REGRESSION MODELING STRATEGIES

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

VANDERBILT UNIVERSITY 3–6 MARCH 2015

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.

See the excellent text *Clinical Prediction Models* by Steyerberg[156].

Chapter 1

Introduction

1.1 Hypothesis Testing, Estimation, and Prediction

Even when only testing H_0 a model based approach has advantages:

- Permutation and rank tests not as useful for estimation
- Cannot readily be extended to cluster sampling or repeated measurements
- Models generalize tests
  - 2-sample t-test, ANOVA $\rightarrow$ multiple linear regression
  - Wilcoxon, Kruskal-Wallis, Spearman $\rightarrow$ proportional odds ordinal logistic model

  - log-rank $\rightarrow$ Cox

- Models not only allow for multiplicity adjustment but for shrinkage of estimates
  - Statisticians comfortable with P-value adjustment but fail to recognize that the difference between the most different treatments is badly biased

Statistical estimation is usually model-based

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

1.2 Examples of Uses of Predictive Multivariable Modeling

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

  - Use e.g. difference in two predicted probabilities

- Cost-effectiveness ratios
  - incremental cost / incremental *ABSOLUTE* benefit
  - most studies use avg. cost difference / avg. benefit, which may apply to no one

1.3 Misunderstandings about Prediction vs. Classification

- Many analysts desire to develop "classifiers" instead of predictions
- Suppose that
  1. response variable is binary
  2. the two levels represent a sharp dichotomy with no gray zone (e.g., complete success vs. total failure with no possibility of a partial success)
  3. one is forced to assign (classify) future observations to only these two choices

4. the cost of misclassification is the same for every future observation, and the ratio of the cost of a false positive to the cost of a false negative equals the (often hidden) ratio implied by the analyst's classification rule

- Then classification is **still suboptimal** for driving the development of a predictive instrument as well as for hypothesis testing and estimation
- Far better is to use the full information in the data to develope a probability model, then develop classification rules on the basis of estimated probabilities
  - $\uparrow$ power, $\uparrow$ precision, $\uparrow$ decision making
- Classification is more problematic if response variable is ordinal or continuous or the groups are not truly distinct (e.g., disease or no disease when severity of disease is on a continuum); dichotomizing it up front for the analysis is not appropriate

- *minimum* loss of information (when dichotomization is at the median) is large
- may require the sample size to increase many–fold to compensate for loss of information[59]

- Two-group classification represents artificial forced choice
  - best option may be "no choice, get more data"
- Unlike prediction (e.g., of absolute risk), classification implicitly uses utility (loss; cost of false positive or false negative) functions
- Hidden problems:
  - Utility function depends on variables not collected (subjects' preferences) that are available only at the decision point
  - Assumes every subject has the same utility function
  - Assumes this function coincides with the analyst's

- Formal decision analysis uses
  - optimum predictions using all available data
  - subject-specific utilities, which are often based on variables not predictive of the outcome
- ROC analysis is misleading except for the special case of mass one-time group decision making with unknowable utilities[a]

See[20,24,56,63,67,174].

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

In summary:

- Classification is a forced choice — a decision.

[a] To make an optimal decision you need to know all relevant data about an individual (used to estimate the probability of an outcome), and the utility (cost, loss function) of making each decision. Sensitivity and specificity do not provide this information. For example, if one estimated that the probability of a disease given age, sex, and symptoms is 0.1 and the "cost"of a false positive equaled the "cost" of a false negative, one would act as if the person does not have the disease. Given other utilities, one would make different decisions. If the utilities are unknown, one gives the best estimate of the probability of the outcome to the decision maker and let her incorporate her own unspoken utilities in making an optimum decision for her.

Besides the fact that cutoffs do not apply to individuals, only to groups, individual decision making does not utilize sensitivity and specificity. For an individual we can compute $\text{Prob}(Y = 1|X = x)$; we don't care about $\text{Prob}(Y = 1|X > c)$, and an individual having $X = x$ would be quite puzzled if she were given $\text{Prob}(X > c|\text{future unknown Y})$ when she already knows $X = x$ so X is no longer a random variable.

Even when group decision making is needed, sensitivity and specificity can be bypassed. For mass marketing, for example, one can rank order individuals by the estimated probability of buying the product, to create a lift curve. This is then used to target the k most likely buyers where k is chosen to meet total program cost constraints.

- Decisions require knowledge of the cost or utility of making an incorrect decision.
- Predictions are made without knowledge of utilities.
- A prediction can lead to better decisions than classification. For example suppose that one has an estimate of the risk of an event, $\hat{P}$. One might make a decision if $\hat{P} < 0.10$ or $\hat{P} > 0.90$ in some situations, even without knowledge of utilities. If on the other hand $\hat{P} = 0.6$ or the confidence interval for P is wide, one might
  - make no decision and instead opt to collect more data
  - make a tentative decision that is revisited later
  - make a decision using other considerations such as the infusion of new resources that allow targeting a larger number of potential customers in a marketing campaign

The Dichotomizing Motorist

- The speed limit is 60.
- I am going faster than the speed limit.
- Will I be caught?

An answer by a dichotomizer:

- Are you going faster than 70?

An answer from a better dichotomizer:

- If you are among other cars, are you going faster than 73?
- If you are exposed are your going faster than 67?

Better:

- How fast are you going and are you exposed?

Analogy to most medical diagnosis research in which +/- diagnosis is a false dichotomy of an underlying disease severity:

- The speed limit is moderately high.
- I am going fairly fast.
- Will I be caught?

1.4 Planning for Modeling

- Chance that predictive model will be used[141]
- Response definition, follow-up
- Variable definitions
- Observer variability
- Missing data
- Preference for continuous variables
- Subjects
- Sites

What can keep a sample of data from being appropriate for modeling:

1. Most important predictor or response variables not collected

2. Subjects in the dataset are ill-defined or not representative of the population to which inferences are needed
3. Data collection sites do not represent the population of sites
4. Key variables missing in large numbers of subjects
5. Data not missing at random
6. No operational definitions for key variables and/or measurement errors severe
7. No observer variability studies done

What else can go wrong in modeling?

1. The process generating the data is not stable.
2. The model is misspecified with regard to non-linearities or interactions, or there are predictors missing.
3. The model is misspecified in terms of the transformation of the response variable or

the model's distributional assumptions.

4. The model contains discontinuities (e.g., by categorizing continuous predictors or fitting regression shapes with sudden changes) that can be gamed by users.
5. Correlations among subjects are not specified, or the correlation structure is misspecified, resulting in inefficient parameter estimates and overconfident inference.
6. The model is overfitted, resulting in predictions that are too extreme or positive associations that are false.
7. The user of the model relies on predictions obtained by extrapolating to combinations of predictor values well outside the range of the dataset used to develop the model.
8. Accurate and discriminating predictions can lead to behavior changes that make future predictions inaccurate.

Iezzoni[92] lists these dimensions to capture, for patient outcome studies:

1. age
2. sex
3. acute clinical stability
4. principal diagnosis
5. severity of principal diagnosis
6. extent and severity of comorbidities
7. physical functional status
8. psychological, cognitive, and psychosocial functioning
9. cultural, ethnic, and socioeconomic attributes and behaviors
10. health status and quality of life
11. patient attitudes and preferences for outcomes

General aspects to capture in the predictors:

1. baseline measurement of response variable
2. current status
3. trajectory as of time zero, or past levels of a key variable
4. variables explaining much of the variation in the response
5. more subtle predictors whose distributions strongly differ between levels of the key variable of interest in an observational study

1.5 Choice of the Model

- In biostatistics and epidemiology and most other areas we usually choose model empirically
- Model must use data efficiently
- Should model overall structure (e.g., acute vs. chronic)

- Robust models are better
- Should have correct mathematical structure (e.g., constraints on probabilities)

1.6 Model uncertainty / Data-driven Model Specification

- Standard errors, C.L., P-values, R^2 wrong if computed as if the model pre-specified
- Stepwise variable selection is widely used and abused
- Bootstrap can be used to repeat all analysis steps to properly penalize variances, etc.
- Ye[192]: "generalized degrees of freedom" (GDF) for any "data mining" or model selection procedure based on least squares
  - Example: 20 candidate predictors, $n = 22$, forward stepwise, best 5-variable model: GDF=14.1
  - Example: CART, 10 candidate predictors, $n = 100$, 19 nodes: GDF=76

- See[118] for an approach involving adding noise to Y to improve variable selection

Chapter 2

General Aspects of Fitting Regression Models

2.1 Notation for Multivariable Regression Models

- Weighted sum of a set of independent or predictor variables
- Interpret parameters and state assumptions by linearizing model with respect to regression coefficients
- Analysis of variance setups, interaction effects, nonlinear effects
- Examining the 2 regression assumptions

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

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

2.2 Model Formulations

General regression model

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

General linear regression model

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

Examples

$$\begin{aligned}
C(Y|X) = \quad & E(Y|X) = \quad X\beta, \\
Y|X \quad & \sim N(X\beta, \sigma^2) \\
C(Y|X) = \ & \text{Prob}\{Y = 1|X\} = \ (1 + \exp(-X\beta))^{-1}
\end{aligned}$$

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[a].

$$\begin{aligned} C'(Y|X) &= X\beta = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p \\ \beta_j &= C'(Y|X_1, X_2, \ldots, X_j + 1, \ldots, X_p) \\ &- C'(Y|X_1, X_2, \ldots, X_j, \ldots, X_p) \end{aligned}$$

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

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

[a]Note that it is not necessary to "hold constant" all other variables to be able to interpret the effect of one predictor. It is sufficient to hold constant the weighted sum of all the variables other than X_j. And in many cases it is not physically possible to hold other variables constant while varying one, e.g., when a model contains X and X^2 (David Hoaglin, personal communication).

$$C(Y|T=M) = \beta_0 + \beta_3.$$

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

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:

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

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

$$- \; [\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2]$$
$$= \; \beta_1 + \beta_3 X_2.$$

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[b].

Model assumes

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

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

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

Interpretations of parameters:

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

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 × 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	$H_0 : \beta_1 = \beta_3 = 0$ $H_a : \beta_1 \neq 0 \text{ or } \beta_3 \neq 0$
sex is not associated with Y sex is associated with Y sex is associated with Y for some value of age	$H_0 : \beta_2 = \beta_3 = 0$ $H_a : \beta_2 \neq 0 \text{ or } \beta_3 \neq 0$
Neither age nor sex is associated with Y Either age or sex is associated with Y	$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ $H_a : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or } \beta_3 \neq 0$

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

2.4 Review of Composite (Chunk) Tests

- In the model

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

  we may want to jointly test the association between all body measurements and response, holding `age` and `sex` constant.

- This 3 d.f. test may be obtained two ways:
  - Remove the 3 variables and compute the change in SSR or SSE
  - Test $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ using matrix algebra (e.g., `anova(fit, weight, waist, tricep)` if `fit` is a fit object created by the R `rms` package)

2.5 Relaxing Linearity Assumption for Continuous Predictors

2.5.1 Avoiding Categorization

- Relationships seldom linear except when predicting one variable from itself measured earlier
- Categorizing continuous predictors into intervals is a disaster; see references
  [2,5,12,26,57,87,106,140,143,161]
  [3,59,65,89,120,124,147,177]
- Some problems caused by this approach:
  1. Estimated values have reduced precision, and associated tests have reduced power
  2. Categorization assumes relationship between predictor and response is flat within intervals; far less reasonable than a linearity assumption in most cases
  3. To make a continuous predictor be more accurately modeled when categorization is used, multiple intervals are required

4. Because of sample size limitations in the very low and very high range of the variable, the outer intervals (e.g., outer quintiles) will be wide, resulting in significant heterogeneity of subjects within those intervals, and residual confounding
5. Categorization assumes that there is a discontinuity in response as interval boundaries are crossed. Other than the effect of time (e.g., an instant stock price drop after bad news), there are very few examples in which such discontinuities have been shown to exist.
6. Categorization only seems to yield interpretable estimates. E.g. odds ratio for stroke for persons with a systolic blood pressure > 160 mmHg compared to persons with a blood pressure ≤ 160 mmHg $\rightarrow$ interpretation of OR depends on distribution of blood pressures in the sample (the proportion of subjects > 170, > 180, etc.). If blood

pressure is modeled as a continuous variable (e.g., using a regression spline, quadratic, or linear effect) one can estimate the ratio of odds for *exact* settings of the predictor, e.g., the odds ratio for 200 mmHg compared to 120 mmHg.

7. Categorization does not condition on full information. When, for example, the risk of stroke is being assessed for a new subject with a known blood pressure (say 162 mmH the subject does not report to her physician "my blood pressure exceeds 160" but rather reports 162 mmHg. The risk for this subject will be much lower than that of a subject with a blood pressure of 200 mmHg.
8. If cutpoints are determined in a way that is not blinded to the response variable, calculation of P-values and confidence intervals requires special simulation techniques; ordinary inferential methods are completely

invalid. E.g.: cutpoints chosen by trial and error utilizing Y, even informally $\rightarrow$ P-values too small and CLs not accurate[c].

9. Categorization not blinded to $Y \rightarrow$ biased effect estimates[5,147]
10. "Optimal" cutpoints do not replicate over studies. Hollander *et al.*[89] state that "...the optimal cutpoint approach has disadvantages. One of these is that in almost every study where this method is applied, another cutpoint will emerge. This makes comparisons across studies extremely difficult or even impossible. Altman et al. point out this problem for studies of the prognostic relevance of the S-phase fraction in breast cancer published in the literature. They identified 19 different cutpoints used in the literature; some of them were solely used because they emerged as the 'optimal' cutpoint in a specific data set. In a meta-analysis on the relationship between cathepsin-D content and disease-free survival in node-negative breast cancer patients, 12 studies were in included with 12 different cutpoints ... Interestingly, neither cathepsin-D nor the S-phase fraction are recommended to be used as prognostic markers in breast cancer in the re-

[c] If a cutpoint is chosen that minimizes the P-value and the resulting P-value is 0.05, the true type I error can easily be above 0.5[89].

cent update of the American Society of Clinical Oncology." Giannoni *et al.*[65] demonstrated that many claimed "optimal cutpoints" are just the observed median values in the sample, which happens to optimize statistical power for detecting a separation in outcomes.

