Contents

1 Introduction

1.1 Hypothesis Testing, Estimation, and Prediction ... ... ... ... ... 3
1.2 Examples of Uses of Predictive Multivariable Modeling ... ... ... ... 5
1.3 Planning for Modeling ... ... ... ... ... ... ... ... ... ... ... ... 6
1.4 Choice of the Model ... ... ... ... ... ... ... ... ... ... ... ... ... ... 8
1.5 Model uncertainty / Data-driven Model Specification ... ... ... ... ... 8

2 General Aspects of Fitting Regression Models ... ... ... ... ... ... 10

2.1 Notation for Multivariable Regression Models ... ... ... ... ... ... 10
2.2 Model Formulations ........................................ 11
2.3 Interpreting Model Parameters .......................... 12
  2.3.1 Nominal Predictors .................................. 13
  2.3.2 Interactions ........................................... 14
  2.3.3 Example: Inference for a Simple Model .......... 15
2.4 Review of Composite (Chunk) Tests ...................... 19
2.5 Relaxing Linearity Assumption for Continuous Predictors ........................................ 19
  2.5.1 Simple Nonlinear Terms ............................. 19
  2.5.2 Splines for Estimating Shape of Regression Function and Determining Predictor Transformations .......................... 20
  2.5.3 Cubic Spline Functions ................................ 23
  2.5.4 Restricted Cubic Splines ............................ 24
  2.5.5 Choosing Number and Position of Knots .......... 27
  2.5.6 Nonparametric Regression .......................... 29
2.5.7 Advantages of Regression Splines over Other Methods

2.6 Recursive Partitioning: Tree-Based Models

2.7 Multiple Degree of Freedom Tests of Association

2.8 Assessment of Model Fit

2.8.1 Regression Assumptions

2.8.2 Modeling and Testing Complex Interactions

2.8.3 Fitting Ordinal Predictors

2.8.4 Distributional Assumptions

3 Missing Data

3.1 Types of Missing Data

3.2 Prelude to Modeling

3.3 Missing Values for Different Types of Response Variables

3.4 Problems With Simple Alternatives to Imputation
3.5 Strategies for Developing Imputation Algorithms ........................................... 51
3.6 Single Conditional Mean Imputation ................................................................. 54
3.7 Multiple Imputation ......................................................................................... 54
3.8 Summary and Rough Guidelines ................................................................... 58

4 Multivariable Modeling Strategies ................................................................. 60

4.1 Prespecification of Predictor Complexity Without Later Simplification ........ 61
4.2 Checking Assumptions of Multiple Predictors Simultaneously .................. 63
4.3 Variable Selection ......................................................................................... 64
4.4 Overfitting and Limits on Number of Predictors ......................................... 67
4.5 Shrinkage ..................................................................................................... 68
4.6 Collinearity ................................................................................................. 71
4.7 Data Reduction ........................................................................................... 73
4.7.1 Variable Clustering .................................................................................. 74
4.7.2 Transformation and Scaling Variables Without Using $Y$ . . . . . . . . 75
4.7.3 Simultaneous Transformation and Imputation . . . . . . . . . . . . . 77
4.7.4 Simple Scoring of Variable Clusters 82
4.7.5 Simplifying Cluster Scores . . . . 83
4.7.6 How Much Data Reduction Is Necessary? . . . . . . . . . . . . . . 83

4.8 Overly Influential Observations . . . . . . 87
4.9 Comparing Two Models . . . . . . . . . . 89
4.10 Summary: Possible Modeling Strategies . 91
   4.10.1 Developing Predictive Models . . . 91
   4.10.2 Developing Models for Effect Estimation . . . . . . . . . . . . . . 94
   4.10.3 Developing Models for Hypothesis Testing . . . . . . . . . . . . . . 95

5 Resampling, Validating, Describing, and Simplifying the Model 96
5.1 The Bootstrap .......................... 96
5.2 Model Validation ...................... 101
  5.2.1 Introduction ...................... 101
  5.2.2 Which Quantities Should Be Used in Validation? ............. 102
  5.2.3 Data-Splitting .................... 104
  5.2.4 Improvements on Data-Splitting: Re-sampling ............... 105
  5.2.5 Validation Using the Bootstrap .... 107
5.3 Describing the Fitted Model ......... 112
5.4 Simplifying the Final Model by Approximating It .............. 113
  5.4.1 Difficulties Using Full Models .... 113
  5.4.2 Approximating the Full Model .... 113

6  S Software ............................. 115
  6.1 The S Modeling Language .......... 116
  6.2 User-Contributed Functions ......... 117
6.3 The Design Library .......................... 119
6.4 Other Functions ............................. 124

10 Binary Logistic Regression ........................ 125

10.1 Model ........................................ 125

10.1.1 Model Assumptions and Interpretation of Parameters .................. 127

10.1.2 Odds Ratio, Risk Ratio, and Risk Difference ......................... 128

10.1.3 Detailed Example ............................ 130

10.1.4 Design Formulations .......................... 137

10.2 Estimation ..................................... 138

10.2.1 Maximum Likelihood Estimates ........................ 138

10.2.2 Estimation of Odds Ratios and Probabilities ......................... 138

10.3 Test Statistics .................................. 138

10.4 Residuals ...................................... 139

10.5 Assessment of Model Fit ........................... 139
10.6 Collinearity ............................................. 154
10.7 Overly Influential Observations ................. 154
10.8 Quantifying Predictive Ability ..................... 154
10.9 Validating the Fitted Model ......................... 155
10.10 Describing the Fitted Model ....................... 158
10.11 S Functions .......................................... 159

11 Logistic Model Case Study: Survival of Ti- 

tanic Passengers ........................................ 165

11.1 Descriptive Statistics ............................... 165
11.2 Exploring Trends with Nonparametric Re-

gression .................................................. 170
11.3 Binary Logistic Model with Casewise Dele-

tion of Missing Values ............................... 171
11.4 Examining Missing Data Patterns ................ 180
11.5 Single Conditional Mean Imputation ............ 183
11.6 Multiple Imputation ................................ 187
11.7 Summarizing the Fitted Model ................... 192
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.
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 → multiple linear regression
– Wilcoxon, Kruskal-Wallis, Spearman → proportional odds ordinal logistic model
– log-rank → Cox

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

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

Statistical estimation is usually model-based

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

• Adjustment depends on how other risk factors relate to hazard

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

- Financial performance, consumer purchasing, loan pay-back
- Ecology
- Product life
- Employment discrimination
- Medicine, epidemiology, health services research
- Probability of diagnosis, time course of a disease
- 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 Planning for Modeling

• Chance that predictive model will be used
• Response definition, follow-up
• Variable definitions
• Observer variability
• Missing data
• Preference for continuous variables
• Subjects
• Sites
Iezzoni 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
1.4 Choice of the Model

- In biostatistics and epidemiology 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)

1.5 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\textsuperscript{81}: “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
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
\begin{align*}
Y & \quad \text{response (dependent) variable} \\
X & \quad X_1, X_2, \ldots, X_p \quad \text{list of predictors} \\
\beta & \quad \beta_0, \beta_1, \ldots, \beta_p \quad \text{regression coefficients} \\
\beta_0 & \quad \text{intercept parameter (optional)} \\
\beta_1, \ldots, \beta_p & \quad \text{weights or regression coefficients} \\
X\beta & \quad \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p, X_0 = 1
\end{align*}

Model: connection between $X$ and $Y$

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

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

2.2 Model Formulations

General regression model

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

General linear regression model

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

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

\[ Y|X \sim N(X\beta, \sigma^2) \]

\[ C(Y|X) = \text{Prob}\{Y = 1|X\} = (1 + \exp(-X\beta))^{-1} \]

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

Example:

\[ C(Y|X) = \text{Prob}\{Y = 1|X\} = (1 + \exp(-X\beta))^{-1} \]

\[ h(u) = \logit(u) = \log\left(\frac{u}{1 - u}\right) \]

\[ h(C(Y|X)) = C'(Y|X) \text{ (link)} \]

General linear regression model:

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

2.3 Interpreting Model Parameters

Suppose that \( X_j \) is linear and doesn’t interact with other \( X \)’s.

\[ C'(Y|X) = X\beta = \beta_0 + \beta_1X_1 + \ldots + \beta_pX_p \]
\[ \beta_j = C'(Y|X_1, X_2, \ldots, X_j + 1, \ldots, X_p) - C'(Y|X_1, X_2, \ldots, X_j, \ldots, X_p) \]

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. \( T = J, K, L, M \):

\[
\begin{align*}
C(Y|T = J) &= \beta_0 \\
C(Y|T = K) &= \beta_0 + \beta_1 \\
C(Y|T = L) &= \beta_0 + \beta_2 \\
C(Y|T = M) &= \beta_0 + \beta_3.
\end{align*}
\]

\[
C(Y|T) = X\beta = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3,
\]

where

\[
\begin{align*}
X_1 &= 1 \text{ if } T = K, \ 0 \text{ otherwise} \\
X_2 &= 1 \text{ if } T = L, \ 0 \text{ otherwise} \\
X_3 &= 1 \text{ if } T = M, \ 0 \text{ otherwise}.
\end{align*}
\]
The test for any differences in the property \( C(Y) \) between treatments is \( H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \).

### 2.3.2 Interactions

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

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

\[
C(Y|X_1 + 1, X_2) - C(Y|X_1, X_2) = \beta_0 + \beta_1(X_1 + 1) + \beta_2X_2 + \beta_3(X_1 + 1)X_2
\]

\[
= [\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2] - [\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2] = \beta_1 + \beta_3X_2.
\]

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

Worse interactions:

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

### 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.

Model assumes

1. age is linearly related to $C(Y)$ for males,
2. age is linearly related to $C(Y)$ for females,
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:

---

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

---
### CHAPTER 2. GENERAL ASPECTS OF FITTING REGRESSION MODELS

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

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

When a high-order effect such as an interaction effect is in the model, be sure to interpret low-order effects by finding out what makes the interaction effect ignorable. In our example, the interaction effect is zero when 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. “$H_a : \text{age is associated with } C(Y)$, powered to detect a linear relationship”
### Most Useful Tests for Linear age $\times$ sex Model

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

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

- In the model
  \[ y \sim \text{age + sex + weight + waist + tricep} \]
  
  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., \( \text{anova(fit, weight, waist, tricep)} \))

2.5 Relaxing Linearity Assumption for Continuous Predictors

2.5.1 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.

2.5.2 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 $\beta_1$, $X \leq a$, $\beta_1 + \beta_2$, $X > a$.
• $\beta_2$ is the slope increment as you pass $a$

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

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

where

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

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

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

\[
X_1 = X \quad X_2 = (X - a)_+ \quad X_3 = (X - b)_+ \quad X_4 = (X - c)_+.
\]

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

### 2.5.3 Cubic Spline Functions

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

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

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

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

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

2.5.4 Restricted Cubic Splines

Stone and Koo\textsuperscript{72}: cubic splines poorly behaved in tails. Constrain function to be linear in tails.

\( k + 3 \rightarrow k - 1 \) parameters\textsuperscript{28}.

To force linearity when \( X < a \): \( X^2 \) and \( X^3 \) terms must be omitted.

To force linearity when \( X > \) last knot: last two \( \beta \)s are redundant, i.e., are just combinations of the other \( \beta \)s.

The restricted spline function with \( k \) knots \( t_1, \ldots, t_k \) is given by\textsuperscript{28}

\[ f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{k-1} X_{k-1} , \]
where $X_1 = X$ and for $j = 1, \ldots, k - 2$,

$$X_{j+1} = (X - t_j)^3_+ - (X - t_{k-1})^3_+(t_k - t_j)/(t_k - t_{k-1})$$

$$+ (X - t_k)^3_+(t_{k-1} - t_j)/(t_k - t_{k-1}).$$

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

Figure 2.2: Restricted cubic spline component variables for $k=5$ and knots at $X = .05, .275, .5, .725,$ and .95. Left panel is a magnification of the right. 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.
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 $\beta$ subject to standard deviations of fitted functions being normalized. See the Web site for a script to create more random spline functions, for $k = 3, \ldots, 7$. 
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

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

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

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_{k-1} = 0.$$
• Fit depends more on choice of \( k \)

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

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

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)\(^2\), \( 75 \) to choose \( k \)
• This chooses \( k \) to maximize model likelihood ratio \( \chi^2 - 2k \).
2.5.6 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

\[X: \quad 1 \quad 2 \quad 3 \quad 5 \quad 8\]
\[Y: \quad 2.1 \quad 3.8 \quad 5.7 \quad 11.1 \quad 17.2\]

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

- 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\textsuperscript{18} moving linear regression smoother \textit{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\textsuperscript{b}
- points near extremes of interval receive almost no weight
- \textit{loess} works much better at extremes of $X$ than moving avg.
- provides an estimate at each observed $X$; other estimates obtained by linear interpolation

\textsuperscript{b}Weight here means something different than regression coefficient. It means how much a point is emphasized in developing the regression coefficients.
– outlier rejection algorithm built-in

• loess works great for binary $Y$ — just turn off outlier detection

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

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

• Another alternative: smoothing splines\footnote{These place knots at all the observed data points but penalize coefficient estimates towards smoothness.}

• Smoothers are very useful for estimating trends in residual plots

2.5.7 Advantages of Regression Splines over Other Methods

Regression splines have several advantages\footnote{These place knots at all the observed data points but penalize coefficient estimates towards smoothness.}:

• Parametric splines can be fitted using any existing regression program

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

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

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

- Extension to non-additive models. Multi-dimensional nonparametric estimators often require burdensome computations.

2.6 Recursive Partitioning: Tree-Based Models

Breiman, Friedman, Olshen, and Stone\textsuperscript{11}: 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\textsuperscript{19}; missing value imputation)
2.7 Multiple Degree of Freedom Tests of Association

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

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

In the 5-knot restricted cubic spline model

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

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

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

Grambsch and O’Brien\(^{36}\) elegantly described the hazards of pretesting

- Studied quadratic regression
- Showed 2 d.f. test of association is nearly
optimal even when regression is linear if non-linearity **entertained**

- Considered ordinary regression model
  \[ E(Y|X) = \beta_0 + \beta_1X + \beta_2X^2 \]

- 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 → 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 $\beta$s

• $SSR$ from quadratic model $> SSR$ from linear model
2.8 Assessment of Model Fit

2.8.1 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. Methods

![Figure 2.4: Regression assumptions for one binary and one continuous predictor](image)

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, $\text{qqnorm}$ plots of overall and stratified residuals

**Advantage:** Simplicity

**Disadvantages:**
- Can only compute standard residuals for uncensored continuous response
- Subjective judgment of non-randomness
- Hard to handle interaction
- Hard to see patterns with large $n$ (trend lines help)
- Seeing patterns does not lead to corrective action

2. Scatterplot of $Y$ vs. $X_2$ using different symbols according to values of $X_1$
**Advantages:** Simplicity, can see interaction

**Disadvantages:**
- Scatterplots cannot be drawn for binary, categorical, or censored $Y$
- Patterns difficult to see if relationships are weak or $n$ large

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

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

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

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

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

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

5. Fit flexible nonlinear parametric model

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

**Disadvantages:**
- Complexity
- Generally difficult to allow for interactions when assessing patterns of effects

Confidence limits, formal inference can be problematic for methods 1-4.
Restricted cubic spline works well for method 5.

