REGRESSION MODELING STRATEGIES WITH APPLICATION TO LOGISTIC REGRESSION

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

INSTITUTE OF BIOMEDICAL INFORMATICS
LJUBLJANA UNIVERSITY 28 NOV.-2 DEC. 2005

Contents

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 $\rightarrow$ multiple linear regression

  - Wilcoxon, Kruskal-Wallis, Spearman $\rightarrow$ proportional odds ordinal logistic model
  - log-rank $\rightarrow$ Cox
- Models not only allow for multiplicity adjustment but for shrinkage of estimates
  - Statisticians comfortable with P-value adjustment but fail to recognize that the difference between the most different treatments is badly biased

Statistical estimation is usually model-based

- Relative effect of increasing cholesterol from 200 to 250 mg/dl on hazard of death, holding other risk factors constant
- Adjustment depends on how other risk factors relate to hazard
- Usually interested in adjusted (partial) effects, not unadjusted (marginal or crude) effects

1.2 Examples of Uses of Predictive Multivariable Modeling

- Financial performance, consumer purchasing, loan pay-back
- Ecology
- Product life
- Employment discrimination
- Medicine, epidemiology, health services research
- Probability of diagnosis, time course of a disease
- Comparing non-randomized treatments
- Getting the correct estimate of relative effects in randomized studies requires covariable adjustment if model is nonlinear
  - Crude odds ratios biased towards 1.0 if sample heterogeneous
- Estimating absolute treatment effect (e.g., risk difference)

  - Use e.g. difference in two predicted probabilities
- Cost-effectiveness ratios
  - incremental cost / incremental *ABSOLUTE* benefit
  - most studies use avg. cost difference / avg. benefit, which may apply to no one

1.3 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[48] 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[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

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

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

2.2 Model Formulations

General regression model

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

General linear regression model

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

Examples

$$\begin{aligned} C(Y|X) = & \quad E(Y|X) = \quad X\beta, \\ Y|X & \sim N(X\beta, \sigma^2) \end{aligned}$$

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

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

Example:

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

General linear regression model:
$C'(Y|X) = X\beta$.

2.3 Interpreting Model Parameters

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

$$C'(Y|X) = X\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$$

$$\begin{aligned}\beta_j &= C'(Y|X_1, X_2, \ldots, X_j + 1, \ldots, X_p) \\ &- C'(Y|X_1, X_2, \ldots, X_j, \ldots, X_p)\end{aligned}$$

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

2.3.1 Nominal Predictors

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

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

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

where

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

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

2.3.2 Interactions

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

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

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

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

If X_1 is binary, the interaction may take the form of a difference in shape (and/or distribution) of X_2 vs. $C(Y)$ depending on whether

$X_1 = 0$ or $X_1 = 1$ (e.g. logarithm vs. square root).

2.3.3 Example: Inference for a Simple Model

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

Model assumes

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

Interpretations of parameters:

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

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

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

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

Hypotheses that are usually inappropriate:

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

More useful hypotheses follow. For any hypothesis need to

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

Most Useful Tests for Linear age $\times$ sex Model

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

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

2.4 Review of Composite (Chunk) Tests

- In the model

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

  we may want to jointly test the association between all body measurements and response, holding `age` and `sex` constant.
- This 3 d.f. test may be obtained two ways:
  - Remove the 3 variables and compute the change in SSR or SSE
  - Test $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ using matrix algebra (e.g., `anova(fit, weight, waist, tricep)`)

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 $\beta_1, X \leq a$, $\beta_1 + \beta_2, X > a$.
- β_2 is the slope increment as you pass a

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

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

where

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

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

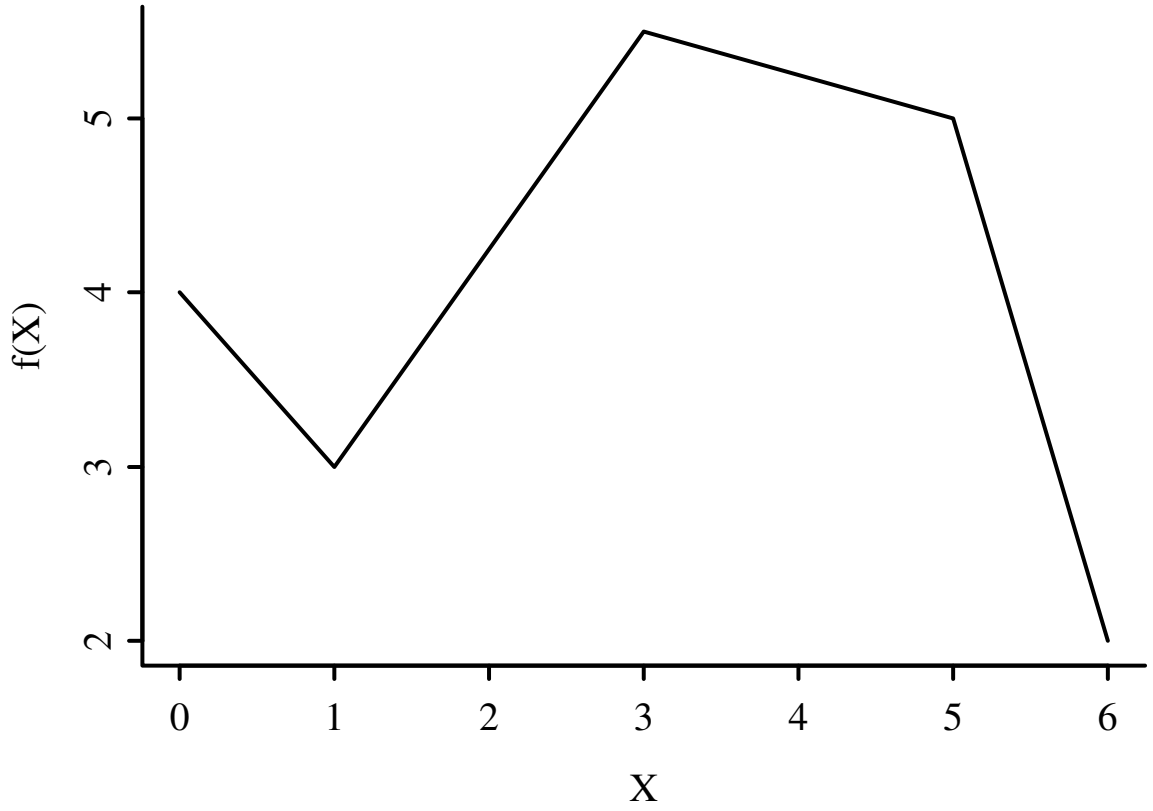

Figure 2.1: A linear spline function with knots at a=1, b=3, c=5

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

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

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

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

2.5.3 Cubic Spline Functions

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

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

$$X_1 = X \qquad X_2 = X^2$$
$$X_3 = X^3 \quad X_4 = (X-a)_+^3$$

$$X_5 = (X - b)_+^3 \; X_6 = (X - c)_+^3.$$

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

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

2.5.4 Restricted Cubic Splines

Stone and Koo[72]: cubic splines poorly behaved in tails. Constrain function to be linear in tails. $k + 3 \rightarrow k - 1$ parameters[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 βs are redundant, i.e., are just combinations of the other βs.

The restricted spline function with k knots $t_1, \ldots, t_k$ is given by[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 $j = 1, \ldots, k-2$,

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

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

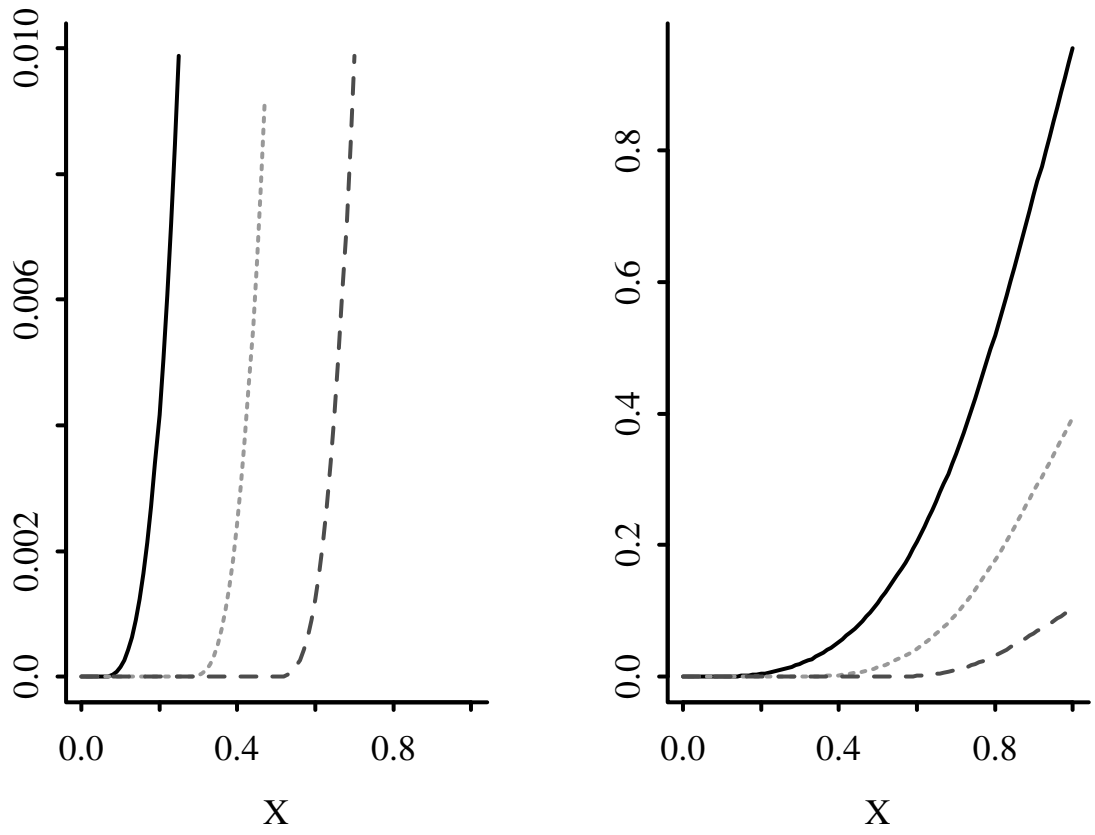

Figure 2.2: Restricted cubic spline component variables for k=5 and knots at $X = .05, .275, .5, .725$, and $.95$. Left panel is a magnification of the right. Fitted functions such as those in Figure 2.3 will be linear combinations of these basis functions as long as knots are at the same locations used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

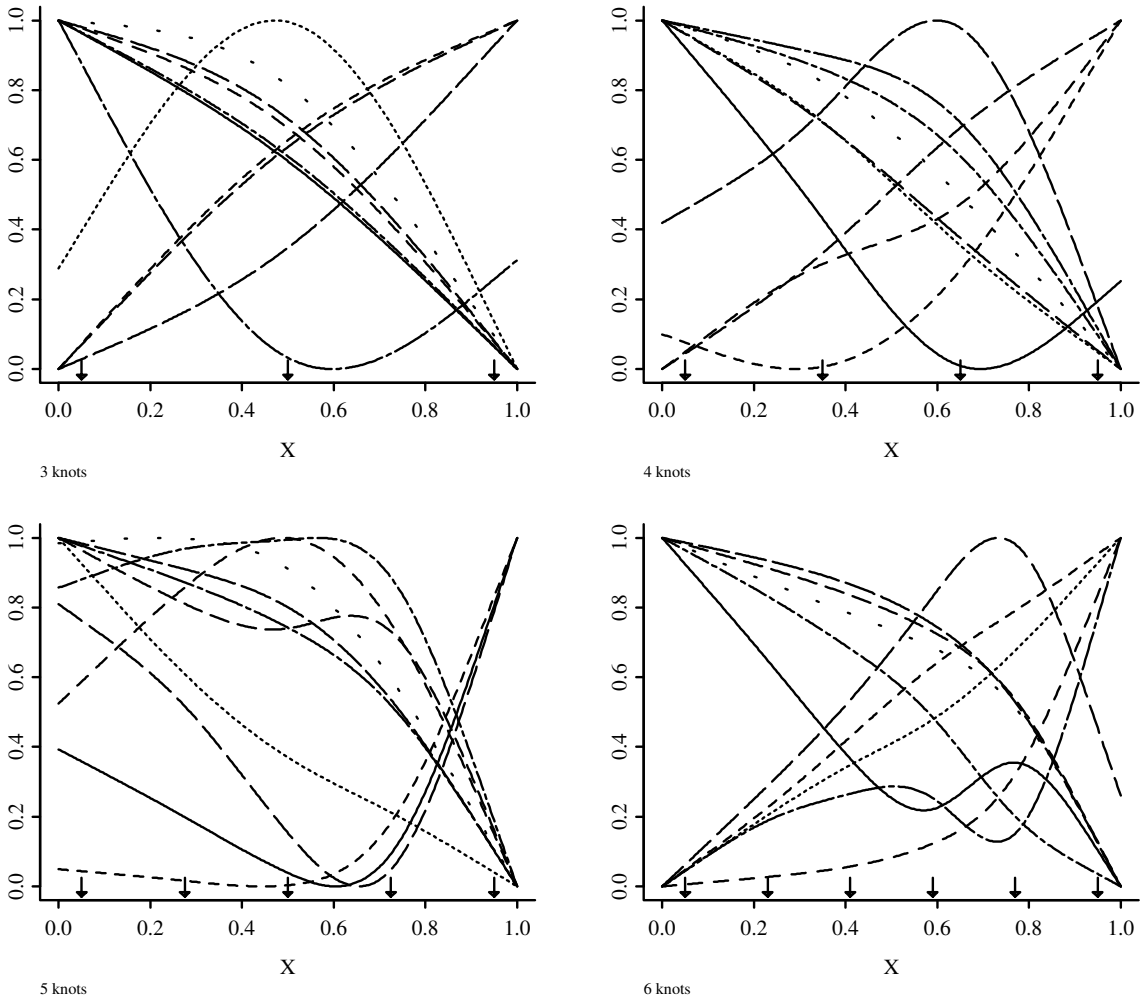

Figure 2.3: Some typical restricted cubic spline functions for $k = 3, 4, 5, 6$. The y-axis is $X\beta$. Arrows indicate knots. These curves were derived by randomly choosing values of β subject to standard deviations of fitted functions being normalized. See the Web site for a script to create more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

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

A test of linearity in X can be obtained by testing

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

2.5.5 Choosing Number and Position of Knots

- Knots are specified in advance in regression splines
- Locations not important in most situations[29, 71]
- Place knots where data exist — fixed quantiles of predictor's marginal distribution

- Fit depends more on choice of k

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

$n < 100$ – replace outer quantiles with 5th smallest and 5th largest X[72].

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

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

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

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

- Moving linear regression far superior to moving avg. (moving flat line)
- Cleveland's[18] moving linear regression smoother *loess* (locally weighted least squares) is the most popular smoother. To estimate central tendency of Y at $X = x$:
  - take all the data having X values within a suitable interval about x (default is $\frac{2}{3}$ of the data)
  - fit weighted least squares linear regression within this neighborhood
  - points near x given the most weight[b]
  - points near extremes of interval receive almost no weight
  - loess works much better at extremes of X than moving avg.
  - provides an estimate at each observed X; other estimates obtained by linear interpolation

[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[c]
- 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[44]:

- 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,

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

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[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[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 $\rightarrow$ dangerous to interpret plot
- What to do if 4 d.f. test insignificant, 3 d.f. test for linearity insig., 1 d.f. test sig. after delete nonlinear terms?

Grambsch and O'Brien [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 $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$
- Two ways to test association between X and Y
- Fit quadratic model and test for linearity ($H_0 : \beta_2 = 0$)
- F-test for linearity significant at $\alpha = 0.05$ level $\rightarrow$ report as the final test of association the 2 d.f. F test of $H_0 : \beta_1 = \beta_2 = 0$
- If the test of linearity insignificant, refit without the quadratic term and final test of association is 1 d.f. test, $H_0 : \beta_1 = 0 | \beta_2 = 0$
- Showed that type I error $> \alpha$
- Fairly accurate P-value obtained by instead testing against F with 2 d.f. even at second stage
- Cause: are retaining the most significant part of F

- **BUT** if test against 2 d.f. can only lose power when compared with original F for testing both βs
- SSR from quadratic model $>$ SSR from linear model

2.8 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

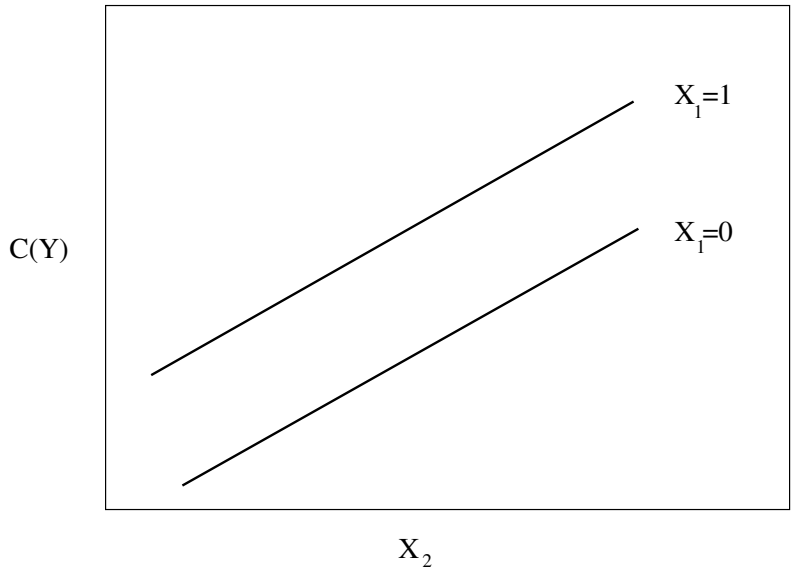

Figure 2.4: Regression assumptions for one binary and one continuous predictor

for checking fit:

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

   Advantage: Simplicity

   Disadvantages:
   - Can only compute standard residuals for uncensored continuous response
   - Subjective judgment of non-randomness
   - Hard to handle interaction
   - Hard to see patterns with large n (trend lines help)
   - Seeing patterns does not lead to corrective action
2. Scatterplot of Y vs. X_2 using different symbols according to values of X_1

Advantages: Simplicity, can see interaction

Disadvantages:

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

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

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

   Disadvantages:

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

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

   Advantages: All regression aspects of the

model can be summarized efficiently with minimal assumptions

Disadvantages:

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

5. Fit flexible nonlinear parametric model

Advantages:

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

Disadvantages:

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

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

Restricted cubic spline works well for method 5.

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

where

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

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

- Plot $\hat{f}(X_2)$ vs. X_2
- n large $\rightarrow$ can fit separate functions by X_1
- Test of linearity: $H_0 : \beta_3 = \beta_4 = 0$
- Nonlinear $\rightarrow$ use transformation suggested by spline fit or keep spline terms
- Tentative transformation $g(X_2) \rightarrow$ check adequacy by expanding $g(X_2)$ in spline function and testing linearity
- Can find transformations by plotting $g(X_2)$ vs. $\hat{f}(X_2)$ for variety of g
- Multiple continuous predictors $\rightarrow$ expand each

using spline

- Example: assess linearity of X_2, X_3

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

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 $X_2 g(X_1)$:

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

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

General spline surface:

- Cover $X_1 \times X_2$ plane with grid and fit patchwise 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[13]
- Interactions of order > 2: more trouble
- 2-way interactions among p predictors: pooled tests
- p tests each with $p - 1$ d.f.