11. Disagreements in cutpoints (which are bound to happen whenever one searches for things that do not exist) cause severe interpretation problems. One study may provide an odds ratio for comparing body mass index (BMI) > 30 with BMI ≤ 30, another for comparing BMI > 28 with BMI ≤ 28. Neither of these has a good definition and the two estimates are not comparable.
12. Cutpoints are arbitrary and manipulatable; cutpoints can be found that can result in both positive and negative associations[177].
13. If a confounder is adjusted for by categorization, there will be residual confounding that can be explained away by inclusion of the continuous form of the predictor in the model in addition to the categories.

- To summarize: The use of a (single) cutpoint c makes many assumptions, including:
  1. Relationship between X and Y is discontinuous at $X = c$ and only $X = c$
  2. c is correctly found as *the* cutpoint
  3. X vs. Y is flat to the left of c
  4. X vs. Y is flat to the right of c
  5. The choice of c does not depend on the values of other predictors

Interactive demonstration of power loss of categorization vs. straight line and quadratic fits in OLS, with varying degree of nonlinearity and noise added to X: `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/BioMod/catgNoise.r` (must run in `RStudio`)

2.5.2 Simple Nonlinear Terms

$$C(Y|X_1) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2.$$

- H_0 : model is linear in X_1 vs. H_a : model is quadratic in $X_1 \equiv H_0 : \beta_2 = 0$.
- Test of linearity may be powerful if true model is not extremely non-parabolic
- Predictions not accurate in general as many phenomena are non-quadratic
- Can get more flexible fits by adding powers higher than 2
- But polynomials do not adequately fit logarithmic functions or "threshold" effects, and have unwanted peaks and valleys.

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

Draftsman'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).

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

where

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

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

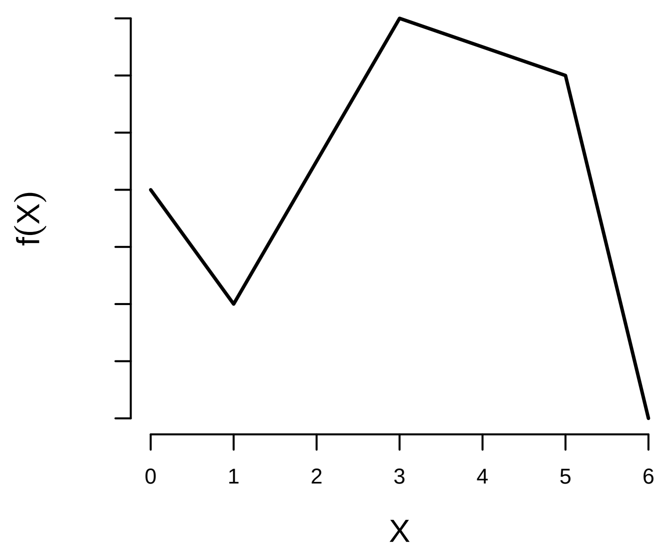


Figure 2.1: A linear spline function with knots at $a = 1, b = 3, c = 5$.

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

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

$$
\begin{aligned}
X_1 = X \quad X_2 = (X - a)_+ \\
X_3 = (X - b)_+ \quad X_4 = (X - c)_+.
\end{aligned}
$$

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

2.5.4 Cubic Spline Functions

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

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

$$
\begin{array}{cc}
X_1 = X & X_2 = X^2 \\
X_3 = X^3 & X_4 = (X-a)_+^3 \\
X_5 = (X-b)_+^3 & X_6 = (X-c)_+^3.
\end{array}
$$

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

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

2.5.5 Restricted Cubic Splines

Stone and Koo[160]: cubic splines poorly behaved in tails. Constrain function to be linear in tails.
$k+3 \rightarrow k-1$ parameters[48].

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[48]

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

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 $X \geq t_k$.

For numerical behavior and to put all basis functions for X on the same scale, R `Hmisc` and `rms` package functions by default divide the terms above by $\tau = (t_k - t_1)^2$.

```
require(Hmisc)
```

```
x ← rcspline.eval(seq(0,1,.01),
                  knots=seq(.05,.95,length=5), inclx=T)
xm ← x
xm[xm > .0106] ← NA
matplot(x[,1], xm, type="l", ylim=c(0,.01),
        xlab=expression(X), ylab='', lty=1)
matplot(x[,1], x,  type="l",
        xlab=expression(X), ylab='', lty=1)
```

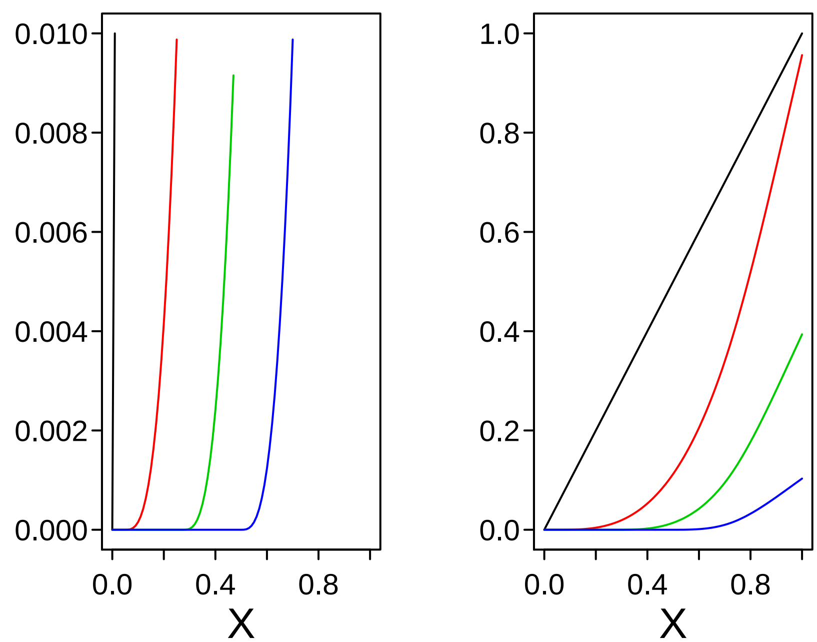


Figure 2.2: Restricted cubic spline component variables for $k = 5$ and knots at $X = .05, .275, .5, .725$, and .95. Nonlinear basis functions are scaled by au. The left panel is a y–magnification of the right panel. Fitted functions such as those in Figure 2.3 will be linear combinations of these basis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

```
x ← seq(0, 1, length=300)
for(nk in 3:6) {
  set.seed(nk)
  knots ← seq(.05, .95, length=nk)
  xx ← rcspline.eval(x, knots=knots, inclx=T)
  for(i in 1 : (nk − 1))
    xx[,i] ← (xx[,i] − min(xx[,i])) /
             (max(xx[,i]) − min(xx[,i]))
```

```
  for(i in 1 : 20) {
    beta  ← 2*runif(nk-1) - 1
    xbeta ← xx %*% beta + 2 * runif(1) - 1
    xbeta ← (xbeta - min(xbeta)) /
             (max(xbeta) - min(xbeta))
    if(i == 1) {
      plot(x, xbeta, type="l", lty=1,
           xlab=expression(X), ylab='', bty="l")
      title(sub=paste(nk,"knots"), adj=0, cex=.75)
      for(j in 1 : nk)
        arrows(knots[j], .04, knots[j], -.03,
               angle=20, length=.07, lwd=1.5)
    }
    else lines(x, xbeta, col=i)
  }
}
```

Once $\beta_0, \ldots, \beta_{k-1}$ are estimated, the restricted cubic spline can be restated in the form

$$\begin{aligned} f(X) = {} & \beta_0 + \beta_1 X + \beta_2 (X - t_1)_+^3 + \beta_3 (X - t_2)_+^3 \\ & + \ldots + \beta_{k+1} (X - t_k)_+^3 \end{aligned}$$

by dividing $\beta_2, \ldots, \beta_{k-1}$ by τ and 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

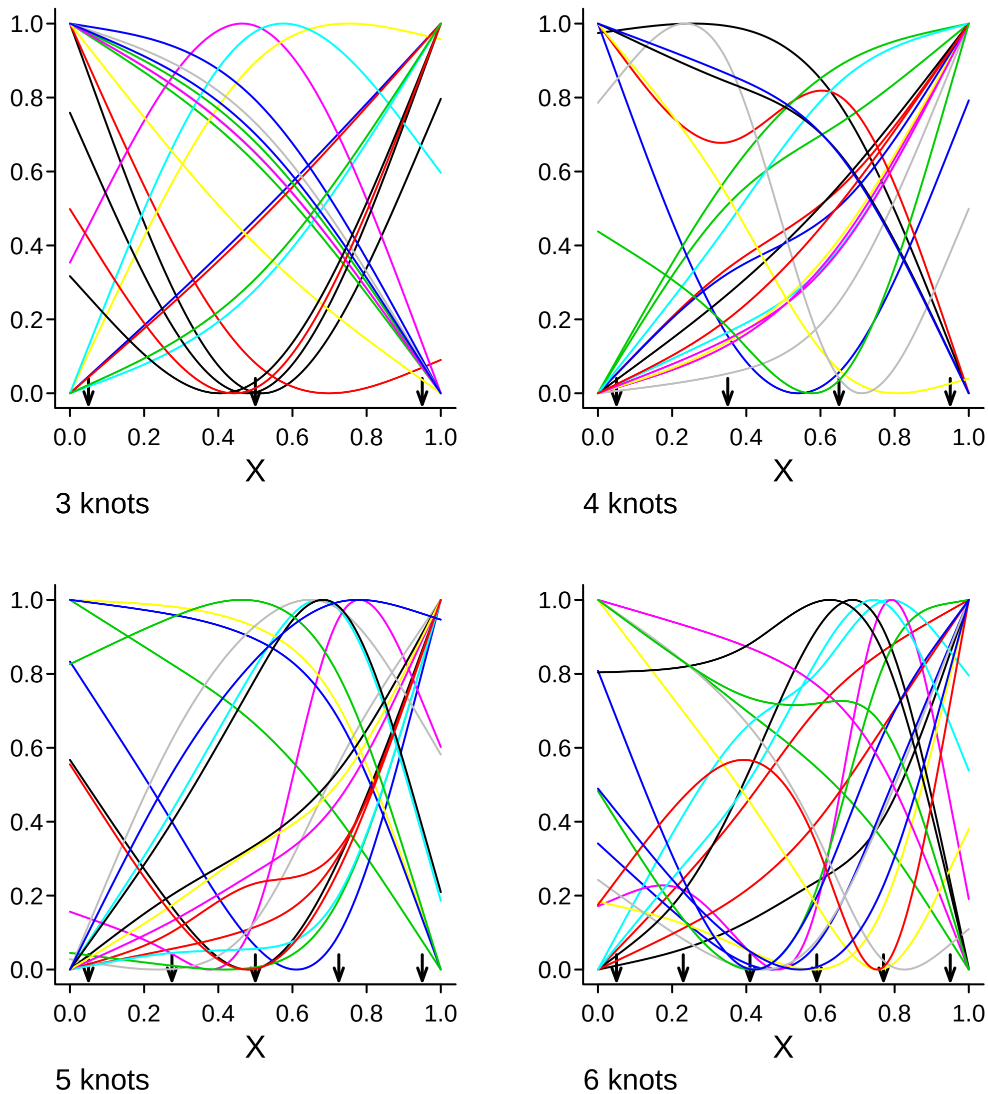


Figure 2.3: Some typical restricted cubic spline functions for $k = 3, 4, 5, 6$. The y–axis is $X\beta$. Arrows indicate knots. These curves were derived by randomly choosing values of β subject to standard deviations of fitted functions being normalized.

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

2.5.6 Choosing Number and Position of Knots

- Knots are specified in advance in regression splines
- Locations not important in most situations[52,159]
- 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[160].

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)[7, 170] to choose k
- This chooses k to maximize model likelihood ratio $\chi^2 - 2k$.

See[69] for a comparison of restricted cubic splines, fractional polynomials, and penalized splines.

2.5.7 Nonparametric Regression

- Estimate tendency (mean or median) of Y as a function of X
- Few assumptions
- Especially handy when there is a single X
- Plotted trend line may be the final result of the analysis

- Simplest smoother: moving average

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

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

  - overlap OK
  - problem in estimating $E(Y)$ at outer X-values
  - estimates very sensitive to bin width
- Moving linear regression far superior to moving avg. (moving flat line)
- Cleveland's[35] 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[d]
- points near extremes of interval receive almost no weight
- loess works much better at extremes of X than moving avg.
- provides an estimate at each observed X; other estimates obtained by linear interpolation
- outlier rejection algorithm built-in

- loess works great for binary Y — just turn off outlier detection
- Other popular smoother: Friedman's "super smoother"
- For loess or supsmu amount of smoothing can be controlled by analyst
- Another alternative: smoothing splines[e]

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

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

- Smoothers are very useful for estimating trends in residual plots

2.5.8 Advantages of Regression Splines over Other Methods

Regression splines have several advantages[82]:

- Parametric splines can be fitted using any existing regression program
- Regression coefficients estimated using standard techniques (ML or least squares), formal tests of no overall association, linearity, and additivity, confidence limits for the estimated regression function are derived by standard theory.
- The fitted function directly estimates transformation predictor should receive to yield linearity in $C(Y|X)$.
- Even when a simple transformation is obvious, spline function can be used to represent the predictor in the final model (and the

d.f. will be correct). Nonparametric methods do not yield a prediction equation.

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

2.6 Recursive Partitioning: Tree-Based Models

Breiman, Friedman, Olshen, and Stone[23]: 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[37]; missing value imputation)

See[9].

2.7 New Directions in Predictive Modeling

The approaches recommended in this course are

- fitting fully pre-specified models without deletion of "insignificant" predictors
- using data reduction methods (masked to Y) to reduce the dimensionality of the predictors and then fitting the number of parameters the data's information content can support

- use shrinkage (penalized estimation) to fit a large model without worrying about the sample size.

The data reduction approach can yield very interpretable, stable models, but there are many decisions to be made when using a two-stage (reduction/model fitting) approach, Newer approaches are evolving, including the following. These new approach handle continuous predictors well, unlike recursive partitioning.