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

where

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

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

- Plot \(\hat{f}(X_2)\) vs. \(X_2\)
- \(n\) large → can fit separate functions by \(X_1\)
- Test of linearity: \(H_0 : \beta_3 = \beta_4 = 0\)
- Nonlinear → use transformation suggested by spline fit or keep spline terms
- Tentative transformation \(g(X_2)\) → 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 → expand each
using spline

- Example: assess linearity of $X_2, X_3$

$$C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2' + \beta_4 X_2'' + \beta_5 X_3 + \beta_6 X_3' + \beta_7 X_3''$$

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

2.8.2 Modeling and Testing Complex Interactions

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

$$C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2' + \beta_4 X_2'' + \beta_5 X_1 X_2 + \beta_6 X_1 X_2' + \beta_7 X_1 X_2''$$

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

- 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 $X_2 g(X_1)$:

$$C(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X'_1 + \beta_3 X''_1$$
$$+ \beta_4 X_2 + \beta_5 X'_2 + \beta_6 X''_2$$
$$+ \beta_7 X_1 X_2 + \beta_8 X_1 X'_2 + \beta_9 X_1 X''_2$$
$$+ \beta_{10} X_2 X'_1 + \beta_{11} X_2 X''_1$$

- Test of additivity is $H_0 : \beta_7 = \beta_8 = \ldots = \beta_{11} = 0$ with 5 d.f.
- Test of lack of fit for the simple product interaction with $X_2$ is $H_0 : \beta_8 = \beta_9 = 0$
- Test of lack of fit for the simple product interaction with $X_1$ is $H_0 : \beta_{10} = \beta_{11} = 0$

General spline surface:

- Cover $X_1 \times X_2$ plane with grid and fit patch-wise cubic polynomial in two variables
- Restrict to be of form $aX_1 + bX_2 + cX_1 X_2$ in corners
- Uses all $(k - 1)^2$ cross-products of restricted cubic spline terms
• See Gray [37, 38, Section 3.2] for penalized splines allowing control of effective degrees of freedom

Other issues:

• $Y$ non-censored (especially continuous) $\rightarrow$ multi-dimensional scatterplot smoother

• 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 \text{disease}
• Calendar time \times \text{treatment}
• Quality \times \text{quantity of a symptom}

2.8.3 Fitting Ordinal Predictors

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

2.8.4 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): 
  \[ C(Y|X) = C(Y = y|X) = d(y) + X \beta \]
  \( C' = \text{log hazard} \)

• Check form of \( d(y) \)

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

Missing Data

3.1 Types of Missing Data

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Informative missing (non-ignorable non-response)

3.2 Prelude to Modeling

- Quantify extent of missing data
- Characterize types of subjects with missing
data

- Find sets of variables missing on same subjects

### 3.3 Missing Values for Different Types of Response Variables

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

### 3.4 Problems With Simple Alternatives to Imputation

Deletion of records—
• Badly biases parameter estimates when missingness is related to $Y$ in a way that is unexplained by non-missing $X$s
• Deletion because of a subset of $X$ being missing always results in inefficient estimates
• Deletion of records with missing $Y$ may result in serious biases $^{23}$
• Only discard obs. when
  – Rarely missing predictor of overriding importance that can’t be imputed from other data
  – Fraction of obs. with missings small and $n$ is large
• No advantage of deletion except savings of analyst time
• Making up missing data better than throwing away real data

Adding extra categories of categorical predictors—
• Including missing data but adding a category ‘missing’ causes serious biases

• Problem acute when values missing because subject too sick

• Difficult to interpret

3.5 Strategies for Developing Imputation Algorithms

Exactly how are missing values estimated?

• Could ignore all other information — random or grand mean fill-in

• Can use external info not used in response model (e.g., zip code for income)

• Need to utilize reason for non-response if possible

• Use statistical model with sometimes-missing $X$ as response variable

• Model to estimate the missing values should include all variables that are either
1. related to the missing data mechanism;
2. have distributions that differ between subjects that have the target variable missing and those that have it measured;
3. associated with the sometimes-missing variable when it is not missing; or
4. included in the final response model

- Ignoring imputation results in biased $\hat{V}(\hat{\beta})$
- `transcan` function in Hmisc library: “optimal” transformations of all variables to make residuals more stable and to allow non-monotonic transformations
- `aregImpute` function in Hmisc: good approximation to full Bayesian multiple imputation procedure using the bootstrap
- `aregImpute` and `transcan` work with `fit.mult.impute` to make final analysis of response variable relatively easy
- Predictive mean matching: replace missing value with observed value of subject having
ing closest predicted value to the predicted value of the subject with the NA

- PMM can result in some donor observations being used repeatedly
- Causes lumpy distribution of imputed values
- Address by sampling from multinomial distribution, probabilities = scaled distance of all predicted values to predicted value ($y^*$) of observation needing imputing
- Tukey’s tricube function is a good weighting function (used in loess): $w_i = (1 - \min(d_i/s, 1)^3)^3$, $d_i = |\hat{y}_i - y^*|$, $s = 0.2 \times \text{mean}(|\hat{y}_i - y^*|)$ is a good default scale factor
  scale so that $\sum w_i = 1$

- Recursive partitioning with surrogate splits — handles case where a predictor of a variable needing imputation is missing itself
3.6 Single Conditional Mean Imputation

- Can fill-in using unconditional mean or median if number of missings low and $X$ is unrelated to other $X$s
- Otherwise, first approximation to good imputation uses other $X$s to predict a missing $X$
- This is a single “best guess” conditional mean
- $\hat{X}_j = Z\hat{\theta}, Z = X_j$
  Cannot include $Y$ in $Z$ without adding random errors to imputed values (would steal info from $Y$)
- Recursive partitioning is very helpful for non-parametrically estimating conditional means

3.7 Multiple Imputation

- Single imputation using a random draw from the conditional distribution for an individual
  $\hat{X}_j = Z\hat{\theta} + \hat{\epsilon}, Z = [X_j, Y]$
\[ \hat{e} = n(0, \hat{\sigma}) \] or a random draw from the calculated residuals

- bootstrap
- approximate Bayesian bootstrap \(^{63}\): sample with replacement from sample with replacement of residuals

- **Multiple imputations** (\(M\)) with random draws
  - Draw sample of \(M\) residuals for each missing value to be imputed
  - Average \(M\) \(\hat{\beta}\)
  - In general can provide least biased estimates of \(\beta\)
  - Simple formula for imputation-corrected var(\(\hat{\beta}\))
    Function of average “apparent” variances and between-imputation variances of \(\hat{\beta}\)
  - **BUT** full multiple imputation needs to account for uncertainty in the imputation models by refitting these models for each of the \(M\) draws
  - transcan **does not do that**; aregImpute **does**
• `aregImpute` algorithm
  – Takes all aspects of uncertainty into account using the bootstrap
  – Different bootstrap resamples used for each imputation by fitting a flexible additive model on a sample with replacement from the original data
  – This model is used to predict all of the original missing and non-missing values for the target variable for the current imputation
  – Uses `ace` or `avas` semiparametric regression models to impute
  – For continuous variables, monotonic transformations of the target variable are assumed when `avas` used
  – For `ace`, the default allows nonmonotonic transformations of target variables
  – Uses predictive mean matching for imputation; no residuals required
  – By default uses weighted PMM; option for
just using closest match

- When a predictor of the target variable is missing, it is first imputed from its last imputation when it was a target variable
- First 3 iterations of process are ignored (“burn-in”)
- Compares favorably to S MICE approach
- Example:

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

See Barzi and Woodward\(^3\) for a nice review of multiple imputation with detailed comparison of results (point estimates and confidence limits for the effect of the sometimes-missing predictor) for various imputation methods.
Table 3.1: Summary of Methods for Dealing with Missing Values

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

3.8 Summary and Rough Guidelines

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

Proportion of missings $\leq 0.05$ : Method of imputing and computing variances doesn’t matter much

Proportion of missings $0.05 - 0.15$ : Constant fill-in if predictor unrelated to other $X$'s. Single “best guess” imputation probably OK. Multiple imputation doesn’t hurt.

Proportion of missings $> 0.15$ : Multiple im-
putation, adjust variances for imputation

Multiple predictors frequently missing More important to do multiple imputation and also to be cautious that imputation might be ineffective.

Reason for missings more important than number of missing values.
“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

4.1 Prespecification of Predictor Complexity Without Later Simplification

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

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

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

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

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

• Specify number of knots for continuous $X$, combine infrequent categories of categorical $X$ based on $\rho^2$

• Allocating d.f. based on sorting $\rho^2$ fair procedure because
  
  – already decided to keep variable in model no matter what $\rho^2$

$^a$This test statistic does not inform the analyst of which groups are different from one another.
– $\rho^2$ does not reveal degree of nonlinearity; high value may be due solely to strong linear effect
– low $\rho^2$ for a categorical variable might lead to collapsing the most disparate categories

• Initial simulations show the procedure to be conservative
• Can move from simpler to more complex models but not the other way round

4.2 Checking Assumptions of Multiple Predictors Simultaneously

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

- Series of potential predictors with no prior knowledge
- \( \uparrow \) exploration \( \rightarrow \uparrow \) shrinkage (overfitting)
- Summary of problem: \( E(\hat{\beta}|\hat{\beta} \text{ “significant”} ) \neq \beta \)
- Biased \( R^2, \hat{\beta}, \) standard errors, \( P \)-values too small
- \( F \) and \( \chi^2 \) statistics do not have the claimed distribution
- Will result in residual confounding if use variable selection to find confounders
- Derksen and Keselman found that in step-wise 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$^{41}$:

- Forward selection, backward elimination
- Stopping rule: “residual $\chi^2$” with d.f. = no. candidates remaining at current step
• Test for significance or use Akaike’s information criterion (AIC\(^2\)), 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 two pre-specified models [9, Section 1.3]
• Roecker\(^{62}\) 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\(^{55}\) and can be used efficiently with maximum likelihood estimation\(^{51}\)
• 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 \textsuperscript{1, 9, 47}

- Data reduction (grouping variables) can help
- Using the bootstrap to select important variables for inclusion in the final model \textsuperscript{65} is problematic
- It is not logical that a population regression coefficient would be exactly zero just because its estimate was “insignificant”

4.4 Overfitting and Limits on Number of Predictors

- Concerned with avoiding overfitting
- $p$ should be $< \frac{m}{15}$ \textsuperscript{42, 43, 59, 60, 67}
- $p =$ number of parameters in full model or number of candidate parameters in a stepwise analysis

\textsuperscript{1}If one considers the power of a two-sample binomial test compared with a Wilcoxon test if the response could be made continuous and the proportional odds assumption holds, the effective sample size for a binary response is $3n_1n_2/n \approx 3 \min(n_1, n_2)$ if $\frac{n_1}{n_2}$ is near 0 or 1 \textsuperscript{[79, Eq. 10, 15]}. Here $n_1$ and $n_2$ are the marginal frequencies of the two response levels \textsuperscript{[60]}.  

\textsuperscript{2}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) \textsuperscript{[79, Eq. 3]}. If all cell sizes are equal, the relative efficiency of having $k$ response categories compared to a continuous response is $1 – \frac{1}{k}$ \textsuperscript{[79, Eq. 14]}, e.g., a 5-level response is almost as efficient as a continuous one if proportional odds holds across category cutoffs.

\textsuperscript{3}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\textsuperscript{6}.  

\textsuperscript{4}
Table 4.1: Limiting Sample Sizes for Various Response Variables

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

- Narrowly distributed predictor $\rightarrow$ even higher $n$
- $p$ includes all variables screened for association with response, including interactions
- Univariable screening (graphs, crosstabs, etc.) in no way reduces multiple comparison problems of model building $^{73}$

4.5 Shrinkage

- Slope of calibration plot; regression to the mean
- Statistical estimation procedure — “pre-shrunken” models
- Aren’t regression coefficients OK because
they’re unbiased?

- Problem is in how we use coefficient estimates

- Consider 20 samples of size \( n = 50 \) from \( U(0, 1) \)

- Compute group means and plot in ascending order

- Equivalent to fitting an intercept and 19 dummies using least squares

- Result generalizes to general problems in plotting \( Y \) vs. \( X \hat{\beta} \)

- Prevent shrinkage by using pre-shrinkage

- Spiegelhalter\(^{70}\): 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\(^{20}\)

- Ridge regression\(^{52, 75}\)
Figure 4.1: Sorted means from 20 samples of size 50 from a uniform \([0, 1]\) distribution. The reference line at 0.5 depicts the true population value of all of the means.
• Penalized MLE $^{37, 45, 76}$

• Heuristic shrinkage parameter of van Houwelingen and le Cessie [75, Eq. 77]

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

• OLS: \[ \hat{\gamma} = \frac{n-p-1}{n-1} \frac{R^2_{adj}}{R^2} \]

\[ R^2_{adj} = 1 - (1 - R^2) \frac{n-1}{n-p-1} \]

• $p$ close to no. candidate variables

• Copas [20, Eq. 8.5] adds 2 to numerator

4.6 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 [16, 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 [57, 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=diagonal of inverse 26, 77

• See Belsley [4, pp. 28-30] for problems with VIF

• Easy approach: SAS \texttt{VARCLUS} procedure 64, S-PLUS \texttt{varclus} function, other clustering techniques: group highly correlated variables

• Can score each group (e.g., first principal component, $PC_1^2$); summary scores not collinear

4.7 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.

4.7.1 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 \(^{25, 64}\) (oblique-rotation PC analysis) → separate variables so that first PC is representative of group

• Can also do hierarchical cluster analysis on similarity matrix based on squared Spearman or Pearson correlations, or more generally, Hoeffding’s \(D^{46}\).