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

- Treatment $\times$ severity of disease being treated
- Age $\times$ risk factors
- Age $\times$ type of disease
- Measurement $\times$ state of a subject during measurement

- Race $\times$ disease
- Calendar time $\times$ treatment
- Quality $\times$ quantity of a symptom

2.8.3 Fitting Ordinal Predictors

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

2.8.4 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[22]):
  $C(Y|X) = C(Y = y|X) = d(y) + X\beta$
  $C =$ log hazard
- Check form of $d(y)$
- Show $d(y)$ does not interact with X

Chapter 3

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 Xs
- 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[74]
- 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[3]

- 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[54]: replace missing value with observed value of subject hav-

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 $\Sigma\, w_i = 1$

- 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 Xs
- Otherwise, first approximation to good imputation uses other Xs to predict a missing X
- This is a single "best guess" conditional mean
- $\hat{X}_j = Z\hat{\theta}, Z = X_{\bar{j}}$
  Cannot include Y in Z without adding random errors to imputed values (would steal info from Y)
- Recursive partitioning is very helpful for nonparametrically estimating conditional means

3.7 Multiple Imputation

- Single imputation using a random draw from the conditional distribution for an individual
  $\hat{X}_j = Z\hat{\theta} + \hat{\epsilon}, Z = [X\bar{j}, Y]$

$\hat{\epsilon} = n(0, \hat{\sigma})$ 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 β
  - 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:

```
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

Method	Deletion	Single	Multiple
Allows non-random missing		x	x
Reduces sample size	x		
Apparent S.E. of $\hat{\beta}$ too low		x	
Increases real S.E. of $\hat{\beta}$	x		
$\hat{\beta}$ biased	if not MCAR	x	

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 ≤ 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 Xs. 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.

Chapter 4

Multivariable Modeling Strategies

- "Spending d.f.": examining or fitting parameters in models, or examining tables or graphs that utilize Y to tell you how to model variables
- If wish to preserve statistical properties, can't retrieve d.f. once they are "spent" (see Grambsch & O'Brien)
- If a scatterplot suggests linearity and you fit a linear model, how many d.f. did you actually spend (i.e., the d.f. that when put into a formula results in accurate confidence limits or P-values)?

- Decide number of d.f. that can be spent
- Decide where to spend them
- Spend them

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 ρ—R^2 based on $rank(X)$ and $rank(X)^2$ vs. $rank(Y)$
- ρ^2 can detect U-shaped relationships
- For categorical X, ρ^2 is R^2 from dummy variables regressed against $rank(Y)$; this is tightly related to the Wilcoxon–Mann–Whitney–Kruskal–Wallis rank test for group differences[a]
- Sort variables by descending order of ρ^2
- Specify number of knots for continuous X, combine infrequent categories of categorical X based on ρ^2
- Allocating d.f. based on sorting ρ^2 fair procedure because
  - already decided to keep variable in model no matter what ρ^2

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

  - ρ^2 does not reveal degree of nonlinearity; high value may be due solely to strong linear effect
  - low ρ^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 $\rightarrow$ assess regression assumptions separately for each predictor

4.3 Variable Selection

- Series of potential predictors with no prior knowledge
- ↑ exploration → ↑ shrinkage (overfitting)
- Summary of problem: $E(\hat{\beta}|\hat{\beta}$ "significant" $) \neq \beta$[15]
- Biased R^2, $\hat{\beta}$, standard errors, P-values too small
- F and χ^2 statistics do not have the claimed distribution[36]
- Will result in residual confounding if use variable selection to find confounders[40]
- Derksen and Keselman[27] found that in stepwise analyses the final model represented noise 0.20-0.74 of time, final model usually contained $< \frac{1}{2}$ actual number of authentic predictors. Also:

  1. "The degree of correlation between the predictor variables affected the frequency

with which authentic predictor variables found their way into the final model.
2. The number of candidate predictor variables affected the number of noise variables that gained entry to the model.
3. The size of the sample was of little practical importance in determining the number of authentic variables contained in the final model.
4. The population multiple coefficient of determination could be faithfully estimated by adopting a statistic that is adjusted by the total number of candidate predictor variables rather than the number of variables in the final model".

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

Variable selection methods[41]:

- Forward selection, backward elimination
- Stopping rule: "residual χ^2" with d.f. = no. candidates remaining at current step

- Test for significance or use Akaike's information criterion (AIC[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 $\mathbf{2}$ 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[1, 9, 47]

- Data reduction (grouping variables) can help
- Using the bootstrap to select important variables for inclusion in the final model [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 $< \frac{m}{15}$ [42, 43, 59, 60, 67]
- p = number of parameters in full model or number of *candidate* parameters in a stepwise analysis

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

[b] Based on the power of a proportional odds model two-sample test when the marginal cell sizes for the response are $n_1, \ldots, n_k$, compared with all cell sizes equal to unity (response is continuous) [79, Eq, 3]. If all cell sizes are equal, the relative efficiency of having k response categories compared to a continuous response is $1 - \frac{1}{k^2}$ [79, Eq. 14], e.g., a 5-level response is almost as efficient as a continuous one if proportional odds holds across category cutoffs.

[c] This is approximate, as the effective sample size may sometimes be boosted somewhat by censored observations, especially for non-proportional hazards methods such as Wilcoxon-type tests[6].

Table 4.1: Limiting Sample Sizes for Various Response Variables

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

- Narrowly distributed predictor $\rightarrow$ even higher n
- p includes *all* variables screened for association with response, including interactions
- 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-shrunk" models
- Aren't regression coefficients OK because

they're unbiased?

- Problem is in how we use coefficient estimates
- Consider 20 samples of size $n = 50$ from $U(0, 1)$
- Compute group means and plot in ascending order
- Equivalent to fitting an intercept and 19 dummies using least squares
- Result generalizes to general problems in plotting Y vs. $X\hat{\beta}$
- 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]

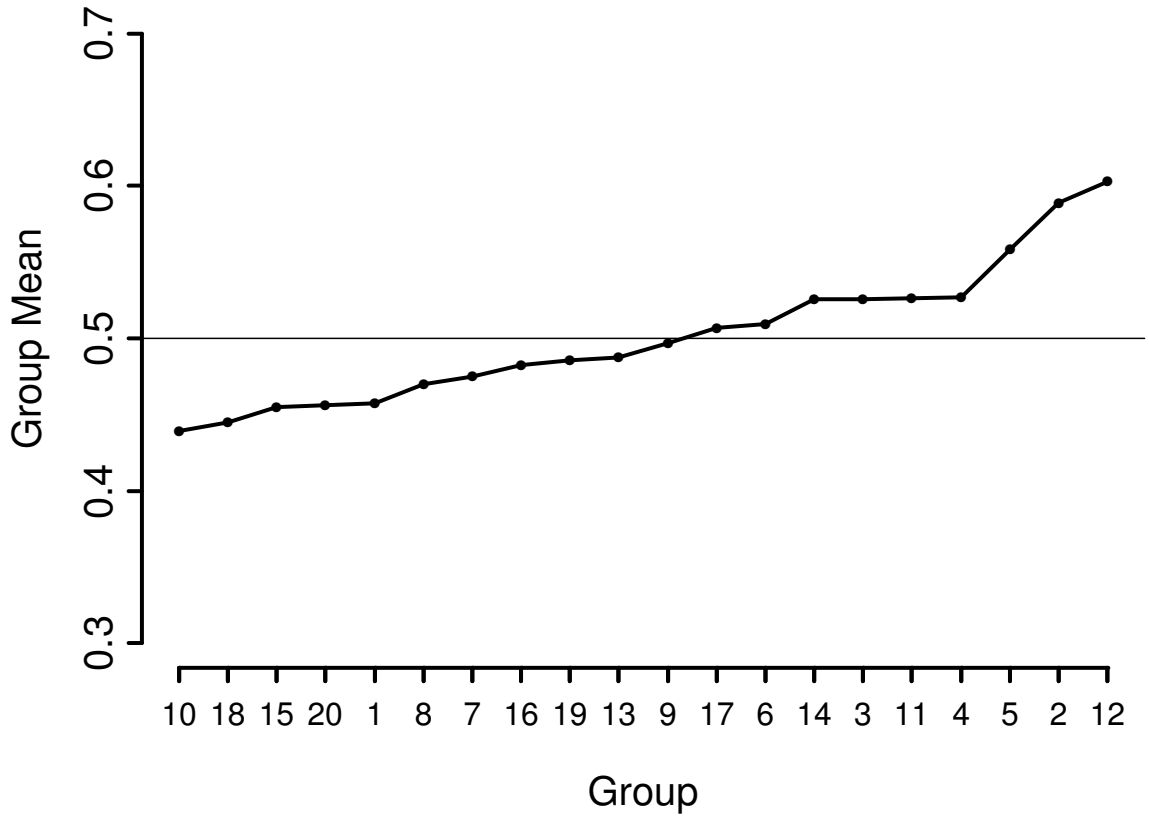

Figure 4.1: Sorted means from 20 samples of size 50 from a uniform $[0, 1]$ distribution. The reference line at 0.5 depicts the true population value of all of the means.

- 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: $\hat{\gamma} = \frac{n-p-1}{n-1} R^2_{\text{adj}} / R^2$
  $R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$
- p close to no. candidate variables
- Copas [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)
- $\uparrow$ s.e. of $\hat{\beta}$, $\downarrow$ power

- This is appropriate $\rightarrow$ 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 `VARCLUS` procedure[64], S-PLUS `varclus` function, other clustering techniques: group highly correlated variables
- Can score each group (e.g., first principal component, PC_1 [25]); 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) $\rightarrow$ separate variables so that first PC is representative of group
- Can also do hierarchical cluster analysis on similarity matrix based on squared Spearman or Pearson correlations, or more generally, Hoeffding's D[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 Xs
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`[50]
  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 Xs
- The imputed values can optionally be shrunk to avoid overfitting for small n or large p
- Constrain imputed values to be in range of non-imputed ones
- Imputations on original scale
  1. Continuous $\rightarrow$ back-solve with linear interpolation
  2. Categorical $\rightarrow$ classification tree (most freq. cat.) or match to category whose canonical score is closest to one predicted

- Multiple imputation — bootstrap or approx. Bayesian boot.
  1. Sample residuals multiple times (default $M = 5$)
  2. Are on "optimally" transformed scale
  3. Back-transform
  4. `fit.mult.impute` works with `aregImpute` and `transcan` output to easily get imputation-corrected variances and avg. $\hat{\beta}$
- Option to insert constants as imputed values (ignored during transformation estimation); helpful when a lab value may be missing because the patient returned to normal
- Imputations and transformed values may be easily obtained for new data
- An S-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

pressure
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[10]

1. Uses nonparametric "super smoother"[35]
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

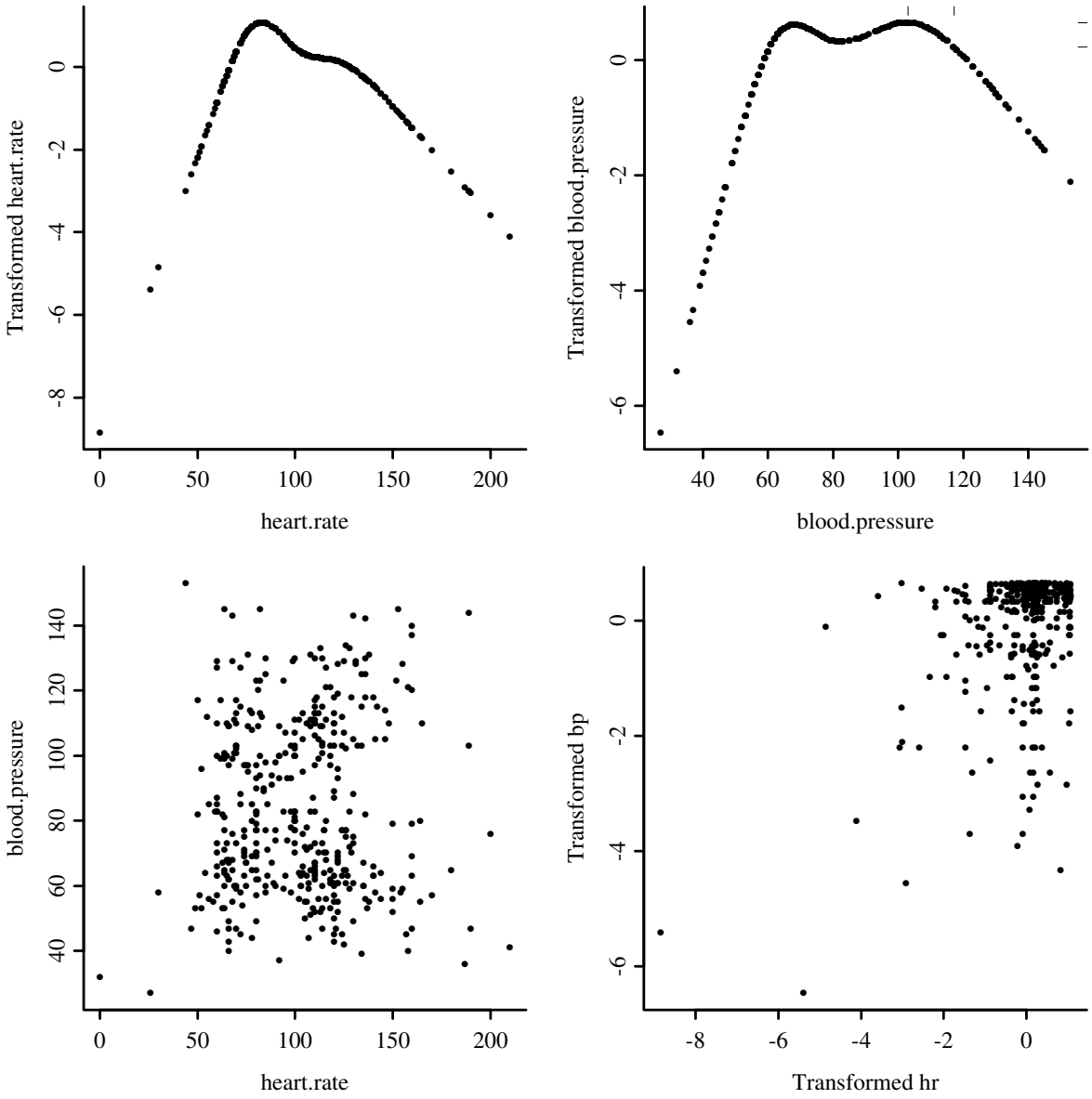

Figure 4.2: Transformations fitted using `transcan`. Tick marks indicate the two imputed values for blood pressure. The lower left plot contains raw data ($D_{xy} = 0.02$); the lower right is a scatterplot of the corresponding transformed values ($D_{xy} = 0.14$). Data courtesy of the SUPPORT study[49].

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

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

Example:

- Binary logistic model, 45 events on 150 subjects
- 10:1 rule $\rightarrow$ analyze 4.5 d.f. total
- Analyst wishes to include age, sex, 10 others
- Not known if age linear or if age and sex additive
- 4 knots $\rightarrow 3 + 1 + 1$ d.f. for age and sex if restrict interaction to be linear
- Full model with 15 d.f. has LR=50
- Expected shrinkage factor $(50 - 15)/50 = 0.7$
- LR$> 15 + 9 = 24 \rightarrow$ reduction may help
- Reduction to $q = (50 - 15)/9 \approx 4$ d.f. necessary

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

Summary of Some Data Reduction Methods

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

4.8 Overly Influential Observations

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

Reasons for influence:

- Too few observations for complexity of model (see Sections 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]
- $h_{ii} > 2(p+1)/n$ 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 $|\text{DFFITS}| > 2\sqrt{(p+1)/(n-p-1)}$[5]
- Others examine entire distribution for "outliers"
- No substitute for careful examination of data[14, 69]
- 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 nonethe less 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., $R^{2\,58}$, model χ^2, Somers' D_{xy}, Spearman's ρ, area under ROC curve)
- Rank measures may not give enough credit to extreme predictions $\rightarrow$ model χ^2, R^2, examine extremes of distribution of $\hat{Y}$
- Examine differences in predicted values from the two models

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[23]. 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[45]
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 Xs 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[70]. 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 [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

F

2. Compute statistic of interest
3. Study behavior over B repetitions

- Example: 1000 samples, 1000 sample medians, compute their sample variance
- F unknown $\rightarrow$ estimate by empirical dist. fctn.

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

where $I(w)$ is 1 if w is true, 0 otherwise.

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

- 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-

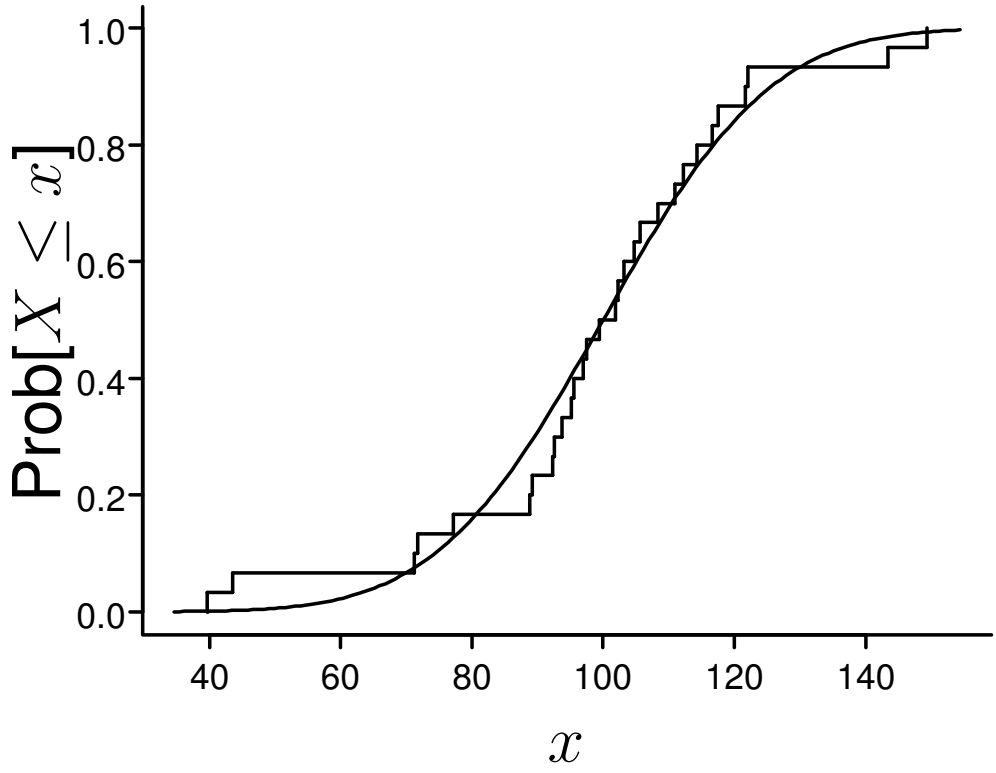

Figure 5.1: Empirical and population cumulative distribution functions

nal data points in each bootstrap sample at least once

- Some observations not selected, others selected more than once
- Efron's *bootstrap* $\rightarrow$ general-purpose technique for estimating properties of estimators without assuming or knowing distribution of data F
- Take B samples of size n with replacement, choose B so that summary measure of indi-

vidual statistics $\approx$ summary if $B = \infty$

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

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

First 20 samples:

Bootstrap Sample	Sample Median
1 6 6 6 9 9	6.0
5 5 6 7 8 8	6.5
1 1 1 5 8 9	3.0
1 1 1 5 8 9	3.0
1 6 8 8 8 9	8.0
1 6 7 8 9 9	7.5
6 6 8 8 9 9	8.0
1 1 7 8 8 9	7.5
1 5 7 8 9 9	7.5
5 6 6 6 7 7	6.0
1 6 8 8 9 9	8.0
1 5 6 6 9 9	6.0
1 6 7 8 8 9	7.5
1 6 7 7 9 9	7.0
1 5 7 8 9 9	7.5
5 6 7 9 9 9	8.0
5 5 6 7 8 8	6.5
6 6 6 7 8 8	6.5
1 1 1 1 6 9	1.0
1 5 7 7 9 9	7.0

- 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])

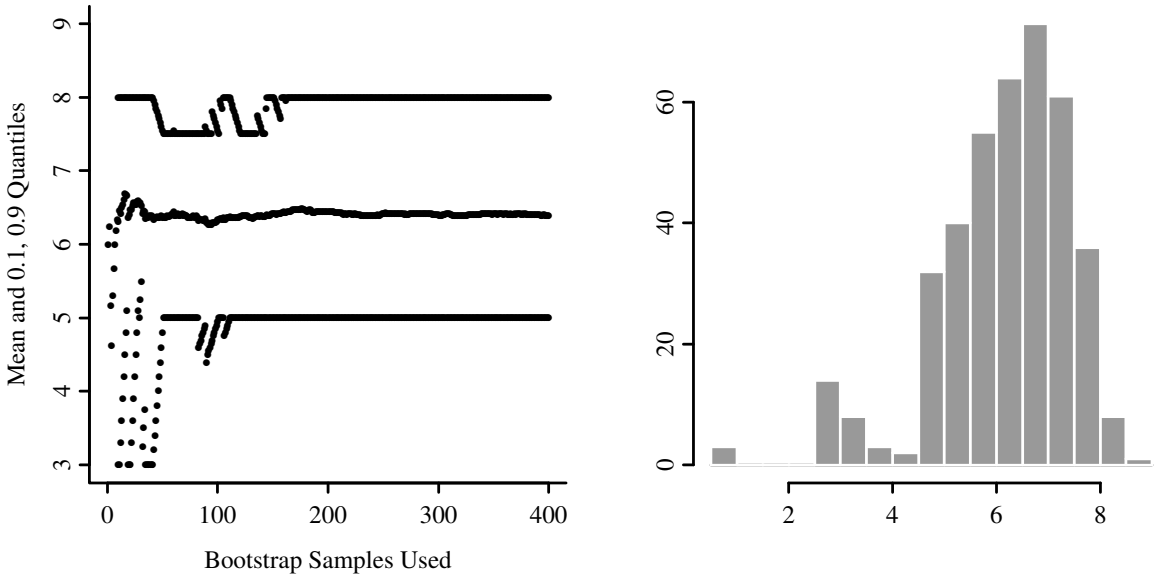

Figure 5.2: Estimating properties of sample median using the bootstrap

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

  - 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 *training* and *test* sets
- Interesting to compare index of accuracy in training and test
- Freeze parameters from training
- Make sure you allow $R^2 = 1 - SSE/SST$ for test sample to be < 0
- Don't compute ordinary R^2 on $X\hat{\beta}$ vs. Y; this allows for linear recalibration $aX\hat{\beta} + b$ vs. Y
- Test sample must be large enough to obtain very accurate assessment of accuracy
- Training sample is what's left
- Example: overall sample $n = 300$, training sample $n = 200$, develop model, freeze $\hat{\beta}$, predict on test sample ($n = 100$), $R^2 = 1 - \frac{\Sigma(Y_i - X_i\hat{\beta})^2}{\Sigma(Y_i - \bar{Y})^2}$.
- Disadvantages of data splitting:
  1. Costly in $\downarrow n$ [9, 62]

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 re-combined 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[65]
- Cross-validation solves many problems of data splitting[30, 66, 75, 80]
- 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^2_{app}) =0.4
  - how inflated is R^2_{app}?

  - get resamples of size 1000 with replacement from original 1000
  - for each resample compute R^2_{boot} = apparent R^2 in bootstrap sample
  - freeze these coefficients (call them $\hat{\beta}_{boot}$), apply to original (whole) sample (X_{orig}, Y_{orig}) to get $R^2_{orig} = R^2(X_{orig}\hat{\beta}_{boot}, Y_{orig})$
  - optimism = $R^2_{boot} - R^2_{orig}$
  - average over $B = 100$ optimisms to get $\overline{optimism}$
  - $R^2_{overfitting\ corrected} = R^2_{app} - \overline{optimism}$
- Is estimating unconditional (not conditional on X) distribution of R^2, etc. [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 γ_0, γ_1
- $\hat{\gamma}_1 < 1$ if overfit, $\hat{\gamma}_0 > 0$ to compensate
- $\hat{\gamma}_1$ quantifies overfitting and useful for improving calibration[70]
- Use Efron's method to estimate optimism in $(0, 1)$, estimate (γ_0, γ_1) by subtracting optimism from $(0, 1)$
- See also Copas[21] 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[70]
- Stepwise modeling often reduces optimism but this is not offset by loss of information from deleting marginal var.

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

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

1. The stepwise selection deleted several variables which were almost significant.
2. These marginal variables have *some* real predictive value, even if it's slight.
3. There is no small set of extremely dominant

variables that would be easily found by stepwise selection.

Other issues:

- See [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
- $\mathbf{Not}$ 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 ↑ when predict for a low-frequency category
- Collinearity
- Can average predictions over categories to marginalize, ↓ 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 σ from full fit
  2. Stopping rule less arbitrary
  3. Inheritance of shrinkage
- If estimates from full model are $\hat{\beta}$ and approx. model is based on a subset T of predictors X, coef. of approx. model are $W\hat{\beta}$, where $W = (T'T)^{-1}T'X$
- Variance matrix of reduced coef.: WVW'

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

The formula for a regression model is given to a modeling function, e.g.