- lasso (shrinkage using L1 norm favoring zero regression coefficients)[157,163]
- elastic net (combination of L1 and L2 norms that handles the $p > n$ case better than the lasso)[197]
- adaptive lasso[179,195]
- more flexible lasso to differentially penalize for variable selection and for regression coefficient estimation[139]
- group lasso to force selection of all or none

of a group of related variables (e.g., dummy variables representing a polytomous predictor)

- group lasso-like procedures that also allow for variables within a group to be removed[180]
- sparse-group lasso using L1 and L2 norms to achieve spareness on groups and within groups of variables[150]
- adaptive group lasso (Wang & Leng)
- Breiman's nonnegative garrote[191]
- "preconditioning", i.e., model simplification after developing a "black box" predictive model[128]
- sparse principal components analysis to achieve parsimony in data reduction[110,111,188,196]
- bagging, boosting, and random forests[84]

One problem prevents most of these methods from being ready for everyday use: they require scaling predictors before fitting the model. When a predictor is represented by nonlinear basis functions, the scaling recommendations in the literature are not sensible. There are also computational issues and difficulties obtaining hypothesis tests and confidence intervals.

When data reduction is not required, generalized additive models[85,189] should also be considered.

2.8 Multiple Degree of Freedom Tests of Association

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

$H_0 : \beta_2 = \beta_3 = 0$ with 2 d.f. to assess association between X_2 and outcome.

In the 5-knot restricted cubic spline model

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

$H_0 : \beta_1 = \ldots = \beta_4 = 0$

- Test of association: 4 d.f.

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

Grambsch and O'Brien[70] 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 nonlinearity **entertained**
- Considered ordinary regression model $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$
- Two ways to test association between X and Y
- Fit quadratic model and test for linearity ($H_0 : \beta_2 = 0$)
- F-test for linearity significant at $\alpha = 0.05$ level $\rightarrow$ report as the final test of association the 2 d.f. F test of $H_0 : \beta_1 = \beta_2 = 0$

- If the test of linearity insignificant, refit without the quadratic term and final test of association is 1 d.f. test, $H_0: \beta_1 = 0 | \beta_2 = 0$
- Showed that type I error $> \alpha$
- Fairly accurate P-value obtained by instead testing against F with 2 d.f. even at second stage
- Cause: are retaining the most significant part of F
- **BUT** if test against 2 d.f. can only lose power when compared with original F for testing both βs
- SSR from quadratic model $> SSR$ from linear model

2.9 Assessment of Model Fit

2.9.1 Regression Assumptions

The general linear regression model is $C(Y|X) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$. Verify linearity and additivity. Special case:

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

where X_1 is binary and X_2 is continuous.

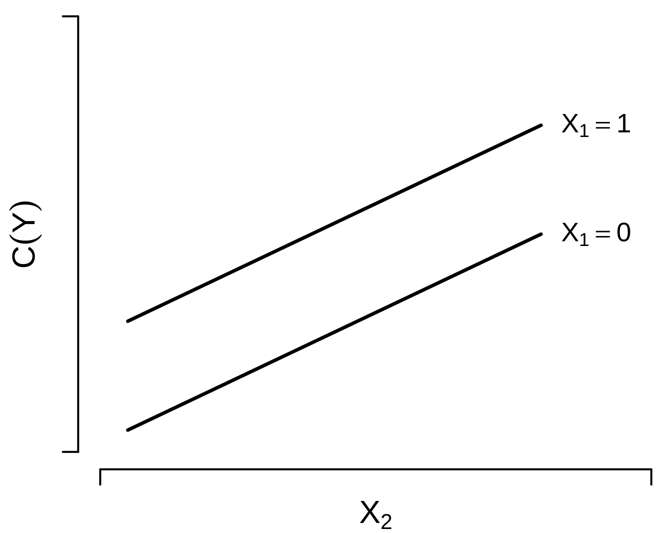


Figure 2.4: Regression assumptions for one binary and one continuous predictor

Methods for checking fit:

1. Fit simple linear additive model and check

examine residual plots for patterns

- For OLS: box plots of e stratified by X_1, scatterplots of e vs. X_2 and $\hat{Y}$, with trend curves (want flat central tendency, constant variability)
- For normality, `qqnorm` plots of overall and stratified residuals

Advantage: Simplicity

Disadvantages:

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

2. Scatterplot of Y vs. X_2 using different symbols according to values of X_1

   Advantages: Simplicity, can see interaction

Disadvantages:

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

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

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

   Disadvantages:

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

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

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

Disadvantages:

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

5. Fit flexible nonlinear parametric model

   Advantages:

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

   Disadvantages:

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

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

Restricted cubic spline works well for method

5.

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

where

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

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

- Plot $\hat{f}(X_2)$ vs. X_2
- n large $\rightarrow$ can fit separate functions by X_1
- Test of linearity: $H_0 : \beta_3 = \beta_4 = 0$
- Few good reasons to do the test other than to demonstrate that linearity is not a good default assumption
- Nonlinear $\rightarrow$ use transformation suggested by spline fit or keep spline terms
- Tentative transformation $g(X_2) \rightarrow$ check adequacy by expanding $g(X_2)$ in spline function and testing linearity
- Can find transformations by plotting $g(X_2)$

vs. $\hat{f}(X_2)$ for variety of g

- Multiple continuous predictors $\rightarrow$ expand each using spline
- Example: assess linearity of X_2, X_3

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

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

2.9.2 Modeling and Testing Complex Interactions

X_1 binary or linear, X_2 continuous:

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

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

- 2 continuous variables: could transform separately and form simple product

- **But** transformations depend on whether interaction terms adjusted for, so it is usually not possible to estimate transformations and interaction effects other than simultaneously
- Compromise: 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 $aX_1 + bX_2 + cX_1X_2$ in corners
- Uses all $(k-1)^2$ cross-products of restricted cubic spline terms
- See Gray [71, 72, Section 3.2] for penalized splines allowing control of effective degrees of freedom. See Berhane *et al.*[16] for a good discussion of tensor splines.

Other issues:

- Y non-censored (especially continuous) $\rightarrow$ multi-dimensional scatterplot smoother[29]
- Interactions of order > 2: more trouble
- 2-way interactions among p predictors: pooled tests
- p tests each with $p-1$ d.f.

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

- Treatment $\times$ severity of disease being treated
- Age $\times$ risk factors
- Age $\times$ type of disease
- Measurement $\times$ state of a subject during measurement
- Race $\times$ disease
- Calendar time $\times$ treatment
- Quality $\times$ quantity of a symptom

2.9.3 Fitting Ordinal Predictors

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

2.9.4 Distributional Assumptions

- Some models (e.g., logistic): all assumptions in $C(Y|X) = X\beta$ (implicitly assuming no omitted variables!)
- Linear regression: $Y \sim X\beta + \epsilon, \epsilon \sim n(0, \sigma^2)$
- Examine distribution of residuals
- Some models (Weibull, Cox[41]):
  $C(Y|X) = C(Y = y|X) = d(y) + X\beta$
  $C =$ 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)[a]
- Informative missing
  (non-ignorable non-response)

See[1,27,50,77,185] for an introduction to missing data and imputation concepts.

3.2 Prelude to Modeling

- Quantify extent of missing data

[a] "Although missing at random (MAR) is a non-testable assumption, it has been pointed out in the literature that we can get very close to MAR if we include enough variables in the imputation models"[77].

- 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[b]
- Y=time to event, follow-up curtailed: covered under survival analysis[c]
- Often discard observations with completely missing Y but sometimes wasteful[d]
- Characterize missings in Y before dropping obs.

[b]Twist *et al.*[165] found instability in using multiple imputation of longitudinal data, and advantages of using instead full likelihood models.

[c]White and Royston[184] provide a method for multiply imputing missing covariate values using censored survival time data.

[d]Y is so valuable that if one is only missing a Y value, imputation is not worthwhile, and imputation of Y is not advised if MCAR or MAR.

3.4 Problems With Simple Alternatives to Imputation

Deletion of records—

- Badly biases parameter estimates when the probability of a case being incomplete is related to Y and not just X[115].
- Deletion because of a subset of X being missing always results in inefficient estimates
- Deletion of records with missing Y can result in biases[42] but is the preferred approach under MCAR[e]
- However von Hippel[176] found advantages to a "use all variables to impute all variables then drop observations with missing Y" approach
- Only discard obs. when
  - MCAR can be justified
  - Rarely missing predictor of overriding importance that can't be imputed from other

[e] Multiple imputation of Y in that case does not improve the analysis and assumes the imputation model is correct.

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
- See[99]

Adding extra categories of categorical predictors—

- Including missing data but adding a category 'missing' causes serious biases[1,94,166]
- Problem acute when values missing because subject too sick
- Difficult to interpret
- Fails even under MCAR[1,50,94,99,168]
- May be OK if values are "missing" because of "not applicable"[f]

[f] E.g. you have a measure of marital happiness, dichotomized as high or low, but your sample contains some unmarried people. OK to have a 3-category variable with values high, low, and unmarried—Paul Allison, IMPUTE list, 4Jul09.

Likewise, serious problems are caused by setting missing continuous predictors to a constant (e.g., zero) and adding an indicator variable to try to estimate the effect of missing values.

Two examples from Donder *et al.*[50] using binary logistic regression, $N = 500$.

Results of 1000 Simulations With $\beta_1 = 1.0$ with MAR and Two Types of Imputation

Imputation Method	$\hat{\beta}_1$	S.E.	Coverage of 0.90 C.I.
Single	0.989	0.09	0.64
Multiple	0.989	0.14	0.90

Now consider a simulation with $\beta_1 = 1, \beta_2 = 0$, X_2 correlated with $X_1 (r = 0.75)$ but redundant in predicting Y, use missingness indicator when X_1 is MCAR in 0.4 of 500 subjects. This is also compared with grand mean fill-in imputation.

Results of 1000 Simulations Adding a Third Predictor Indicating Missing for X_1

Imputation Method	$\hat{\beta}_1$	$\hat{\beta}_2$
Indicator	0.55	0.51
Overall mean	0.55	

In the incomplete observations the constant X_1 is uncorrelated with X_2.

3.5 Strategies for Developing an Imputation Model

The goal of imputation is to preserve the information and meaning of the non-missing data.

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[11, 77]
- 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

- `transcan` and `aregImpute` use the following for fitting imputation models:
  1. initialize `NA`s to median (mode for categoricals)
  2. expand all categorical predictors using dummy variables
  3. expand all continuous predictors using restricted cubic splines
  4. optionally optimally transform the variable being predicted by expanding it with restricted cubic splines and using the first canonical variate (multivariate regression) as the optimum transformation (maximizing R^2)
  5. one-dimensional scoring of categorical variables being predicted using canonical variates on dummy variables representing the categories (Fisher's optimum scoring algorithm); when imputing categories, solve for which category yields a score that is

closest to the predicted score

- `aregImpute` and `transcan` work with `fit.mult.impute` to make final analysis of response variable relatively easy
- Predictive mean matching[115]: replace missing value with observed value of subject having closest predicted value to the predicted value of the subject with the `NA`. Key considerations are how to
  1. model the target when it is not `NA`
  2. match donors on predicted values
  3. avoid overuse of "good" donors to disallow excessive ties in imputed data
  4. account for all uncertainties
- Predictive model for each target uses any outcomes, all predictors in the final model other than the target, plus auxiliary variables not in the outcome model
- No distributional assumptions
- Predicted values need only be monotonically

related to real predictive values

- 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
- [185] discusses an alternative method based on choosing a donor observation at random

from the q closest matches ($q = 3$, for example)

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
- $\hat{X}_j = Z\hat{\theta}, Z = X_{\bar{j}}$ plus possibly auxiliary variables that precede X_j in the causal chain that are not intended to be in the outcome model.
  Cannot include Y in Z without adding random errors to imputed values as done with multiple imputation (would steal info from Y)
- Recursive partitioning can sometimes be helpful for nonparametrically estimating conditional

means

3.7 Multiple Imputation

- Single imputation could use a random draw from the conditional distribution for an individual
  $\hat{X}_j = Z\hat{\theta} + \hat{\epsilon}, Z = [X\bar{j}, Y]$ plus auxiliary variables
  $\hat{\epsilon} = n(0, \hat{\sigma})$ or a random draw from the calculated residuals
  - bootstrap
  - approximate Bayesian bootstrap[77,144]: 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

- Note that multiple imputation can and should use the response variable for imputing predictors[123]
- `aregImpute` algorithm[123]
  - 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 flexible parametric additive regression models to impute
- There is an option to allow target variables to be optimally transformed, even non-monotonically (but this can overfit)
- By default uses predictive mean matching for imputation; no residuals required (can also do more parametric regression imputation)
- By default uses weighted PMM; many other matching options
- Uses by default van Buuren's "Type 1" matching [27, Section 3.4.2] to capture the right amount of uncertainty by computing predicted values for missing values using a regression fit on the bootstrap sample, and finding donor observations by matching those

predictions to predictions from potential donors using the regression fit from the original sample of complete observations