4.7.2 Transformation and Scaling Variables Without Using \(Y\)

• Reduce \(p\) by estimating transformations using associations with other predictors

• Purely categorical predictors – correspondence analysis \(^{17, 24, 39, 53, 56}\)

• Mixture of qualitative and continuous variables: qualitative principal components

• Maximum total variance (MTV) of Young, Takane, de Leeuw \(^{56, 82}\)
1. Compute $PC_1$ of variables using correlation matrix

2. Use regression (with splines, dummies, etc.) to predict $PC_1$ from each $X$ — expand each $X_j$ and regress it separately on $PC_1$ to get working transformations

3. Recompute $PC_1$ on transformed $X$s

4. Repeat 3-4 times until variation explained by $PC_1$ plateaus and transformations stabilize

- Maximum generalized variance (MGV) method of Sarle [50, 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

1. Allows flexible transformations including monotonic splines

2. Does not allow restricted cubic splines, so may be unstable unless monotonicity assumed

3. Allows simultaneous imputation but often yields wild estimates

4.7.3 Simultaneous Transformation and Imputation

S-PLUS transcan Function for Data Reduction & Imputation
• Initialize missings to medians (or most frequent category)
• Initialize transformations to original variables
• Take each variable in turn as $Y$
• Exclude obs. missing on $Y$
• Expand $Y$ (spline or dummy variables)
• Score (transform $Y$) using first canonical variate
• Missing $Y \rightarrow$ predict canonical variate from $X$s
• The imputed values can optionally be shrunk to avoid 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-PLUS function `Function` will create a series of S-PLUS functions that transform each predictor

• Example: $n = 415$ acutely ill patients
  1. Relate heart rate to mean arterial blood
pressures
2. Two blood pressures missing
3. Heart rate not monotonically related to blood pressure
4. See Figure 4.2

ACE (Alternating Conditional Expectation) of Breiman and Friedman

1. Uses nonparametric “super smoother”
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. $Y$)
  3. Expand original variable using spline, test additional predictive information over original transformation
Figure 4.2: Transformations fitted using transcan. Tick marks indicate the two imputed values for blood pressure. The lower left plot contains raw data ($D_{xy} = 0.02$); the lower right is a scatterplot of the corresponding transformed values ($D_{xy} = 0.14$). Data courtesy of the SUPPORT study\textsuperscript{49}. 

4.7.4 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 $X_2$
4.7.5 Simplifying Cluster Scores

4.7.6 How Much Data Reduction Is Necessary?

Using Expected Shrinkage to Guide Data Reduction

- Fit full model with all candidates, $p$ d.f., LR likelihood ratio $\chi^2$
- Compute $\hat{\gamma}$
- If $\hat{\gamma} < 0.9$, consider shrunken estimator from whole model, or data reduction (again not using $Y$)
- $q$ regression d.f. for reduced model
- Assume best case: discarded dimensions had no association with $Y$
- Expected loss in LR is $p - q$
- New shrinkage $[LR - (p - q) - q] / [LR - (p - q)]$
- Solve for $q \rightarrow q \leq (LR - p) / 9$
- Under these assumptions, no hope unless original LR $> p + 9$
• No $\chi^2$ lost by dimension reduction $\rightarrow q \leq LR/10$

Example:

• Binary logistic model, 45 events on 150 subjects
• 10:1 rule $\rightarrow$ analyze 4.5 d.f. total
• Analyst wishes to include age, sex, 10 others
• Not known if age linear or if age and sex additive
• 4 knots $\rightarrow 3 + 1 + 1$ d.f. for age and sex if restrict interaction to be linear
• Full model with 15 d.f. has LR=50
• Expected shrinkage factor $(50 - 15)/50 = 0.7$
• $LR > 15 + 9 = 24 \rightarrow$ reduction may help
• Reduction to $q = (50 - 15)/9 \approx 4$ d.f. necessary
• Have to assume age linear, reduce other 10 to 1 d.f.
• Separate hypothesis tests intended → use full model, adjust for multiple comparisons
<table>
<thead>
<tr>
<th><strong>Goals</strong></th>
<th><strong>Reasons</strong></th>
<th><strong>Methods</strong></th>
</tr>
</thead>
</table>
| Group predictors so that each group represents a single dimension that can be summarized with a single score | - ↓ d.f. arising from multiple predictors | Variable clustering  
- Subject matter knowledge  
- Group predictors to maximize proportion of variance explained by $PC_1$ of each group  
- Hierarchical clustering using a matrix of similarity measures between predictors |
| Transform predictors | - ↓ d.f. due to non-linear and dummy variable components  
- Allows predictors to be optimally combined  
- Make $PC_1$ more reasonable summary  
- Use in customized model for imputing missing values on each predictor | - Maximum total variance on a group of related predictors  
- Canonical variates on the total set of predictors |
| Score a group of predictors | ↓ d.f. for group to unity | - $PC_1$  
- Simple point scores |
| Multiple dimensional scoring of all predictors | ↓ d.f. for all predictors combined | Principal components $1, 2, \ldots, k, k < p$ computed from all transformed predictors |
4.8 Overly Influential Observations

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

Reasons for influence:

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

5. Sensitive to WBC $> 60000$ ($n = 16$)

6. Predictions stable if truncate WBC to 40000 ($n = 46$ above 40000)

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

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

Statistical Measures:

- Leverage: capacity to be influential (not necessarily infl.)
  Diagonals of “hat matrix” $H = X(X'X)^{-1}X'$ — measures how an obs. predicts its own response $^5$

  - $h_{ii} > \frac{2(p + 1)}{n}$ may signal a high leverage point $^5$

- 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 $|\text{DFFITS}| > 2\sqrt{(p + 1)/(n - p - 1)}$
- Others examine entire distribution for “outliers”
- No substitute for careful examination of data
- Maximum likelihood estimation requires 1-step approximations

4.9 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., $R^2$, model $\chi^2$, Somers’ $D_{xy}$, Spearman’s $\rho$, area under ROC curve)
- Rank measures may not give enough credit to extreme predictions $\rightarrow$ model $\chi^2$, $R^2$, examine extremes of distribution of $\hat{Y}$
- Examine differences in predicted values from the two models
4.10 Summary: Possible Modeling Strategies

Strategy in a nutshell:

- Decide how many d.f. can be spent
- Decide where to spend them
- Spend them
- Don’t reconsider, especially if inference needed

4.10.1 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. For certain models it is effective to multiply impute $Y$. 
4. Characterize and impute missing $X$. In most cases use multiple imputation based on $X$ and $Y$.

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

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

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

9. Check linearity assumptions and make transformations in $X$'s as needed but beware.
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. 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 over-
fitting 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\textsuperscript{33} to penalize for the randomness inherent in the multiple steps.

20. Develop simplifications to the final model as needed.

4.10.2 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

4.10.3 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 5

Resampling, Validating, Describing, and Simplifying the Model

5.1 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
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

- $F_n$ corresponds to density function placing probability $\frac{1}{n}$ at each observed data point ($\frac{k}{n}$ if point duplicated $k$ times)
- Pretend that $F \equiv F_n$
- Sampling from $F_n \equiv$ sampling with replacement from observed data $Y_1, \ldots, Y_n$
- Large $n \rightarrow$ selects $1 - e^{-1} \approx 0.632$ of origi-
nal data points in each bootstrap sample at least once

- Some observations not selected, others selected more than once

- Efron’s *bootstrap* → general-purpose technique for estimating properties of estimators without assuming or knowing distribution of data $F$

- Take $B$ samples of size $n$ with replacement, choose $B$ so that summary measure of indi-
individual statistics \approx \text{summary if } B = \infty

- Bootstrap based on distribution of \textit{observed} differences between a resampled parameter estimate and the original estimate telling us about the distribution of \textit{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).

First 20 samples:
<table>
<thead>
<tr>
<th>Bootstrap Sample</th>
<th>Sample Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 6 6 6 9 9</td>
<td>6.0</td>
</tr>
<tr>
<td>5 5 6 7 8 8</td>
<td>6.5</td>
</tr>
<tr>
<td>1 1 1 5 8 9</td>
<td>3.0</td>
</tr>
<tr>
<td>1 1 1 5 8 9</td>
<td>3.0</td>
</tr>
<tr>
<td>1 6 8 8 8 9</td>
<td>8.0</td>
</tr>
<tr>
<td>1 6 7 8 9 9</td>
<td>7.5</td>
</tr>
<tr>
<td>6 6 8 8 9 9</td>
<td>8.0</td>
</tr>
<tr>
<td>1 1 7 8 8 9</td>
<td>7.5</td>
</tr>
<tr>
<td>1 5 7 8 9 9</td>
<td>7.5</td>
</tr>
<tr>
<td>5 6 6 6 7 7</td>
<td>6.0</td>
</tr>
<tr>
<td>1 6 8 8 9 9</td>
<td>8.0</td>
</tr>
<tr>
<td>1 5 6 6 9 9</td>
<td>6.0</td>
</tr>
<tr>
<td>1 6 7 8 8 9</td>
<td>7.5</td>
</tr>
<tr>
<td>1 6 7 7 9 9</td>
<td>7.0</td>
</tr>
<tr>
<td>1 5 7 8 9 9</td>
<td>7.5</td>
</tr>
<tr>
<td>5 6 7 9 9 9</td>
<td>8.0</td>
</tr>
<tr>
<td>5 5 6 7 8 8</td>
<td>6.5</td>
</tr>
<tr>
<td>6 6 6 7 8 8</td>
<td>6.5</td>
</tr>
<tr>
<td>1 1 1 1 6 9</td>
<td>1.0</td>
</tr>
<tr>
<td>1 5 7 7 9 9</td>
<td>7.0</td>
</tr>
</tbody>
</table>

- Histogram tells us whether we can assume normality for the bootstrap medians or need to use quantiles of medians to construct C.L.
- Need high $B$ for quantiles, low for variance (but see [8])
5.2 Model Validation

5.2.1 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
\textbf{CHAPTER 5. RESAMPLING, VALIDATING, DESCRIBING, AND SIMPLIFYING THE MODEL}

- 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

5.2.2 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: $R^2$; binary logistic model: ROC area, equivalent to rank correlation between predicted probability of event and 0/1 event

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

- Split data into \textit{training} and \textit{test} sets
- Interesting to compare index of accuracy in training and test
- Freeze parameters from training
- Make sure you allow $R^2 = 1 - \frac{SSE}{SST}$ for test sample to be $< 0$
- Don’t compute ordinary $R^2$ on $X \hat{\beta}$ vs. $Y$; this allows for linear recalibration $aX \hat{\beta} + b$ vs. $Y$
- Test sample must be large enough to obtain very accurate assessment of accuracy
- Training sample is what’s left
- Example: overall sample $n = 300$, training sample $n = 200$, develop model, freeze $\hat{\beta}$, predict on test sample ($n = 100$), $R^2 = 1 - \frac{\sum(Y_i - X_i \hat{\beta})^2}{\sum(Y_i - Y)^2}$.
- Disadvantages of data splitting:
  1. Costly in $n^9$,
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

5.2.4 Improvements on Data-Splitting: Resampling

- No sacrifice in sample size
- Work when modeling process automated
- Bootstrap excellent for studying arbitrariness of variable selection
- Cross-validation solves many problems of data splitting
- 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

• Randomization method
  1. Randomly permute $Y$
  2. Optimism = performance of fitted model compared to what expect by chance
5.2.5 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 $R^2$ ($R_{app}^2$) = 0.4
  - how inflated is $R_{app}^2$?
– get resamples of size 1000 with replacement from original 1000
– for each resample compute $R_{boot}^2 = \text{apparent } R^2 \text{ in bootstrap sample}$
– freeze these coefficients (call them $\hat{\beta}_{boot}$), apply to original (whole) sample $(X_{\text{orig}}, Y_{\text{orig}})$ to get $R_{orig}^2 = R^2(X_{\text{orig}}\hat{\beta}_{boot}, Y_{\text{orig}})$
– optimism = $R_{boot}^2 - R_{orig}^2$
– average over $B = 100$ optimisms to get $\text{optimism} = \overline{\text{optimism}}$
– $R_{\text{overfitting corrected}}^2 = R_{\text{app}}^2 - \overline{\text{optimism}}$

• Is estimating unconditional (not conditional on $X$) distribution of $R^2$, etc. [33, 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 $n^{30}$, [31, p. 253], 32

Bootstrap useful for assessing calibration in addition to discrimination:
• Fit $C(Y|X) = X\beta$ on bootstrap sample
• Re-fit $C(Y|X) = \gamma_0 + \gamma_1 X\hat{\beta}$ on same data
• $\hat{\gamma}_0 = 0$, $\hat{\gamma}_1 = 1$
• Test data (original dataset): re-estimate $\gamma_0$, $\gamma_1$
• $\hat{\gamma}_1 < 1$ if overfit, $\hat{\gamma}_0 > 0$ to compensate
• $\hat{\gamma}_1$ quantifies overfitting and useful for improving calibration
• Use Efron’s method to estimate optimism in $(0, 1)$, estimate $(\gamma_0, \gamma_1)$ by subtracting optimism from $(0, 1)$
• See also Copas and van Houwelingen and le Cessie [75, p. 1318]

See [34] for warnings about the bootstrap, and [30] 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
• Stepwise modeling often reduces optimism but this is not offset by loss of information from deleting marginal var.

<table>
<thead>
<tr>
<th>Method</th>
<th>Apparent Rank Correlation of Predicted vs. Observed</th>
<th>Over-Bias-Corrected Optimism</th>
<th>Bias-Corrected Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Model</td>
<td>0.50</td>
<td>0.06</td>
<td>0.44</td>
</tr>
<tr>
<td>Stepwise Model</td>
<td>0.47</td>
<td>0.05</td>
<td>0.42</td>
</tr>
</tbody>
</table>

In this example, stepwise modeling lost a possible \(0.50 - 0.47 = 0.03\) predictive discrimination. The full model fit will especially be an improvement when

1. The stepwise selection deleted several variables which were almost significant.
2. These marginal variables have *some* real predictive value, even if it’s slight.
3. There is no small set of extremely dominant
variables that would be easily found by step-wise selection.

Other issues:

- See [75] for many interesting ideas
- Faraway $^{33}$ shows how bootstrap is used to penalize for choosing transformations for $Y$, outlier and influence checking, variable selection, etc. simultaneously
- Brownstone [12, 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 $^7$ and Chatfield $^{15}$ for more interesting examples of problems resulting from data-driven analyses.
5.3 Describing the Fitted Model

- 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
- Nomogram
• Use regression tree to approximate the full model

5.4 Simplifying the Final Model by Approximating It

5.4.1 Difficulties Using Full Models

• Predictions are conditional on all variables, standard errors \(\uparrow\) when predict for a low-frequency category
• Collinearity
• Can average predictions over categories to marginalize, \(\downarrow\) s.e.