```
lrm(y ~ rcs(x,4))
```

is read "use a logistic regression model to model y as a function of x, representing x by a restricted cubic spline with 4 default knots"[a].

`update` function: re-fit model with changes in terms or data:

```
f  ← lrm(y ~ rcs(x,4) + x2 + x3)
f2 ← update(f, subset=sex=="male")
f3 ← update(f, .~.-x2)          # remove x2 from model
f4 ← update(f, .~. + rcs(x5,5))# add rcs(x5,5) to model
f5 ← update(f, y2 ~ .)          # same terms, new response var.
```

6.2 User-Contributed Functions

- S is high-level object-oriented language.
- S-PLUS (UNIX, Linux, Microsoft Windows)
- R (UNIX, Linux, Mac, Windows)
- Multitude of user-contributed functions on `StatLib`

[a] `lrm` and `rcs` are in the `Design` library.

- International community of users through `S-news`

Some S functions:

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

6.3 The Design Library

- datadist function to compute predictor distribution summaries

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

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

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

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

Table 6.1: Design Fitting Functions

Function	Purpose	Related S Functions
ols	Ordinary least squares linear model	lm
lrm	Binary and ordinal logistic regression model Has options for penalized MLE	glm
psm	Accelerated failure time parametric survival models	survreg
cph	Cox proportional hazards regression	coxph
bj	Buckley-James censored least squares model	survreg,lm
glmD	Design version of glm	glm

Table 6.2: `Design` Transformation Functions

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

Example:

- `treat`: categorical variable with levels `"a","b","c"`
- `num.diseases`: ordinal variable, 0-4
- `age`: continuous
  Restricted cubic spline
- `cholesterol`: continuous
  (3 missings; use median)
  `log(cholesterol+10)`
- Allow `treat` $\times$ `cholesterol` interaction
- Program to fit logistic model, test all effects in design, estimate effects (e.g. inter-quartile range odds ratios), plot estimated transformations

Function	Purpose	Related Functions
print	Print parameters and statistics of fit	
coef	Fitted regression coefficients	
formula	Formula used in the fit	
specs	Detailed specifications of fit	
robcov	Robust covariance matrix estimates	
bootcov	Bootstrap covariance matrix estimates and bootstrap distributions of estimates	
pentrace	Find optimum penalty factors by tracing effective AIC for a grid of penalties	
effective.df	Print effective d.f. for each type of variable in model, for penalized fit or pentrace result	
summary	Summary of effects of predictors	
plot.summary	Plot continuously shaded confidence bars for results of summary	
anova	Wald tests of most meaningful hypotheses	
plot.anova	Graphical depiction of anova	
contrast	General contrasts, C.L., tests	
plot	Plot effects of predictors	
gendata	Easily generate predictor combinations	
predict	Obtain predicted values or design matrix	
fastbw	Fast backward step-down variable selection	step
residuals	(or resid) Residuals, influence stats from fit	
sensuc	Sensitivity analysis for unmeasured confounder	
which.influence	Which observations are overly influential	residuals
latex	LaTeX representation of fitted model	Function
Dialog	Create a menu to enter predictor values and obtain predicted values from fit	Function nomogram
Function	S function analytic representation of $X\hat{\beta}$ from a fitted regression model	
Hazard	S function analytic representation of a fitted hazard function (for psm)	
Survival	S function analytic representation of fitted survival function (for psm, cph)	
Quantile	S function analytic representation of fitted function for quantiles of survival time (for psm, cph)	
Mean	S function analytic representation of fitted function for mean survival time	
nomogram	Draws a nomogram for the fitted model	latex, plot
survest	Estimate survival probabilities (psm, cph)	survfit
survplot	Plot survival curves (psm, cph)	plot.survfit
validate	Validate indexes of model fit using resampling	
calibrate	Estimate calibration curve using resampling	val.prob
vif	Variance inflation factors for fitted model	
naresid	Bring elements corresponding to missing data back into predictions and residuals	
naprint	Print summary of missing values	
impute	Impute missing values	aregImpute
fit.mult.impute		

```
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()
specs(fit)                          # Describe the design characteristics
anova(fit)
anova(fit, treat, cholesterol)      # Test these 2 by themselves
plot(anova(fit))                    # Summarize anova graphically
summary(fit)                        # Estimate effects using default ranges
plot(summary(fit))                  # Graphical display of effects with C.I.
summary(fit, treat="b", age=60)     # Specify reference cell and adjustment val
summary(fit, age=c(50,70))          # Estimate effect of increasing age from
                                    # 50 to 70
summary(fit, age=c(50,60,70))       # Increase age from 50 to 70, adjust to
                                    # 60 when estimating effects of other
                                    # factors
# If had not defined datadist, would have to define ranges for all var.

# Estimate and test treatment (b-a) effect averaged over 3 cholesterols
contrast(fit, list(treat='b', cholesterol=c(150,200,250)),
              list(treat='a', cholesterol=c(150,200,250)),
         type='average')

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

```
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
- In $C(Y|X)$ C is $\text{Prob}\{Y = 1\}$
- $g(u)$ is $\frac{1}{1+e^{-u}}$

10.1 Model

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

$$P = [1 + \exp(-x)]^{-1}$$

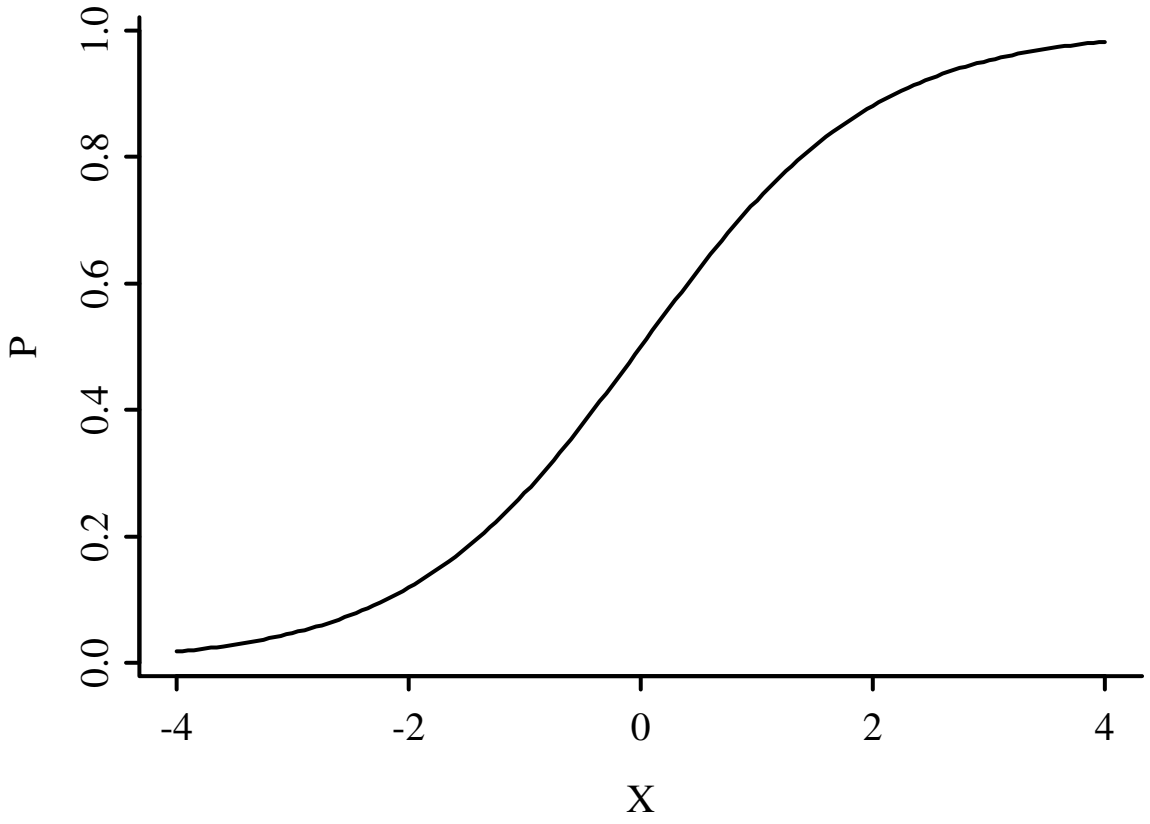

Figure 10.1: Logistic function

- $O = \frac{P}{1-P}$
- $P = \frac{O}{1+O}$
- $X\beta = \log \frac{P}{1-P}$
- $e^{X\beta} = O$

10.1.1 Model Assumptions and Interpretation of Parameters

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

- Increase X_j by $d \rightarrow$ increase odds $Y = 1$ by $\exp(\beta_j d)$, increase log odds by $\beta_j d$.
- If there is only one predictor X and that predictor is binary, the model can be written

$$\begin{aligned} \text{logit}\{Y = 1|X = 0\} &= \beta_0 \\ \text{logit}\{Y = 1|X = 1\} &= \beta_0 + \beta_1. \end{aligned}$$

- One continuous predictor:

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

- Two treatments (indicated by $X_1 = 0$ or 1) and one continuous covariable (X_2).