- 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 R `MICE` approach
- Example:

```
a ← aregImpute(∼ age + sex + bp + death + heart.attack.before.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[11] 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. Barnes *et al.*[10] have a good overview of imputation methods and a comparison of bias and confidence interval coverage for the methods when applied to longitudinal data with a small number of subjects. Horton and Kleinman[90] have a good review of several software packages for dealing with missing data, and a comparison of them with `aregImpute`. Harel and Zhou[77] provide a nice overview of multiple imputation and discuss some of the available software. White and Carlin[183] studied bias of multiple imputation vs. complete-case analysis. White *et al.*[185] provide much practical guidance.

Caution: Methods can generate imputations having very reasonable distributions but still not having the property that final response model regression coefficients have nominal confidence interval coverage. It is worth checking that imputations generate the correct collinearities among covariates.

- With `MICE` and `aregImpute` we are using the chained equation approach[185]
- Chained equations handles a wide variety of target variables to be imputed and allows for multiple variables to be missing on the same subject
- Iterative process cycles through all target variables to impute all missing values[167]
- Does not attempt to use the full Bayesian multivariate model for all target variables, mak-

ing it more flexible and easy to use

- Possible to create improper imputations, e.g., imputing conflicting values for different target variables
- However, simulation studies[167] demonstrate very good performance of imputation based on chained equations

3.8 Diagnostics

- MCAR can be partially assessed by comparing distribution of non-missing Y for those subjects with complete X vs. those subjects having incomplete X[115]
- Yucel and Zaslavsky[194] (see also[86])
- Interested in reasonableness of imputed values for a sometimes-missing predictor X_j
- Duplicate entire dataset
- In the duplicated observations set all non-

missing values of X_j to missing; let w denote this set of observations set to missing

- Develop imputed values for the missing values of X_j
- In the observations in w compare the distribution of imputed X_j to the original values of X_j

3.9 Summary and Rough Guidelines

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	

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

$f < 0.03$**:** It doesn't matter very much how you impute missings or whether you adjust variance of regression coefficient estimates for having imputed data in this case. For continuous variables imputing missings with the median non-missing value is adequate; for categorical predictors the most frequent category can be used. Complete case analysis is also an option here. Multiple imputation may be needed to check that the simple approach "worked."

$f \geq 0.03$**:** Use multiple imputation with number of imputations[g] equal to $\max(5, 100f)$. Fewer imputations may be possible with very large sample sizes. Type 1 predictive mean matching is usually preferred, with weighted selection of donors. Account for imputation in estimating the covariance matrix for final parameter estimates. Use the t distribution instead of the Gaussian distribution for tests

[g] White *et al.*[185] recommend choosing M so that the key inferential statistics are very reproducible should the imputation analysis be repeated. They suggest the use of $100f$ imputations. See also [27, Section 2.7].

and confidence intervals, if possible, using the estimated d.f. for the parameter estimates.

Multiple predictors frequently missing: More imputations may be required. Perform a "sensitivity to order" analysis by creating multiple imputations using different orderings of sometimes missing variables. It may be beneficial to initially sort variables so that the one with the most `NA`s will be imputed first.

Reason for missings more important than number of missing values.

Extreme amount of missing data does not prevent one from using multiple imputation, because alternatives are worse[93].

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
- General references: [66, 80, 127, 158]

There are many choices to be made when deciding upon a global modeling strategy, including choice between

- parametric and nonparametric procedures
- parsimony and complexity
- parsimony and good discrimination ability
- interpretable models and black boxes.

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"
- Measure needs to mask analyst to true form of regression to preserve statistical properties

4.1.1 Learning From a Saturated Model

When the effective sample size available is sufficiently large so that a saturated main effects model may be fitted, a good approach to gauging predictive potential is the following.

- Let all continuous predictors be represented as restricted cubic splines with k knots, where k is the maximum number of knots the analyst entertains for the current problem.

- Let all categorical predictors retain their original categories except for pooling of very low prevalence categories (e.g., ones containing < 6 observations).
- Fit this general main effects model.
- Compute the partial χ^2 statistic for testing the association of each predictor with the response, adjusted for all other predictors. In the case of ordinary regression convert partial F statistics to χ^2 statistics or partial R^2 values.
- Make corrections for chance associations to “level the playing field” for predictors having greatly varying d.f., e.g., subtract the d.f. from the partial χ^2 (the expected value of χ^2_p is p under H_0).
- Make certain that tests of nonlinearity are not revealed as this would bias the analyst.

- Sort the partial association statistics in descending order.

Commands in the `rms` package can be used to plot only what is needed. Here is an example for a logistic model.

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

4.1.2 Using Marginal Generalized Rank Correlations

When collinearities or confounding are not problematic, a quicker approach based on pairwise measures of association can be useful. This approach will not have numerical problems (e.g., singular covariance matrix) and is based on:

- 2 d.f. generalization of Spearman ρ—R^2 based on $rank(X)$ and $rank(X)^2$ vs. $rank(Y)$
- ρ^2 can detect U-shaped relationships
- For categorical X, ρ^2 is R^2 from dummy variables regressed against $rank(Y)$; this is tightly

related to the Wilcoxon–Mann–Whitney–Kruskal–Wallis rank test for group differences[a]

- Sort variables by descending order of ρ^2
- Specify number of knots for continuous X, combine infrequent categories of categorical X based on ρ^2

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

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

Initial simulations show the procedure to be conservative. Note that one can move from

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

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
- $\uparrow$ exploration $\rightarrow$ $\uparrow$ shrinkage (overfitting)

- Summary of problem: $E(\hat{\beta}|\hat{\beta}$ "significant" $) \neq \beta$[31]
- Biased R^2, $\hat{\beta}$, standard errors, P-values too small
- F and χ^2 statistics do not have the claimed distribution[b][70]
- Will result in residual confounding if use variable selection to find confounders[74]
- Derksen and Keselman[47] 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 vari-

[b] Lockhart *et al.*[117] provide an example with $n = 100$ and 10 orthogonal predictors where all true βs are zero. The test statistic for the first variable to enter has type I error of 0.39 when the nominal α is set to 0.05.

ables that gained entry to the model.

3. The size of the sample was of little practical importance in determining the number of authentic variables contained in the final model.
4. The population multiple coefficient of determination could be faithfully estimated by adopting a statistic that is adjusted by the total number of candidate predictor variables rather than the number of variables in the final model".

- Global test with p d.f. insignificant $\rightarrow$ **stop**

Simulation experiment, true $\sigma^2 = 6.25$, 8 candidate variables, 4 of them related to Y in the population. Select best model using AIC.

```
require(MASS)
```

```
sim ← function(n, sigma=2.5, pr=FALSE, prcor=FALSE) {
  x1 ← rnorm(n)
  x2 ← x1 + 0.5 * rnorm(n)
  x3 ← rnorm(n)
  x4 ← x3 + 1.5 * rnorm(n)
  x5 ← x1 + rnorm(n)/1.3
  x6 ← x2 + rnorm(n)/1.3
  x7 ← x3 + x4 + rnorm(n)
  x8 ← x7 + 0.5 * rnorm(n)
  if(prcor) return(round(cor(cbind(x1,x2,x3,x4,x5,x6,x7,x8)),2))
```

```
  lp ← x1 + x2 + .5*x3 + .4*x7
  y ← lp + sigma*rnorm(n)
  f ← lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8)
  g ← stepAIC(f, trace=0)
  p ← g$rank - 1
  xs ← if(p == 0) 'none' else
   gsub('[ \\+x]','', as.character(formula(g))[3])
  if(pr) print(formula(g), showEnv=FALSE)
  ssesw ← sum(resid(g)^2)
  s2s ← ssesw/g$df.residual
  # Set SSEsw / (n - gdf - 1) = true sigma^2
  gdf ← n - 1 - ssesw/(sigma^2)
  # Compute root mean squared error against true linear predictor
  rmse.full ← sqrt(mean((fitted(f) - lp) ^ 2))
  rmse.step ← sqrt(mean((fitted(g) - lp) ^ 2))
  list(stats=c(n=n, vratio=s2s/(sigma^2),
         gdf=gdf, apparentdf=p, rmse.full=rmse.full, rmse.step=rmse.step),
       xselected=xs)
}

rsim ← function(B, n) {
  xs ← character(B)
  r ← matrix(NA, nrow=B, ncol=6)
  for(i in 1:B) {
    w      ← sim(n)
    r[i,] ← w$stats
    xs[i] ← w$xselected
  }
  colnames(r) ← names(w$stats)
  s ← apply(r, 2, median)
  p ← r[, 'apparentdf']
  s['apparentdf'] ← mean(p)
  print(round(s, 2))
  print(table(p))
  cat('Prob[correct model]=', round(sum(xs == '1237')/B, 2), '\n')
}
```

Show the correlation matrix being assumed for the Xs:

```
sim(50000, prcor=TRUE)
```

```
     x1   x2   x3   x4   x5   x6   x7   x8
x1 1.00 0.89 0.01 0.00 0.79 0.74 0.00 0.00
x2 0.89 1.00 0.01 0.00 0.71 0.82 0.00 0.00
```

```
x3 0.01 0.01 1.00 0.55 0.01 0.00 0.74 0.73
x4 0.00 0.00 0.55 1.00 0.00 0.00 0.88 0.86
x5 0.79 0.71 0.01 0.00 1.00 0.58 0.00 0.00
x6 0.74 0.82 0.00 0.00 0.58 1.00 0.00 0.00
x7 0.00 0.00 0.74 0.88 0.00 0.00 1.00 0.98
x8 0.00 0.00 0.73 0.86 0.00 0.00 0.98 1.00
```

Simulate to find the distribution of the number of variables selected, the proportion of simulations in which the true model (X_1, X_2, X_3, X_7) was found, the median value of $\hat{\sigma}^2/\sigma^2$, the median effective d.f., and the mean number of apparent d.f., for varying sample sizes.

```
set.seed(11)
rsim(100, 20)  # actual model not selected once
```

```
          n      vratio         gdf  apparentdf   rmse.full
      20.00        0.70        8.09        3.65        1.62
  rmse.step
       1.56
p
 1  2  3  4  5  6  7
 4 16 32 22 14  9  3
Prob[correct model]= 0
```

```
rsim(100, 40)
```

```
          n      vratio         gdf  apparentdf   rmse.full
      40.00        0.87        7.34        3.06        1.21
  rmse.step
       1.15
p
 1  2  3  4  5  6  7
 2 34 33 21  8  1  1
Prob[correct model]= 0
```

```
rsim(100, 150)
```

```
          n      vratio         gdf  apparentdf   rmse.full
     150.00        0.97        9.08        3.81        0.59
  rmse.step
       0.62
p
 2  3  4  5  6
10 24 44 19  3
Prob[correct model]= 0.13
```

```
rsim(100, 300)
```

```
          n      vratio         gdf  apparentdf   rmse.full
     300.00        0.98        9.26        4.21        0.43
  rmse.step
       0.41
p
 3  4  5  6
12 60 23  5
Prob[correct model]= 0.38
```

```
rsim(100, 2000)
```

```
          n      vratio         gdf  apparentdf   rmse.full
    2000.00        1.00        6.30        4.58        0.17
  rmse.step
       0.15
p
 4  5  6  7
54 35 10  1
Prob[correct model]= 0.52
```

As $n \uparrow$ the mean number of variables selected increased. The proportion of simulations in which the correct model was found increased from 0 to 0.52. σ^2 is underestimated in non-large samples by a factor of 0.70, resulting in the d.f. needed to de-bias $\hat{\sigma}^2$ being 8.1 when the apparent d.f. was only 3.65 on the average, when $n = 20$. Variable selection did increase closeness to the true $X\beta$ for some sample sizes.

Variable selection methods[78]:

- 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[7]), 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 a limited number of pre-specified models [21, Section 1.3]
- Roecker[142] 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 un-

necessary

- Step-down is usually better than forward[119] and can be used efficiently with maximum likelihood estimation[107]
- Fruitless to try different stepwise methods to look for agreement[187]
- 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[4, 21, 91]
- Data reduction (grouping variables) can help
- Using the bootstrap to select important variables for inclusion in the final model[146] is problematic[8]
- 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
- Assume typical problem in medicine, epidemiology, and the social sciences in which the signal:noise ratio is small (higher ratios allow for more aggressive modeling)
- p should be $< \frac{m}{15}$[79, 81, 129, 130, 152, 169, 175]
- $p =$ number of parameters in full model or number of *candidate* parameters in a stepwise analysis
- Derived from simulations to find minimum sample size so that apparent discrimination = validated discrimination
- Applies to typical signal:noise ratios found outside of tightly controlled experiments
- If true R^2 is high, many parameters can be estimated from smaller samples

Table 4.1: Limiting Sample Sizes for Various Response Variables

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

- 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[162]

4.5 Shrinkage

- Slope of calibration plot; regression to the mean

[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 [186, Eq. 10, 15]. Here n_1 and n_2 are the marginal frequencies of the two response levels [130].

[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) [186, Eq, 3]. If all cell sizes are equal, the relative efficiency of having k response categories compared to a continuous response is $1 - \frac{1}{k^2}$ [186, 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[15].

- Statistical estimation procedure — “pre-shrunk” models
- Aren’t regression coefficients OK because they’re unbiased?
- Problem is in how we use coefficient estimates
- Consider 20 samples of size $n = 50$ from $U(0,1)$
- Compute group means and plot in ascending order
- Equivalent to fitting an intercept and 19 dummies using least squares
- Result generalizes to general problems in plotting Y vs. $X\hat{\beta}$