5.4.2 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 $\sigma$ from full fit
  2. Stopping rule less arbitrary
  3. Inheritance of shrinkage

• If estimates from full model are $\hat{\beta}$ and approx. model is based on a subset $T$ of predictors $X$, coef. of approx. model are $W\hat{\beta}$, where $W = (T'T)^{-1}T'X$

• Variance matrix of reduced coef.: $WVW'$
Chapter 6

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
expansions

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

6.1 The S Modeling Language

S statistical modeling language:

response ~ terms

\[ y \sim \text{age} + \text{sex} \]  # age + sex main effects
\[ y \sim \text{age} + \text{sex} + \text{age:sex} \]  # add second-order interaction
\[ y \sim \text{age:sex} \]  # second-order interaction +
\[ y \sim (\text{age} + \text{sex} + \text{pressure})^2 \]  # all main effects
\[ y \sim (\text{age} + \text{sex} + \text{pressure})^2 - \text{sex:pressure} \]  # all main effects and all 2nd order
\[ y \sim (\text{age} + \text{race})\text{*sex} \]  # age+race+sex+age:sex+race:sex
\[ y \sim \text{treatment}\text{*}(\text{age*race + age*sex}) \]  # no interact. with race,sex
\[ \sqrt{y} \sim \text{sex*sqrt(age)} + \text{race} \]  # functions, with dummy variables generated if
\[ \text{sqrt}(y) \sim \text{sex*sqrt(age)} + \text{race} \]  # race is an S factor (classification) variable
\[ y \sim \text{sex} + \text{poly(age,2)} \]  # poly generates orthogonal polynomials
\[ \text{race.sex} \leftarrow \text{interaction(race,sex)} \]  # for when you want dummy variables for
\[ y \sim \text{age} + \text{race.sex} \]  # all combinations of the factors
The formula for a regression model is given to a modeling function, e.g.

\[
\text{lrm}(y \sim \text{rcs}(x,4))
\]

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:

\[
\begin{align*}
  f & \leftarrow \text{lrm}(y \sim \text{rcs}(x,4) + x2 + x3) \\
  f2 & \leftarrow \text{update}(f, \text{subset}=\text{sex}=="\text{male}") \\
  f3 & \leftarrow \text{update}(f, . \sim . - x2) \quad \# \text{remove } x2 \text{ from model} \\
  f4 & \leftarrow \text{update}(f, . \sim . + \text{rcs}(x5,5)) \quad \# \text{add } \text{rcs}(x5,5) \text{ to model} \\
  f5 & \leftarrow \text{update}(f, y2 \sim .) \quad \# \text{same terms, new response var.}
\end{align*}
\]

6.2 User-Contributed Functions

- S is high-level object-oriented language.
- S-P\textsc{L}US (UNIX, Linux, Microsoft Windows)
- R (UNIX, Linux, Mac, Windows)
- Multitude of user-contributed functions on StatLib

\(^a\)\text{lrm} \text{and } \text{rcs} \text{ are in the Design library.}
• International community of users through S-news

Some S functions:

• See Venables and Ripley
• Hierarchical clustering: \texttt{hclust}
• Principal components: \texttt{princomp}, \texttt{prcomp}
• Canonical correlation: \texttt{cancor}
• ACE: \texttt{ace}
• \texttt{areg.boot} \textbf{(Harrell)}
• Rank correlation methods:
  \texttt{rcorr}, \texttt{hoeffd}, \texttt{spearman2} \textbf{(Harrell)}
• Variable clustering: \texttt{varclus} \textbf{(Harrell)}
• \texttt{transcan}, \texttt{aregImpute} \textbf{(Harrell)}
• Correspondence analysis: see Web page
• Restricted cubic spline design matrix:
  \texttt{rcspline.eval} \textbf{(Harrell)}
• Re-state restricted spline in simpler form: \texttt{rcspline}
6.3 The Design Library

- `datadist` function to compute predictor distribution summaries

\[ y \sim \text{sex } + \text{lsp(age,c(20,30,40,50,60)}) + \text{sex } \%\text{ia}\% \text{lsp(age,c(20,30,40,50,60)}) \]

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

\[ y \sim \text{lsp(age,30)} + \text{rcs(cholesterol,4)} + \text{lsp(age,30)} \%\text{ia}\% \text{rcs(cholesterol,4)} \]

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

<table>
<thead>
<tr>
<th>Function</th>
<th>Purpose</th>
<th>Related S Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>ols</code></td>
<td>Ordinary least squares linear model</td>
<td><code>lm</code></td>
</tr>
<tr>
<td><code>lrm</code></td>
<td>Binary and ordinal logistic regression model</td>
<td><code>glm</code></td>
</tr>
<tr>
<td></td>
<td>Has options for penalized MLE</td>
<td></td>
</tr>
<tr>
<td><code>psm</code></td>
<td>Accelerated failure time parametric survival models</td>
<td><code>survreg</code></td>
</tr>
<tr>
<td><code>cph</code></td>
<td>Cox proportional hazards regression</td>
<td><code>coxph</code></td>
</tr>
<tr>
<td><code>bj</code></td>
<td>Buckley-James censored least squares model</td>
<td><code>survreg,lm</code></td>
</tr>
<tr>
<td><code>glmD</code></td>
<td>Design version of <code>glm</code></td>
<td><code>glm</code></td>
</tr>
</tbody>
</table>
Table 6.2: Design Transformation Functions

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

Example:

- **treat**: categorical variable with levels "a", "b", "c"
- **num.diseases**: ordinal variable, 0-4
- **age**: continuous
  - Restricted cubic spline
- **cholesterol**: continuous
  - (3 missings; use median)
  - \( \log(\text{cholesterol}+10) \)
- **Allow** \( \text{treat} \times \text{cholesterol} \) interaction
- Program to fit logistic model, test all effects in design, estimate effects (e.g. inter-quartile range odds ratios), plot estimated transformations
<table>
<thead>
<tr>
<th>Function</th>
<th>Purpose</th>
<th>Related Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>print</td>
<td>Print parameters and statistics of fit</td>
<td></td>
</tr>
<tr>
<td>coef</td>
<td>Fitted regression coefficients</td>
<td></td>
</tr>
<tr>
<td>formula</td>
<td>Formula used in the fit</td>
<td></td>
</tr>
<tr>
<td>specs</td>
<td>Detailed specifications of fit</td>
<td></td>
</tr>
<tr>
<td>robcov</td>
<td>Robust covariance matrix estimates</td>
<td></td>
</tr>
<tr>
<td>bootcov</td>
<td>Bootstrap covariance matrix estimates</td>
<td></td>
</tr>
<tr>
<td>pentrace</td>
<td>Find optimum penalty factors by tracing effective AIC for a grid of penalties</td>
<td></td>
</tr>
<tr>
<td>effective.df</td>
<td>Print effective d.f. for each type of variable in model, for penalized fit or pentrace result</td>
<td></td>
</tr>
<tr>
<td>summary</td>
<td>Summary of effects of predictors</td>
<td></td>
</tr>
<tr>
<td>plot.summary</td>
<td>Plot continuously shaded confidence bars for results of summary</td>
<td></td>
</tr>
<tr>
<td>anova</td>
<td>Wald tests of most meaningful hypotheses</td>
<td></td>
</tr>
<tr>
<td>plot.anova</td>
<td>Graphical depiction of anova</td>
<td></td>
</tr>
<tr>
<td>contrast</td>
<td>General contrasts, C.L., tests</td>
<td></td>
</tr>
<tr>
<td>plot</td>
<td>Plot effects of predictors</td>
<td></td>
</tr>
<tr>
<td>gendata</td>
<td>Easily generate predictor combinations</td>
<td></td>
</tr>
<tr>
<td>predict</td>
<td>Obtain predicted values or design matrix</td>
<td></td>
</tr>
<tr>
<td>fastbw</td>
<td>Fast backward step-down variable selection</td>
<td>step</td>
</tr>
<tr>
<td>residuals</td>
<td>(of resid) Residuals, influence stats from fit</td>
<td>resid</td>
</tr>
<tr>
<td>sensuc</td>
<td>Sensitivity analysis for unmeasured confounder</td>
<td></td>
</tr>
<tr>
<td>which.influence</td>
<td>Which observations are overly influential</td>
<td>residuals</td>
</tr>
<tr>
<td>latex</td>
<td>( \LaTeX ) representation of fitted model</td>
<td>Function</td>
</tr>
<tr>
<td>Dialog</td>
<td>Create a menu to enter predictor values and obtain predicted values from fit</td>
<td>latex, plot</td>
</tr>
<tr>
<td>Function</td>
<td>S function analytic representation of ( X \hat{\beta} ) from a fitted regression model</td>
<td>nomogram</td>
</tr>
<tr>
<td>Hazard</td>
<td>S function analytic representation of a fitted hazard function (for psm)</td>
<td></td>
</tr>
<tr>
<td>Survival</td>
<td>S function analytic representation of fitted survival function (for psm, cph)</td>
<td></td>
</tr>
<tr>
<td>Quantile</td>
<td>S function analytic representation of fitted function for quantiles of survival time (for psm, cph)</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>S function analytic representation of fitted function for mean survival time</td>
<td></td>
</tr>
<tr>
<td>nomogram</td>
<td>Draws a nomogram for the fitted model</td>
<td>latex, plot</td>
</tr>
<tr>
<td>survest</td>
<td>Estimate survival probabilities (psm, cph)</td>
<td>survfit</td>
</tr>
<tr>
<td>survplot</td>
<td>Plot survival curves (psm, cph)</td>
<td>plot.survfit</td>
</tr>
<tr>
<td>validate</td>
<td>Validate indexes of model fit using resampling</td>
<td></td>
</tr>
<tr>
<td>calibrate</td>
<td>Estimate calibration curve using resampling</td>
<td>val.prob</td>
</tr>
<tr>
<td>vif</td>
<td>Variance inflation factors for fitted model</td>
<td></td>
</tr>
<tr>
<td>naresid</td>
<td>Bring elements corresponding to missing data back into predictions and residuals</td>
<td></td>
</tr>
<tr>
<td>naprint</td>
<td>Print summary of missing values</td>
<td>aregImpute</td>
</tr>
<tr>
<td>impute</td>
<td>Impute missing values</td>
<td></td>
</tr>
<tr>
<td>fit.mult.impute</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
library(Design, T)  # 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 Design
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()
# Describe the design characteristics
specs(fit)
anova(fit)
plot(anova(fit))
summary(fit)
plot(summary(fit))
summary(fit, treat="b", age=60)
summary(fit, age=c(50,70))
summary(fit, age=c(50,60,70))

# 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')

plot(fit, age=seq(20,80,length=100), treat=NA, conf.int=F)
# Plot relationship between age and log
# odds, separate curve for each treat,
# no C.I.
plot(fit, age=NA, cholesterol=NA)# 3-dimensional perspective plot for age,
# cholesterol, and log odds using default
# ranges for both variables
plot(fit, num.diseases=NA, fun=function(x) 1/(1+exp(-x)) ,
ylab="Prob", conf.int=.9)  # 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 ← plot(f, age=ag, ...)$x.xbeta[,2]
# 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
# plot 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
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))
6.4 Other Functions

- **supsmu**: Friedman’s “super smoother”
- **lowess**: Cleveland’s scatterplot smoother
- **glm**: generalized linear models (see `glmD`)
- **gam**: Generalized additive models
- **rpart**: Like original CART with surrogate splits for missings, censored data extension (Atkinson & Therneau)
- **tree**: classification and regression trees

- **validate.tree** in Design

- **loess**: multi-dimensional scatterplot smoother

```r
f ← loess(y ~ age * pressure)
plot(f)  # cross-sectional plots
ages ← seq(20,70,length=40)
pressures ← seq(80,200,length=40)
pred ← predict(f, expand.grid(age=ages, pressure=pressures))
persp(ages, pressures, pred)  # 3-d plot
```
Chapter 10

Binary Logistic Regression

• $Y = 0, 1$
• Time of event not important
• $\ln C(Y|X)$ $C$ is $\text{Prob}\{Y = 1\}$
• $g(u)$ is $\frac{1}{1 + e^{-u}}$

10.1 Model

$$\text{Prob}\{Y = 1|X\} = \left[1 + \exp(-X\beta)\right]^{-1}.$$  
$$P = \left[1 + \exp(-x)\right]^{-1}$$
Figure 10.1: Logistic function
\[ O = \frac{P}{1-P} \]
\[ P = \frac{O}{1+O} \]
\[ X\beta = \log \frac{P}{1-P} \]
\[ e^{X\beta} = O \]

10.1.1 Model Assumptions and Interpretation of Parameters

\[
\text{logit}\{Y = 1|X\} = \text{logit}(P) = \log[P/(1-P)] = X\beta, 
\]

- Increase \( X_j \) by \( d \) → increase odds \( Y = 1 \) by \( \exp(\beta_jd) \), increase log odds by \( \beta_jd \).

- If there is only one predictor \( X \) and that predictor is binary, the model can be written

\[
\text{logit}\{Y = 1|X = 0\} = \beta_0 \\
\text{logit}\{Y = 1|X = 1\} = \beta_0 + \beta_1. 
\]

- One continuous predictor:

\[
\text{logit}\{Y = 1|X\} = \beta_0 + \beta_1X, 
\]
• Two treatments (indicated by $X_1 = 0$ or 1) and one continuous covariable ($X_2$).

\[
\logit\{Y = 1|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2,
\]

\[
\logit\{Y = 1|X_1 = 0, X_2\} = \beta_0 + \beta_2 X_2
\]

\[
\logit\{Y = 1|X_1 = 1, X_2\} = \beta_0 + \beta_1 + \beta_2 X_2.
\]

10.1.2 Odds Ratio, Risk Ratio, and Risk Difference

• Odds ratio capable of being constant

• Ex: risk factor doubles odds of disease

<table>
<thead>
<tr>
<th>Without Risk Factor</th>
<th>With Risk Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Probability</td>
</tr>
<tr>
<td></td>
<td>.2</td>
</tr>
<tr>
<td></td>
<td>.5</td>
</tr>
<tr>
<td></td>
<td>.8</td>
</tr>
<tr>
<td></td>
<td>.9</td>
</tr>
<tr>
<td></td>
<td>.98</td>
</tr>
</tbody>
</table>

Let $X_1$ be a binary risk factor and let $A = \{X_2, \ldots, X_p\}$ be the other factors. Then the
Figure 10.2: Absolute benefit as a function of risk of the event in a control subject and the relative effect (odds ratio) of the risk factor. The odds ratios are given for each curve.
estimate of $\text{Prob}\{Y = 1|X_1 = 1, A\} - \text{Prob}\{Y = 1|X_1 = 0, A\}$ is

$$\frac{1}{1 + \exp[-(\beta_0 + \beta_1 + \beta_2 X_2 + \ldots + \beta_p X_p)]} - \frac{1}{1 + \exp[-(\beta_0 + \beta_2 X_2 + \ldots + \beta_p X_p)]}$$

$$= \frac{1}{1 + \left(\frac{1-R}{R}\right)\exp(-\beta_1)} - \hat{R},$$

where $R = \text{Prob}[Y = 1|X_1 = 0, A]$.