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

$$\begin{aligned}\text{logit}\{Y = 1|X_1 = 0, X_2\} &= \beta_0 + \beta_2 X_2 \\ \text{logit}\{Y = 1|X_1 = 1, X_2\} &= \beta_0 + \beta_1 + \beta_2 X_2.\end{aligned}$$

10.1.2 Odds Ratio, Risk Ratio, and Risk Difference

- Odds ratio capable of being constant
- Ex: risk factor doubles odds of disease

Without Risk Factor		With Risk Factor	
Probability	Odds	Odds	Probability
.2	.25	.5	.33
.5	1	2	.67
.8	4	8	.89
.9	9	18	.95
.98	49	98	.99

Let X_1 be a binary risk factor and let $A = \{X_2, \ldots, X_p\}$ be the other factors. Then the

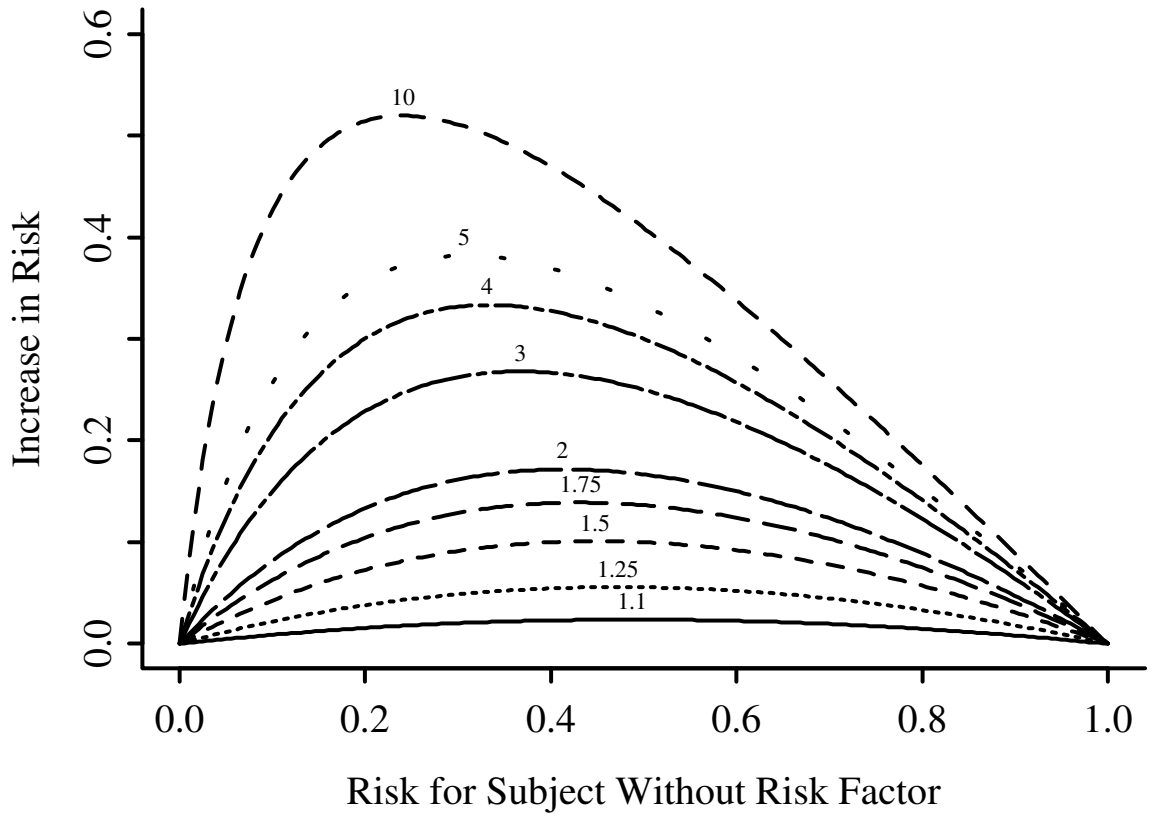

Figure 10.2: Absolute benefit as a function of risk of the event in a control subject and the relative effect (odds ratio) of the risk factor. The odds ratios are given for each curve.

estimate of $\text{Prob}\{Y = 1|X_1 = 1, A\} - \text{Prob}\{Y = 1|X_1 = 0, A\}$ is

$$\begin{aligned}
&\frac{1}{1 + \exp -[\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p]} \\
&- \frac{1}{1 + \exp -[\hat{\beta}_0 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p]} \\
&= \frac{1}{1 + (\frac{1-\hat{R}}{\hat{R}}) \exp(-\hat{\beta}_1)} - \hat{R},
\end{aligned}$$

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

SEX	RESPONSE			
Frequency				
Row Pct	0	1	Total	Odds/Log
F	14	6	20	6/14=.429
	70.00	30.00		-.847
M	6	14	20	14/6=2.33
	30.00	70.00		.847
Total	20	20	40	

M:F odds ratio = (14/6)/(6/14) = 5.44, log=1.695

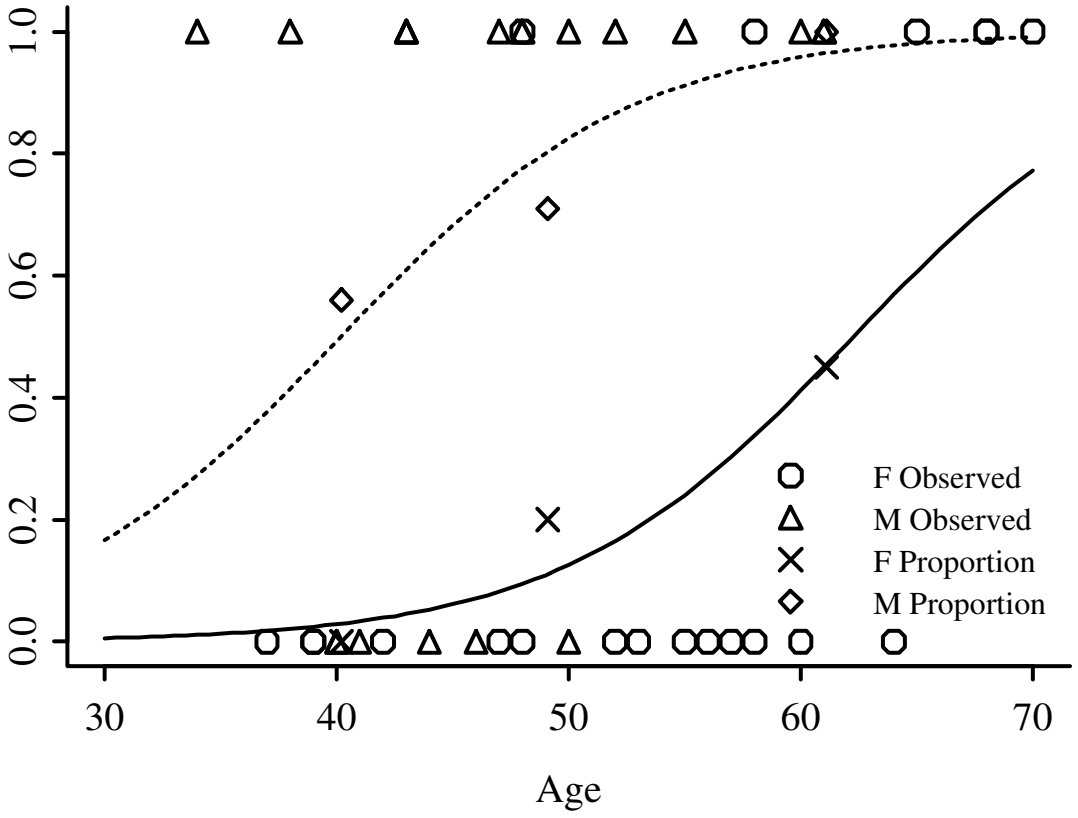

Figure 10.3: Data, subgroup proportions, and fitted logistic model

STATISTICS FOR TABLE OF SEX BY RESPONSE

Statistic	DF	Value	Prob
Chi Square	1	6.400	0.011
Likelihood Ratio Chi-Square	1	6.583	0.010

Fitted Logistic Model

Parameter	Estimate	Std Err	Wald χ^2	P
β_0	-0.8472978	0.48795	3.015237	
β_1	1.6945956	0.69007	6.030474	0.0141

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 OF AGE BY RESPONSE

AGE	RESPONSE			
Frequency				
Row Pct	0	1	Total	Odds/Log
<45	8	5	13	5/8=.625
	61.5	38.4		-.47
45-54	6	6	12	6/6=1
	50.0	50.0		0
55+	6	9	15	9/6=1.5
	40.0	60.0		.405
Total	20	20	40	

55+ : <45 odds ratio = (9/6)/(5/8) = 2.4, log=.875

Fitted Logistic Model

Parameter	Estimate	Std Err	Wald χ^2	P
β_0	-2.7338405	1.83752	2.213422	0.1368
β_1	0.0539798	0.03578	2.276263	0.1314

The estimate of β_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:

Log likelihood ($\beta_1 = 0$) : -27.727
Log likelihood (max) : -26.511
LR $\chi^2(H_0 : \beta_1 = 0)$: -2(-27.727- -26.511) = 2.4

(Compare 2.432 with the Wald statistic 2.28.)

Next we consider the simultaneous association of age and sex with response.

SEX=F

AGE	RESPONSE		
Frequency			
Row Pct	0	1	Total
<45	4	0	4
	100.0	0.0	
45-54	4	1	5
	80.0	20.0	
55+	6	5	11
	54.6	45.4	
Total	14	6	20

SEX=M

AGE	RESPONSE		
Frequency			
Row Pct	0	1	Total
<45	4	5	9
	44.4	55.6	
45-54	2	5	7
	28.6	71.4	
55+	0	4	4
	0.0	100.0	
Total	6	14	20

A logistic model for relating sex and age simultaneously to response is given below.

Fitted Logistic Model

Parameter	Estimate	Std Err	Wald χ^2	P
β_0	-9.8429426	3.67576	7.17057	0.0074
β_1 (sex)	3.4898280	1.19917	8.46928	0.0036
β_2 (age)	0.1580583	0.06164	6.57556	0.0103

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 χ^2 ($H_0 : \beta_1 = \beta_2 = 0$)	:	-2(-27.727- -19.458)= 16.538
LR χ^2 ($H_0 : \beta_1 = 0$) sex|age	:	-2(-26.511- -19.458) = 14.106
LR χ^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

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

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 χ^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 Ps vary; $\hat{\beta}$ computed by trial and error using an iterative maximization technique

10.2.2 Estimation of Odds Ratios and Probabilities

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

10.3 Test Statistics

- Likelihood ratio test best

- Score test second best (score $\chi^2 \equiv$ Pearson χ^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

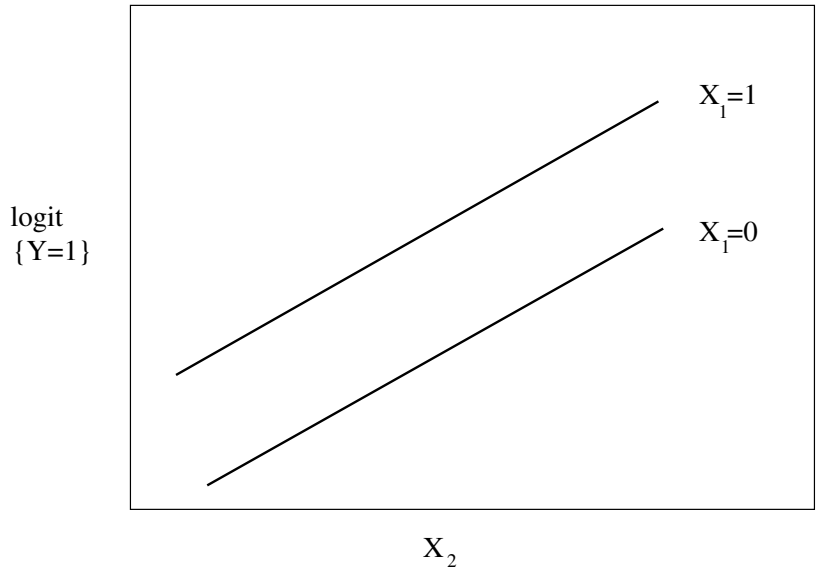

Figure 10.4: Logistic regression assumptions for one binary and one continuous predictor

- $\hat{P}$ = 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 Xs $\rightarrow$ nonparametric smoother
- `plsmo` function makes it easy to use loess to compute logits of nonparametric estimates (`fun=qlogis`)
- General: restricted cubic spline expansion

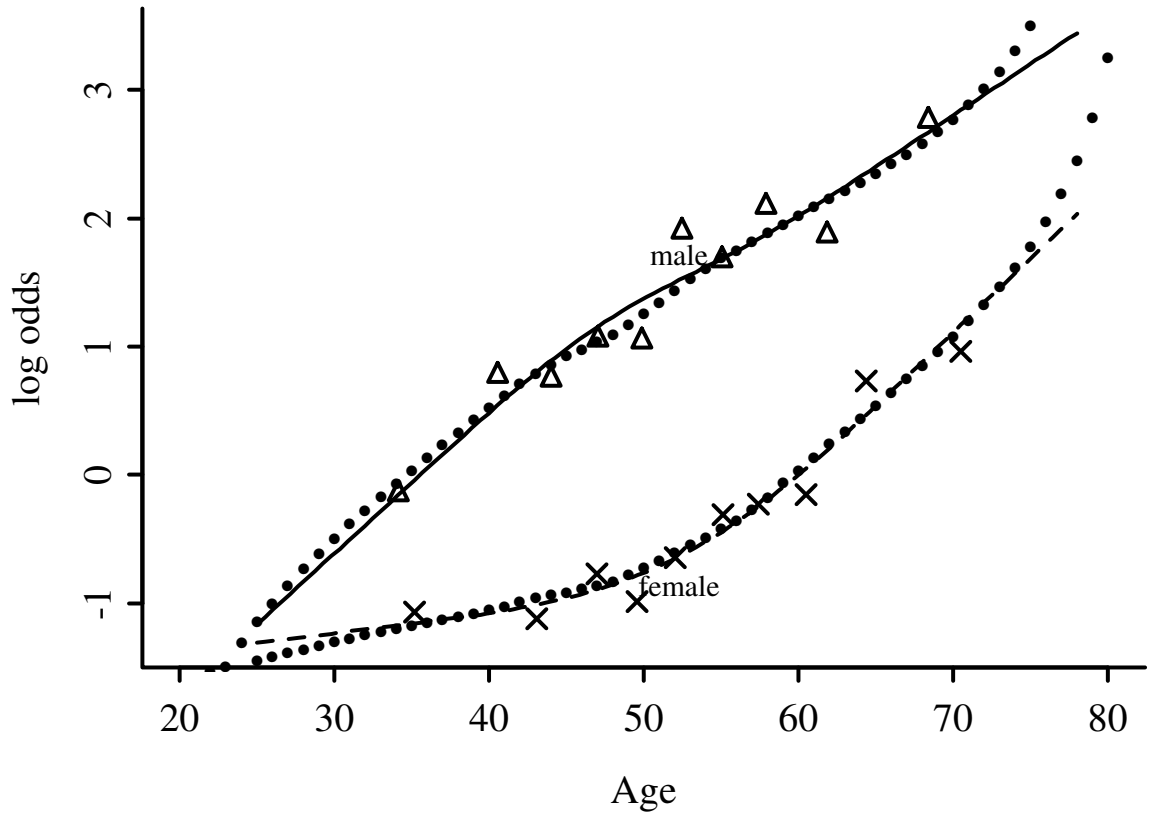

Figure 10.5: Logit proportions of significant coronary artery disease by sex and deciles of age for n=3504 patients, with spline fits (smooth curves). Spline fits are for $k = 4$ knots at age=36, 48, 56, and 68 years, and interaction between age and sex is allowed. Smooth nonparametric estimates are shown as dotted curves. Data courtesy of the Duke Cardiovascular Disease Databank.

of one or more predictors

$$\begin{aligned}\text{logit}\{Y=1|X\} &= \hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_2+\hat{\beta}_3X_2'+\hat{\beta}_4X_2''\\ &= \hat{\beta}_0+\hat{\beta}_1X_1+f(X_2),\end{aligned}$$

$$\begin{aligned}\text{logit}\{Y=1|X\} &= \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_2'+\beta_4X_2''\\ &\quad +\beta_5X_1X_2+\beta_6X_1X_2'+\beta_7X_1X_2''\end{aligned}$$

Model / Hypothesis	Likelihood Ratio χ^2	d.f.	P	Formula
a: sex, age (linear, no interaction)	766.0	2		
b: sex, age, age $\times$ sex	768.2	3		
c: sex, spline in age	769.4	4		
d: sex, spline in age, interaction	782.5	7		
H_0 : no age $\times$ sex interaction given linearity	2.2	1	.14	$(b-a)$
H_0 : age linear | no interaction	3.4	2	.18	$(c-a)$
H_0 : age linear, no interaction	16.6	5	.005	$(d-a)$
H_0 : age linear, product form interaction	14.4	4	.006	$(d-b)$
H_0 : no interaction, allowing for nonlinearity in age	13.1	3	.004	$(d-c)$

- 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

k	Model χ^2	AIC
0	99.23	97.23
3	112.69	108.69
4	121.30	115.30
5	123.51	115.51
6	124.41	114.51

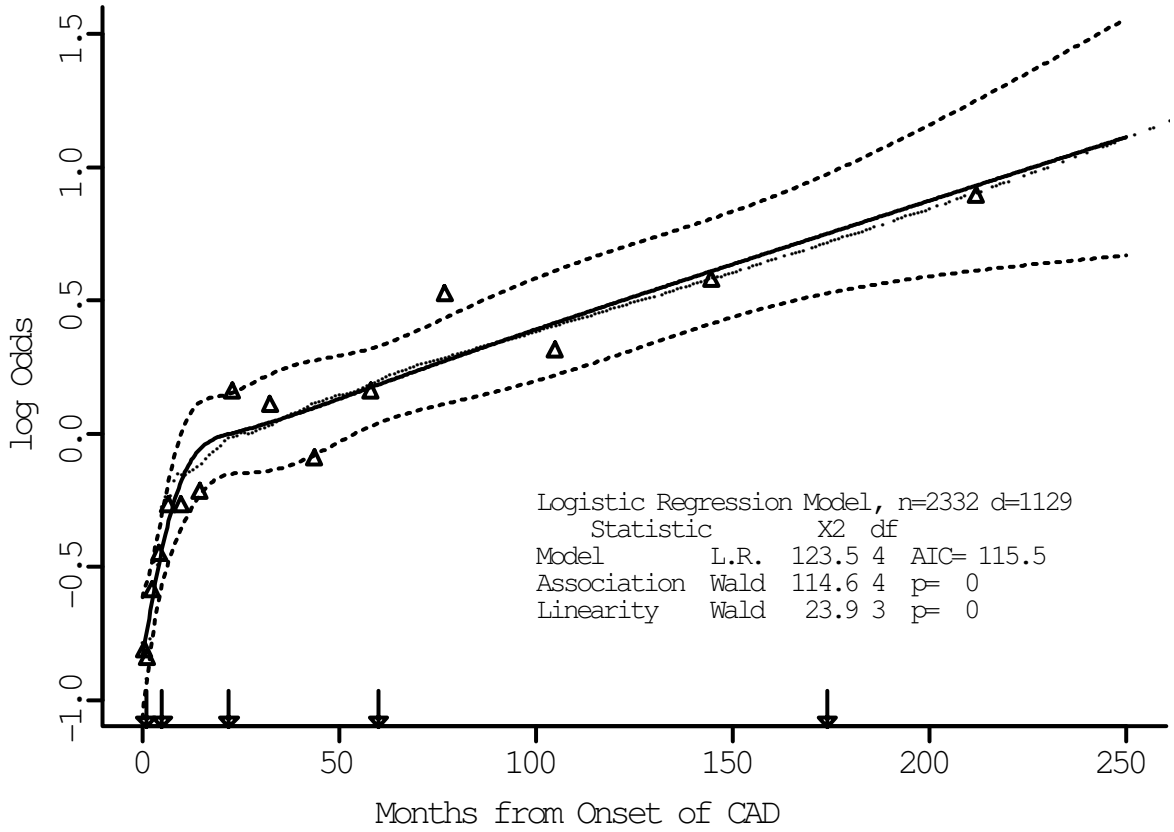

Figure 10.6: Estimated relationship between duration of symptoms and the log odds of severe coronary artery disease for $k = 5$. Knots are marked with arrows. Solid line is spline fit; dotted line is a nonparametric "super-smoothed" estimate.

- Sample of 2258 pts

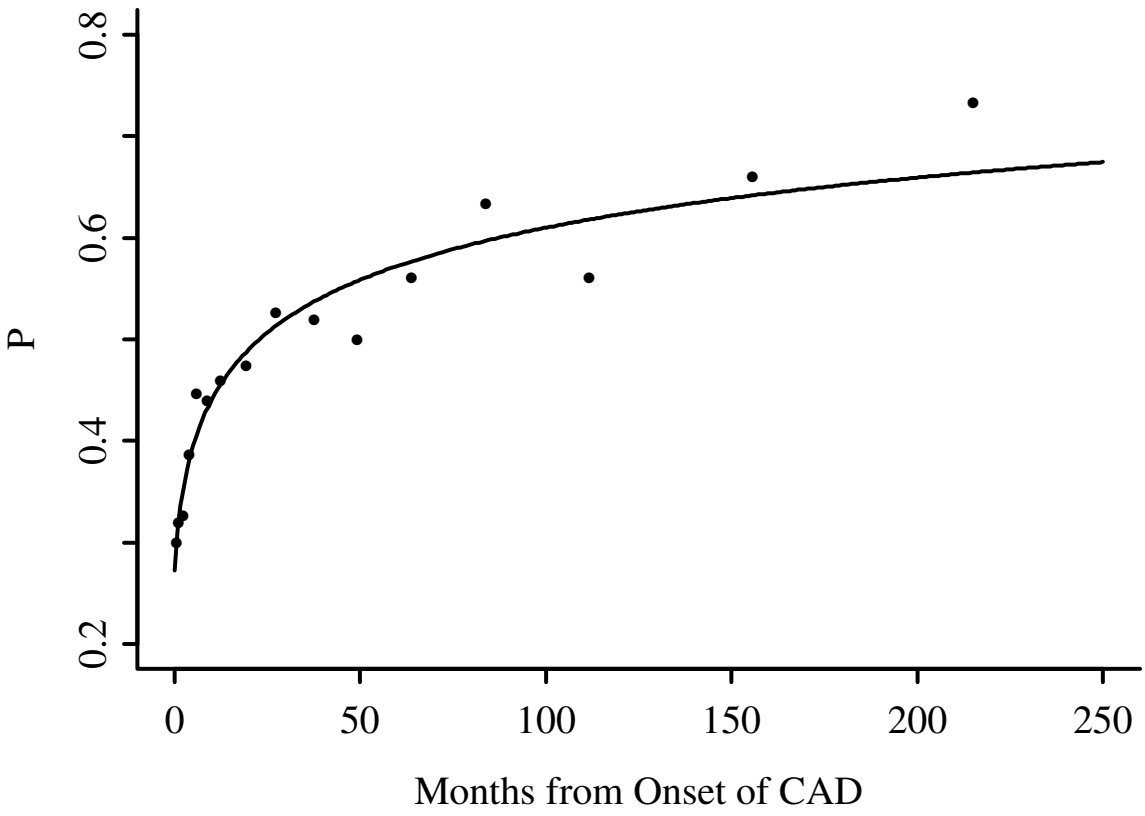

Figure 10.7: Fitted linear logistic model in $\log_{10}(\text{duration}+1)$, with subgroup estimates using groups of 150 patients. Fitted equation is logit(`tvdlm`) $= -.9809 + .7122 \log_{10}(\text{months} + 1)$.

- 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

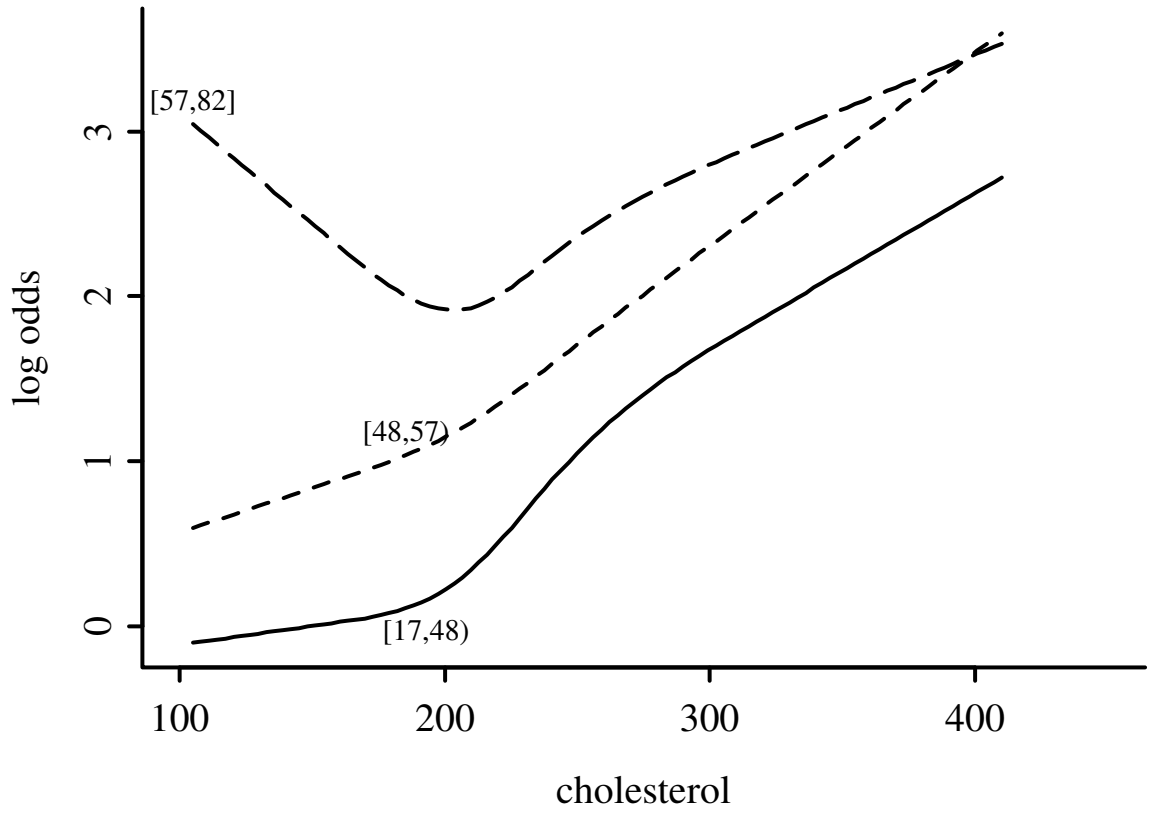

Figure 10.8: Log odds of significant coronary artery disease modeling age with two dummy variables