```
set.seed(123)
n ← 50
y ← runif(20*n)
group ← rep(1:20,each=n)
ybar ← tapply(y, group, mean)
ybar ← sort(ybar)
plot(1:20, ybar, type='n', axes=FALSE, ylim=c(.3,.7),
     xlab='Group', ylab='Group Mean')
lines(1:20, ybar)
points(1:20, ybar, pch=20, cex=.5)
axis(2)
axis(1, at=1:20, labels=FALSE)
for(j in 1:20) axis(1, at=j, labels=names(ybar)[j])
abline(h=.5, col=gray(.85))
```

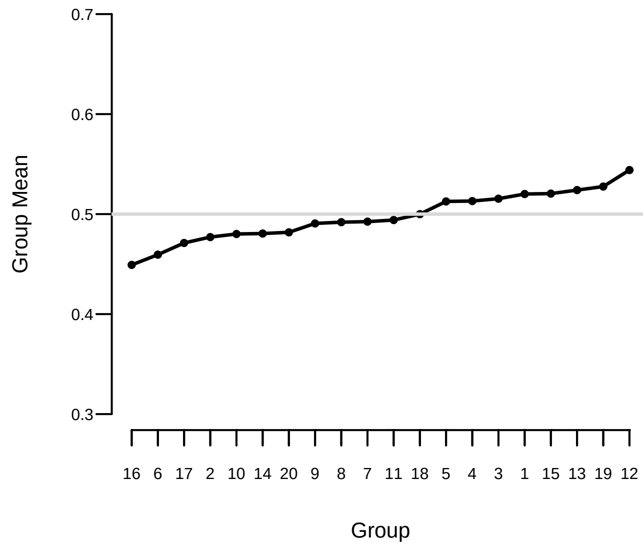


Figure 4.1: Sorted means from 20 samples of size 50 from a uniform $[0, 1]$ distribution. The reference line at 0.5 depicts the true population value of all of the means.

- Prevent shrinkage by using pre-shrinkage
- Spiegelhalter[155]: 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[39]
- Ridge regression[108, 170]
- Penalized MLE[71, 83, 173]
- Heuristic shrinkage parameter of van Houwelin-

gen and le Cessie [170, Eq. 77]

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

- OLS[f]: $\hat{\gamma} = \frac{n-p-1}{n-1} R^2_{\text{adj}}/R^2$
  $R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$
- p close to no. candidate variables
- Copas [39, 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 [32, Chap. 9]

[f] An excellent discussion about such indexes may be found in `http://r.789695.n4.nabble.com/Adjusted-R-squared-formula-in-lm-td4656857.html`

- 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 [125, 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[45, 181]

- See Belsley [13, pp. 28-30] for problems with VIF
- Easy approach: SAS `VARCLUS` procedure[145], S `varclus` function, other clustering techniques: group highly correlated variables
- Can score each group (e.g., first principal component, PC_1[44]); summary scores not collinea

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 Redundancy Analysis

- Remove variables that have poor distributions
  - E.g., categorical variables with fewer than 2 categories having at least 20 observations
- Use flexible additive parametric additive models to determine how well each variable can be predicted from the remaining variables

- Variables dropped in stepwise fashion, removing the most predictable variable at each step
- Remaining variables used to predict
- Process continues until no variable still in the list of predictors can be predicted with an R^2 or adjusted R^2 greater than a specified threshold or until dropping the variable with the highest R^2 (adjusted or ordinary) would cause a variable that was dropped earlier to no longer be predicted at the threshold from the now smaller list of predictors
- R/S function `redun` in `Hmisc` package
- Related to *principal variables*[121] but faster

4.7.2 Variable Clustering

- Goal: Separate variables into groups
  - variables within group correlated with each other

  - variables not correlated with non-group members

- Score each dimension, stop trying to separate effects of factors measuring same phenomenon
- Variable clustering[44, 145] (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[88].
- See [75] for a method related to variable clustering and sparse principal components.
- [33] implement many more variable clustering methods

4.7.3 Transformation and Scaling Variables Without Using Y

- Reduce p by estimating transformations using associations with other predictors

- Purely categorical predictors – correspondence analysis[34,43,73,109,122]
- Mixture of qualitative and continuous variables: qualitative principal components
- Maximum total variance (MTV) of Young, Takane, de Leeuw[122,193]
  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 [104, 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`[104]

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

4.7.4 Simultaneous Transformation and Imputation

S `transcan` Function for Data Reduction & Imputation

- Initialize missings to medians (or most frequent category)
- Initialize transformations to original variables
- Take each variable in turn as Y
- Exclude obs. missing on Y
- Expand Y (spline or dummy variables)
- Score (transform Y) using first canonical variate
- Missing $Y \rightarrow$ predict canonical variate from Xs
- The imputed values can optionally be shrunk to avoid overfitting for small n or large p

- Constrain imputed values to be in range of non-imputed ones
- Imputations on original scale
  1. Continuous $\rightarrow$ back-solve with linear interpolation
  2. Categorical $\rightarrow$ classification tree (most freq. cat.) or match to category whose canonical score is closest to one predicted
- Multiple imputation — bootstrap or approx. Bayesian boot.
  1. Sample residuals multiple times (default $M = 5$)
  2. Are on "optimally" transformed scale
  3. Back-transform
  4. `fit.mult.impute` works with `aregImpute` and `transcan` output to easily get imputation-corrected variances and avg. $\hat{\beta}$
- Option to insert constants as imputed values (ignored during transformation estimation); helpful when a lab value may be miss-

ing because the patient returned to normal

- Imputations and transformed values may be easily obtained for new data
- An S function `Function` will create a series of S functions that transform each predictor
- Example: $n = 415$ acutely ill patients
  1. Relate heart rate to mean arterial blood pressure
  2. Two blood pressures missing
  3. Heart rate not monotonically related to blood pressure
  4. See Figures 4.2 and 4.3

```
require(Hmisc)
```

```
getHdata(support)    # Get data frame from web site
heart.rate      ← support$hrt
blood.pressure ← support$meanbp
blood.pressure[400:401]
```

```
Mean Arterial Blood Pressure Day 3
[1] 151 136
```

```
blood.pressure[400:401] ← NA  # Create two missings
d ← data.frame(heart.rate, blood.pressure)
par(pch=46)   # Figure 4.2
w ← transcan(∼ heart.rate + blood.pressure, transformed=TRUE,
             imputed=TRUE, show.na=TRUE, data=d)
```

```
Convergence criterion:2.901 0.035
```

```
0.007
Convergence in 4 iterations
R² achieved in predicting each variable:

    heart.rate blood.pressure
         0.259          0.259

Adjusted R²:

    heart.rate blood.pressure
         0.254          0.253
```

```
w$imputed$blood.pressure
```

```
     400      401
132.4057 109.7741
```

```
t ← w$transformed
spe ← round(c(spearman(heart.rate, blood.pressure),
              spearman(t[,'heart.rate'],
                       t[,'blood.pressure'])), 2)
```

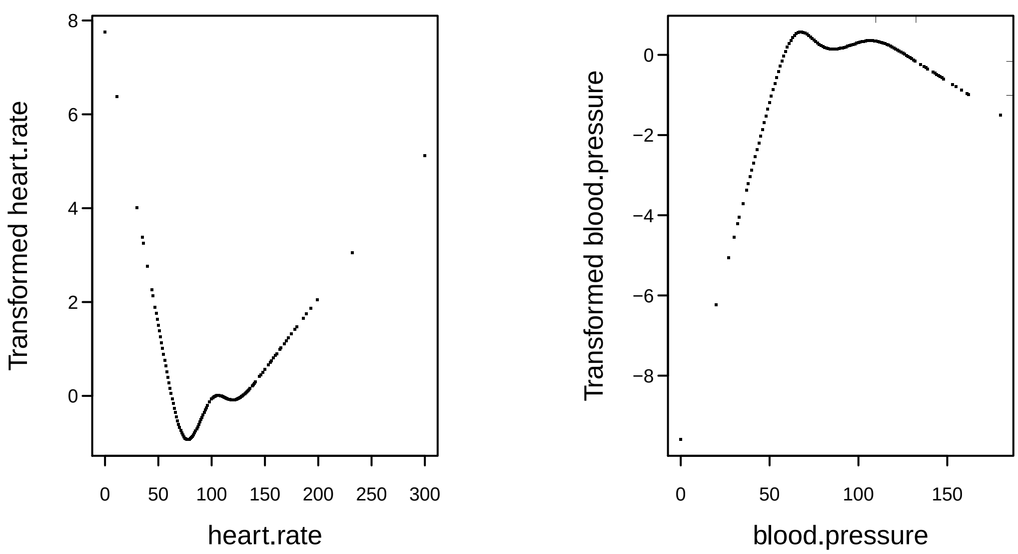


Figure 4.2: Transformations fitted using `transcan`. Tick marks indicate the two imputed values for blood pressure.

```
plot(heart.rate, blood.pressure)   # Figure 4.3
plot(t[,'heart.rate'], t[,'blood.pressure'],
     xlab='Transformed hr', ylab='Transformed bp')
```

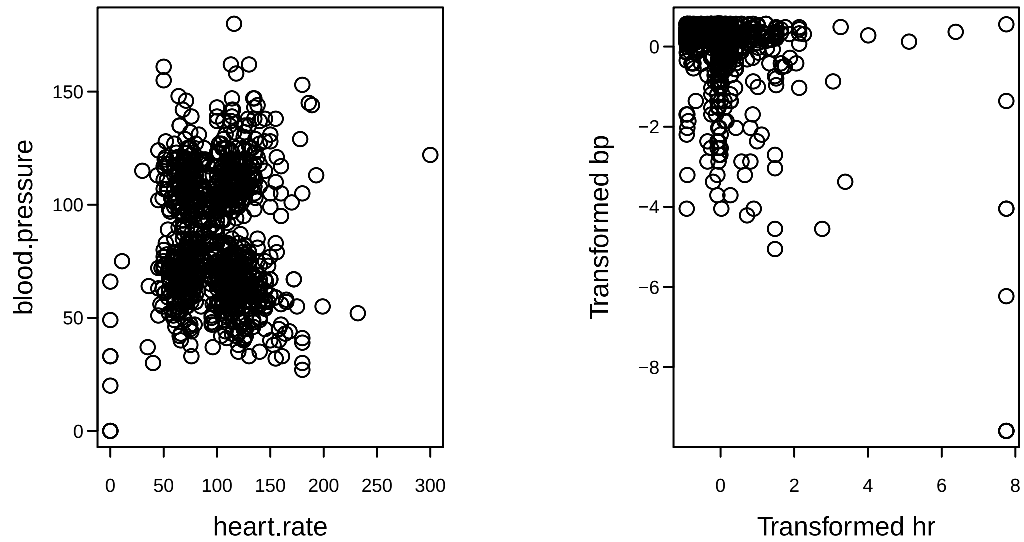


Figure 4.3: The lower left plot contains raw data (Spearman $\rho = -0.02$); the lower right is a scatterplot of the corresponding transformed values ($\rho = -0.13$). Data courtesy of the SUPPORT study[98].

ACE (Alternating Conditional Expectation) of Breiman and Friedman[22]

1. Uses nonparametric "super smoother"[62]
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

4.7.5 Simple Scoring of Variable Clusters

- Try to score groups of transformed variables with PC_1
- Reduces d.f. by pre-transforming var. and by combining multiple var.
- Later may want to break group apart, but delete all variables in groups whose summary scores do not add significant information
- Sometimes simplify cluster score by finding a subset of its constituent variables which predict it with high R^2.

Series of dichotomous variables:

- Construct X_1 = 0-1 according to whether any variables positive

- Construct X_2 = number of positives
- Test whether original variables add to X_1 or X_2

4.7.6 Simplifying Cluster Scores

4.7.7 How Much Data Reduction Is Necessary?

Using Expected Shrinkage to Guide Data Reduction

- Fit full model with all candidates, p d.f., LR likelihood ratio χ^2
- Compute $\hat{\gamma}$
- If < 0.9, consider shrunken estimator from whole model, or data reduction (again not using Y)
- q regression d.f. for reduced model
- Assume best case: discarded dimensions had no association with Y
- Expected loss in LR is $p - q$

- New shrinkage $[\mathrm{LR}-(p-q)-q]/[\mathrm{LR}-(p-q)]$
- Solve for $q \rightarrow q \leq (\mathrm{LR}-p)/9$
- Under these assumptions, no hope unless original LR $> p+9$
- No χ^2 lost by dimension reduction $\rightarrow q \leq \mathrm{LR}/10$

Example:

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

- LR$> 15 + 9 = 24 \rightarrow$ reduction may help
- Reduction to $q = (50 - 15)/9 \approx 4$ d.f. necessary
- Have to assume age linear, reduce other 10 to 1 d.f.
- Separate hypothesis tests intended $\rightarrow$ use full model, adjust for multiple comparisons

Summary of Some Data Reduction Methods

Goals	Reasons	Methods
Group predictors so that each group represents a single dimension that can be summarized with a single score	• $\downarrow$ 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	• $\downarrow$ 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	$\downarrow$ d.f. for group to unity	• PC_1 • Simple point scores
Multiple dimensional scoring of all predictors	$\downarrow$ 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[14]
- $h_{ii} > 2(p + 1)/n$ may signal a high leverage point[14]
- DFBETAS: change in $\hat{\beta}$ upon deletion of each

obs, scaled by s.e.

- DFFIT: change in $X\hat{\beta}$ upon deletion of each obs
- DFFITS: DFFIT standardized by s.e. of $\hat{\beta}$
- Some classify obs as overly influential when $|\mathrm{DFFITS}| > 2\sqrt{(p+1)/(n-p-1)}$[14]
- Others examine entire distribution for "outliers"
- No substitute for careful examination of data[30,154]
- 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[126], model χ^2, Somers' D_{xy}, Spearman's ρ, area under ROC curve)
- Worthwhile to compare models on a measure not used to optimize either model, e.g., mean absolute error, median absolute error if using OLS
- Rank measures may not give enough credit

to extreme predictions $\rightarrow$ model χ^2, R^2, examine extremes of distribution of $\hat{Y}$

- Examine differences in predicted values from the two models
- See[131–134] for discussions and examples of low power for testing differences in ROC areas, and for other approaches.

4.10 Summary: Possible Modeling Strategies

Greenland[74] discusses many important points:

- Stepwise variable selection on confounders leaves important confounders uncontrolled
- Shrinkage is far superior to variable selection
- Variable selection does more damage to confidence interval widths than to point estimates
- Claims about unbiasedness of ordinary MLEs are misleading because they assume the model

is correct and is the only model entertained

- “models need to be complex to capture uncertainty about the relations . . . an honest uncertainty assessment requires parameters for all effects that we know may be present. This advice is implicit in an antiparsimony principle often attributed to L. J. Savage ‘All models should be as big as an elephant’ (see Draper, 1995)”

Global Strategies

- Use a method known not to work well (e.g., stepwise variable selection without penalization; recursive partitioning), document how poorly the model performs (e.g. using the bootstrap), and use the model anyway
- Develop a black box model that performs poorly and is difficult to interpret (e.g., does not incorporate penalization)

- Develop a black box model that performs well and is difficult to interpret
- Develop interpretable approximations to the black box
- Develop an interpretable model (e.g. give priority to additive effects) that performs well and is likely to perform equally well on future data from the same stream

Preferred Strategy in a Nutshell

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

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[42]. 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[83]
7. Use the entire sample in model development
8. Can do highly structured testing to simplify

"initial" model

(a) Test entire group of predictors with a single P-value
(b) Make each continuous predictor have same number of knots, and select the number that optimizes AIC
(c) Test the combined effects of all nonlinear terms with a single P-value

9. Make tests of linearity of effects in the model only to demonstrate to others that such effects are often statistically significant. Don't remove individual insignificant effects from the model.
10. Check additivity assumptions by testing prespecified 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[155]. But confidence limits, etc., must account for variable selection (e.g., bootstrap).
14. This is the “final” model.
15. Interpret the model graphically and by computing predicted values and appropriate test statistics. Compute pooled tests of association for collinear predictors.
16. Validate this model for calibration and discrimination ability, preferably using bootstrapping.
17. Shrink parameter estimates if there is overfitting but no further data reduction is desired (unless shrinkage built-in to estimation)
18. When missing values were imputed, adjust final variance-covariance matrix for imputation. Do this as early as possible because it will affect other findings.
19. When all steps of the modeling strategy can

be automated, consider using Faraway's method[58] 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

Describing, Resampling, Validating, and Simplifying the Model

5.1 Describing the Fitted Model

5.1.1 Interpreting Effects

- Regression coefficients if 1 d.f. per factor, no interaction
- **Not** standardized regression coefficients
- Many programs print meaningless estimates such as effect of increasing age^2 by one unit, holding age constant
- Need to account for nonlinearity, interaction, and use meaningful ranges
- For monotonic relationships, estimate $X\hat{\beta}$ at

quartiles of continuous variables, separately for various levels of interacting factors

- Subtract estimates, anti-log, e.g., to get inter-quartile-range odds or hazards ratios. Base C.L. on s.e. of difference.
- Plot effect of each predictor on $X\beta$ or some transformation of $X\beta$. See also [96].
- Nomogram
- Use regression tree to approximate the full model

5.1.2 Indexes of Model Performance

Error Measures

- Central tendency of prediction errors
  - Mean absolute prediction error: mean $|Y - \hat{Y}|$
  - Mean squared prediction error
    - Binary Y: Brier score (quadratic proper scoring rule)

- Logarithmic proper scoring rule (avg. log-likelihood)

- Discrimination measures
  - Pure discrimination: rank correlation of $(\hat{Y}, Y)$
    * Spearman ρ, Kendall τ, Somers' D_{xy}
    * Y binary $\rightarrow D_{xy} = 2 \times (C - \frac{1}{2})$
      C = concordance probability = area under receiver operating characteristic curve $\propto$ Wilcoxon-Mann-Whitney statistic
  - Mostly discrimination: R^2
    * R^2_{adj}—overfitting corrected if model pre-specified
  - Brier score can be decomposed into discrimination and calibration components
  - Discrimination measures based on variation in $\hat{Y}$
    * regression sum of squares
    * g–index

- Calibration measures

– calibration–in–the–large: average $\hat{Y}$ vs. average Y
– high-resolution calibration curve (calibration–in–the–small)
– calibration slope and intercept
– maximum absolute calibration error
– mean absolute calibration error
– 0.9 quantile of calibration error

g–Index

- Based on Gini's mean difference
  – mean over all possible $i \neq j$ of $|Z_i - Z_j|$
  – interpretable, robust, highly efficient measure of variation
- $g =$ Gini's mean difference of $X_i\hat{\beta} = \hat{Y}$
- Example: $Y =$ systolic blood pressure; $g = 11$mmHg is typical difference in $\hat{Y}$
- Independent of censoring etc.

- For models in which anti-log of difference in $\hat{Y}$ represent meaningful ratios (odds ratios, hazard ratios, ratio of medians):
  $g_r = \exp(g)$
- For models in which $\hat{Y}$ can be turned into a probability estimate (e.g., logistic regression):
  $g_p =$ Gini's mean difference of $\hat{P}$
- These g–indexes represent e.g. "typical" odds ratios, "typical" risk differences
- Can define partial g

5.2 The Bootstrap

- If know population model, use simulation or analytic derivations to study behavior of statistical estimator
- Suppose Y has a cumulative dist. fctn. $F(y) = \mathrm{Prob}\{Y \leq y\}$

- 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.

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} [Y_i \leq y].$$

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

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

- F_n corresponds to density function placing

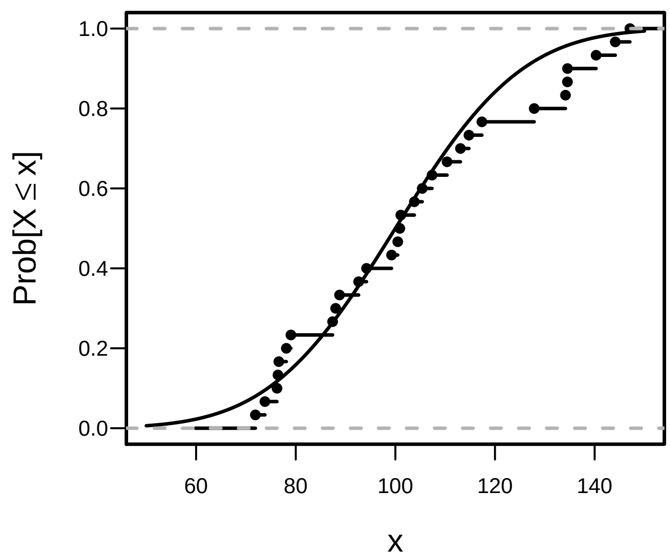


Figure 5.1: Empirical and population cumulative distribution function

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 original data points in each bootstrap sample at least once
- Some observations not selected, others selected more than once
- Efron's *bootstrap* $\rightarrow$ general-purpose technique for estimating properties of estimators

without assuming or knowing distribution of data F

- Take B samples of size n with replacement, choose B so that summary measure of individual statistics $\approx$ summary if $B = \infty$
- Bootstrap based on distribution of *observed* differences between a resampled parameter estimate and the original estimate telling us about the distribution of *unobservable* differences between the original estimate and the unknown parameter

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