- Risk ratio is $\frac{1+e^{-X_2/\beta}}{1+e^{-X_1/\beta}}$
- Does not simplify like odds ratio, which is $\frac{e^{X_1/\beta}}{e^{X_2/\beta}} = e^{(X_1 - X_2)/\beta}$

### 10.1.3 Detailed Example

**TABLE OF SEX BY RESPONSE**

<table>
<thead>
<tr>
<th>SEX</th>
<th>RESPONSE</th>
<th>Frequency</th>
<th>Row Pct</th>
<th>Odds/Log</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0</td>
<td>1</td>
<td>Total</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td>14</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>70.00</td>
<td>30.00</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td></td>
<td>6</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30.00</td>
<td>70.00</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>20</td>
<td>20</td>
<td>40</td>
</tr>
</tbody>
</table>

M:F odds ratio = (14/6)/(6/14) = 5.44, log=-1.695
Figure 10.3: Data, subgroup proportions, and fitted logistic model
CHAPTER 10. BINARY LOGISTIC REGRESSION

STATISTICS FOR TABLE OF SEX BY RESPONSE

<table>
<thead>
<tr>
<th>Statistic</th>
<th>DF</th>
<th>Value</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi Square</td>
<td>1</td>
<td>6.400</td>
<td>0.011</td>
</tr>
<tr>
<td>Likelihood Ratio Chi-Square</td>
<td>1</td>
<td>6.583</td>
<td>0.010</td>
</tr>
</tbody>
</table>

Fitted Logistic Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std Err</th>
<th>Wald $\chi^2$</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>-0.8472978</td>
<td>0.48795</td>
<td>3.015237</td>
<td></td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>1.6945956</td>
<td>0.69007</td>
<td>6.030474</td>
<td>0.0141</td>
</tr>
</tbody>
</table>

Log likelihood ($\beta_1 = 0$) : -27.727
Log likelihood (max) : -24.435
LR $\chi^2 (H_0 : \beta_1 = 0)$ : -2(-27.727- -24.435) = 6.584

Next, consider the relationship between age and response, ignoring sex.

<table>
<thead>
<tr>
<th>AGE</th>
<th>RESPONSE</th>
<th>Frequency</th>
<th>Row Pct</th>
<th>0</th>
<th>1</th>
<th>Total</th>
<th>Odds/Log</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;45</td>
<td>I</td>
<td>8</td>
<td>61.5</td>
<td>5</td>
<td>5</td>
<td>13</td>
<td>5/8=.625</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>38.4</td>
<td>-.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45-54</td>
<td>I</td>
<td>6</td>
<td>50.0</td>
<td>6</td>
<td>6</td>
<td>12</td>
<td>6/6=1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50.0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>55+</td>
<td>I</td>
<td>6</td>
<td>40.0</td>
<td>9</td>
<td>9</td>
<td>15</td>
<td>9/6=1.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>60.0</td>
<td>.405</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>20</td>
<td>100.0</td>
<td>20</td>
<td>20</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

55+ : <45 odds ratio = (9/6)/(5/8) = 2.4, log=.875
The estimate of $\beta_1$ is in rough agreement with that obtained from the frequency table. The 55+:<45 log odds ratio is .875, and since the respective mean ages in the 55+ and <45 age groups are 61.1 and 40.2, an estimate of the log odds ratio increase per year is .875/(61.1–40.2)=.875/20.9=.042.

The likelihood ratio test for $H_0$: no association between age and response is obtained as follows:

\[
\begin{align*}
\text{Log likelihood (}{\beta_1}=0) & : -27.727 \\
\text{Log likelihood (max)} & : -26.511 \\
\text{LR } \chi^2(H_0 : \beta_1 = 0) & : -2(-27.727- -26.511) = 2.432
\end{align*}
\]

(Compare 2.432 with the Wald statistic 2.28.)

Next we consider the simultaneous association of age and sex with response.
A logistic model for relating sex and age simultaneously to response is given below.

### Fitted Logistic Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Std Err</th>
<th>Wald $\chi^2$</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>-9.8429426</td>
<td>3.67576</td>
<td>7.17057</td>
<td>0.0074</td>
</tr>
<tr>
<td>$\beta_1$ (sex)</td>
<td>3.4898280</td>
<td>1.19917</td>
<td>8.46928</td>
<td>0.0036</td>
</tr>
<tr>
<td>$\beta_2$ (age)</td>
<td>0.1580583</td>
<td>0.06164</td>
<td>6.57556</td>
<td>0.0103</td>
</tr>
</tbody>
</table>

### SEX=F

<table>
<thead>
<tr>
<th>AGE</th>
<th>RESPONSE</th>
<th>Frequency</th>
<th>Row Pct</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;45</td>
<td></td>
<td>4</td>
<td>100.0</td>
<td>0</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>45-54</td>
<td></td>
<td>4</td>
<td>80.0</td>
<td>1</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>55+</td>
<td></td>
<td>6</td>
<td>54.6</td>
<td>5</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>14</td>
<td></td>
<td>6</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

### SEX=M

<table>
<thead>
<tr>
<th>AGE</th>
<th>RESPONSE</th>
<th>Frequency</th>
<th>Row Pct</th>
<th>0</th>
<th>1</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;45</td>
<td></td>
<td>4</td>
<td>44.4</td>
<td>5</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>45-54</td>
<td></td>
<td>2</td>
<td>28.6</td>
<td>5</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>55+</td>
<td></td>
<td>0</td>
<td>0.0</td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>6</td>
<td></td>
<td>14</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>
Likelihood ratio tests are obtained from the information below.
Log likelihood ($\beta_1 = 0, \beta_2 = 0$) : -27.727
Log likelihood (max) : -19.458
Log likelihood ($\beta_1 = 0$) : -26.511
Log likelihood ($\beta_2 = 0$) : -24.435
LR $\chi^2$ ($H_0 : \beta_1 = \beta_2 = 0$) : $-2(-27.727 - -19.458) = 16.538$
LR $\chi^2$ ($H_0 : \beta_1 = 0$) sex|age : $-2(-26.511 - -19.458) = 14.106$
LR $\chi^2$ ($H_0 : \beta_2 = 0$) age|sex : $-2(-24.435 - -19.458) = 9.954$

The 14.1 should be compared with the Wald statistic of 8.47, and 9.954 should be compared with 6.58. The fitted logistic model is plotted separately for females and males in Figure 10.3. The fitted model is

$$
\text{logit}\{\text{Response} = 1|\text{sex, age}\} = -9.84 + 3.49 \times \text{sex} + .158 \times \text{age}
$$

where as before sex=0 for females, 1 for males. For example, for a 40 year old female, the predicted logit is $-9.84 + .158(40) = -3.52$. The predicted probability of a response is $1/[1 + \exp(3.52)] = .029$. For a 40 year old male, the predicted logit is $-9.84 + 3.49 + .158(40) = -.03$, with a probability of .492.
10.1.4 Design Formulations

- Can do ANOVA using \( k - 1 \) dummies for a \( k \)-level predictor
- Can get same \( \chi^2 \) statistics as from a contingency table
- Can go farther: covariable adjustment
- Simultaneous comparison of multiple variables between two groups: Turn problem backwards to predict group from all the dependent variables
- This is more robust than a parametric multivariate test
- Propensity scores for adjusting for nonrandom treatment selection: Predict treatment from all baseline variables
- Adjusting for the predicted probability of getting a treatment adjusts adequately for confounding from all of the variables
- In a randomized study, using logistic model to adjust for covariables, even with perfect
balance, will improve the treatment effect estimate

10.2 Estimation

10.2.1 Maximum Likelihood Estimates

Like binomial case but \( P \)'s vary; \( \hat{\beta} \) computed by trial and error using an iterative maximization technique

10.2.2 Estimation of Odds Ratios and Probabilities

\[
\hat{P}_i = \left[ 1 + \exp(-X_i\hat{\beta}) \right]^{-1}.
\]
\[
\left\{ 1 + \exp[-(X_i\hat{\beta} \pm zs)] \right\}^{-1}.
\]

10.3 Test Statistics

- Likelihood ratio test best
• Score test second best (score $\chi^2$ $\equiv$ Pearson $\chi^2$)
• Wald test may misbehave but is quick

10.4 Residuals

Partial residuals (to check predictor transformations)

$$r_{ij} = \hat{\beta}_j X_{ij} + \frac{Y_i - \hat{P}_i}{\hat{P}_i(1 - \hat{P}_i)};$$

10.5 Assessment of Model Fit

$$\text{logit}\{Y = 1|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2,$$

• Can verify by plotting stratified proportions
\[ \logit \{Y=1\} = X \]

Figure 10.4: Logistic regression assumptions for one binary and one continuous predictor

- \( \hat{P} \) = number of events divided by stratum size
- \( \hat{O} = \frac{\hat{P}}{1-\hat{P}} \)
- Plot \( \log \hat{O} \) (scale on which linearity is assumed)
- Stratified estimates are noisy
- 1 or 2 \( X \)s \( \rightarrow \) nonparametric smoother
- `plsmo` function makes it easy to use loess to compute logits of nonparametric estimates (\( \text{fun=qlogis} \))
- General: restricted cubic spline expansion
Figure 10.5: Logit proportions of significant coronary artery disease by sex and deciles of age for n=3504 patients, with spline fits (smooth curves). Spline fits are for $k = 4$ knots at age=36, 48, 56, and 68 years, and interaction between age and sex is allowed. Smooth nonparametric estimates are shown as dotted curves. Data courtesy of the Duke Cardiovascular Disease Databank.
of one or more predictors

\[
\text{logit}\{Y = 1|X\} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X'_2 + \hat{\beta}_4 X''_2
\]

\[
= \hat{\beta}_0 + \hat{\beta}_1 X_1 + f(X_2),
\]

\[
\text{logit}\{Y = 1|X\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X'_2 + \beta_4 X''_2
\]

\[
+ \beta_5 X_1 X_2 + \beta_6 X_1 X'_2 + \beta_7 X_1 X''_2
\]

<table>
<thead>
<tr>
<th>Model / Hypothesis</th>
<th>Likelihood Ratio (\chi^2)</th>
<th>d.f.</th>
<th>(P)</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>a: sex, age (linear, no interaction)</td>
<td>766.0</td>
<td>2</td>
<td></td>
<td>((b - a))</td>
</tr>
<tr>
<td>b: sex, age, age (\times) sex</td>
<td>768.2</td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>c: sex, spline in age</td>
<td>769.4</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>d: sex, spline in age, interaction</td>
<td>782.5</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(H_0: ) no age (\times) sex interaction</td>
<td>2.2</td>
<td>1</td>
<td>.14</td>
<td>((b - a))</td>
</tr>
<tr>
<td>(H_0: ) age linear (\mid) no interaction</td>
<td>3.4</td>
<td>2</td>
<td>.18</td>
<td>((c - a))</td>
</tr>
<tr>
<td>(H_0: ) age linear, no interaction</td>
<td>16.6</td>
<td>5</td>
<td>.005</td>
<td>((d - a))</td>
</tr>
<tr>
<td>(H_0: ) age linear, product form interaction</td>
<td>14.4</td>
<td>4</td>
<td>.006</td>
<td>((d - b))</td>
</tr>
<tr>
<td>(H_0: ) no interaction, allowing for nonlinearity in age</td>
<td>13.1</td>
<td>3</td>
<td>.004</td>
<td>((d - c))</td>
</tr>
</tbody>
</table>

- Example of finding transform. of a single continuous predictor
- Duration of symptoms vs. odds of severe coronary disease
- Look at AIC to find best \# knots for the money
<table>
<thead>
<tr>
<th>$k$</th>
<th>Model $\chi^2$</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>99.23</td>
<td>97.23</td>
</tr>
<tr>
<td>3</td>
<td>112.69</td>
<td>108.69</td>
</tr>
<tr>
<td>4</td>
<td>121.30</td>
<td>115.30</td>
</tr>
<tr>
<td>5</td>
<td>123.51</td>
<td>115.51</td>
</tr>
<tr>
<td>6</td>
<td>124.41</td>
<td>114.51</td>
</tr>
</tbody>
</table>

Figure 10.6: Estimated relationship between duration of symptoms and the log odds of severe coronary artery disease for $k = 5$. Knots are marked with arrows. Solid line is spline fit; dotted line is a nonparametric “super-smoothed” estimate.

- Sample of 2258 pts
Figure 10.7: Fitted linear logistic model in $\log_{10}(\text{duration}+1)$, with subgroup estimates using groups of 150 patients. Fitted equation is $\text{logit}(tvdln) = -0.9809 + 0.7122 \log_{10}(\text{months} + 1)$. 
• Predict significant coronary disease
• For now stratify age into tertiles to examine interactions simply
• Model has 2 dummies for age, sex, age $\times$ sex, 4-knot restricted cubic spline in cholesterol, age tertile $\times$ cholesterol

Figure 10.8: Log odds of significant coronary artery disease modeling age with two dummy variables

anova(fit)
Wald Statistics

<table>
<thead>
<tr>
<th>Factor</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>age.tertile (Main+Interactions)</td>
<td>112.62</td>
<td>10</td>
<td>0.0000</td>
</tr>
<tr>
<td>All Interactions</td>
<td>22.37</td>
<td>8</td>
<td>0.0043</td>
</tr>
<tr>
<td>sex (Main+Interactions)</td>
<td>328.90</td>
<td>3</td>
<td>0.0000</td>
</tr>
<tr>
<td>All Interactions</td>
<td>9.61</td>
<td>2</td>
<td>0.0082</td>
</tr>
<tr>
<td>cholesterol (Main+Interactions)</td>
<td>94.01</td>
<td>9</td>
<td>0.0000</td>
</tr>
<tr>
<td>All Interactions</td>
<td>10.03</td>
<td>6</td>
<td>0.1234</td>
</tr>
<tr>
<td>Nonlinear (Main+Interactions)</td>
<td>10.30</td>
<td>6</td>
<td>0.1124</td>
</tr>
<tr>
<td>age.tertile * sex</td>
<td>9.61</td>
<td>2</td>
<td>0.0082</td>
</tr>
<tr>
<td>age.tertile * cholesterol</td>
<td>10.03</td>
<td>6</td>
<td>0.1232</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A, B)$ vs. $AB$</td>
<td>2.40</td>
<td>4</td>
<td>0.6635</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>10.30</td>
<td>6</td>
<td>0.1124</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>22.37</td>
<td>8</td>
<td>0.0043</td>
</tr>
<tr>
<td>TOTAL NONLINEAR+INTERACTION</td>
<td>30.12</td>
<td>10</td>
<td>0.0008</td>
</tr>
<tr>
<td>TOTAL</td>
<td>404.94</td>
<td>14</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

- Now model age as continuous predictor
- Start with nonparametric surface using $Y = 0/1$
- Next try parametric fit using linear spline in age, chol. (3 knots each), all product terms
- Next try smooth spline surface, include all cross-products
Figure 10.9: Local regression fit for the probability of significant coronary disease vs. age and cholesterol for males, based on the S-PLUS loess function
Figure 10.10: Linear spline surface for males, with knots for age at 46, 52, 59 and knots for cholesterol at 196, 224, and 259 (quartiles)
Figure 10.11: Restricted cubic spline surface in two variables, each with $k = 4$ knots
### Wald Statistics

<table>
<thead>
<tr>
<th>Factor</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>age * cholesterol</td>
<td>12.95</td>
<td>9</td>
<td>0.1649</td>
</tr>
<tr>
<td>Nonlinear Interaction: $f(A, B)$ vs. $AB$</td>
<td>7.27</td>
<td>8</td>
<td>0.5078</td>
</tr>
<tr>
<td>$f(A, B)$ vs. $Af(B) + Bg(A)$</td>
<td>5.41</td>
<td>4</td>
<td>0.2480</td>
</tr>
<tr>
<td>Nonlinear Interaction in age vs. $Af(B)$</td>
<td>6.44</td>
<td>6</td>
<td>0.3753</td>
</tr>
<tr>
<td>Nonlinear Interaction in cholesterol vs. $Bg(A)$</td>
<td>6.27</td>
<td>6</td>
<td>0.3931</td>
</tr>
</tbody>
</table>

- Now restrict surface by excluding doubly non-linear terms

![Restricted cubic spline fit with age × spline(cholesterol) and cholesterol × spline(age)](image_url)
### Wald Statistics

<table>
<thead>
<tr>
<th>Factor</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>age * cholesterol</td>
<td>10.83</td>
<td>5</td>
<td>0.0548</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A, B)$ vs. $AB$</td>
<td>3.12</td>
<td>4</td>
<td>0.5372</td>
</tr>
<tr>
<td>Nonlinear Interaction in age vs. $Af(B)$</td>
<td>1.60</td>
<td>2</td>
<td>0.4496</td>
</tr>
<tr>
<td>Nonlinear Interaction in cholesterol vs. $Bg(A)$</td>
<td>1.64</td>
<td>2</td>
<td>0.4399</td>
</tr>
</tbody>
</table>

- Finally restrict the interaction to be a simple product. The Wald test for age $\times$ cholesterol interaction yields $\chi^2 = 7.99$ with 1 d.f., $p = .005$.