```
anova(fit)
```

Wald Statistics

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

- 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

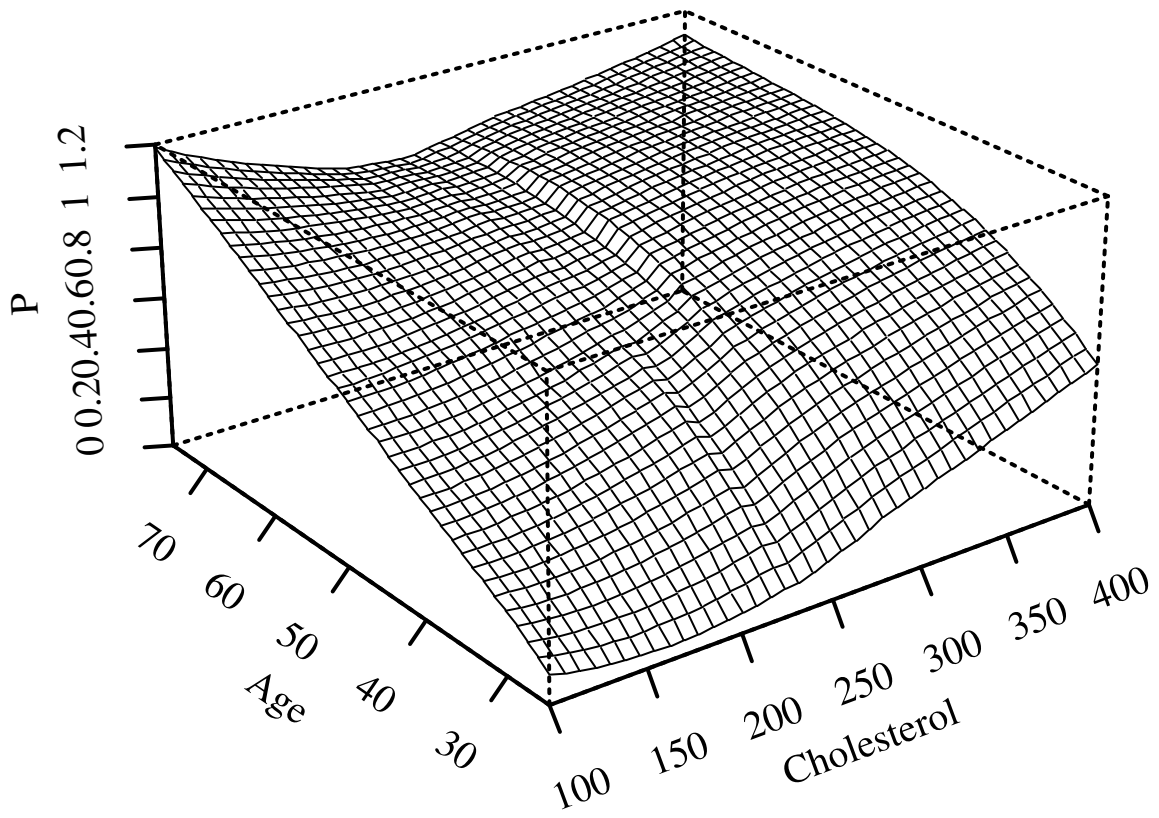

Figure 10.9: Local regression fit for the probability of significant coronary disease vs. age and cholesterol for males, based on the S-PLUS `loess` function

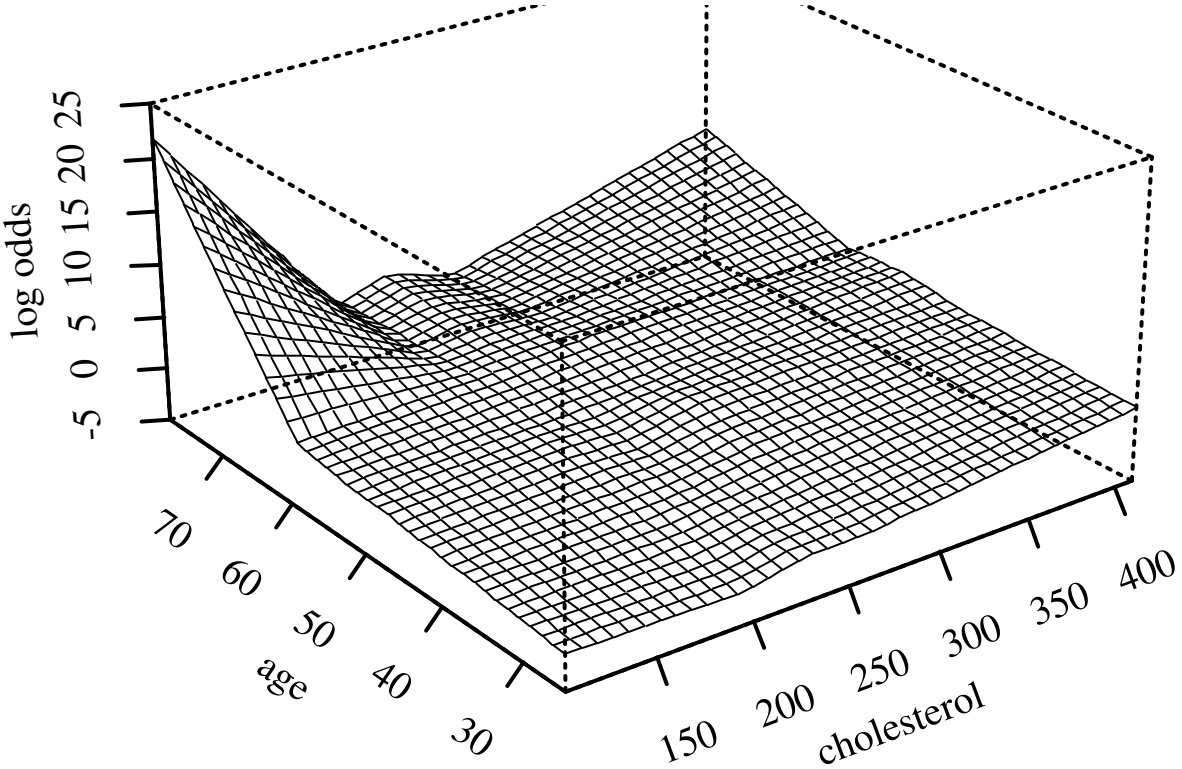

Figure 10.10: Linear spline surface for males, with knots for age at 46, 52, 59 and knots for cholesterol at 196, 224, and 259 (quartiles)

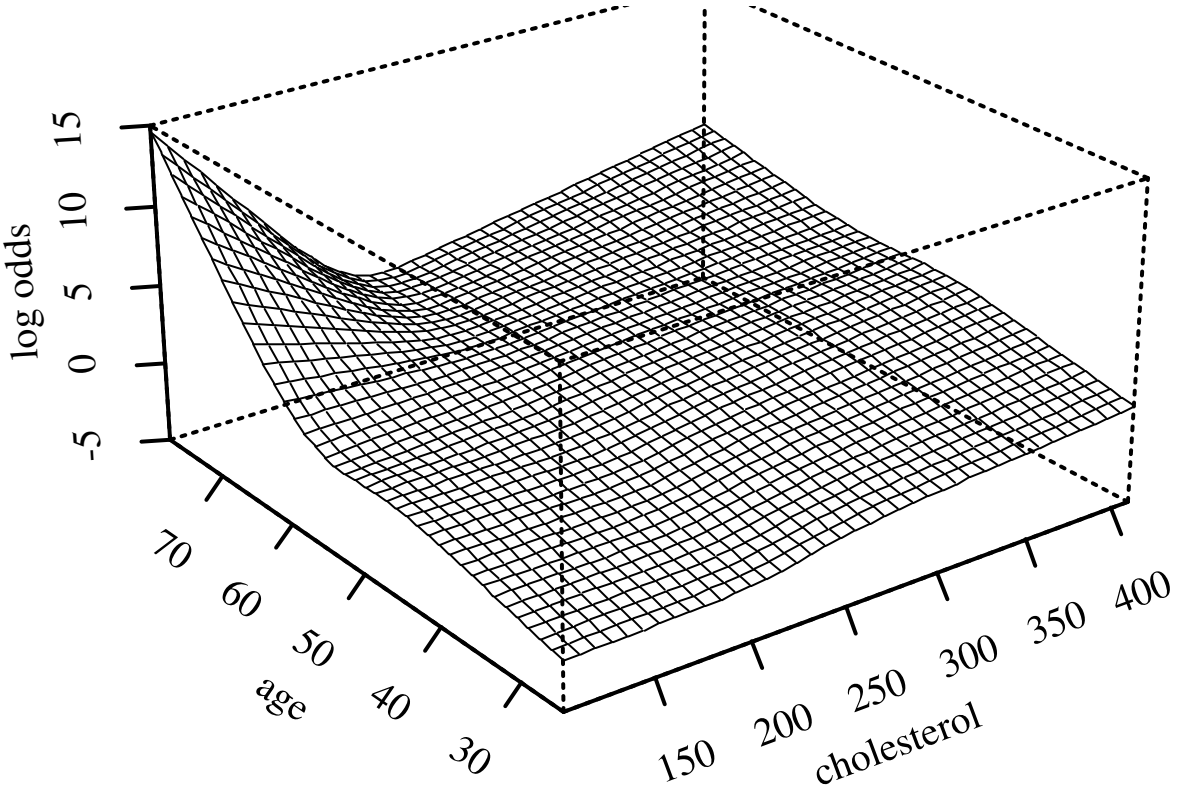

Figure 10.11: Restricted cubic spline surface in two variables, each with $k = 4$ knots

Wald Statistics

Factor	χ^2	d.f.	P
age * cholesterol	12.95	9	0.1649
Nonlinear Interaction : $f(A,B)$ vs. AB	7.27	8	0.5078
$f(A,B)$ vs. $Af(B)+Bg(A)$	5.41	4	0.2480
Nonlinear Interaction in age vs. $Af(B)$	6.44	6	0.3753
Nonlinear Interaction in cholesterol vs. $Bg(A)$	6.27	6	0.3931

- Now restrict surface by excluding doubly nonlinear terms

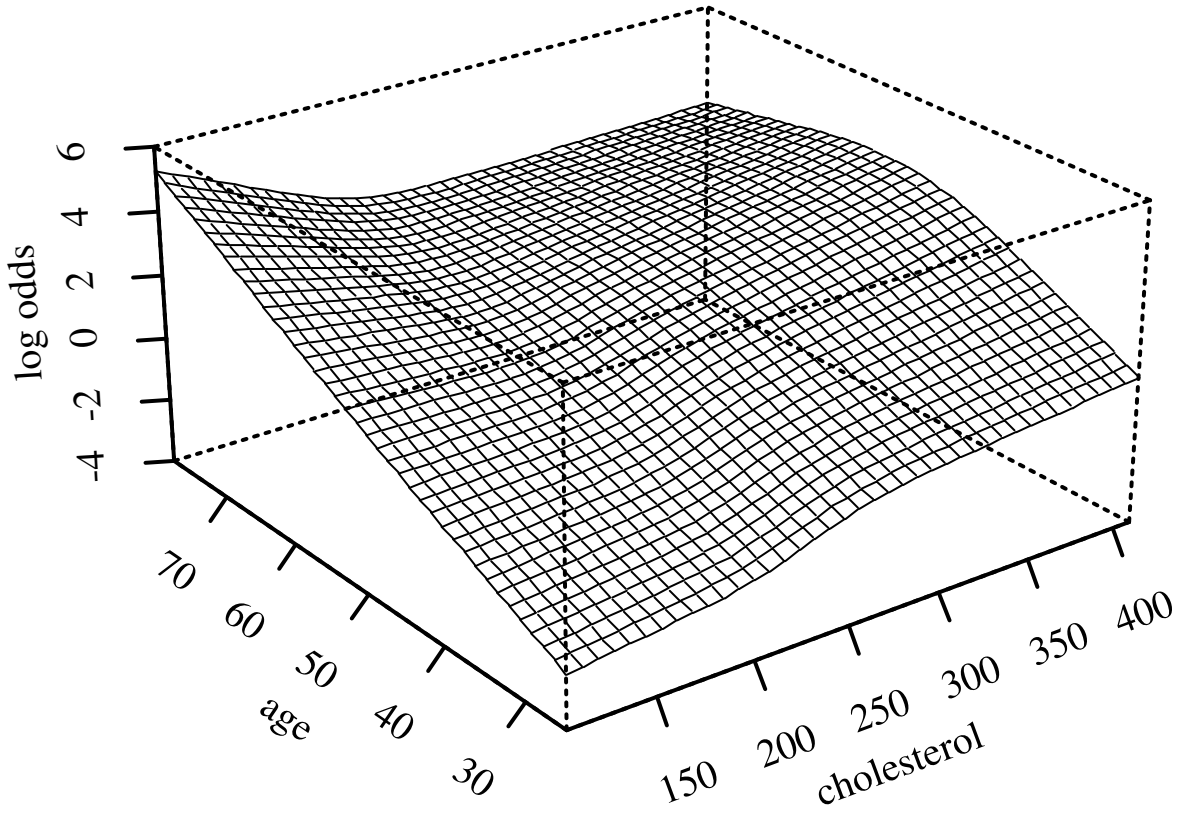

Figure 10.12: Restricted cubic spline fit with age $\times$ spline(cholesterol) and cholesterol $\times$ spline(age)

Wald Statistics

Factor	χ^2	d.f.	P
age * cholesterol	10.83	5	0.0548
Nonlinear Interaction : $f(A,B)$ vs. AB	3.12	4	0.5372
Nonlinear Interaction in age vs. $Af(B)$	1.60	2	0.4496
Nonlinear Interaction in cholesterol vs. $Bg(A)$	1.64	2	0.4399

- 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.

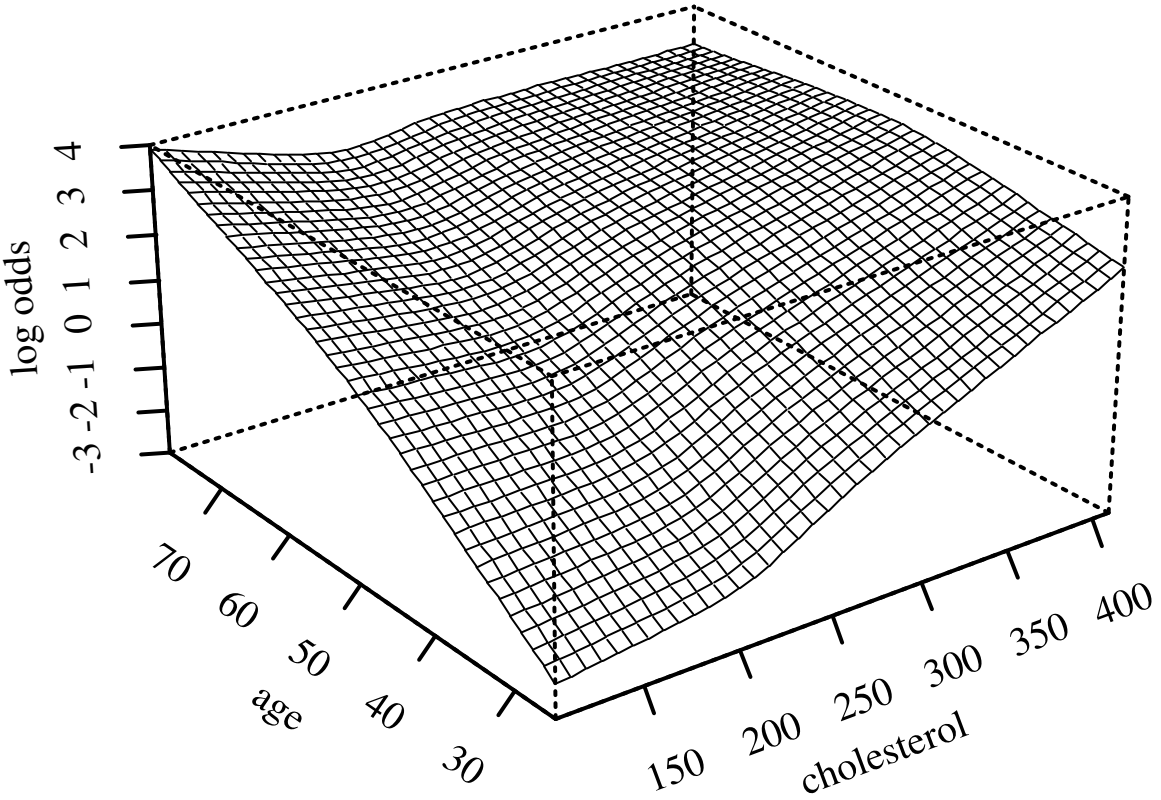

Figure 10.13: Spline fit with nonlinear effects of cholesterol and age and a simple product interaction

- 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

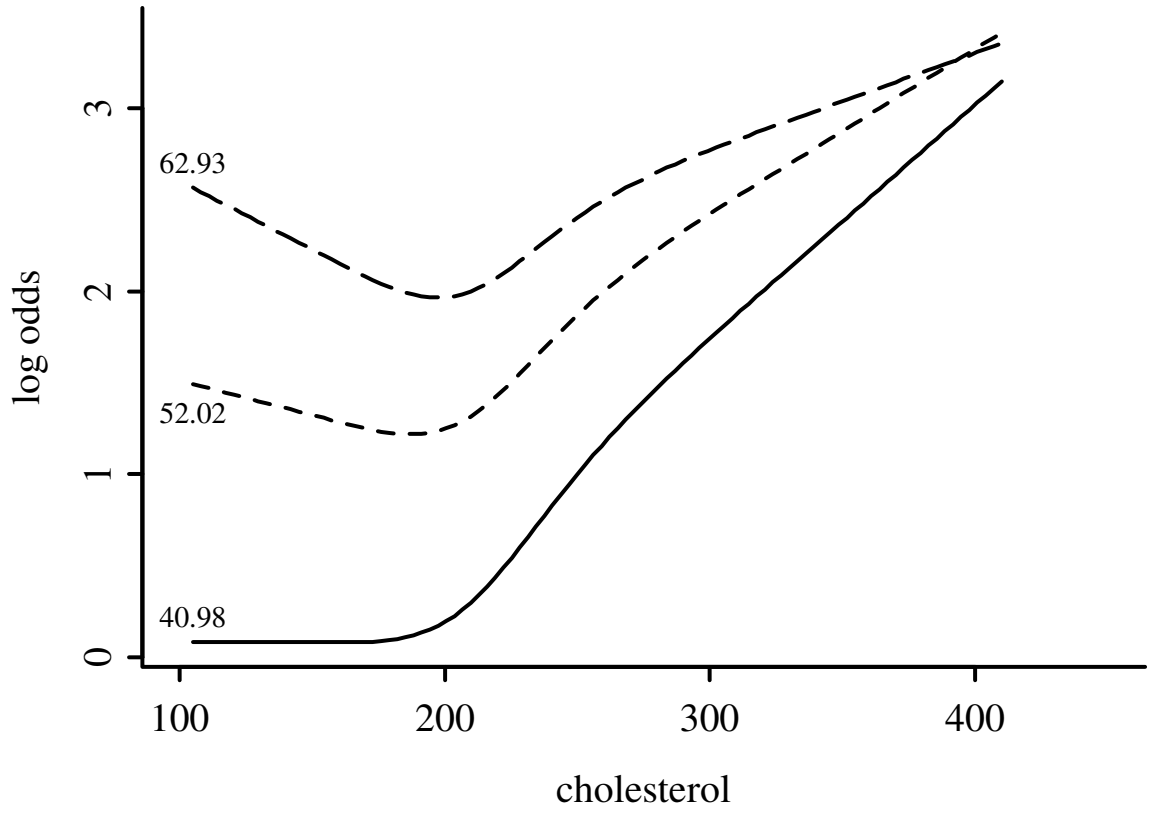

Figure 10.14: Predictions from linear interaction model with mean age in tertiles indicated.

- Using residuals for "duration of symptoms" example
- Relative merits of strat., nonparametric, splines

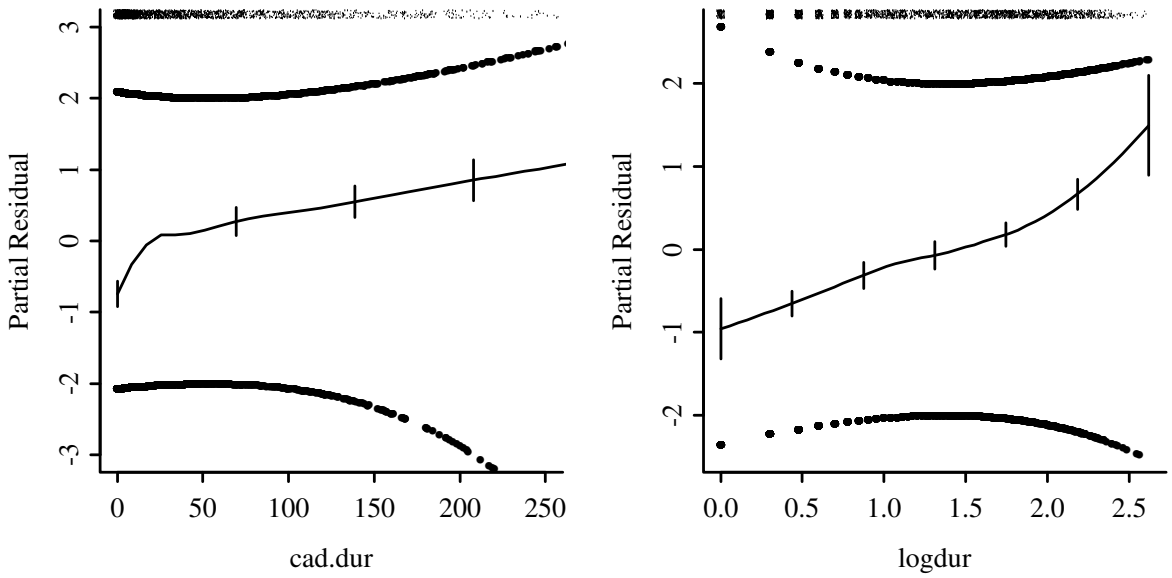

Figure 10.15: Partial residuals for duration and $\log_{10}$(duration+1). Data density shown at top of each plot.

for checking fit

Method	Choice Required	Assumes Additivity	Uses Ordering of X	Low Variance	Good Resolution on X
Stratification	Intervals				
Smoother on X_1 stratifying on X_2	Bandwidth		x (not on X_2)	x (if min. strat.)	x (X_1)
Smooth partial residual plot	Bandwidth	x	x	x	x
Spline model for all Xs	Knots	x	x	x	x

- 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:

$$R^2_{\mathrm{N}} = \frac{1 - \exp(-\mathrm{LR}/n)}{1 - \exp(-L^0/n)},$$

- 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 $\frac{1}{1+e^{-X_{new}\hat{\beta}_{old}}}$ against estimated prob. that $Y = 1$ on new data
  - Discrimination: C or D_{xy}
  - R^2 or B

- Use bootstrap to estimate calibration equation

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

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

- Bootstrap validation of age-sex-response data, 80 samples
- 2 predictors forced into every model

Table 10.1: Validation of 2-variable Logistic Model

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

- 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

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

- Try adding 5 noise candidate variables

Table 10.3: Validation of Model with 5 Noise Variables

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

Number of Factors Selected	0	1	2	3	4	5
Frequency	38	3	6	9	5	4

- The first 15 patterns of factors selected are:

```
age sex x1 x2 x3 x4 x5


*   *
*   *           *
*


*   *
*   *   *       *  *


*   *           *

*   *         *
```

10.10 Describing the Fitted Model

Factor	Low	High	Diff.	Effect	S.E.	Lower 0.95	Upper 0.95
age	46	59	13	0.90	0.21	0.49	1.32
Odds Ratio	46	59	13	2.47	NA	1.63	3.74
cholesterol	196	259	63	0.79	0.18	0.44	1.15
Odds Ratio	196	259	63	2.21	NA	1.55	3.17
sex - female:male	1	2	NA	-2.46	0.15	-2.75	-2.16
Odds Ratio	1	2	NA	0.09	NA	0.06	0.12

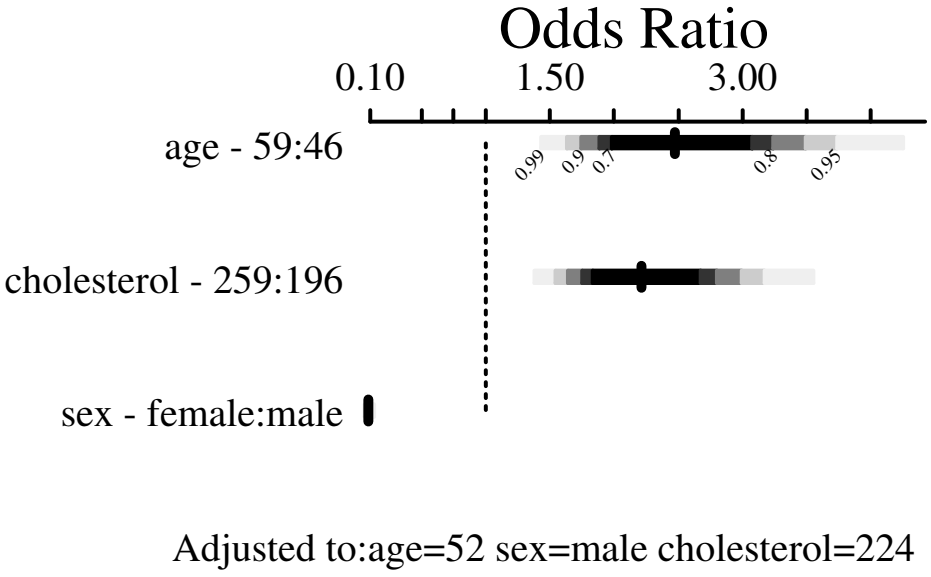

Figure 10.16: Odds ratios and confidence bars, using quartiles of age and cholesterol for assessing their effects on the odds of coronary disease.

10.11 S Functions

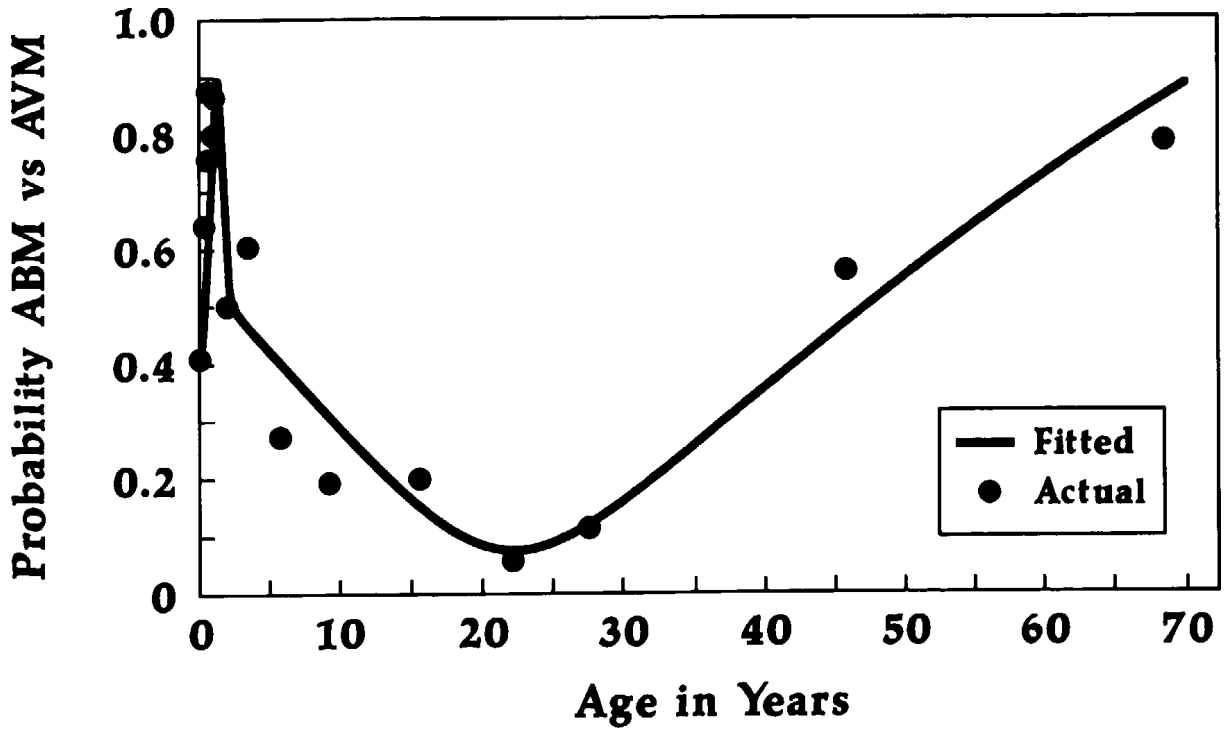

Figure 10.17: Linear spline fit for probability of bacterial vs. viral meningitis as a function of age at onset [68]. Copyrighted 1989, American Medical Association. Reprinted by permission.

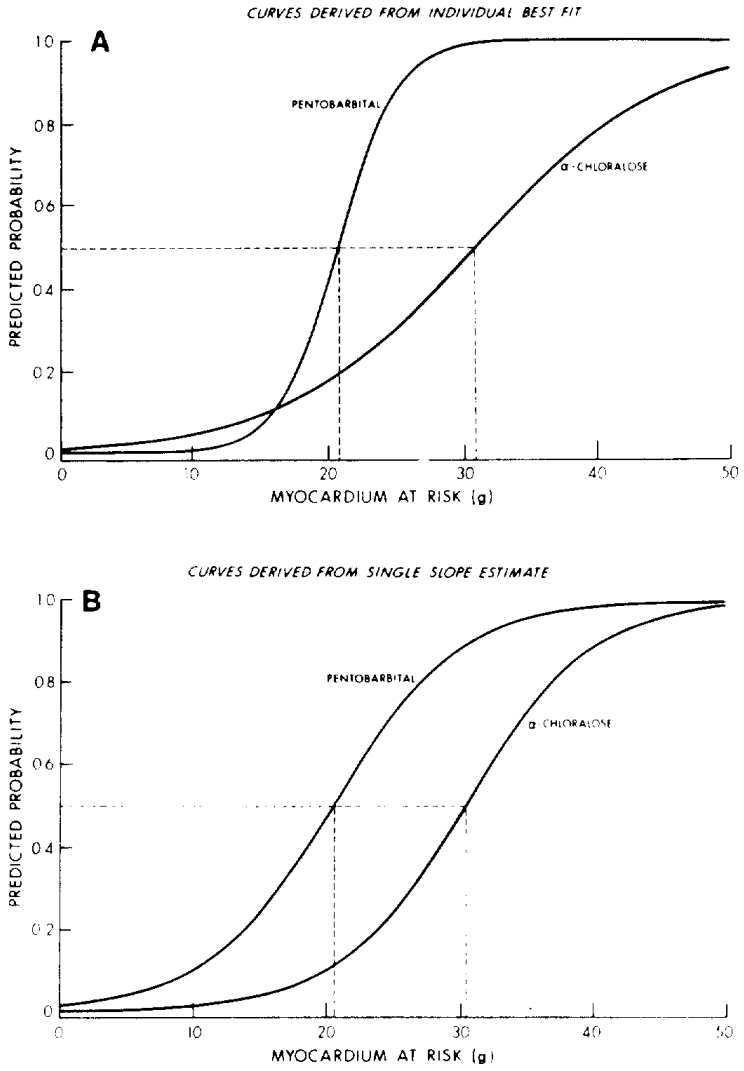

Figure 10.18: (A) Relationship between myocardium at risk and ventricular fibrillation, based on the individual best fit equations for animals anesthetized with pentobarbital and α-chloralose. The amount of myocardium at risk at which 0.5 of the animals are expected to fibrillate (MAR_{50}) is shown for each anesthetic group. (B) Relationship between myocardium at risk and ventricular fibrillation, based on equations derived from the single slope estimate. Note that the MAR_{50} describes the overall relationship between myocardium at risk and outcome when either the individual best fit slope or the single slope estimate is used. The shift of the curve to the right during α-chloralose anesthesia is well described by the shift in MAR_{50}. Test for interaction had P=0.10^{78}. Reprinted by permission, NRC Research Press.

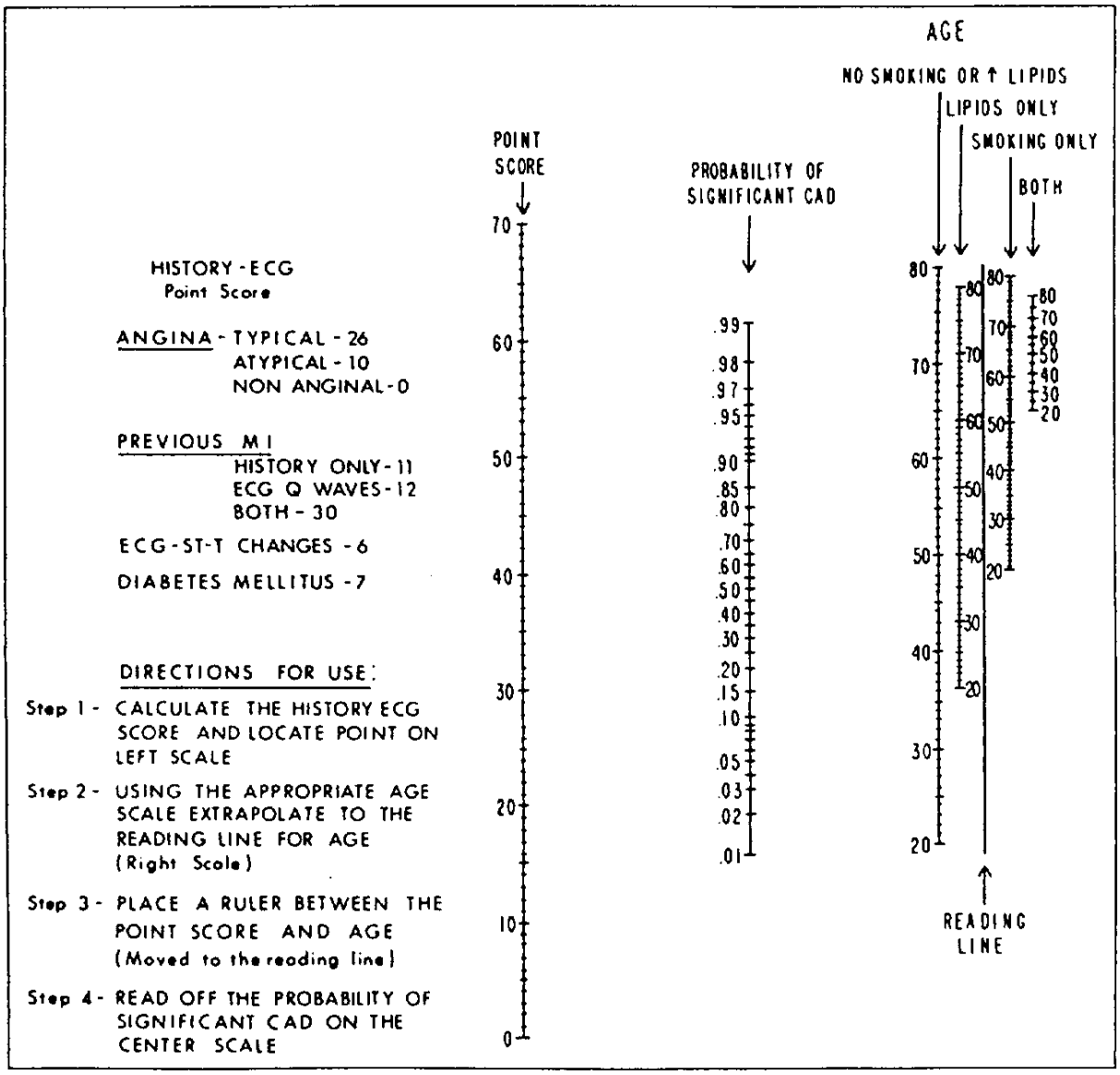

Figure 10.19: A nomogram for estimating the likelihood of significant coronary artery disease (CAD) in women. ECG = electrocardiographic; MI = myocardial infarction [61]. Reprinted from American Journal of Medicine, Vol 75, Pryor DB et al., "Estimating the likelihood of significant coronary artery disease", p. 778, Copyright 1983, with permission from Excerpta Medica, Inc.

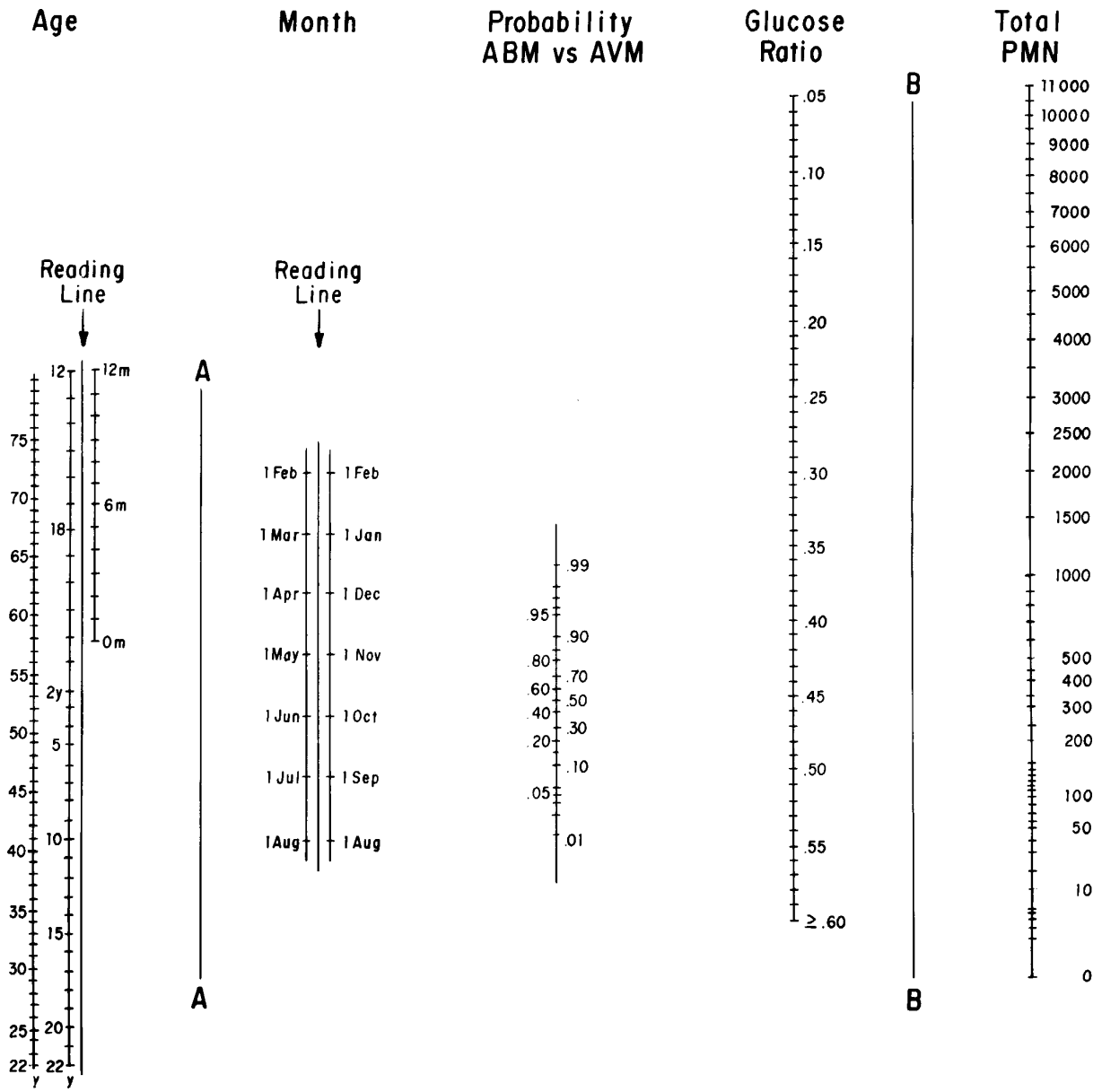

Figure 10.20: Nomogram for estimating probability of bacterial (ABM) vs. viral (AVM) meningitis. Step 1, place ruler on reading lines for patient's age and month of presentation and mark intersection with line A; step 2, place ruler on values for glucose ratio and total polymorphonuclear leukocyte (PMN) count in cerbrospinal 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[68]. Copyrighted 1989, American Medical Association. Reprinted by permission.

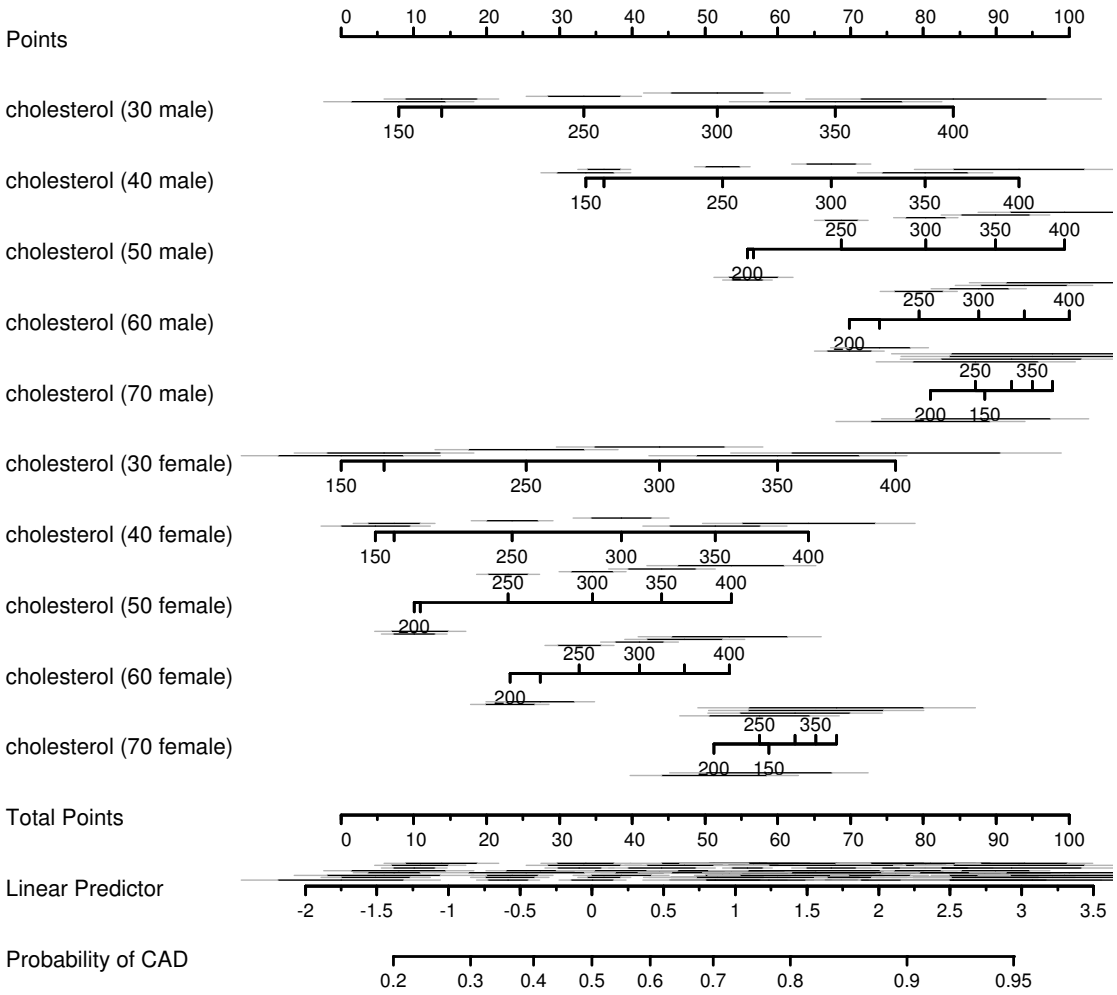

Figure 10.21: Nomogram relating age, sex, and cholesterol to the log odds and to the probability of significant coronary artery disease. Select one axis corresponding to sex and to age $\in \{30, 40, 50, 60, 70\}$. There was linear interaction between age and sex and between age and cholesterol. 0.70 and 0.90 confidence intervals are shown (0.90 in gray). Note that for the "Linear Predictor" scale there are various lengths of confidence intervals near the same value of $X\hat{\beta}$, demonstrating that the standard error of $X\hat{\beta}$ depends on the individual X values. Also note that confidence intervals corresponding to smaller patient groups (e.g., females) are wider.

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

```
   n missing unique
1309       0      3

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

survived : Survived

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

age : Age (Year)

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

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

sex : Sex

```
   n missing unique
1309       0      2

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

sibsp : Number of Siblings/Spouses Aboard

```
   n missing unique   Mean
1309 0       7      0.4989

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

parch : Number of Parents/Children Aboard

```
   n missing unique  Mean
1309 0       8      0.385

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

```
> dd ← datadist(titanic3[,v])
> # describe distributions of variables to Design
> options(datadist='dd')
> attach(titanic3[,v])
```

```
> options(digits=2)
> s ← summary(survived ∼ age + sex + pclass +
+               cut2(sibsp,0:3) + cut2(parch,0:3))
> s   # usual print
> w ← latex(s)
> # create LaTeX code for Table 11.1
> plot(s)
> # convert table to dot plot (Figure 11.1)
```

Table 11.1: Survived $N = 1309$

	N	survived
Age		
[0.167,21.000)	249	0.46
[21.000,28.000)	255	0.38
[28.000,39.000)	277	0.40
[39.000,80.000]	265	0.39
Missing	263	0.28
Sex		
female	466	0.73
male	843	0.19
Passenger Class		
1st	323	0.62
2nd	277	0.43
3rd	709	0.26
Number of Siblings/Spouses Aboard		
0	891	0.35
1	319	0.51
2	42	0.45
[3,8]	57	0.16
Number of Parents/Children Aboard		
0	1002	0.34
1	170	0.59
2	113	0.50
[3,9]	24	0.29
Overall		
	1309	0.38

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

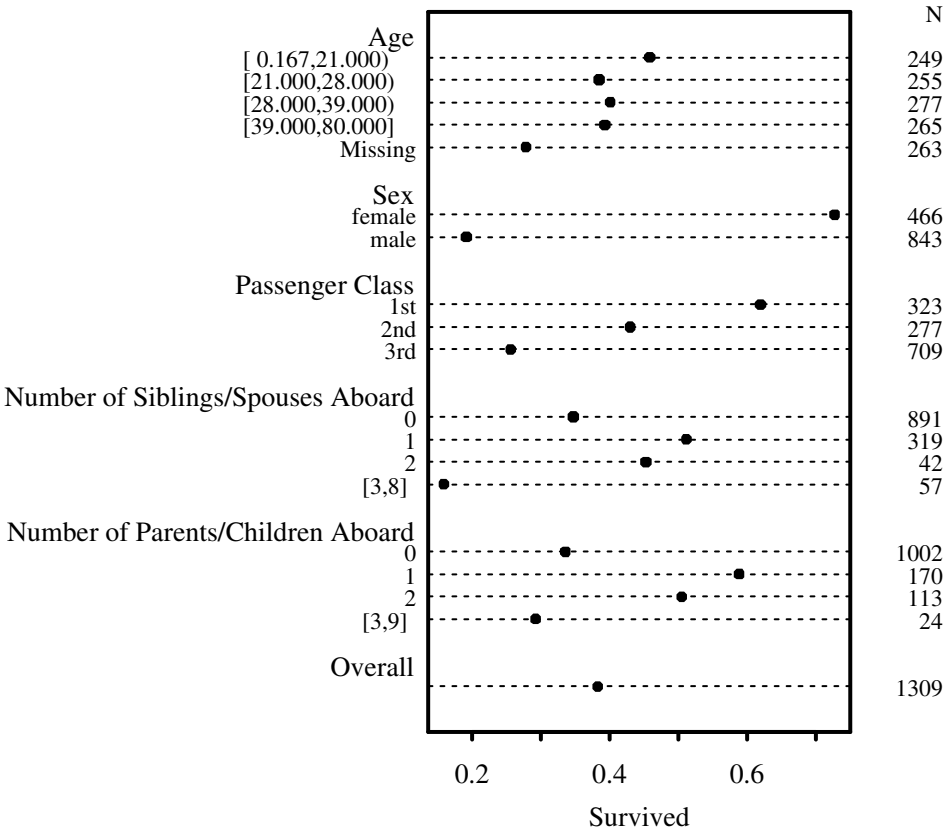

Figure 11.1: Univariable summaries of Titanic survival

```
> 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)
```

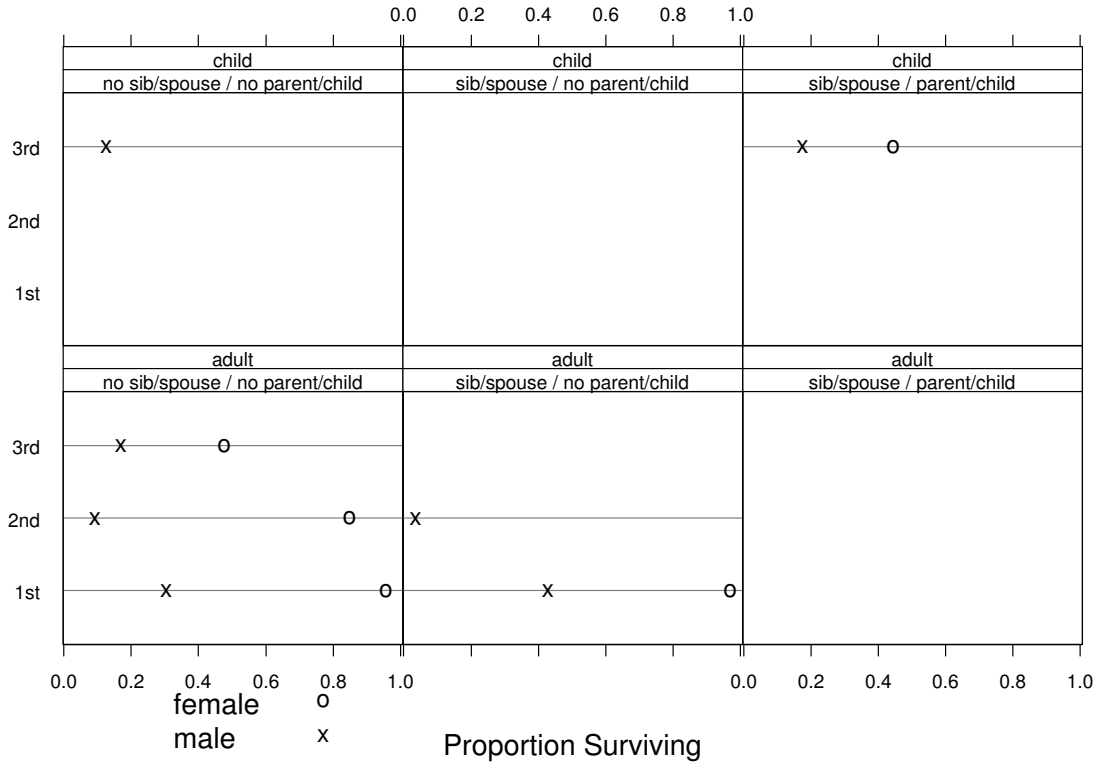

Figure 11.2: Multi-way summary of Titanic survival

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)
```

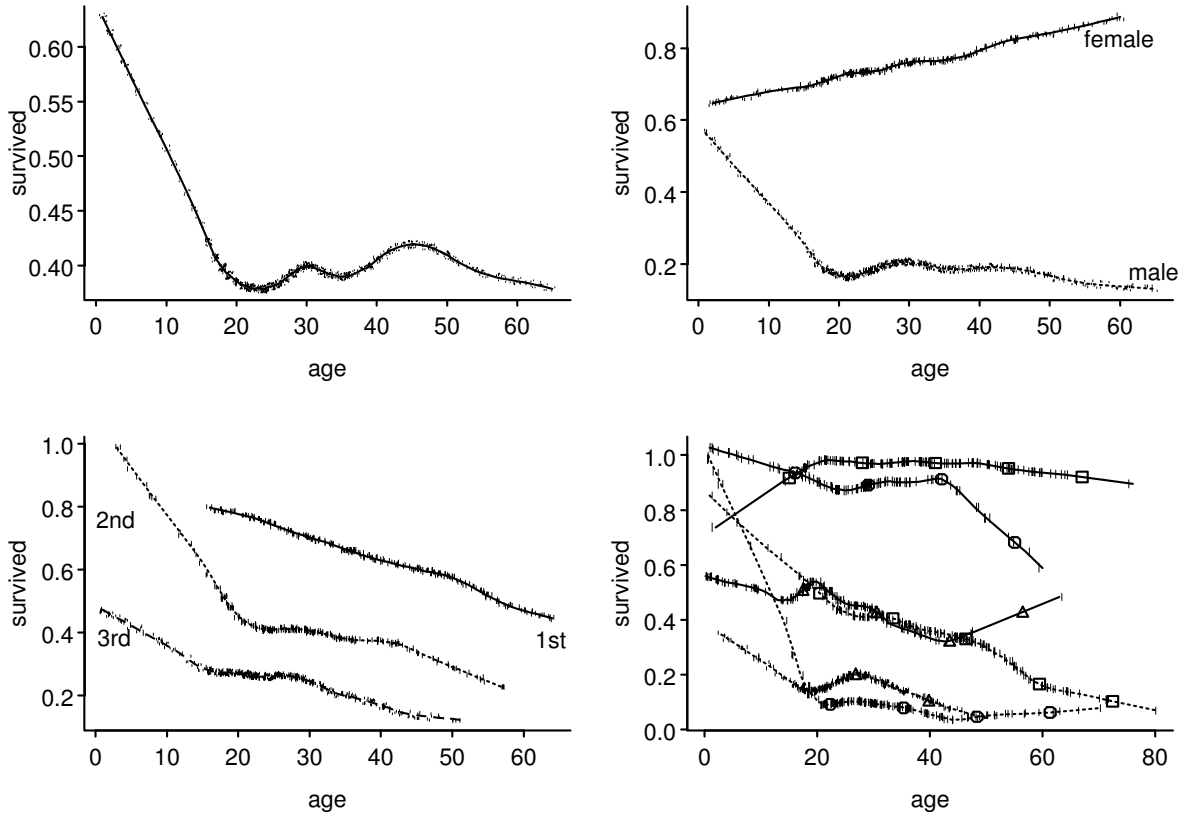

Figure 11.3: Nonparametric regression (`loess`) estimates of the relationship between age and the probability of surviving the Titanic. The top left panel shows unstratified estimates. The top right panel depicts relationships stratified by sex. The bottom left and right panels show respectively estimates 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)
```

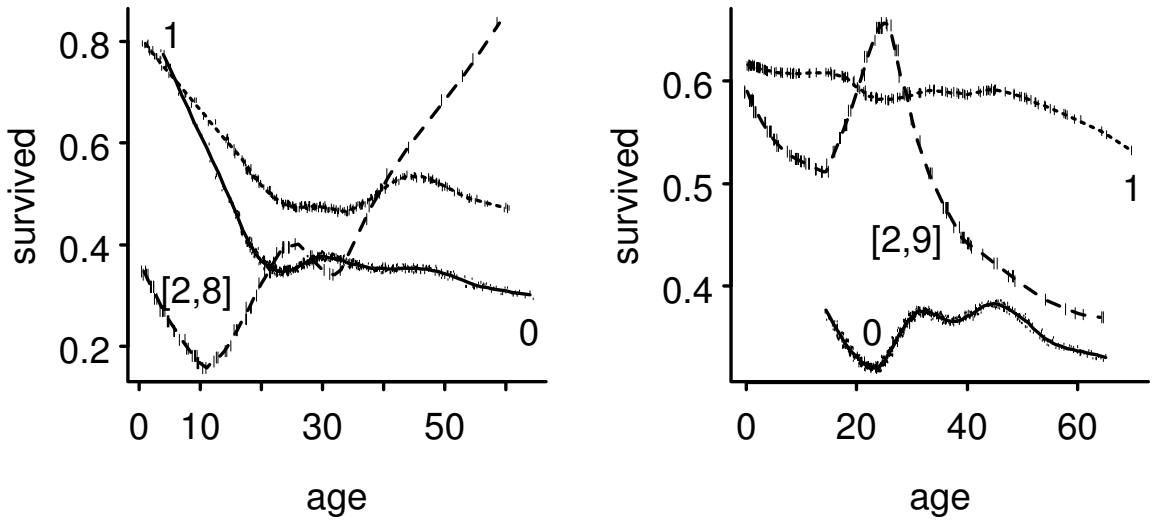

Figure 11.4: Relationship between age and survival stratified by the number of siblings or spouses on board (left panel) or by the number of parents or children of the passenger on board (right panel)

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.