```
options(digits=3)
y ← c(2,5,6,7,8,9,10,11,12,13,14,19,20,21)
y ← c(1,5,6,7,8,9)
set.seed(17)
n ← length(y)
n2 ← n/2
n21 ← n2+1
B ← 400
M ← double(B)
plot(0, 0, xlim=c(0,B), ylim=c(3,9),
```

```
     xlab="Bootstrap Samples Used",
     ylab="Mean and 0.1, 0.9 Quantiles", type="n")
for(i in 1:B) {
  s ← sample(1:n, n, replace=T)
  x ← sort(y[s])
  m ← .5*(x[n2]+x[n21])
  M[i] ← m
  if(i ≤ 20) {
    w ← as.character(x)
    cat(w, "& &", sprintf('%.1f',m),
        if(i < 20) "\\\\\n" else "\\\\ \\hline\n",
        file='~/doc/rms/validate/tab.tex', append=i > 1)
  }
  points(i, mean(M[1:i]), pch=46)
  if(i ≥ 10)     {
    q ← quantile(M[1:i], c(.1,.9))
    points(i, q[1], pch=46, col='blue')
    points(i, q[2], pch=46, col='blue')
  }
}
table(M)
```

```
M
  1   3 3.5   4 4.5   5 5.5   6 6.5   7 7.5   8 8.5   9
  6  10   7   8   2  23  43  75  59  66  47  42  11   1
```

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

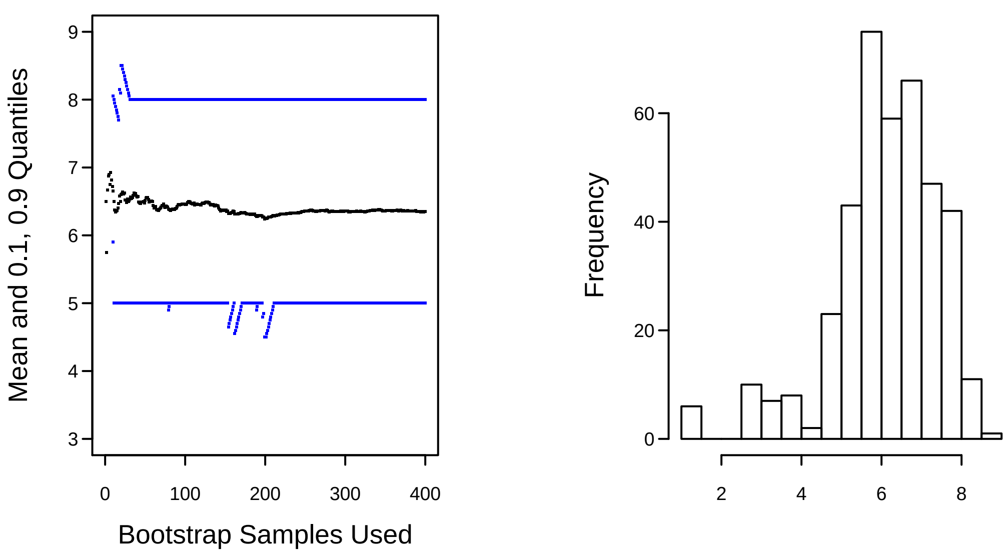


Figure 5.2: Estimating properties of sample median using the bootstrap

First 20 samples:

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

- Histogram tells us whether we can assume normality for the bootstrap medians or need to use quantiles of medians to construct C.L.
- Need high B for quantiles, low for variance (but see [19])

5.3 Model Validation

5.3.1 Introduction

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

[a]But in many cases it is better to combine data and include country or calendar time as a predictor.

Working definition of external validation: Validation of a prediction tool on a sample that was not available at publication time. **Alternate**: Validation of a prediction tool by an independent research team.

One suggested hierarchy of the quality of various validation methods is as follows, ordered from worst to best.

1. Attempting several validations (internal or external) and reporting only the one that "worked"
2. Reporting apparent performance on the training dataset (no validation)
3. Reporting predictive accuracy on an undersized independent test sample
4. Internal validation using data splitting where at least one of the training and test samples is not huge and the investigator is not aware of the arbitrariness of variable selection done on a single sample
5. Strong internal validation using 100 repeats

of 10-fold cross-validation or several hundred bootstrap resamples, repeating *all* analysis steps involving Y afresh at each re-sample and the arbitrariness of selected “important variables” is reported (if variable selection is used)

6. External validation on a large test sample, done by the original research team
7. Re-analysis by an independent research team using strong internal validation of the original dataset
8. External validation using new test data, done by an independent research team
9. External validation using new test data generated using different instruments/technology, done by an independent research team

5.3.2 Which Quantities Should Be Used in Validation?

- OLS: R^2 is one good measure for quantifying drop-off in predictive ability

- Example: $n = 10, p = 9$, apparent $R^2 = 1$ but R^2 will be close to zero on new subjects
- Example: $n = 20, p = 10$, apparent $R^2 = .9$, R^2 on new data 0.7, $R^2_{adj} = 0.79$
- Adjusted R^2 solves much of the bias problem assuming p in its formula is the largest number of parameters ever examined against Y
- Few other adjusted indexes exist
- Also need to validate models with phantom d.f.
- Cross-validation or bootstrap can provide unbiased estimate of any index; bootstrap has higher precision
- Two main types of quantities to validate
  1. Calibration or reliability: ability to make unbiased estimates of response ($\hat{Y}$ vs. Y)
  2. Discrimination: ability to separate responses OLS: R^2; g–index; binary logistic model: ROC area, equivalent to rank correlation

between predicted probability of event and 0/1 event

- Unbiased validation nearly always necessary, to detect overfitting

5.3.3 Data-Splitting

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

predict on test sample ($n = 100$), $R^2 = 1 - \frac{\Sigma(Y_i - X_i\hat{\beta})^2}{\Sigma(Y_i - \bar{Y})^2}$.

- Disadvantages of data splitting:
  1. Costly in $\downarrow n$[21,142]
  2. Requires *decision* to split at beginning of analysis
  3. Requires larger sample held out than cross-validation
  4. Results vary if split again
  5. Does not validate the final model (from recombined data)
  6. Not helpful in getting CL corrected for var. selection

5.3.4 Improvements on Data-Splitting: Resampling

- No sacrifice in sample size
- Work when modeling process automated
- Bootstrap excellent for studying arbitrariness of variable selection[146]

- Cross-validation solves many problems of data splitting[53, 149, 170, 190]
- Example of $\times$-validation:
  1. Split data at random into 10 tenths
  2. Leave out $\frac{1}{10}$ of data at a time
  3. Develop model on $\frac{9}{10}$, including any variable selection, pre-testing, etc.
  4. Freeze coefficients, evaluate on $\frac{1}{10}$
  5. Average R^2 over 10 reps
- Drawbacks:
  1. Choice of number of groups and repetitions
  2. Doesn't show full variability of var. selection
  3. Does not validate full model
  4. Lower precision than bootstrap
  5. Need to do 50 repeats of 10-fold cross-validation to ensure adequate precision
- Randomization method

1. Randomly permute Y
2. Optimism = performance of fitted model compared to what expect by chance

5.3.5 Validation Using the Bootstrap

- Estimate optimism of *final whole sample fit* without holding out data
- From original X and Y select sample of size n with replacement
- Derive model from bootstrap sample
- Apply to original sample
- Simple bootstrap uses average of indexes computed on original sample
- Estimated optimism = difference in indexes
- Repeat about $B = 100$ times, get average expected optimism
- Subtract average optimism from apparent index in final model

- Example: $n = 1000$, have developed a final model that is hopefully ready to publish. Call estimates from this final model $\hat{\beta}$.
  - final model has apparent R^2 (R^2_{app}) =0.4
  - how inflated is R^2_{app}?
  - get resamples of size 1000 with replacement from original 1000
  - for each resample compute R^2_{boot} = apparent R^2 in bootstrap sample
  - freeze these coefficients (call them $\hat{\beta}_{boot}$), apply to original (whole) sample (X_{orig}, Y_{orig}) to get $R^2_{orig} = R^2(X_{orig}\hat{\beta}_{boot}, Y_{orig})$
  - optimism = $R^2_{boot} - R^2_{orig}$
  - average over $B = 100$ optimisms to get $\overline{optimism}$
  - $R^2_{overfitting\ corrected} = R^2_{app} - \overline{optimism}$
- Is estimating unconditional (not conditional on X) distribution of R^2, etc. [58, 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[53], [54, p. 253],[55]

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[155]
- Use Efron's method to estimate optimism in $(0, 1)$, estimate (γ_0, γ_1) by subtracting optimism from $(0, 1)$
- See also Copas[40] and van Houwelingen and le Cessie [170, p. 1318]

See [61] for warnings about the bootstrap, and [53] 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[155]
- 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 [170] for many interesting ideas
- Faraway[58] shows how bootstrap is used to penalize for choosing transformations for Y, outlier and influence checking, variable selection, etc. simultaneously
- Brownstone [25, 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[18] and Chatfield[31]

for more interesting examples of problems resulting from data-driven analyses.

5.4 Bootstrapping Ranks of Predictors

- Order of importance of predictors not pre-specified
- Researcher interested in determining "winners" and "losers"
- Bootstrap useful in documenting the difficulty of this task
- Get confidence limits of the rank of each predictor in the scale of partial χ^2 - d.f.
- Example using OLS