![Figure 10.13: Spline fit with nonlinear effects of cholesterol and age and a simple product interaction](image)
• See how well this simple interaction model compares with initial model using 2 dummies for age

• Request predictions to be made at mean age within tertiles

Figure 10.14: Predictions from linear interaction model with mean age in tertiles indicated.

• Using residuals for “duration of symptoms” example

• Relative merits of strat., nonparametric, splines
Stratification for checking fit

<table>
<thead>
<tr>
<th>Method</th>
<th>Choice Required</th>
<th>Assumes Additivity</th>
<th>Uses Ordering of $X$</th>
<th>Low Variance</th>
<th>Good Resolution on $X$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratification</td>
<td>Intervals</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoother on $X_1$</td>
<td>Bandwidth</td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>stratifying on $X_2$</td>
<td>(not on $X_2$)</td>
<td>(if min. strat.)</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Smooth partial</td>
<td>Bandwidth</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>residual plot</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spline model for all $X$s</td>
<td>Knots</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Hosmer-Lemeshow test is a commonly used test of goodness-of-fit of a binary logistic model.
  - Compares proportion of events with mean predicted probability within deciles of $\hat{P}$
  - Arbitrary (number of groups, how to form groups)
– Low power (too many d.f.)
– Does not reveal the culprits

• A new omnibus test based of SSE has more power and requires no grouping; still does not lead to corrective action.

• Any omnibus test lacks power against specific alternatives such as nonlinearity or interaction

10.6 Collinearity

10.7 Overly Influential Observations

10.8 Quantifying Predictive Ability

• Generalized $R^2$: equals ordinary $R^2$ in normal case:

$$R_N^2 = \frac{1 - \exp(-LR/n)}{1 - \exp(-L^0/n)},$$
• Brier score (calibration + discrimination):
  \[ B = \frac{1}{n} \sum_{i=1}^{n} (\hat{P}_i - Y_i)^2, \]

• \( c = \) “concordance probability” = ROC area

• Related to Wilcoxon-Mann-Whitney stat and Somers’ \( D_{xy} \)
  \[ D_{xy} = 2(c - .5). \]

• “Percent classified correctly” has lots of problems

10.9 Validating the Fitted Model

• Possible indexes
  – Accuracy of \( \hat{P} \): calibration
    Plot \( \frac{1}{1+e^{-X_{new}\beta_{old}}} \) against estimated prob.
    that \( Y = 1 \) on new data
  – Discrimination: \( C \) or \( D_{xy} \)
  – \( R^2 \) or \( B \)
• Use bootstrap to estimate calibration equation

\[ P_c = \text{Prob}\{Y = 1 | X\hat{\beta}\} = [1 + \exp - (\gamma_0 + \gamma_1 X\hat{\beta})]^{-1}, \]

\[ E_{\text{max}}(a, b) = \max_{a \leq \hat{P} \leq b} |\hat{P} - \hat{P}_c|, \]

• Bootstrap validation of age-sex-response data, 80 samples

• 2 predictors forced into every model

<table>
<thead>
<tr>
<th>Index</th>
<th>Original Sample</th>
<th>Training Sample</th>
<th>Test Sample</th>
<th>Optimism</th>
<th>Corrected Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D_{xy})</td>
<td>0.70</td>
<td>0.70</td>
<td>0.67</td>
<td>0.03</td>
<td>0.66</td>
</tr>
<tr>
<td>(R^2)</td>
<td>0.45</td>
<td>0.47</td>
<td>0.43</td>
<td>0.04</td>
<td>0.41</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Slope</td>
<td>1.00</td>
<td>1.00</td>
<td>0.91</td>
<td>0.09</td>
<td>0.91</td>
</tr>
<tr>
<td>(E_{\text{max}})</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>D</td>
<td>0.39</td>
<td>0.42</td>
<td>0.36</td>
<td>0.06</td>
<td>0.33</td>
</tr>
<tr>
<td>U</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.02</td>
<td>-0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>Q</td>
<td>0.44</td>
<td>0.47</td>
<td>0.35</td>
<td>0.12</td>
<td>0.32</td>
</tr>
<tr>
<td>B</td>
<td>0.16</td>
<td>0.15</td>
<td>0.17</td>
<td>-0.02</td>
<td>0.18</td>
</tr>
</tbody>
</table>

• Allow for step-down at each re-sample

• Use individual tests at \(\alpha = 0.10\)

• Both age and sex selected in 76 of 80, neither in 1 sample
Table 10.2: Validation of 2-variable Stepwise Model

<table>
<thead>
<tr>
<th>Index</th>
<th>Original Sample</th>
<th>Training Sample</th>
<th>Test Sample</th>
<th>Optimism</th>
<th>Corrected Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{xy}$</td>
<td>0.70</td>
<td>0.71</td>
<td>0.66</td>
<td>0.05</td>
<td>0.64</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.45</td>
<td>0.50</td>
<td>0.42</td>
<td>0.07</td>
<td>0.38</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>-0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>Slope</td>
<td>1.00</td>
<td>1.00</td>
<td>0.86</td>
<td>0.14</td>
<td>0.86</td>
</tr>
<tr>
<td>$E_{max}$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>D</td>
<td>0.39</td>
<td>0.45</td>
<td>0.36</td>
<td>0.09</td>
<td>0.30</td>
</tr>
<tr>
<td>U</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.02</td>
<td>-0.07</td>
<td>0.02</td>
</tr>
<tr>
<td>Q</td>
<td>0.44</td>
<td>0.50</td>
<td>0.34</td>
<td>0.16</td>
<td>0.27</td>
</tr>
<tr>
<td>B</td>
<td>0.16</td>
<td>0.15</td>
<td>0.18</td>
<td>-0.03</td>
<td>0.19</td>
</tr>
</tbody>
</table>

- Try adding 5 noise candidate variables
Table 10.3: Validation of Model with 5 Noise Variables

<table>
<thead>
<tr>
<th>Index</th>
<th>Original Sample</th>
<th>Training Sample</th>
<th>Test Sample</th>
<th>Optimism</th>
<th>Corrected Index</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D_{xy}$</td>
<td>0.70</td>
<td>0.32</td>
<td>0.26</td>
<td>0.05</td>
<td>0.64</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.45</td>
<td>0.23</td>
<td>0.17</td>
<td>0.06</td>
<td>0.39</td>
</tr>
<tr>
<td>Intercept</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.03</td>
<td>0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td>Slope</td>
<td>1.00</td>
<td>1.00</td>
<td>0.85</td>
<td>0.15</td>
<td>0.85</td>
</tr>
<tr>
<td>$E_{max}$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>D</td>
<td>0.39</td>
<td>0.21</td>
<td>0.13</td>
<td>0.07</td>
<td>0.32</td>
</tr>
<tr>
<td>U</td>
<td>-0.05</td>
<td>-0.05</td>
<td>0.03</td>
<td>-0.08</td>
<td>0.03</td>
</tr>
<tr>
<td>Q</td>
<td>0.44</td>
<td>0.26</td>
<td>0.10</td>
<td>0.15</td>
<td>0.29</td>
</tr>
<tr>
<td>B</td>
<td>0.16</td>
<td>0.20</td>
<td>0.23</td>
<td>-0.03</td>
<td>0.19</td>
</tr>
</tbody>
</table>

- Number of Factors Selected

<table>
<thead>
<tr>
<th>Frequency</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>38</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

The first 15 patterns of factors selected are:

```
age sex x1 x2 x3 x4 x5
*  *
*  *  *
*  *
*  *
*  *  *  *
*  *
*  *
*  *
*  *
*  *
```

10.10 Describing the Fitted Model
### Table 10.11: Effect of Factors on the Odds of Coronary Disease

<table>
<thead>
<tr>
<th>Factor</th>
<th>Low</th>
<th>High</th>
<th>Diff.</th>
<th>Effect</th>
<th>S.E.</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>46</td>
<td>59</td>
<td>13</td>
<td>0.90</td>
<td>0.21</td>
<td>0.49</td>
<td>1.32</td>
</tr>
<tr>
<td>Odds Ratio</td>
<td>46</td>
<td>59</td>
<td>13</td>
<td>2.47</td>
<td>NA</td>
<td>1.63</td>
<td>3.74</td>
</tr>
<tr>
<td>cholesterol</td>
<td>196</td>
<td>259</td>
<td>63</td>
<td>0.79</td>
<td>0.18</td>
<td>0.44</td>
<td>1.15</td>
</tr>
<tr>
<td>Odds Ratio</td>
<td>196</td>
<td>259</td>
<td>63</td>
<td>2.21</td>
<td>NA</td>
<td>1.55</td>
<td>3.17</td>
</tr>
<tr>
<td>sex - female:male</td>
<td>1</td>
<td>2</td>
<td>NA</td>
<td>-2.46</td>
<td>0.15</td>
<td>-2.75</td>
<td>-2.16</td>
</tr>
<tr>
<td>Odds Ratio</td>
<td>1</td>
<td>2</td>
<td>NA</td>
<td>0.09</td>
<td>NA</td>
<td>0.06</td>
<td>0.12</td>
</tr>
</tbody>
</table>

### Figure 10.16: Odds Ratios and Confidence Bars

- **Odds Ratio**
  - 0.10
  - 1.50
  - 3.00
- **Age (59:46)**
- **Cholesterol (259:196)**
- **Sex (Female:Male)**

Adjusted to: age=52, sex=male, cholesterol=224

Figure 10.16: Odds ratios and confidence bars, using quartiles of age and cholesterol for assessing their effects on the odds of coronary disease.

### 10.11 S Functions
Figure 10.17: Linear spline fit for probability of bacterial vs. viral meningitis as a function of age at onset. Copyrighted 1989, American Medical Association. Reprinted by permission.
Figure 10.18: (A) Relationship between myocardium at risk and ventricular fibrillation, based on the individual best fit equations for animals anesthetized with pentobarbital and α-chloralose. The amount of myocardium at risk at which 0.5 of the animals are expected to fibrillate (MAR\textsubscript{50}) is shown for each anesthetic group. (B) Relationship between myocardium at risk and ventricular fibrillation, based on equations derived from the single slope estimate. Note that the MAR\textsubscript{50} describes the overall relationship between myocardium at risk and outcome when either the individual best fit slope or the single slope estimate is used. The shift of the curve to the right during α-chloralose anesthesia is well described by the shift in MAR\textsubscript{50}. Test for interaction had P=0.10\textsuperscript{78}. Reprinted by permission, NRC Research Press.
Figure 10.19: A nomogram for estimating the likelihood of significant coronary artery disease (CAD) in women. ECG = electrocardiographic; MI = myocardial infarction\textsuperscript{61}. Reprinted from American Journal of Medicine, Vol 75, Pryor DB et al., “Estimating the likelihood of significant coronary artery disease”, p. 778, Copyright 1983, with permission from Excerpta Medica, Inc.
Figure 10.20: Nomogram for estimating probability of bacterial (ABM) vs. viral (AVM) meningitis. Step 1, place ruler on reading lines for patient's age and month of presentation and mark intersection with line A; step 2, place ruler on values for glucose ratio and total polymorphonuclear leukocyte (PMN) count in cerebrospinal fluid and mark intersection with line B; step 3, use ruler to join marks on lines A and B, then read off the probability of ABM vs. AVM. Copyrighted 1989, American Medical Association. Reprinted by permission.
Figure 10.21: Nomogram relating age, sex, and cholesterol to the log odds and to the probability of significant coronary artery disease. Select one axis corresponding to sex and to age ∈ {30, 40, 50, 60, 70}. There was linear interaction between age and sex and between age and cholesterol. 0.70 and 0.90 confidence intervals are shown (0.90 in gray). Note that for the “Linear Predictor” scale there are various lengths of confidence intervals near the same value of $X\hat{\beta}$, demonstrating that the standard error of $X\hat{\beta}$ depends on the individual $X$ values. Also note that confidence intervals corresponding to smaller patient groups (e.g., females) are wider.
Chapter 11

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 of the University of Virginia. It is available in R, S-PLUS, and Excel formats from hesweb1.med.virginia.edu/biostat/s/data under the name titanic3.