```
> 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`

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

```
    > 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.

```
> 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
```

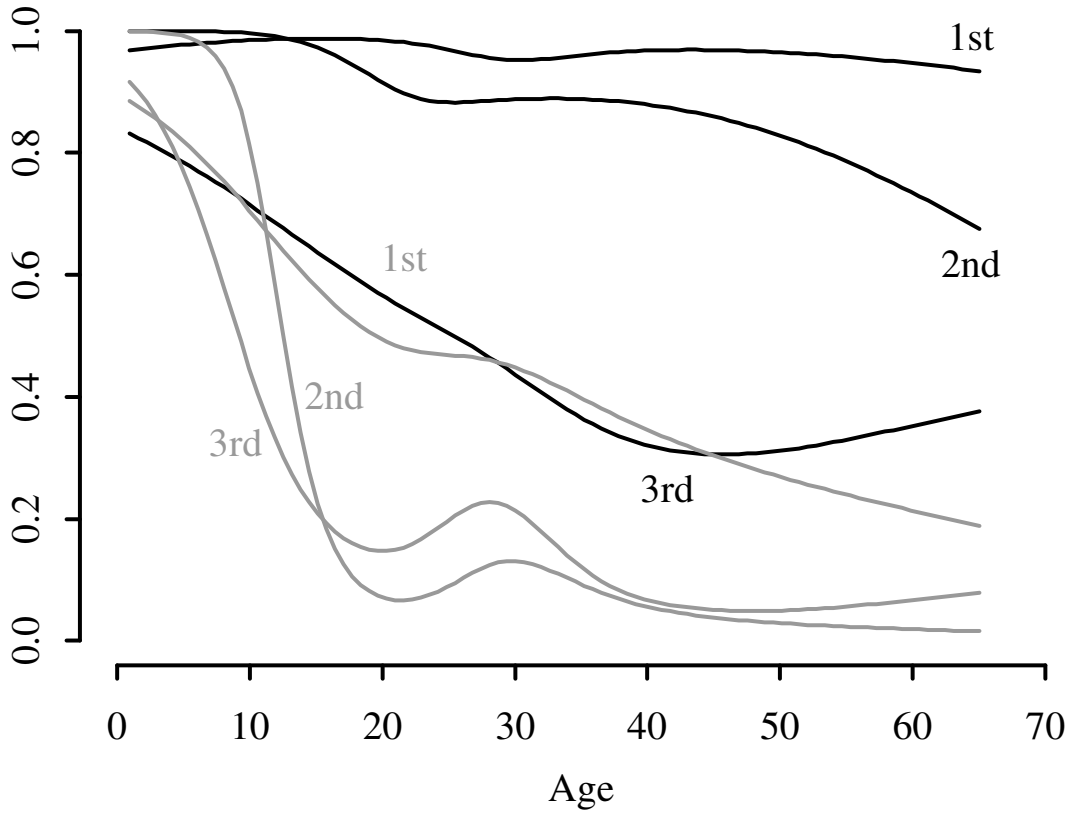

Figure 11.5: Effects of predictors on probability of survival of Titanic passengers, estimated for zero siblings or spouses. Lines for females are black; males are drawn using gray scale.

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`

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

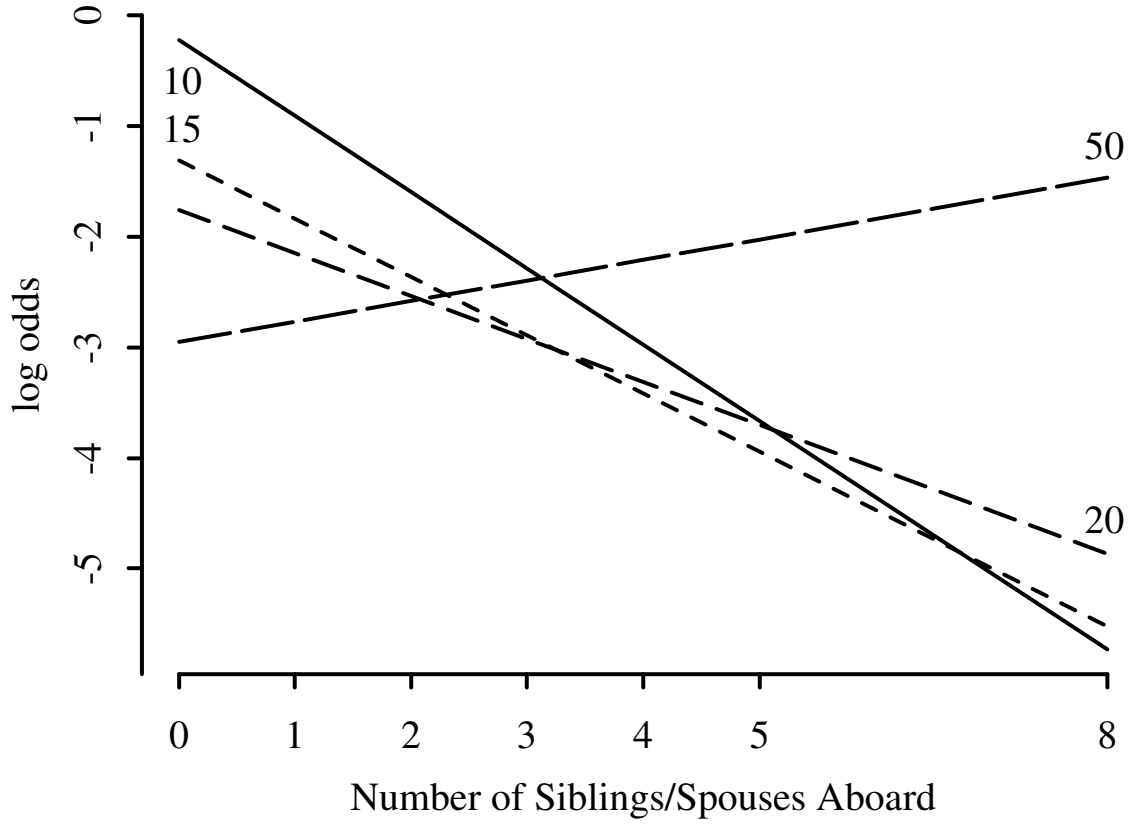

Figure 11.6: Effect of number of siblings and spouses on the log odds of surviving, for third class males. Numbers next to lines are ages in years.

Validate the model using the bootstrap to check overfitting. Ignoring two very insignificant pooled tests.

```
> 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)
```

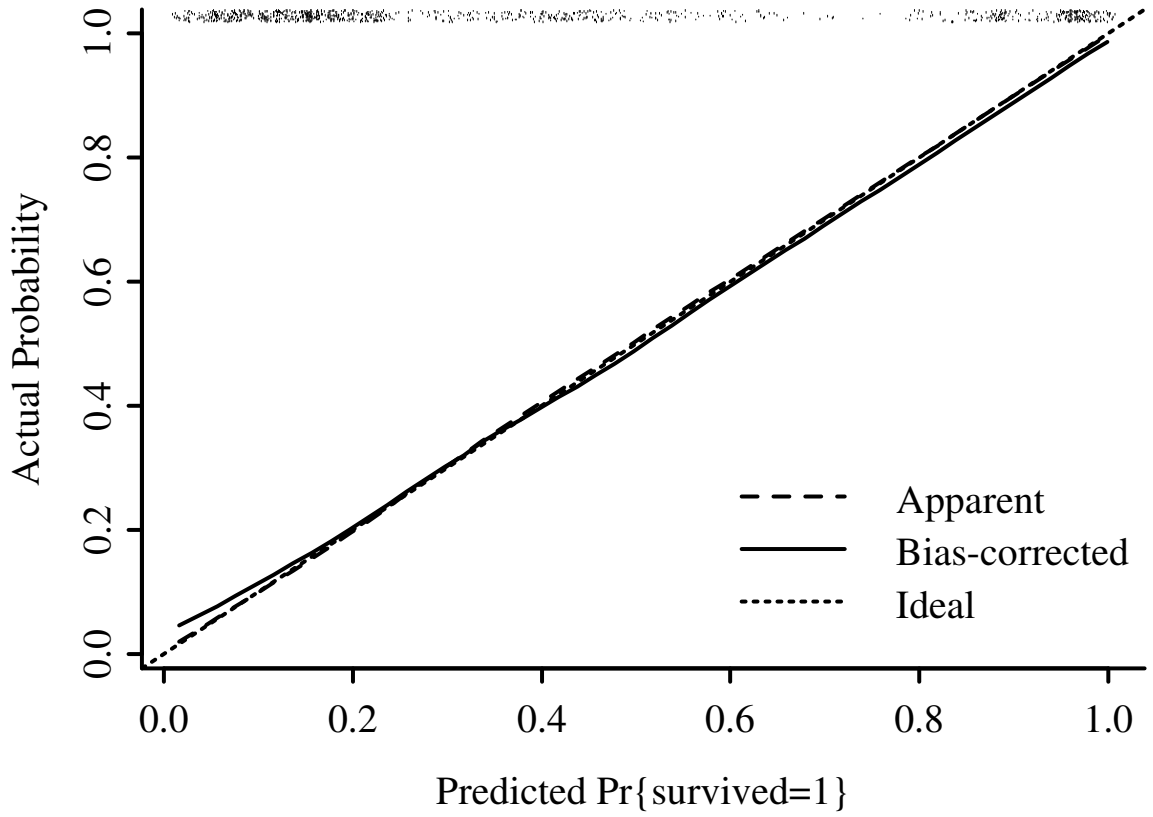

Figure 11.7: Bootstrap overfitting-corrected loess nonparametric calibration curve for casewise deletion model

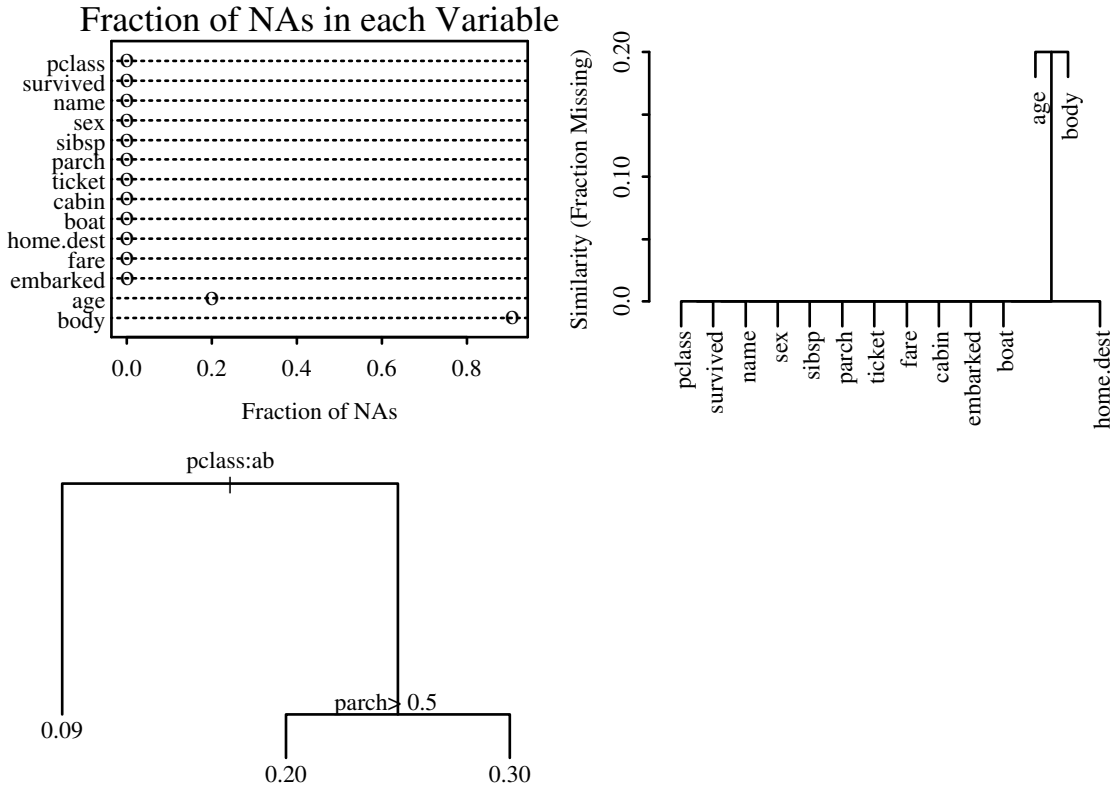

Figure 11.8: Patterns of missing data. Upper left panel shows the fraction of observations missing on each predictor. Upper right panel depicts a hierarchical cluster analysis of missingness combinations. The similarity measure shown on the Y-axis is the fraction of observations for which both variables are missing. Lower left panel shows the result of recursive partitioning for predicting `is.na(age)`. The `rpart` function found only strong patterns according to passenger class.

But moderate problem with missing data

11.4 Examining Missing Data Patterns

```
> 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
```

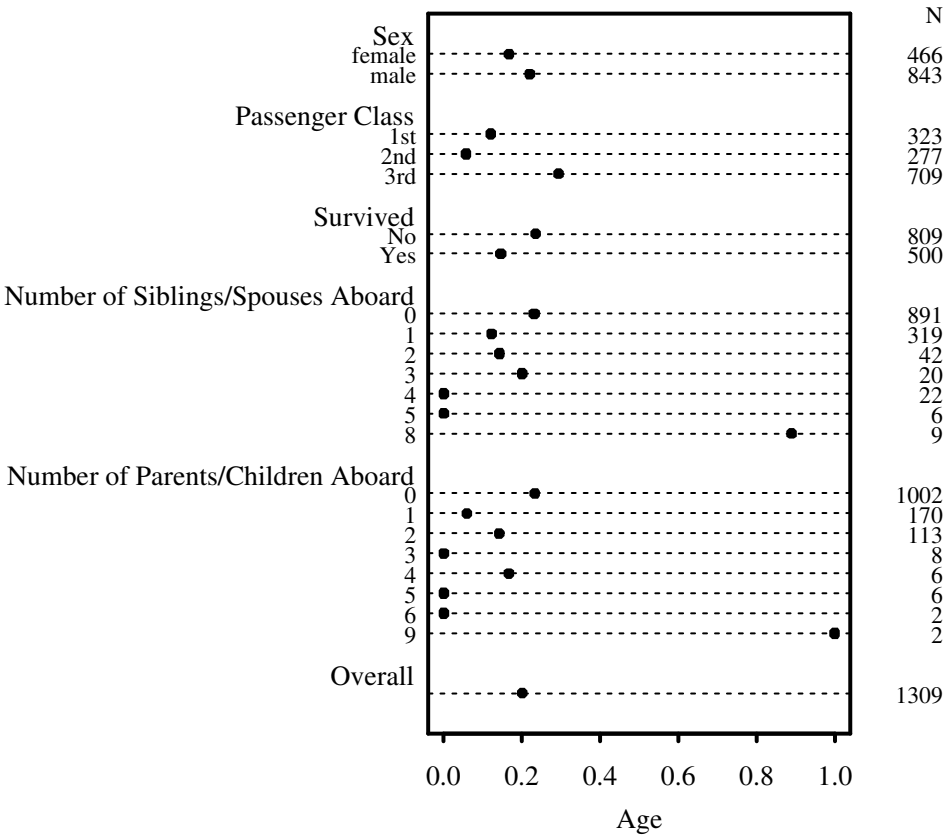

Figure 11.9: Univariable descriptions of proportion of passengers with missing age

Table 11.4: Wald Statistics for `is.na(age)`

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

```
    > m ← lrm(is.na(age) ∼ sex * pclass + survived + sibsp + parch)
    > m

Frequencies of Responses
 FALSE TRUE
  1046  263


  Obs Max Deriv Model L.R. d.f. P   C  Dxy Gamma Tau-a   R2 Brier
 1309    5e-006        115    8 0 0.7 0.41  0.45  0.13 0.13  0.15

                          Coef    S.E. Wald Z      P
            Intercept -2.20299 0.36412 -6.05  0.0000
             sex=male  0.64395 0.39526  1.63  0.1033
           pclass=2nd -1.00793 0.66578 -1.51  0.1300
           pclass=3rd  1.61242 0.35962  4.48  0.0000
             survived -0.18058 0.18280 -0.99  0.3232
                sibsp  0.04353 0.07372  0.59  0.5548
                parch -0.35261 0.12529 -2.81  0.0049
sex=male * pclass=2nd  0.13467 0.75451  0.18  0.8583
sex=male * pclass=3rd -0.85628 0.42144 -2.03  0.0422

    > 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

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

Iterations:

R-squared achieved in predicting each variable:

   age   sex pclass sibsp parch
 0.258 0.078  0.244 0.241 0.288

Adjusted R-squared:

   age   sex pclass sibsp parch
 0.254 0.074   0.24 0.238 0.285


Coefficients of canonical variates for predicting each (row) variable

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

Summary of imputed values

age
   n missing unique  Mean   .05   .10   .25   .50   .75   .90   .95
```

```
 263 0       24     28.41 16.76 21.66 26.17 28.04 28.04 42.92 42.92

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

Starting estimates for imputed values:

 age sex pclass sibsp parch
  28   2      3     0     0


    > # Look at mean imputed values by sex,pclass and observed means
    > # age.i is age, filled in with conditional mean estimates

    > age.i ← impute(xtrans, age)



    > 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 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)
```

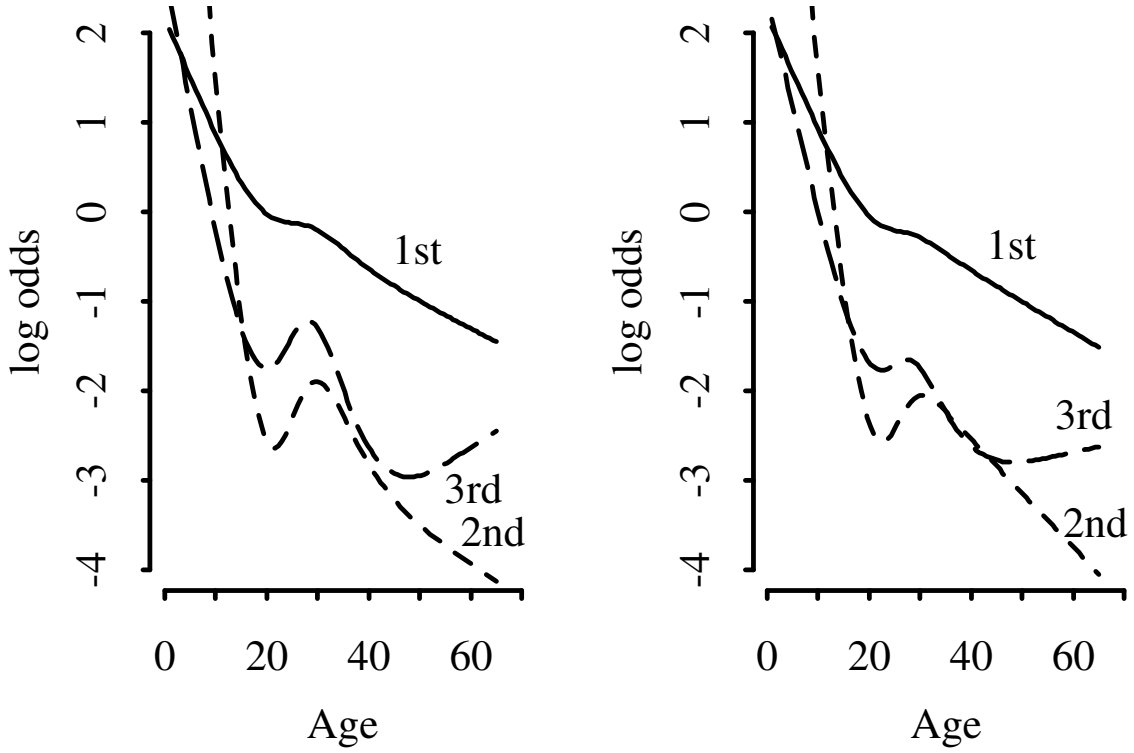

Figure 11.10: Predicted log odds of survival for males from fit using casewise deletion (left panel) and single conditional mean imputation (right panel). `sibsp` is set to zero for these predicted values.

```
> anova(f.si)   # Table 11.5
```

Table 11.5: Wald Statistics for `survived`

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

11.6 Multiple Imputation

Again needed to force a linear age relationship.