```
# Use the plot method for anova, with pl=FALSE to suppress actual
# plotting of chi-square - d.f. for each bootstrap repetition.
# Rank the negative of the adjusted chi-squares so that a rank of
# 1 is assigned to the highest.  It is important to tell
# plot.anova.rms not to sort the results, or every bootstrap
# replication would have ranks of 1,2,3,... for the stats.
require(rms)
n ← 300
set.seed(1)
d ← data.frame(x1=runif(n), x2=runif(n), x3=runif(n), x4=runif(n),
               x5=runif(n), x6=runif(n), x7=runif(n), x8=runif(n),
            x9=runif(n), x10=runif(n), x11=runif(n), x12=runif(n))
d$y ← with(d, 1*x1 + 2*x2 +  3*x3  +  4*x4  + 5*x5 + 6*x6 + 7*x7 +
```

```
                8*x8 + 9*x9 + 10*x10 + 11*x11 + 12*x12 + 9*rnorm(n))

f ← ols(y ~ x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11+x12, data=d)
B ← 1000
ranks ← matrix(NA, nrow=B, ncol=12)
rankvars ← function(fit)
  rank(plot(anova(fit), sort='none', pl=FALSE))
Rank ← rankvars(f)
for(i in 1:B) {
  j ← sample(1:n, n, TRUE)
  bootfit ← update(f, data=d, subset=j)
  ranks[i,] ← rankvars(bootfit)
  }
lim ← t(apply(ranks, 2, quantile, probs=c(.025,.975)))
predictor ← factor(names(Rank), names(Rank))
w ← data.frame(predictor, Rank, lower=lim[,1], upper=lim[,2])
require(ggplot2)
ggplot(w, aes(x=predictor, y=Rank)) + geom_point() + coord_flip() +
  scale_y_continuous(breaks=1:12) +
  geom_errorbar(aes(ymin=lim[,1], ymax=lim[,2]), width=0)
```

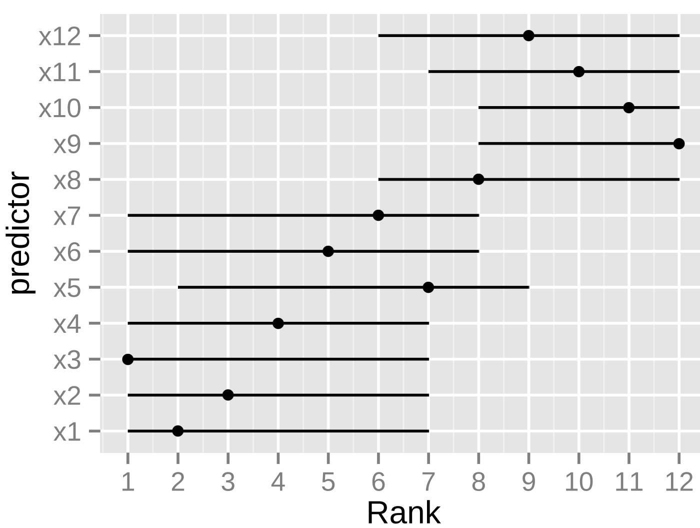


Figure 5.3: Bootstrap percentile 0.95 confidence limits for ranks of predictors in an OLS model. Ranking is on the basis of partial χ^2 minus d.f. Point estimates are original ranks

5.5 Simplifying the Final Model by Approximating It

5.5.1 Difficulties Using Full Models

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

5.5.2 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'

5.6 How Do We Break Bad Habits?

- Insist on validation of predictive models and discoveries

- Show collaborators that split-sample validation is not appropriate unless the number of subjects is huge
  - Split more than once and see volatile results
  - Calculate a confidence interval for the predictive accuracy in the test dataset and show that it is very wide
- Run simulation study with no real associations and show that associations are easy to find
- Analyze the collaborator's data after randomly permuting the Y vector and show some positive findings
- Show that alternative explanations are easy to posit
  - Importance of a risk factor may disappear if 5 "unimportant" risk factors are added back to the model
  - Omitted main effects can explain apparent

interactions

- *Uniqueness analysis*: attempt to predict the predicted values from a model derived by data torture from all of the features not used in the model

Chapter 6

R Software

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

R is comprehensive:

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

- Single R 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 R Modeling Language

R 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 R 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

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

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

Some R functions:

- See Venables and Ripley
- Hierarchical clustering: `hclust`
- Principal components: `princomp, prcomp`
- Canonical correlation: `cancor`
- Nonparametric transform-both-sides additive models:
  `ace, avas`
- Parametric transform-both-sides additive models:
  `areg`, `areg.boot` (`Hmisc` package in R))
- Rank correlation methods:
  `rcorr`, `hoeffd`, `spearman2` (`Hmisc`)
- Variable clustering: `varclus` (`Hmisc`)
- Single imputation: `transcan` (`Hmisc`)
- Multiple imputation: `aregImpute` (`Hmisc`)
- Restricted cubic splines:
  `rcspline.eval` (`Hmisc`)

- Re-state restricted spline in simpler form: `rcspline.restate` (`Hmisc`)

6.3 The rms Package

- `datadist` function to compute predictor distribution summaries

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

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

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

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

Table 6.1: `rms` Fitting Functions

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

Table 6.2: `rms` Transformation Functions

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

Function	Purpose	Related Functions
`print`	Print parameters and statistics of fit	
`coef`	Fitted regression coefficients	
`formula`	Formula used in the fit	
`specs`	Detailed specifications of fit	
`vcov`	Fetch covariance matrix	
`logLik`	Fetch maximized log-likelihood	
`AIC`	Fetch AIC with option to put on chi-square basis	
`lrtest`	Likelihood ratio test for two nested models	
`univarLR`	Compute all univariable LR χ^2	
`robcov`	Robust covariance matrix estimates	
`bootcov`	Bootstrap covariance matrix estimates and bootstrap distributions of estimates	
`pentrace`	Find optimum penalty factors by tracing effective AIC for a grid of penalties	
`effective.df`	Print effective d.f. for each type of variable in model, for penalized fit or `pentrace` result	
`summary`	Summary of effects of predictors	
`plot.summary`	Plot continuously shaded confidence bars for results of `summary`	
`anova`	Wald tests of most meaningful hypotheses	
`plot.anova`	Graphical depiction of anova	
`contrast`	General contrasts, C.L., tests	
`gendata`	Easily generate predictor combinations	
`predict`	Obtain predicted values or design matrix	
`Predict`	Obtain predicted values and confidence limits easily varying a subset of predictors and others set at default values	
`plot.Predict`	Plot the result of `Predict` using `lattice`	
`ggplot.Predict`	Plot the result of `Predict` using `ggplot2`	
`fastbw`	Fast backward step-down variable selection	`step`
`residuals`	(or `resid`) Residuals, influence stats from fit	
`sensuc`	Sensitivity analysis for unmeasured confounder	
`which.influence`	Which observations are overly influential	`residuals`
`latex`	LaTeX representation of fitted model	`Function`

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

Example:

- `treat`: categorical variable with levels `"a","b","c"`
- `num.diseases`: ordinal variable, 0-4
- `age`: continuous
  Restricted cubic spline
- `cholesterol`: continuous
  (3 missings; use median)
  `log(cholesterol+10)`

- Allow `treat` × `cholesterol` interaction
- Program to fit logistic model, test all effects in design, estimate effects (e.g. inter-quartile range odds ratios), plot estimated transformations

```
require(rms)                         # make new functions available
ddist ← datadist(cholesterol, treat, num.diseases, age)
# Could have used ddist ← datadist(data.frame.name)
options(datadist="ddist")            # defines data dist. to rms
cholesterol ← impute(cholesterol)
fit ← lrm(y ∼ treat + scored(num.diseases) + rcs(age) +
                log(cholesterol+10) + treat:log(cholesterol+10))
describe(y ∼ treat + scored(num.diseases) + rcs(age))
# or use describe(formula(fit)) for all variables used in fit
# describe function (in Hmisc) gets simple statistics on variables
# fit ← robcov(fit)                  # Would make all statistics that follow
                                     # use a robust covariance matrix
                                     # would need x=T, y=T in lrm()
specs(fit)                           # Describe the design characteristics
anova(fit)
anova(fit, treat, cholesterol)       # Test these 2 by themselves
plot(anova(fit))                     # Summarize anova graphically
summary(fit)                         # Estimate effects using default ranges
plot(summary(fit))                   # Graphical display of effects with C.I.
summary(fit, treat="b", age=60)      # Specify reference cell and adjustment val
summary(fit, age=c(50,70))           # Estimate effect of increasing age from
                                     # 50 to 70
summary(fit, age=c(50,60,70))        # Increase age from 50 to 70, adjust to
                                     # 60 when estimating effects of other
                                     # factors
# If had not defined datadist, would have to define ranges for all var.

# Estimate and test treatment (b-a) effect averaged over 3 cholesterols
contrast(fit, list(treat='b', cholesterol=c(150,200,250)),
              list(treat='a', cholesterol=c(150,200,250)),
         type='average')
# See the help file for contrast.rms for several examples of
# how to obtain joint tests of multiple contrasts and how to get
# double differences (interaction contrasts)
```

```
p ← Predict(fit, age=seq(20,80,length=100), treat, conf.int=FALSE)
plot(p)                           # Plot relationship between age and log
# or ggplot(p)                    # odds, separate curve for each treat,
                                  # no C.I.
plot(p, ~ age | treat)            # Same but 2 panels
ggplot(p, groups=FALSE)
bplot(Predict(fit, age, cholesterol, np=50))
                                  # 3-dimensional perspective plot for age,
                                  # cholesterol, and log odds using default
                                  # ranges for both variables
plot(Predict(fit, num.diseases, fun=function(x) 1/(1+exp(−x)), conf.int=.9),
     ylab="Prob")                 # Plot estimated probabilities instead of
                                  # log odds (or use ggplot())
# Again, if no datadist were defined, would have to tell plot all limits
logit ← predict(fit, expand.grid(treat="b",num.dis=1:3,age=c(20,40,60),
                 cholesterol=seq(100,300,length=10)))
# Could also obtain list of predictor settings interactively}
logit ← predict(fit, gendata(fit, nobs=12))

# Since age doesn't interact with anything, we can quickly and
# interactively try various transformations of age, taking the spline
# function of age as the gold standard. We are seeking a linearizing
# transformation.

ag ← 10:80
logit ← predict(fit, expand.grid(treat="a", num.dis=0, age=ag,
                 cholesterol=median(cholesterol)), type="terms")[,"age"]
# Note: if age interacted with anything, this would be the age
#       "main effect" ignoring interaction terms
# Could also use
#    logit ← Predict(f, age=ag, ...)$yhat,
# which allows evaluation of the shape for any level of interacting
#  factors.  When age does not interact with anything, the result from
# predict(f, ..., type="terms") would equal the result from
# Predict if all other terms were ignored

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

plot(ag^.5, logit)                # try square root vs. spline transform.
plot(ag^1.5, logit)               # try 1.5 power

latex(fit)                        # invokes latex.lrm, creates fit.tex
# Draw a nomogram for the model fit
plot(nomogram(fit))
```

```
# Compose R 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 `Glm`)
- `gam`: Generalized additive models
- `rpart`: Like original CART with surrogate splits for missings, censored data extension (Atkinson & Therneau)
- `validate.rpart`: in `rms`; validates recursive partitioning with respect to certain accuracy indexes

- `loess`: multi-dimensional scatterplot smoother

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

Chapter 7

Modeling Longitudinal Responses using Generalized Least Squares

7.1 Notation

- N subjects
- Subject i ($i = 1, 2, \ldots, N$) has n_i responses measured at times $t_{i1}, t_{i2}, \ldots, t_{in_i}$
- Response at time t for subject i: Y_{it}
- Subject i has baseline covariates X_i
- Generally the response measured at time $t_{i1} = 0$ is a covariate in X_i instead of being the first measured response Y_{i0}
- Time trend in response is modeled with k parameters so that the time "main effect" has

k d.f.