11.1 Descriptive Statistics

> library(Hmisc,T); library(Design,T)
> # List of names of variables to analyze
> v ← c('pclass','survived','age','sex','sibsp','parch')
> describe(titanic3[,v])

titanic3

6 Variables 1309 Observations

pclass : Passenger Class

<table>
<thead>
<tr>
<th>n missing</th>
<th>unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>1309</td>
<td>0</td>
</tr>
</tbody>
</table>

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

survived : Survived

<table>
<thead>
<tr>
<th>n missing</th>
<th>unique</th>
<th>Sum</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1309</td>
<td>0</td>
<td>2</td>
<td>0.382</td>
</tr>
</tbody>
</table>

age : Age (Year)

<table>
<thead>
<tr>
<th>n missing</th>
<th>unique</th>
<th>Mean</th>
<th>.05</th>
<th>.10</th>
<th>.25</th>
<th>.50</th>
<th>.75</th>
<th>.90</th>
<th>.95</th>
</tr>
</thead>
<tbody>
<tr>
<td>1046</td>
<td>263</td>
<td>98</td>
<td>29.88</td>
<td>5</td>
<td>14</td>
<td>21</td>
<td>28</td>
<td>39</td>
<td>50</td>
</tr>
</tbody>
</table>

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

sex : Sex

<table>
<thead>
<tr>
<th>n missing</th>
<th>unique</th>
</tr>
</thead>
<tbody>
<tr>
<td>1309</td>
<td>0</td>
</tr>
</tbody>
</table>

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

sibsp : Number of Siblings/Spouses Aboard

<table>
<thead>
<tr>
<th>n missing</th>
<th>unique</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1309</td>
<td>0</td>
<td>0.4989</td>
</tr>
</tbody>
</table>

Frequency 891 319 42 20 22 2 6 9
% 68 24 3 2 2 0 1

parch : Number of Parents/Children Aboard

<table>
<thead>
<tr>
<th>n missing</th>
<th>unique</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1309</td>
<td>0</td>
<td>0.385</td>
</tr>
</tbody>
</table>

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 Design
> options(datadist='dd')
> attach(titanic3[,v])
Show 4-way relationships after collapsing levels. Suppress estimates based on < 25 passengers.
Figure 11.1: Univariable summaries of Titanic survival
CHAPTER 11. LOGISTIC MODEL CASE STUDY: SURVIVAL OF TITANIC PASSENGERS

```r
> 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 library
> Dotplot(pclass ~ survived | sibsp.parch*agec,
+ groups=sex, data=s, pch=1:2,
+ xlab='Proportion Surviving') # Figure 11.2
> Key(.09,-.065)
```

Figure 11.2: Multi-way summary of Titanic survival
11.2 Exploring Trends with Nonparametric Regression

> par(mfrow=c(2,2))  # To create Figure 11.3
> plsmo(age, survived, datadensity=T)
> plsmo(age, survived, group=sex, datadensity=T)
> plsmo(age, survived, group=pclass, datadensity=T)
> plsmo(age, survived, group=interaction(pclass,sex),
+       datadensity=T)

Figure 11.3: 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.

> par(mfrow=c(1,2))  # Figure 11.4
> plsmo(age, survived, group=cut2(sibsp,0:2), datadensity=T)
> plsmo(age, survived, group=cut2(parch,0:2), datadensity=T)
CHAPTER 11. LOGISTIC MODEL CASE STUDY: SURVIVAL OF TITANIC PASSENGERS

11.3 Binary Logistic Model with Casewise Deletion of Missing Values

First fit a model that is saturated with respect to age, sex, pclass. Insufficient variation in sibsp, parch to fit complex interactions or nonlinearities.

```r
> f1 <- lrm(survived ~ sex*pclass*rcs(age,5) + rcs(age,5)*(sibsp + parch))
> anova(f1)  # actually used latex(anova(f1)) to get Table 11.2
```

3-way interactions, parch clearly insignificant, so drop
### Table 11.2: Wald Statistics for survived

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

Frequencies of Responses
0 1
619 427

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

Obs Max Deriv Model L.R. d.f. P C Dxy Gamma Tau-a R2 Brier
1046 6e-006 554 26 0 0.88 0.76 0.76 0.37 0.56 0.13

Coef S.E. Wald Z P
Intercept 3.30746 1.84266 1.79 0.0727
sex=male -1.14781 1.08782 -1.06 0.2914
pclass=2nd 6.73087 3.96166 1.70 0.0893
pclass=3rd -1.64369 1.82988 -0.90 0.3691
age 0.08860 0.13455 0.66 0.5102
age’ -0.74099 0.65128 -1.14 0.2552
age’’ 4.92642 4.00468 1.23 0.2186
age’’’ -6.61296 5.41003 -1.22 0.2216
sibsp -1.04461 0.34414 -3.04 0.0024
sex=male * pclass=2nd -0.76822 0.70827 -1.08 0.2781
sex=male * pclass=3rd 2.15200 0.62140 3.46 0.0005
sex=male * age -0.21911 0.07215 -3.04 0.0024
sex=male * age’ 1.08422 0.38862 2.79 0.0053
sex=male * age’’ -6.55781 2.65108 -2.47 0.0134
sex=male * age’’’ 8.37161 3.85324 2.17 0.0298
pclass=2nd * age -0.54459 0.26526 -2.05 0.0401
pclass=3rd * age -0.16335 0.13083 -1.25 0.2118
pclass=2nd * age’ 1.91559 1.01892 1.88 0.0601
pclass=3rd * age’ 0.82045 0.60914 1.35 0.1780
pclass=2nd * age’’ -8.95448 5.50269 -1.63 0.1037
pclass=3rd * age’’ -5.42760 3.64751 -1.49 0.1367
pclass=2nd * age’’’ 9.39265 6.95595 1.35 0.1769
pclass=3rd * age’’’ 7.54036 4.85185 1.55 0.1202
age * sibsp 0.03571 0.03398 1.05 0.2933
age’ * sibsp -0.04665 0.22126 -0.21 0.8330
age’’ * sibsp 0.55743 1.66797 0.33 0.7382
age’’’ * sibsp -1.19370 2.57112 -0.46 0.6425

> anova(f)  # Table 11.3
Show the many effects of predictors.

```r
> for(sx in c('female','male'))
+   plot(f, age=NA, pclass=NA, sex=sx, # Fig. 11.5
+       fun=plogis, ylim=c(0,1), add=sx=='male',
+       conf.int=F, col=if(sx=='female')1 else 2,
+       adj.subtitle=F, lty=1)
> plot(f, sibsp=NA, age=c(10,15,20,50), conf.int=F) # Fig. 11.6
```

Figure 11.5: 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.

Note that children having many siblings apparently had lower survival. Married adults had slightly higher survival than unmarried ones.
## Table 11.3: Wald Statistics for survived

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>199.42</td>
<td>7</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>56.14</td>
<td>6</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>pclass (Factor+Higher Order Factors)</td>
<td>108.73</td>
<td>12</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>42.83</td>
<td>10</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>age (Factor+Higher Order Factors)</td>
<td>47.04</td>
<td>20</td>
<td>0.0006</td>
</tr>
<tr>
<td>All Interactions</td>
<td>24.51</td>
<td>16</td>
<td>0.0789</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>22.72</td>
<td>15</td>
<td>0.0902</td>
</tr>
<tr>
<td>sibsp (Factor+Higher Order Factors)</td>
<td>19.95</td>
<td>5</td>
<td>0.0013</td>
</tr>
<tr>
<td>All Interactions</td>
<td>10.99</td>
<td>4</td>
<td>0.0267</td>
</tr>
<tr>
<td>sex $\times$ pclass (Factor+Higher Order Factors)</td>
<td>35.40</td>
<td>2</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>sex $\times$ age (Factor+Higher Order Factors)</td>
<td>10.08</td>
<td>4</td>
<td>0.0391</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>8.17</td>
<td>3</td>
<td>0.0426</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A,B)$ vs. $AB$</td>
<td>8.17</td>
<td>3</td>
<td>0.0426</td>
</tr>
<tr>
<td>pclass $\times$ age (Factor+Higher Order Factors)</td>
<td>6.86</td>
<td>8</td>
<td>0.5516</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>6.11</td>
<td>6</td>
<td>0.4113</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A,B)$ vs. $AB$</td>
<td>6.11</td>
<td>6</td>
<td>0.4113</td>
</tr>
<tr>
<td>age $\times$ sibsp (Factor+Higher Order Factors)</td>
<td>10.99</td>
<td>4</td>
<td>0.0267</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>1.81</td>
<td>3</td>
<td>0.6134</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A,B)$ vs. $AB$</td>
<td>1.81</td>
<td>3</td>
<td>0.6134</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>22.72</td>
<td>15</td>
<td>0.0902</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>67.58</td>
<td>18</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>70.68</td>
<td>21</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>253.18</td>
<td>26</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>
Figure 11.6: 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.
Validate the model using the bootstrap to check overfitting. Ignoring two very insignificant pooled tests.

```r
> f <- update(f, x=T, y=T)
> # x=T,y=T adds raw data to fit object so can bootstrap
> set.seed(131) # so can replicate re-samples
> options(digits=2)

> validate(f, B=80)

    index.orig training test optimism index.corrected n
    Dxy     0.7560    0.7742 0.7417  0.0325 0.7235     80
    R2      0.5545    0.5800 0.5293  0.0507 0.5039     80
   Intercept 0.0000    0.0000 -0.0505  0.0505 -0.0505     80
    Slope   1.0000    1.0000 0.8806  0.1194  0.8806     80
     Emax   0.0000    0.0000 0.0368  0.0368  0.0368     80
         B  0.1303    0.1251 0.1340 -0.0089 0.1392     80

> cal ← calibrate(f, B=80) # Figure 11.7
> plot(cal)
```
Figure 11.7: Bootstrap overfitting-corrected loess nonparametric calibration curve for casewise deletion model
Figure 11.8: 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 \texttt{is.na(age)}. The \texttt{rpart} function found only strong patterns according to passenger class.
But moderate problem with missing data

11.4 Examining Missing Data Patterns

```r
> na.patterns <- naclus(titanic3)
> library(rpart)
> who.na <- rpart(is.na(age) ~ sex + pclass + survived +
    + sibsp + parch, minbucket=15)
> par(mfrow=c(2,2))
> naplot(na.patterns, 'na per var')
> plot(na.patterns)
> plot(who.na); text(who.na)              # Figure 11.8
> par(mfrow=c(1,1))                      # Reset to 1x1 plot setup

> plot(summary(is.na(age) ~ sex + pclass + survived +
    + sibsp + parch))  # Figure 11.9
```
Figure 11.9: Univariable descriptions of proportion of passengers with missing age
Table 11.4: Wald Statistics for `is.na(age)`

<table>
<thead>
<tr>
<th></th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>5.61</td>
<td>3</td>
<td>0.1324</td>
</tr>
<tr>
<td>All Interactions</td>
<td>5.58</td>
<td>2</td>
<td>0.0614</td>
</tr>
<tr>
<td>pclass (Factor+Higher Order Factors)</td>
<td>68.43</td>
<td>4</td>
<td>$&lt; 0.0001$</td>
</tr>
<tr>
<td>All Interactions</td>
<td>5.58</td>
<td>2</td>
<td>0.0614</td>
</tr>
<tr>
<td>survived</td>
<td>0.98</td>
<td>1</td>
<td>0.3232</td>
</tr>
<tr>
<td>sibsp</td>
<td>0.35</td>
<td>1</td>
<td>0.5548</td>
</tr>
<tr>
<td>parch</td>
<td>7.92</td>
<td>1</td>
<td>0.0049</td>
</tr>
<tr>
<td>sex $\times$ pclass (Factor+Higher Order Factors)</td>
<td>5.58</td>
<td>2</td>
<td>0.0614</td>
</tr>
<tr>
<td>TOTAL</td>
<td>82.90</td>
<td>8</td>
<td>$&lt; 0.0001$</td>
</tr>
</tbody>
</table>

> m ← lrm(is.na(age) ~ sex $\times$ pclass + survived + sibsp + parch)
> m

Frequencies of Responses
FALSE TRUE
1046 263

<table>
<thead>
<tr>
<th>Obs</th>
<th>Max Deriv Model L.R. d.f. P</th>
<th>C</th>
<th>Dxy</th>
<th>Gamma</th>
<th>Tau-a</th>
<th>R2</th>
<th>Brier</th>
</tr>
</thead>
<tbody>
<tr>
<td>1309</td>
<td>5e-006</td>
<td>115</td>
<td>0</td>
<td>0.7</td>
<td>0.45</td>
<td>0.13</td>
<td>0.13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coef</th>
<th>S.E.</th>
<th>Wald Z</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.20299</td>
<td>0.36412</td>
<td>-6.05</td>
</tr>
<tr>
<td>sex=male</td>
<td>0.64395</td>
<td>0.39526</td>
<td>1.63</td>
</tr>
<tr>
<td>pclass=2nd</td>
<td>-1.00793</td>
<td>0.66578</td>
<td>-1.51</td>
</tr>
<tr>
<td>pclass=3rd</td>
<td>1.61242</td>
<td>0.35962</td>
<td>4.48</td>
</tr>
<tr>
<td>survived</td>
<td>-0.18058</td>
<td>0.18280</td>
<td>-0.99</td>
</tr>
<tr>
<td>sibsp</td>
<td>0.04353</td>
<td>0.07372</td>
<td>0.59</td>
</tr>
<tr>
<td>parch</td>
<td>-0.35261</td>
<td>0.12529</td>
<td>-2.81</td>
</tr>
<tr>
<td>sex=male $\times$ pclass=2nd</td>
<td>0.13467</td>
<td>0.75451</td>
<td>0.18</td>
</tr>
<tr>
<td>sex=male $\times$ pclass=3rd</td>
<td>-0.85628</td>
<td>0.42144</td>
<td>-2.03</td>
</tr>
</tbody>
</table>