```
> 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

           age   sex pclass sibsp parch survived
     age        0.37  6.76  -2.17 -2.48 -5.37
     sex  0.00       -0.10   0.15  0.21  1.95
  pclass  0.07 -0.09         0.12  0.16  1.26
   sibsp -0.03  0.18  0.15         0.84 -0.65
   parch -0.03  0.22  0.18   0.75        0.29
survived -0.01  0.23  0.16  -0.06  0.03

Summary of imputed values

age
    n missing unique  Mean   .05   .10   .25   .50   .75   .90   .95
 1315 0       1006   28.96  6.88 12.27 19.38 27.80 37.94 47.73 52.95

lowest :  0.1667  0.5826  0.6028  0.6567  0.9990
highest: 68.9805 70.5000 72.2900 73.8221 74.0000
```

```
Starting estimates for imputed values:

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

```
   > # 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 60 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 sex=male pclass=2nd pclass=3rd age age' age'' age''' sibsp
         1      1.1          1          1   1    1   1.1    1.1   1.3
 sex=male * pclass=2nd sex=male * pclass=3rd sex=male * age
                    1                     1           1.2
 sex=male * age' sex=male * age'' sex=male * age''' pclass=2nd * age
             1.1              1.1               1.1                 1
 pclass=3rd * age pclass=2nd * age' pclass=3rd * age' pclass=2nd * age''
                1                  1                 1                  1
 pclass=3rd * age'' pclass=2nd * age''' pclass=3rd * age''' age * sibsp
                1.1                    1                 1.1         1.3
 age' * sibsp age'' * sibsp age''' * sibsp
```

```
          1.3           1.3           1.3
```

```
    > f.mi

Frequencies of Responses
   0   1
 809 500


  Obs Max Deriv Model L.R. d.f. P    C  Dxy Gamma Tau-a   R2 Brier
 1309    2e-007        662   26 0 0.87 0.74  0.75  0.35 0.54  0.13
```

```
. . . . .
```

```
    > anova(f.mi)   # Table 11.6
```

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

Show estimated effects of age by classes.

```
    > par(mfrow=c(1,2))     # Figure 11.11
    > plot(f.si, age.i=NA, pclass=NA, ylim=c(-4,2),
    +      conf.int=F, adj.subtitle=F)
    > plot(f.mi, age=NA,   pclass=NA, ylim=c(-4,2),
    +      conf.int=F, adj.subtitle=F)
    > par(mfrow=c(1,1))
```

Table 11.6: Wald Statistics for `survived`

	χ^2	$d.f.$	P
sex (Factor+Higher Order Factors)	237.33	7	< 0.0001
All Interactions	53.50	6	< 0.0001
pclass (Factor+Higher Order Factors)	116.29	12	< 0.0001
All Interactions	37.18	10	0.0001
age (Factor+Higher Order Factors)	46.58	20	0.0007
All Interactions	23.78	16	0.0944
Nonlinear (Factor+Higher Order Factors)	20.12	15	0.1673
sibsp (Factor+Higher Order Factors)	17.81	5	0.0032
All Interactions	8.03	4	0.0905
sex $\times$ pclass (Factor+Higher Order Factors)	32.15	2	< 0.0001
sex $\times$ age (Factor+Higher Order Factors)	10.06	4	0.0394
Nonlinear	7.73	3	0.0520
Nonlinear Interaction : f(A,B) vs. AB	7.73	3	0.0520
pclass $\times$ age (Factor+Higher Order Factors)	5.68	8	0.6828
Nonlinear	5.16	6	0.5240
Nonlinear Interaction : f(A,B) vs. AB	5.16	6	0.5240
age $\times$ sibsp (Factor+Higher Order Factors)	8.03	4	0.0905
Nonlinear	1.31	3	0.7256
Nonlinear Interaction : f(A,B) vs. AB	1.31	3	0.7256
TOTAL NONLINEAR	20.12	15	0.1673
TOTAL INTERACTION	64.94	18	< 0.0001
TOTAL NONLINEAR + INTERACTION	67.84	21	< 0.0001
TOTAL	291.66	26	< 0.0001

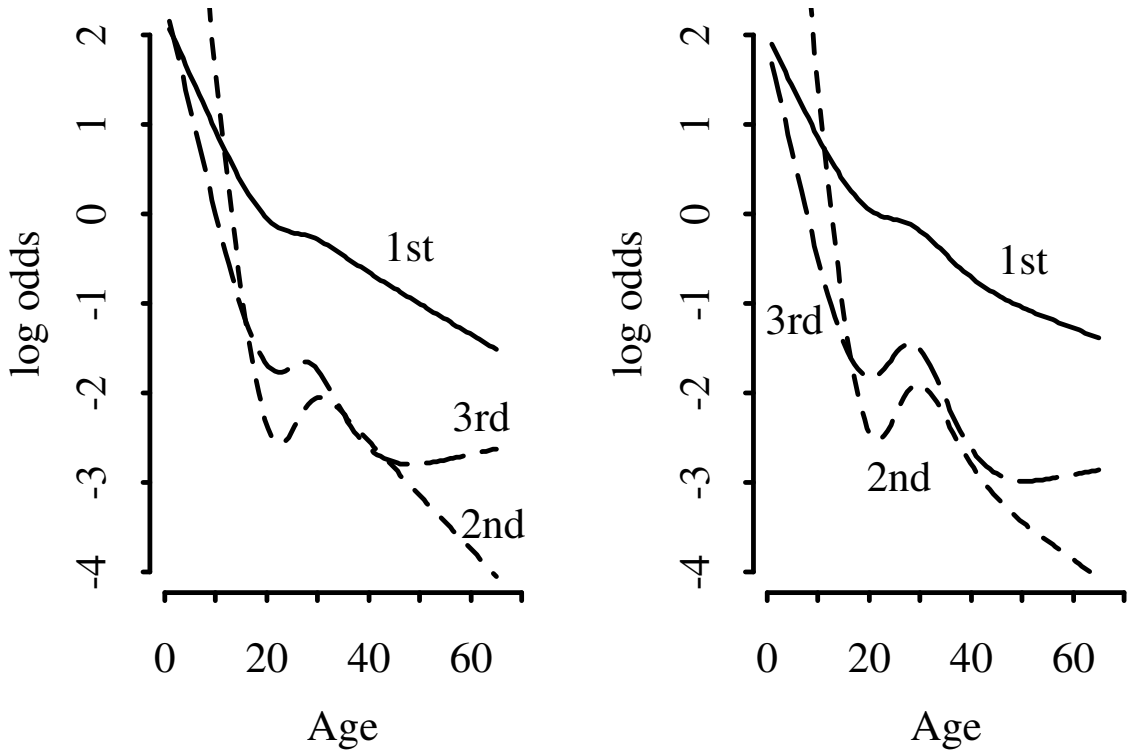

Figure 11.11: Predicted log odds of survival for males from fit using single conditional mean imputation again (left panel) and multiple random draw imputation (right panel). Both sets of predictions are for `sibsp`=0.

11.7 Summarizing the Fitted Model

Show odds ratios for changes in predictor values

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

```
   age    sex pclass sibsp phat
 1   2 female    1st     0 0.97
 2  21 female    1st     0 0.98
 3  50 female    1st     0 0.97
 4   2   male    1st     0 0.85
 5  21   male    1st     0 0.50
 6  50   male    1st     0 0.26
 7   2 female    2nd     0 1.00
 8  21 female    2nd     0 0.90
 9  50 female    2nd     0 0.83
10   2   male    2nd     0 1.00
11  21   male    2nd     0 0.07
12  50   male    2nd     0 0.03
13   2 female    3rd     0 0.78
14  21 female    3rd     0 0.58
15  50 female    3rd     0 0.36
16   2   male    3rd     0 0.80
17  21   male    3rd     0 0.14
18  50   male    3rd     0 0.05
```

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 *
```

```
        pmax(age - 56, 0)^3 - 0.92513167 * sibsp + (sex == "male") *
        (-0.58477831 * (pclass == "2nd") + 1.9845306 * (pclass ==
        "3rd")) + (sex == "male") * (-0.19813189 * age +
        0.00035992445 * pmax(age - 5, 0)^3 - 0.0022232117 * pmax(
        age - 21, 0)^3 + 0.002898667 * pmax(age - 28, 0)^3 -
        0.0011424428 * pmax(age - 37, 0)^3 + 0.00010706304 * pmax(
        age - 56, 0)^3) + (pclass == "2nd") * (-0.52315032 * age +
        0.00066960929 * pmax(age - 5, 0)^3 - 0.0030561012 * pmax(
        age - 21, 0)^3 + 0.0031229941 * pmax(age - 28, 0)^3 -
        0.0007700192 * pmax(age - 37, 0)^3 + 3.3517068e-005 * pmax(
        age - 56, 0)^3) + (pclass == "3rd") * (-0.13015825 * age +
        0.00024102399 * pmax(age - 5, 0)^3 - 0.0015255584 * pmax(
        age - 21, 0)^3 + 0.002044236 * pmax(age - 28, 0)^3 -
        0.00084927821 * pmax(age - 37, 0)^3 + 8.9576665e-005 *
        pmax(age - 56, 0)^3) + sibsp * (0.031980533 * age -
        1.679181e-005 * pmax(age - 5, 0)^3 + 0.00015089349 * pmax(
        age - 21, 0)^3 - 0.00030324003 * pmax(age - 28, 0)^3 +
        0.0002139911 * pmax(age - 37, 0)^3 - 4.4852752e-005 * pmax(
        age - 56, 0)^3)
}
    > # Note: if don't define sibsp to pred.logit, default to 0
    > plogis(pred.logit(age=c(2,21,50), sex='male', pclass='3rd'))

[1] 0.80 0.14 0.05
```

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.

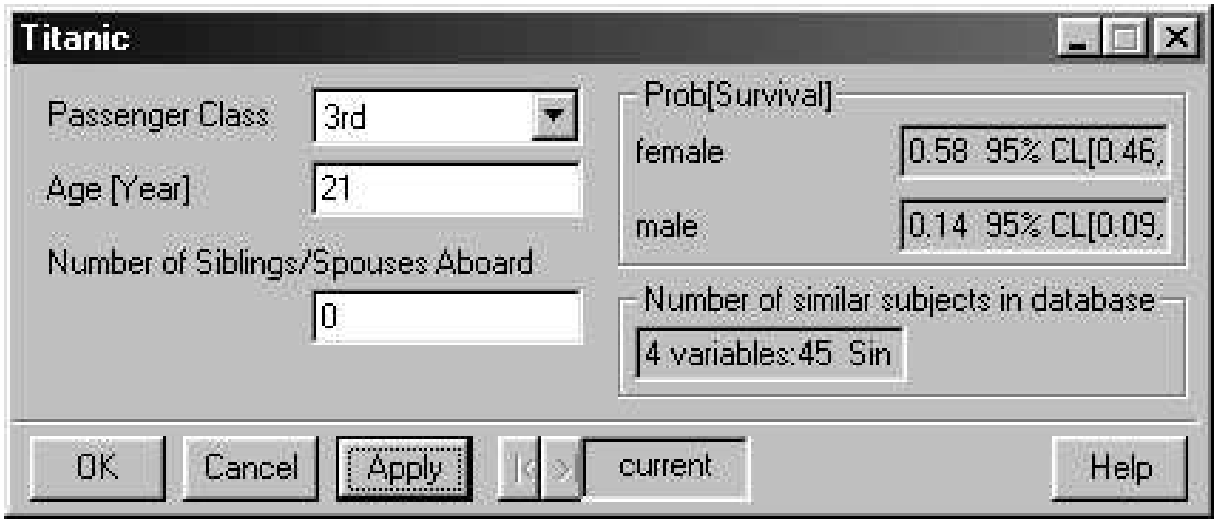

Figure 11.13: Menu for deriving and plotting $\hat{P}$ and 0.95 C.L.

```
> 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()
```

R / S-PLUS Software Used

Library	Purpose	Functions
`Hmisc` Harrell	Miscellaneous functions	`summary,plsmo,naclus,llist,latex summarize,Dotplot,transcan,impute fit.mult.impute,describe,dataRep`
`Design` Harrell	Modeling, validation, Model presentation graphics	`datadist,lrm,nomogram, validate,calibrate, Function,rcs,Dialog`
`rpart` Atkinson&Therneau	Recursive partitioning	`rpart`

Bibliography

[1] D. G. Altman and P. K. Andersen. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, 8:771–783, 1989. [67]

[2] A. C. Atkinson. A note on the generalized information criterion for choice of a model. *Biometrika*, 67:413–418, 1980. [28, 66]

[3] F. Barzi and M. Woodward. Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *American Journal of Epidemiology*, 160:34–45, 2004. [52, 57]

[4] D. A. Belsley. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. Wiley, New York, 1991. [73]

[5] D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York, 1980. [88, 89]

[6] J. K. Benedetti, P. Liu, H. N. Sather, J. Seinfeld, and M. A. Epton. Effective sample size for tests of censored survival data. *Biometrika*, 69:343–349, 1982. [67]

[7] M. Blettner and W. Sauerbrei. Influence of model-building strategies on the results of a case-control study. *Statistics in Medicine*, 12:1325–1338, 1993. [111]

[8] J. G. Booth and S. Sarkar. Monte Carlo approximation of bootstrap variances. *American Statistician*, 52:354–357, 1998. [100]

[9] L. Breiman. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, 87:738–754, 1992. [66, 67, 104]

[10] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, 80:580–619, 1985. [80]

[11] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1984. [32]

[12] D. Brownstone. Regression strategies. In *Proceedings of the 20th Symposium on the Interface between Computer Science and Statistics*, pages 74–79, Washington, DC, 1988. American Statistical Association. [111]

[13] J. M. Chambers and T. J. Hastie, editors. *Statistical Models in S*. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1992. [45]

[14] C. Chatfield. Avoiding statistical pitfalls (with discussion). *Statistical Science*, 6:240–268, 1991. [89]

[15] C. Chatfield. Model uncertainty, data mining and statistical inference (with discussion). *Journal of the Royal Statistical Society A*, 158:419–466, 1995. [64, 111]

[16] S. Chatterjee and B. Price. *Regression Analysis by Example*. Wiley, New York, second edition, 1991. [72]

[17] A. Ciampi, J. Thiffault, J.-P. Nakache, and B. Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition. *Computational Statistics and Data Analysis*, 1986:185–204, 1986. [75]

[18] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836, 1979. [30]

[19] E. F. Cook and L. Goldman. Asymmetric stratification: An outline for an efficient method for controlling confounding in cohort studies. *American Journal of Epidemiology*, 127:626–639, 1988. [34]

[20] J. B. Copas. Regression, prediction and shrinkage (with discussion). *Journal of the Royal Statistical Society B*, 45:311–354, 1983. [69, 71]

[21] J. B. Copas. Cross-validation shrinkage of regression predictors. *Journal of the Royal Statistical Society B*, 49:175–183, 1987. [109]

[22] D. R. Cox. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society B*, 34:187–220, 1972. [47]

[23] S. L. Crawford, S. L. Tennstedt, and J. B. McKinlay. A comparison of analytic methods for non-random missingness of outcome data. *Journal of Clinical Epidemiology*, 48:209–219, 1995. [50, 91]

[24] N. J. Crichton and J. P. Hinde. Correspondence analysis as a screening method for indicants for clinical diagnosis. *Statistics in Medicine*, 8:1351–1362, 1989. [75]

[25] R. B. D'Agostino, A. J. Belanger, E. W. Markson, M. Kelly-Hayes, and P. A. Wolf. Development of health risk appraisal functions in the presence of multiple indicators: The Framingham Study nursing home institutionalization model. *Statistics in Medicine*, 14:1757–1770, 1995. [73, 75]

[26] C. E. Davis, J. E. Hyde, S. I. Bangdiwala, and J. J. Nelson. An example of dependencies among variables in a conditional logistic regression. In S. Moolgavkar and R. Prentice, editors, *Modern Statistical Methods in Chronic Disease Epidemiology*, pages 140–147. Wiley, New York, 1986. [73]

[27] S. Derksen and H. J. Keselman. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, 45:265–282, 1992. [64]

[28] T. F. Devlin and B. J. Weeks. Spline functions for logistic regression modeling. In *Proceedings of the Eleventh Annual SAS Users Group International Conference*, pages 646–651, Cary, NC, 1986. SAS Institute, Inc. [24]

[29] S. Durrleman and R. Simon. Flexible regression models with cubic splines. *Statistics in Medicine*, 8:551–561, 1989. [27]

[30] B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78:316–331, 1983. [105, 108, 109]

[31] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993. [108]

[32] B. Efron and R. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *Journal of the American Statistical Association*, 92:548–560, 1997. [108]

[33] J. J. Faraway. The cost of data analysis. *Journal of Computational and Graphical Statistics*, 1:213–229, 1992. [94, 108, 111]

[34] D. Freedman, W. Navidi, and S. Peters. *On the Impact of Variable Selection in Fitting Regression Equations*, pages 1–16. Lecture Notes in Economics and Mathematical Systems. Springer-Verlag, New York, 1988. [109]

[35] J. H. Friedman. A variable span smoother. Technical Report 5, Laboratory for Computational Statistics, Department of Statistics, Stanford University, 1984. [80]

[36] P. M. Grambsch and P. C. O'Brien. The effects of transformations and preliminary tests for non-linearity in regression. *Statistics in Medicine*, 10:697–709, 1991. [35, 64]

[37] R. J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, 87:942–951, 1992. [45, 71]

[38] R. J. Gray. Spline-based tests in survival analysis. *Biometrics*, 50:640–652, 1994. [45]

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

[40] S. Greenland. When should epidemiologic regressions use random coefficients? *Biometrics*, 56:915–921, 2000. [64]

[41] F. E. Harrell. The LOGIST Procedure. In *SUGI Supplemental Library Users Guide*, pages 269–293. SAS Institute, Inc., Cary, NC, Version 5 edition, 1986. [65]

[42] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati. Regression modeling strategies for improved prognostic prediction. *Statistics in Medicine*, 3:143–152, 1984. [67]

[43] F. E. Harrell, K. L. Lee, D. B. Matchar, and T. A. Reichert. Regression models for prognostic prediction: Advantages, problems, and suggested solutions. *Cancer Treatment Reports*, 69:1071–1077, 1985. [67]

[44] F. E. Harrell, K. L. Lee, and B. G. Pollock. Regression models in clinical studies: Determining relationships between predictors and response. *Journal of the National Cancer Institute*, 80:1198–1202, 1988. [31]

[45] F. E. Harrell, P. A. Margolis, S. Gove, K. E. Mason, E. K. Mulholland, D. Lehmann, L. Muhe, S. Gatchalian, and H. F. Eichenwald. Development of a clinical prediction model for an ordinal outcome: The World Health Organization ARI Multicentre Study of clinical signs and etiologic agents of pneumonia, sepsis, and meningitis in young infants. *Statistics in Medicine*, 17:909–944, 1998. [71, 92]

[46] W. Hoeffding. A non-parametric test of independence. *Annals of Mathematical Statistics*, 19:546–557, 1948. [75]

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

[48] L. I. Iezzoni. Dimensions of risk. In L. I. Iezzoni, editor, *Risk Adjustment for Measuring Health Outcomes*, chapter 2, pages 29–118. Foundation of the American College of Healthcare Executives, Ann Arbor, MI, 1994. [7]

[49] W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, P. Layde, R. K. Oye, P. E. Bellamy, R. B. Hakim, and D. P. Wagner. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122:191–203, 1995. [81]

[50] W. F. Kuhfeld. The PRINQUAL procedure. In *SAS/STAT User's Guide*, volume 2, chapter 34, pages 1265–1323. SAS Institute, Inc., Cary NC, fourth edition, 1990. [76, 77]

[51] J. F. Lawless and K. Singhal. Efficient screening of nonnormal regression models. *Biometrics*, 34:318–327, 1978. [66]

[52] S. le Cessie and J. C. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41:191–201, 1992. [69]

[53] A. Leclerc, D. Luce, F. Lert, J. F. Chastang, and P. Logeay. Correspondance analysis and logistic modelling: Complementary use in the analysis of a health survey among nurses. *Statistics in Medicine*, 7:983–995, 1988. [75]

[54] R. J. A. Little. Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6:287–296, 1988. [52]

[55] N. Mantel. Why stepdown procedures in variable selection. *Technometrics*, 12:621–625, 1970. [66]

[56] G. Michailidis and J. de Leeuw. The Gifi system of descriptive multivariate analysis. *Statistical Science*, 13:307–336, 1998. [75, 75]

[57] R. H. Myers. *Classical and Modern Regression with Applications*. PWS-Kent, Boston, 1990. [72]

[58] N. J. D. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78:691–692, 1991. [90]

[59] P. Peduzzi, J. Concato, A. R. Feinstein, and T. R. Holford. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *Journal of Clinical Epidemiology*, 48:1503–1510, 1995. [67]

[60] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49:1373–1379, 1996. [67, 67]

[61] D. B. Pryor, F. E. Harrell, K. L. Lee, R. M. Califf, and R. A. Rosati. Estimating the likelihood of significant coronary artery disease. *American Journal of Medicine*, 75:771–780, 1983. [162]

[62] E. B. Roecker. Prediction error and its estimation for subset-selected models. *Technometrics*, 33:459–468, 1991. [66, 104]

[63] D. Rubin and N. Schenker. Multiple imputation in health-care data bases: An overview and some applications. *Statistics in Medicine*, 10:585–598, 1991. [55]

[64] W. S. Sarle. The VARCLUS procedure. In *SAS/STAT User's Guide*, volume 2, chapter 43, pages 1641–1659. SAS Institute, Inc., Cary NC, fourth edition, 1990. [73, 75]

[65] W. Sauerbrei and M. Schumacher. A bootstrap resampling procedure for model building: Application to the Cox regression model. *Statistics in Medicine*, 11:2093–2109, 1992. [67, 105]

[66] J. Shao. Linear model selection by cross-validation. *Journal of the American Statistical Association*, 88:486–494, 1993. [105]

[67] L. R. Smith, F. E. Harrell, and L. H. Muhlbaier. Problems and potentials in modeling survival. In M. L. Grady and H. A. Schwartz, editors, *Medical Effectiveness Research Data Methods (Summary Report), AHCPR Pub. No. 92-0056*, pages 151–159. US Dept. of Health and Human Services, Agency for Health Care Policy and Research, Rockville, MD, 1992. [67]

[68] A. Spanos, F. E. Harrell, and D. T. Durack. Differential diagnosis of acute meningitis: An analysis of the predictive value of initial observations. *Journal of the American Medical Association*, 262:2700–2707, 1989. [160, 163]

[69] I. Spence and R. F. Garrison. A remarkable scatterplot. *American Statistician*, 47:12–19, 1993. [89]

[70] D. J. Spiegelhalter. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5:421–433, 1986. [69, 93, 109, 110]

[71] C. J. Stone. Comment: Generalized additive models. *Statistical Science*, 1:312–314, 1986. [27]

[72] C. J. Stone and C. Y. Koo. Additive splines in statistics. In *Proceedings of the Statistical Computing Section ASA*, pages 45–48, Washington, DC, 1985. [24, 28]

[73] G. Sun, T. L. Shook, and G. L. Kay. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *Journal of Clinical Epidemiology*, 49:907–916, 1996. [68]

[74] W. Vach and M. Blettner. Missing data in epidemiologic studies. In *Encyclopedia of Biostatistics*, pages 2641–2654. Wiley, New York, 1998. [51]

[75] J. C. van Houwelingen and S. le Cessie. Predictive value of statistical models. *Statistics in Medicine*, 8:1303–1325, 1990. [28, 69, 71, 105, 109, 111]

[76] P. Verweij and H. C. van Houwelingen. Penalized likelihood in Cox regression. *Statistics in Medicine*, 13:2427–2436, 1994. [71]

[77] Y. Wax. Collinearity diagnosis for a relative risk regression analysis: An application to assessment of diet-cancer relationship in epidemiological studies. *Statistics in Medicine*, 11:1273–1287, 1992. [73]

[78] T. L. Wenger, F. E. Harrell, K. K. Brown, S. Lederman, and H. C. Strauss. Ventricular fibrillation following canine coronary reperfusion: Different outcomes with pentobarbital and α-chloralose. *Canadian Journal of Physiology and Pharmacology*, 62:224–228, 1984. [161]

[79] J. Whitehead. Sample size calculations for ordered categorical data. *Statistics in Medicine*, 12:2257–2271, 1993. [67]

[80] C. F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, 14(4):1261–1350, 1986. [105]

[81] J. Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93:120–131, 1998. [9]

[82] F. W. Young, Y. Takane, and J. de Leeuw. The principal components of mixed measurement level multivariate data: An alternating least squares method with optimal scaling features. *Psychometrika*, 43:279–281, 1978. [75]

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`.