation of diagnostic tests, prediction models, and molecular markers. *Am Statistician*, 62(4):314–320, 2008. [8]

[114] E. Vittinghoff and C. E. McCulloch. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epi*, 165:710–718, 2006. [69]

[115] H. Wainer. Finding what is not there through the unfortunate binning of results: The Mendel effect. *Chance*, 19(1):49–56, 2006. [26, 29]

[116] H. Wang and C. Leng. Unified LASSO estimation by least squares approximation. *J Am Stat Assoc*, 102:1039–1048, 2007. [46]

[117] S. Wang, B. Nan, N. Zhou, and J. Zhu. Hierarchically penalized Cox regression with grouped variables. *Biometrika*, 96(2):307–322, 2009. [46]

[118] Y. Wax. Collinearity diagnosis for a relative risk regression analysis: An application to assessment of diet-cancer relationship in epidemiological studies. *Stat Med*, 11:1273–1287, 1992. [74]

[119] J. Whitehead. Sample size calculations for ordered categorical data. *Stat Med*, 12:2257–2271, 1993. [69]

[120] R. E. Wiegand. Performance of using multiple stepwise algorithms for variable selection. *Stat Med*, 29:1647–1659, 2010. [68]

[121] D. M. Witten and R. Tibshirani. Testing significance of features by lassoed principal components. *Ann Appl Stat*, 2(3):986–1012, 2008. [47]

[122] S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton, FL, 2006. ISBN 9781584884743. [47]

[123] C. F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann Stat*, 14(4):1261–1350, 1986. [112]

[124] S. Xiong. Some notes on the nonnegative garrote. *Technometrics*, 2010. [47]

[125] J. Ye. On measuring and correcting the effects of data mining and model selection. *J Am Stat Assoc*, 93:120–131, 1998. [15]

[126] 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. [77]

[127] H. H. Zhang and W. Lu. Adaptive lasso for Cox's proportional hazards model. *Biometrika*, 94:691–703, 2007. [46]

[128] H. Zhou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *J Comp Graph Stat*, 15:265–286, 2006. [47]

[129] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J Roy Stat Soc B*, 67(2):301–320, 2005. [46]

R packages written by FE Harrell are freely available from CRAN.

To obtain a 588-page book with detailed examples and case studies and notes on the theory and applications of survival analysis, logistic regression, and linear models, order Regression Modeling Strategies *with Applications to Linear Models, Logistic Regression, and Survival Analysis* by FE Harrell from Springer NY (2001). Steyerberg[104] and Dupont[38] are excellent texts for accompanying the book.

To obtain a glossary of statistical terms and other handouts related to diagnostic and prognostic modeling, point your Web browser to `biostat.mc.vanderbilt.edu/ClinStat`.