> anova(m)  # Table 11.4

`pclass` and `parch` are the important predictors of missing age.
11.5 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

```r
> xtrans ← transcan(~ I(age) + sex + pclass + sibsp + parch,
+       imputed=T, pl=F)
> summary(xtrans)

Iterations:
R-squared achieved in predicting each variable:

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>sex</th>
<th>pclass</th>
<th>sibsp</th>
<th>parch</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>0.258</td>
<td>0.078</td>
<td>0.244</td>
<td>0.241</td>
<td>0.288</td>
</tr>
</tbody>
</table>

Adjusted R-squared:

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>sex</th>
<th>pclass</th>
<th>sibsp</th>
<th>parch</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>0.254</td>
<td>0.074</td>
<td>0.24</td>
<td>0.238</td>
<td>0.285</td>
</tr>
</tbody>
</table>

Coefficients of canonical variates for predicting each (row) variable

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>sex</th>
<th>pclass</th>
<th>sibsp</th>
<th>parch</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>-0.89</td>
<td>6.13</td>
<td>-1.81</td>
<td>-2.77</td>
<td></td>
</tr>
<tr>
<td>value</td>
<td>-0.02</td>
<td>0.56</td>
<td>0.10</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>value</td>
<td>0.08</td>
<td>0.26</td>
<td>0.07</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>value</td>
<td>-0.02</td>
<td>0.04</td>
<td>0.07</td>
<td>0.87</td>
<td></td>
</tr>
<tr>
<td>value</td>
<td>-0.03</td>
<td>0.29</td>
<td>0.22</td>
<td>0.75</td>
<td></td>
</tr>
</tbody>
</table>

Summary of imputed values

<table>
<thead>
<tr>
<th></th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>.05</td>
</tr>
</tbody>
</table>
CHAPTER 11. LOGISTIC MODEL CASE STUDY: SURVIVAL OF TITANIC PASSENGERS

263 0 24 28.41 16.76 21.66 26.17 28.04 28.04 42.92 42.92

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)

> i ← is.imputed(age.i)
> tapply(age.i[i], list(sex[i], pclass[i]), mean)

1st 2nd 3rd
female 38.1 27.1 22.4
male 40.7 29.7 25.0

> tapply(age, list(sex, pclass), mean, na.rm=T)

1st 2nd 3rd
female 37 27.5 22.2
male 41 30.8 26.0

> dd ← datadist(dd, age.i)
> f.si ← lrm(survived ~ (sex + pclass + rcs(age.i,5))^2 +
+ rcs(age.i,5)*sibsp)
> f.si

Frequencies of Responses
0 1
809 500

Obs Max Deriv Model L.R. d.f. P C Dxy Gamma Tau-a R2 Brier
1309 0.0004 641 26 0.86 0.72 0.73 0.34 0.53 0.13

........
> par(mfrow=c(1,2)) # Figure 11.10
> plot(f, age=NA, pclass=NA, ylim=c(-4,2),
+     conf.int=F, adj.subtitle=F)
> plot(f.si, age.i=NA, pclass=NA, ylim=c(-4,2),
+     conf.int=F, adj.subtitle=F)

Figure 11.10: Predicted log odds 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.

> anova(f.si) # Table 11.5
<table>
<thead>
<tr>
<th>Term</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>245.53</td>
<td>7</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>52.80</td>
<td>6</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>pclass (Factor+Higher Order Factors)</td>
<td>112.02</td>
<td>12</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>36.77</td>
<td>10</td>
<td>0.0001</td>
</tr>
<tr>
<td>age.i (Factor+Higher Order Factors)</td>
<td>49.25</td>
<td>20</td>
<td>0.0003</td>
</tr>
<tr>
<td>All Interactions</td>
<td>25.53</td>
<td>16</td>
<td>0.0610</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>19.86</td>
<td>15</td>
<td>0.1772</td>
</tr>
<tr>
<td>sibsp (Factor+Higher Order Factors)</td>
<td>21.74</td>
<td>5</td>
<td>0.0006</td>
</tr>
<tr>
<td>All Interactions</td>
<td>12.25</td>
<td>4</td>
<td>0.0156</td>
</tr>
<tr>
<td>sex $\times$ pclass (Factor+Higher Order Factors)</td>
<td>30.25</td>
<td>2</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>sex $\times$ age.i (Factor+Higher Order Factors)</td>
<td>8.95</td>
<td>4</td>
<td>0.0622</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>5.63</td>
<td>3</td>
<td>0.1308</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A,B)$ vs. $AB$</td>
<td>5.63</td>
<td>3</td>
<td>0.1308</td>
</tr>
<tr>
<td>pclass $\times$ age.i (Factor+Higher Order Factors)</td>
<td>6.04</td>
<td>8</td>
<td>0.6427</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>5.44</td>
<td>6</td>
<td>0.4882</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A,B)$ vs. $AB$</td>
<td>5.44</td>
<td>6</td>
<td>0.4882</td>
</tr>
<tr>
<td>age.i $\times$ sibsp (Factor+Higher Order Factors)</td>
<td>12.25</td>
<td>4</td>
<td>0.0156</td>
</tr>
<tr>
<td>Nonlinear</td>
<td>2.04</td>
<td>3</td>
<td>0.5639</td>
</tr>
<tr>
<td>Nonlinear Interaction : $f(A,B)$ vs. $AB$</td>
<td>2.04</td>
<td>3</td>
<td>0.5639</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>19.86</td>
<td>15</td>
<td>0.1772</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>66.83</td>
<td>18</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>69.48</td>
<td>21</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>305.58</td>
<td>26</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>
11.6 Multiple Imputation

Again needed to force a linear age relationship.

```r
> set.seed(17)  # so can reproduce the random draws
> xtrans <- transcan(~ I(age) + sex + pclass +
+                    sibsp + parch + survived,
+                    n.impute=5, pl=F,
+                    trantab=T, imputed=T)

> summary(xtrans)

Adjusted R-squared achieved in predicting each variable:

age  sex  pclass  sibsp  parch  survived
0.28  0.3   0.33   0.25   0.29   0.37

Coefficients of canonical variates for predicting each (row) variable

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>sex</th>
<th>pclass</th>
<th>sibsp</th>
<th>parch</th>
<th>survived</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>0.37</td>
<td>6.76</td>
<td>-2.17</td>
<td>-2.48</td>
<td>-5.37</td>
<td></td>
</tr>
<tr>
<td>sex</td>
<td>0.00</td>
<td>-0.10</td>
<td>0.15</td>
<td>0.21</td>
<td>1.95</td>
<td></td>
</tr>
<tr>
<td>pclass</td>
<td>0.07</td>
<td>-0.09</td>
<td>0.12</td>
<td>0.16</td>
<td>1.26</td>
<td></td>
</tr>
<tr>
<td>sibsp</td>
<td>-0.03</td>
<td>0.18</td>
<td>0.15</td>
<td>0.84</td>
<td>-0.65</td>
<td></td>
</tr>
<tr>
<td>parch</td>
<td>-0.03</td>
<td>0.22</td>
<td>0.18</td>
<td>0.75</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>survived</td>
<td>-0.01</td>
<td>0.23</td>
<td>0.16</td>
<td>-0.06</td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>

Summary of imputed values

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>missing</th>
<th>unique</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>1315</td>
<td>0</td>
<td>1006</td>
<td>28.96</td>
</tr>
</tbody>
</table>

lowest: 0.1667 0.5826 0.6028 0.6567 0.9900
highest: 68.9805 70.5000 72.2900 73.8221 74.0000
Starting estimates for imputed values:

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>pclass</th>
<th>sibsp</th>
<th>parch</th>
<th>survived</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

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

1 2 3 4 5 16 37 49 46 48 49 38 45 32 50 40 26 41 42 36 67 29 43 47 48 36 67 30 49 42 60 57 39 38 48 45 70 38 32 31 30 18 71 42 40 62 41 54 75 40 42 34 44 29 81 33 38 46 63 62 107 49 38 47 37 37

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

> f.mi <- fit.mult.impute(survived ~ (sex + pclass + rcs(age,5))^2 +
                        + rcs(age,5)*sibsp,
                        + lrm, xtrans)

Variance Inflation Factors Due to Imputation:

Intercept 1 1.1 1 1 1 1.1 1.1 1.3
sex=male 1.1 1 1 1 1 1.1
pclass=2nd 1 1 1 1.1
pclass=3rd 1 1 1 1 1.1

sex=male * pclass=2nd 1 1 1 1 1
sex=male * age 1 1 1 1 1.1
sex=male * age' 1 1 1.1 1 1
sex=male * age'' 1 1 1 1 1.1
pclass=3rd * age 1 1 1 1 1.1
pclass=2nd * age 1 1 1 1 1
pclass=3rd * age' 1 1 1.1 1 1
pclass=2nd * age' 1 1 1 1 1.1
pclass=3rd * age'' 1 1 1 1 1.1
pclass=2nd * age'' 1 1 1 1 1.1
pclass=3rd * age''' 1 1 1 1 1.1
pclass=2nd * age''' 1 1 1 1 1.1
pclass=3rd * age''' 1 1 1 1 1.3
age * sibsp 1 1 1 1 1.1
age' * sibsp 1 1 1 1 1.3
age'' * sibsp 1 1 1 1 1.3
age''' * sibsp 1 1 1 1 1.3

age * sibsp 1 1 1 1 1.1
age' * sibsp 1 1 1 1 1.3
age'' * sibsp 1 1 1 1 1.3
age''' * sibsp 1 1 1 1 1.3

sex * pclass 1 1 1 1 1
sex * age 1 1 1 1 1.1
sex * age' 1 1 1 1 1.1
sex * age'' 1 1 1 1 1.3
sex * age''' 1 1 1 1 1.3
pclass * age 1 1 1 1 1
pclass * age' 1 1 1 1 1.1
pclass * age'' 1 1 1 1 1.3
pclass * age''' 1 1 1 1 1.3
sibsp * age 1 1 1 1 1
sibsp * age' 1 1 1 1 1.1
sibsp * age'' 1 1 1 1 1.3
sibsp * age''' 1 1 1 1 1.3

VIF 1.1 1 1 1 1.1 1.1 1.3 1 1.2 1 1 1.1 1.1 1.1 1.1 1.1 1.3 1 1 1 1 1.1 1.1 1.3 1.1 1
The Wald $\chi^2$ 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.
### Table 11.6: Wald Statistics for survived

<table>
<thead>
<tr>
<th>Factor</th>
<th>$\chi^2$</th>
<th>d.f.</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sex (Factor+Higher Order Factors)</td>
<td>237.33</td>
<td>7</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>53.50</td>
<td>6</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>pclass (Factor+Higher Order Factors)</td>
<td>116.29</td>
<td>12</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>All Interactions</td>
<td>37.18</td>
<td>10</td>
<td>0.0001</td>
</tr>
<tr>
<td>age (Factor+Higher Order Factors)</td>
<td>46.58</td>
<td>20</td>
<td>0.0007</td>
</tr>
<tr>
<td>All Interactions</td>
<td>23.78</td>
<td>16</td>
<td>0.0944</td>
</tr>
<tr>
<td>Nonlinear (Factor+Higher Order Factors)</td>
<td>20.12</td>
<td>15</td>
<td>0.1673</td>
</tr>
<tr>
<td>sibsp (Factor+Higher Order Factors)</td>
<td>17.81</td>
<td>5</td>
<td>0.0032</td>
</tr>
<tr>
<td>All Interactions</td>
<td>8.03</td>
<td>4</td>
<td>0.0905</td>
</tr>
<tr>
<td>sex × pclass (Factor+Higher Order Factors)</td>
<td>32.15</td>
<td>2</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>sex × age (Factor+Higher Order Factors)</td>
<td>10.06</td>
<td>4</td>
<td>0.0394</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>7.73</td>
<td>3</td>
<td>0.0520</td>
</tr>
<tr>
<td>pcass × age (Factor+Higher Order Factors)</td>
<td>5.68</td>
<td>8</td>
<td>0.6828</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>5.16</td>
<td>6</td>
<td>0.5240</td>
</tr>
<tr>
<td>age × sibsp (Factor+Higher Order Factors)</td>
<td>8.03</td>
<td>4</td>
<td>0.0905</td>
</tr>
<tr>
<td>Nonlinear Interaction : f(A,B) vs. AB</td>
<td>1.31</td>
<td>3</td>
<td>0.7256</td>
</tr>
<tr>
<td>TOTAL NONLINEAR</td>
<td>20.12</td>
<td>15</td>
<td>0.1673</td>
</tr>
<tr>
<td>TOTAL INTERACTION</td>
<td>64.94</td>
<td>18</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TOTAL NONLINEAR + INTERACTION</td>
<td>67.84</td>
<td>21</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>TOTAL</td>
<td>291.66</td>
<td>26</td>
<td>&lt; 0.0001</td>
</tr>
</tbody>
</table>
Figure 11.11: Predicted log odds 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 \( \text{sibsp}=0 \).
11.7 Summarizing the Fitted Model

Show odds ratios for changes in predictor values

```r
> s ← summary(f.mi, age=c(1,30), sibsp=0:1)
> # override default ranges for 3 variables
> plot(s, log=T)  # Figure 11.12
```

Figure 11.12: 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')
> options(digits=1)

> data.frame(combos, phat)

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>pclass</th>
<th>sibsp</th>
<th>phat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>female</td>
<td>1st</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td>21</td>
<td>female</td>
<td>1st</td>
<td>0.98</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>female</td>
<td>1st</td>
<td>0.97</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>male</td>
<td>1st</td>
<td>0.85</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>male</td>
<td>1st</td>
<td>0.50</td>
</tr>
<tr>
<td>6</td>
<td>50</td>
<td>male</td>
<td>1st</td>
<td>0.26</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>female</td>
<td>2nd</td>
<td>1.00</td>
</tr>
<tr>
<td>8</td>
<td>21</td>
<td>female</td>
<td>2nd</td>
<td>0.90</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
<td>female</td>
<td>2nd</td>
<td>0.83</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>male</td>
<td>2nd</td>
<td>1.00</td>
</tr>
<tr>
<td>11</td>
<td>21</td>
<td>male</td>
<td>2nd</td>
<td>0.07</td>
</tr>
<tr>
<td>12</td>
<td>50</td>
<td>male</td>
<td>2nd</td>
<td>0.03</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>female</td>
<td>3rd</td>
<td>0.78</td>
</tr>
<tr>
<td>14</td>
<td>21</td>
<td>female</td>
<td>3rd</td>
<td>0.58</td>
</tr>
<tr>
<td>15</td>
<td>50</td>
<td>female</td>
<td>3rd</td>
<td>0.36</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>male</td>
<td>3rd</td>
<td>0.80</td>
</tr>
<tr>
<td>17</td>
<td>21</td>
<td>male</td>
<td>3rd</td>
<td>0.14</td>
</tr>
<tr>
<td>18</td>
<td>50</td>
<td>male</td>
<td>3rd</td>
<td>0.05</td>
</tr>
</tbody>
</table>

We can also get predicted values by creating an S-PLUS function that will evaluate the model on demand.

> pred.logit ← Function(f.mi)

> pred.logit
function(sex = "male", pclass = "3rd", age = 28, sibsp = 0) {
  3.440147 - 1.4381326 * (sex == "male") + 6.296592 * (pclass == "2nd") - 2.0779809 * (pclass == "3rd") + 0.082134387 * age - 0.00025084004 * pmax(age - 5, 0)^3 + 0.0016482259 * pmax(age - 21, 0)^3 - 0.0022035565 * pmax(age - 28, 0)^3 + 0.00088444824 * pmax(age - 37, 0)^3 - 7.8277628e-005 *
A nomogram could be used to obtain predicted values manually, but this is not feasible when so many interaction terms are present. Dialog function for Win/NT S-PLUS.
> drep ← dataRep(~ roundN(age,10) + sex + pclass +
+    roundN(sibsp, clip=0:1))
> Dialog(fitPar('f.mi', lp=F, fun=list('Prob[Survival]'=plogis)),
+    limits='data', basename='Titanic',
+    vary=list(sex=c('female','male')), datarep=drep)
> runmenu.Titanic()

<table>
<thead>
<tr>
<th>Library</th>
<th>Purpose</th>
<th>Functions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hmisc</td>
<td>Miscellaneous functions</td>
<td>summary, plsmo, naclus, llist, latex</td>
</tr>
<tr>
<td>Harrell</td>
<td></td>
<td>summarize, Dotplot, transcan, impute</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fit.mult.impute, describe, dataRep</td>
</tr>
<tr>
<td>Design</td>
<td>Modeling, validation,</td>
<td>datadist, lrm, nomogram, validate, calibrate,</td>
</tr>
<tr>
<td>Harrell</td>
<td>Model presentation</td>
<td>Function, rcs, Dialog</td>
</tr>
<tr>
<td></td>
<td>graphics</td>
<td></td>
</tr>
<tr>
<td>rpart</td>
<td>Recursive partitioning</td>
<td>rpart</td>
</tr>
<tr>
<td>Atkinson&amp;Therneau</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bibliography


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/twiki/bin/view/Main/RmS?topic=ClinStat.