- Let the basis functions modeling the time effect be $g_1(t), g_2(t), \ldots, g_k(t)$

7.2 Model Specification for Effects on $E(Y)$

7.2.1 Common Basis Functions

- k dummy variables for $k+1$ unique times (assumes no functional form for time but may spend many d.f.)
- $k = 1$ for linear time trend, $g_1(t) = t$
- k–order polynomial in t
- $k+1$–knot restricted cubic spline (one linear term, $k-1$ nonlinear terms)

7.2.2 Model for Mean Profile

- A model for mean time-response profile without interactions between time and any X:

$$E[Y_{it}|X_i] = X_i\beta + \gamma_1 g_1(t) + \gamma_2 g_2(t) + \ldots + \gamma_k g_k(t)$$

- Model with interactions between time and some X's: add product terms for desired interaction effects
- Example: To allow the mean time trend for subjects in group 1 (reference group) to be arbitrarily different from time trend for subjects in group 2, have a dummy variable for group 2, a time "main effect" curve with k d.f. and all k products of these time components with the dummy variable for group 2

7.2.3 Model Specification for Treatment Comparisons

- In studies comparing two or more treatments, a response is often measured at baseline (pre-randomization)
- Analyst has the option to use this measurement as Y_{i0} or as part of X_i

- Jim Rochon (Rho, Inc., Chapel Hill NC) has the following comments about this:

> For RCTs, I draw a sharp line at the point when the intervention begins. The LHS [left hand side of the model equation] is reserved for something that is a response to treatment. Anything before this point can potentially be included as a covariate in the regression model. This includes the "baseline" value of the outcome variable. Indeed, the best predictor of the outcome at the end of the study is typically where the patient began at the beginning. It drinks up a lot of variability in the outcome; and, the effect of other covariates is typically mediated through this variable.
>
> I treat anything after the intervention begins as an outcome. In the western scientific method, an "effect" must follow the "cause" even if by a split second.
>
> Note that an RCT is different than a cohort study. In a cohort study, "Time 0" is not terribly meaningful. If we want to model, say, the trend over time, it would be legitimate, in my view, to include the "baseline" value on the LHS of that regression model.
>
> Now, even if the intervention, e.g., surgery, has an immediate effect, I would include still reserve the LHS for anything that might legitimately be considered as the response to the intervention. So, if we cleared a blocked artery and then measured the MABP, then that would still be included on the LHS.
>
> Now, it could well be that most of the therapeutic effect occurred by the time that the first repeated measure was taken, and then levels off. Then, a plot of the means would essentially be two parallel lines and the treatment effect is the distance between the lines, i.e., the difference in the intercepts.
>
> If the linear trend from baseline to Time 1 continues beyond Time 1, then the lines will have a common intercept but the slopes will diverge. Then, the treatment effect will the difference in slopes.
>
> One point to remember is that the estimated intercept is the value at time 0 that we predict from the set of repeated measures post randomization. In the first case above, the model will predict different intercepts even though randomization would suggest that they would start from the same place. This is because we were asleep at the switch and didn't record the "action" from baseline to time 1. In the second case, the model will predict the same intercept values because the linear trend from baseline to time 1 was continued thereafter.
>
> More importantly, there are considerable benefits to including it as a covariate on the RHS. The baseline value tends to be the best predictor of the outcome post-randomization, and this maneuver increases the precision of the estimated treatment effect. Additionally, any other prognostic factors correlated with the outcome variable will also be correlated with the baseline value of that outcome, and this has two important consequences. First, this greatly reduces the need to enter a large number of prognostic factors as covariates in the linear models. Their effect is already mediated through the baseline value of the outcome variable. Secondly, any imbalances across the treatment arms in important prognostic factors will induce an imbalance across the treatment arms in the baseline value of the outcome. Including the baseline value thereby reduces the need to enter these variables as covariates in the linear models.

Stephen Senn[148] states that temporally and logically, a "baseline cannot be a *response* to treatment", so baseline and response cannot be modeled in an integrated framework.

> . . . one should focus clearly on 'outcomes' as being the only values that can be influenced by treatment and examine critically any schemes that assume that these are linked in some rigid and deterministic view to 'baseline' values. An alternative tradition sees a baseline as being merely one of a number of measurements capable of improving predictions of outcomes and models it in this way.

The final reason that baseline cannot be modeled as the response at time zero is that many studies have inclusion/exclusion criteria that include cutoffs on the baseline variable. In other words, the baseline measurement comes from a truncated distribution. In general it is not appropriate to model the baseline with the same distributional shape as the follow-up measurements. Thus the approach recommended by Liang and Zeger[113] and Liu *et al.*[116] are problematic[a].

7.3 Modeling Within-Subject Dependence

- Random effects and mixed effects models have become very popular
- Disadvantages:
  - Induced correlation structure for Y may be unrealistic
  - Numerically demanding

[a] In addition to this, one of the paper's conclusions that analysis of covariance is not appropriate if the population means of the baseline variable are not identical in the treatment groups is not correct[148]. See[97] for a rebuke of[116].

  - Require complex approximations for distributions of test statistics

- Extended linear model (with no random effects) is a logical extension of the univariate model (e.g., few statisticians use subject random effects for univariate Y)
- This was known as growth curve models and generalized least squares[68, 137] and was developed long before mixed effect models became popular
- Pinheiro and Bates (Section 5.1.2) state that "in some applications, one may wish to avoid incorporating random effects in the model to account for dependence among observations, choosing to use the within-group component Λ_i to directly model variance-covariance structure of the response."
- We will assume that $Y_{it}|X_i$ has a multivariate normal distribution with mean given above and with variance-covariance matrix V_i, an $n_i \times n_i$ matrix that is a function of $t_{i1}, \ldots, t_{in_i}$

- We further assume that the diagonals of V_i are all equal
- Procedure can be generalized to allow for heteroscedasticity over time or with respect to X (e.g., males may be allowed to have a different variance than females)
- This *extended linear model* has the following assumptions:
  - all the assumptions of OLS at a single time point including correct modeling of predictor effects and univariate normality of responses conditional on X
  - the distribution of two responses at two different times for the same subject, conditional on X, is bivariate normal with a specified correlation coefficient
  - the joint distribution of all n_i responses for the i^{th} subject is multivariate normal with the given correlation pattern (which implies the previous two distributional assumptions)
  - responses from any times for any two dif-

ferent subjects are uncorrelated

What Methods To Use for Repeated Measurements / Serial Data? [a,b]

	Repeated Measures ANOVA	GEE	Mixed Effects Model	GLS	LOCF	Summary Statistic[c]
Assumes normality	$\times$		$\times$	$\times$		
Assumes independence of measurements within subject	$\times^d$	$\times^e$				
Assumes a correlation structure[f]	$\times$	$\times^g$	$\times$	$\times$		
Requires same measurement times for all subjects	$\times$				?	
Does not allow smooth modeling of time to save d.f.	$\times$					
Does not allow adjustment for baseline covariates	$\times$					
Does not easily extend to non-continuous Y	$\times$			$\times$		
Loses information by not using intermediate measurements					$\times^h$	$\times$
Does not allow widely varying # of observations per subject	$\times$	$\times^i$			$\times$	$\times^j$
Does not allow for subjects to have distinct trajectories[k]	$\times$	$\times$		$\times$	$\times$	
Assumes subject-specific effects are Gaussian			$\times$			
Badly biased if non-random dropouts	?	$\times$			$\times$	
Biased in general					$\times$	
Hard to get tests & CLs			$\times^l$		$\times^m$	
Requires large # subjects/clusters		$\times$				
SEs are wrong	$\times^n$				$\times$	
Assumptions are not verifiable in small samples	$\times$	N/A	$\times$	$\times$	$\times$	
Does not extend to complex settings such as time-dependent covariates and dynamic[o] models	$\times$		$\times$	$\times$	$\times$	?

[a] Thanks to Charles Berry, Brian Cade, Peter Flom, Bert Gunter, and Leena Choi for valuable input.

[b] GEE: generalized estimating equations; GLS: generalized least squares; LOCF: last observation carried forward.

[c] E.g., compute within-subject slope, mean, or area under the curve over time. Assumes that the summary measure is an adequate summary of the time profile and assesses the relevant treatment effect.

[d] Unless one uses the Huynh-Feldt or Greenhouse-Geisser correction

[e] For full efficiency, if using the working independence model

[f] Or requires the user to specify one

[g] For full efficiency of regression coefficient estimates

[h] Unless the last observation is missing

[i] The cluster sandwich variance estimator used to estimate SEs in GEE does not perform well in this situation, and neither does the working independence model because it does not weight subjects properly.

[j] Unless one knows how to properly do a weighted analysis

[k] Or uses population averages

[l] Unline GLS, does not use standard maximum likelihood methods yielding simple likelihood ratio χ^2 statistics. Requires high-dimensional integration to marginalize random effects, using complex approximations, and if using SAS, unintuitive d.f. for the various tests.

[m] Because there is no correct formula for SE of effects; ordinary SEs are not penalized for imputation and are too small

[n] If correction not applied

[o] E.g., a model with a predictor that is a lagged value of the response variable

Gardiner *et al.*[64] compared several longitudinal data models, especially with regard to assumptions and how regression coefficients are estimated. Peters *et al.*[135] have an empirical study confirming that the "use all available data" approach of likelihood–based longitudinal models makes imputation of follow-up measurements unnecessary.

7.4 Parameter Estimation Procedure

- Generalized least squares
- Like weighted least squares but uses a covariance matrix that is not diagonal
- Each subject can have her own shape of V_i due to each subject being measured at a different set of times
- Maximum likelihood
- Newton-Raphson or other trial-and-error methods used for estimating parameters

- For small number of subjects, advantages in using REML (restricted maximum likelihood) instead of ordinary MLE [49, Section 5.3], [136, Chapter 5],[68] (esp. to get more unbiased estimate of the covariance matrix)
- When imbalances are not severe, OLS fitted ignoring subject identifiers may be efficient
  - But OLS standard errors will be too small as they don't take intra-cluster correlation into account
  - May be rectified by substituting covariance matrix estimated from Huber-White cluster sandwich estimator or from cluster bootstrap
- When imbalances are severe and intra-subject correlations are strong, OLS is not expected to be efficient because it gives equal weight to each observation
  - a subject contributing two distant observations receives $\frac{1}{5}$ the weight of a subject having 10 tightly-spaced observations

7.5 Common Correlation Structures

- Usually restrict ourselves to *isotropic* correlation structures — correlation between responses within subject at two times depends only on a measure of distance between the two times, not the individual times
- We simplify further and assume depends on $|t_1 - t_2|$
- Can speak interchangeably of correlations of residuals within subjects or correlations between responses measured at different times on the same subject, conditional on covariates X
- Assume that the correlation coefficient for Y_{it_1} vs. Y_{it_2} conditional on baseline covariates X_i for subject i is $h(|t_1 - t_2|, \rho)$, where ρ is a vector (usually a scalar) set of fundamental correlation parameters
- Some commonly used structures when times are continuous and are not equally spaced [136,

Section 5.3.3] (`nlme` correlation function names are at the right if the structure is implemented in `nlme`):

Compound symmetry : $h = \rho$ if $t_1 \neq t_2$, 1 if $t_1 = t_2$ `nlme corCompSymm`
(Essentially what two-way ANOVA assumes)

Autoregressive-moving average lag 1 : $h = \rho^{|t_1 - t_2|} = \rho^s$ `corCAR1`
where $s = |t_1 - t_2|$

Exponential : $h = \exp(-s/\rho)$ `corExp`

Gaussian : $h = \exp[-(s/\rho)^2]$ `corGaus`

Linear : $h = (1 - s/\rho)I(s < \rho)$ `corLin`

Rational quadratic : $h = 1 - (s/\rho)^2/[1 + (s/\rho)^2]$ `corRatio`

Spherical : $h = [1 - 1.5(s/\rho) + 0.5(s/\rho)^3]I(s < \rho)$ `corSpher`

Linear exponent AR(1) : $h = \rho^{d_{min} + \delta \frac{s - d_{min}}{d_{max} - d_{min}}}$, 1 if $t_1 = t_2$[151]

The structures 3–7 use ρ as a scaling parameter, not as something restricted to be in $[0, 1]$

7.6 Checking Model Fit

- Constant variance assumption: usual residual plots
- Normality assumption: usual qq residual plots
- Correlation pattern: **Variogram**

- Estimate correlations of all possible pairs of residuals at different time points
- Pool all estimates at same absolute difference in time s
- Variogram is a plot with $y = 1 - \hat{h}(s, \rho)$ vs. s on the x-axis
- Superimpose the theoretical variogram assumed by the model

7.7 R Software

- Nonlinear mixed effects model package of Pinheiro & Bates in S-PLUS and R
- For linear models, fitting functions are
  - `lme` for mixed effects models
  - `gls` for generalized least squares without random effects
- For this version the rms package has `Gls` so that many features of rms can be used:

`anova` : all partial Wald tests, test of linearity, pooled tests

`summary` : effect estimates (differences in $\hat{Y}$) and confidence limits, can be plotted

`plot` : continuous effect plots

`nomogram` : nomogram

`Function` : generate R function code for fitted model

`latex` : LaTeX representation of fitted model

In addition, `Gls` has a bootstrap option (hence you do not use rms's `bootcov` for `Gls` fits).
To get regular `gls` functions named `anova` (for likelihood ratio tests, AIC, etc.) or `summary` use `anova.gls` or `summary.gls`

- `nlme` package has many graphics and fit-checking functions
- Several functions will be demonstrated in the case study

7.8 Case Study

Consider the dataset in Table 6.9 of Davis [46, pp. 161-163] from a multicenter, randomized controlled trial of botulinum toxin type B (BotB) in patients with cervical dystonia from nine U.S. sites.

- Randomized to placebo ($N = 36$), 5000 units of BotB ($N = 36$), 10,000 units of BotB ($N = 37$)
- Response variable: total score on Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS), measuring severity, pain, and disability of cervical dystonia (high scores mean more impairment)
- TWSTRS measured at baseline (week 0) and weeks 2, 4, 8, 12, 16 after treatment began
- Dataset `cdystonia` from web site

7.8.1 Graphical Exploration of Data

```
require(rms)
```

```
getHdata(cdystonia)
attach(cdystonia)

# Construct unique subject ID
uid ← with(cdystonia, factor(paste(site, id)))

# What is the frequency of each pattern of subjects' time points?
table(tapply(week, uid,
             function(w) paste(sort(unique(w)), collapse=' ')))
```

```
            0         0 2 4   0 2 4 12 16       0 2 4 8
            1             1             3             1
  0 2 4 8 12 0 2 4 8 12 16    0 2 4 8 16   0 2 8 12 16
            1            94             1             2
 0 4 8 12 16       0 4 8 16
            4             1
```

```
# Plot raw data, superposing subjects
xl ← xlab('Week'); yl ← ylab('TWSTRS-total score')
ggplot(cdystonia, aes(x=week, y=twstrs, color=factor(id))) +
       geom_line() + xl + yl +
       facet_grid(treat ~ site) + guides(color=FALSE) # Fig. 7.1
```

```
# Show quartiles
ggplot(cdystonia, aes(x=week, y=twstrs)) + xl + yl + ylim(0, 70) +
  stat_summary(fun.data="median_hilow", conf.int=0.5, geom='smooth') +
  facet_wrap(~ treat, nrow=2)  # Fig. 7.2
```

```
# Show means with bootstrap nonparametric CLs
ggplot(cdystonia, aes(x=week, y=twstrs)) + xl + yl + ylim(0, 70) +
  stat_summary(fun.data="mean_cl_boot", geom='smooth') +
  facet_wrap(~ treat, nrow=2)  # Fig. 7.3
```

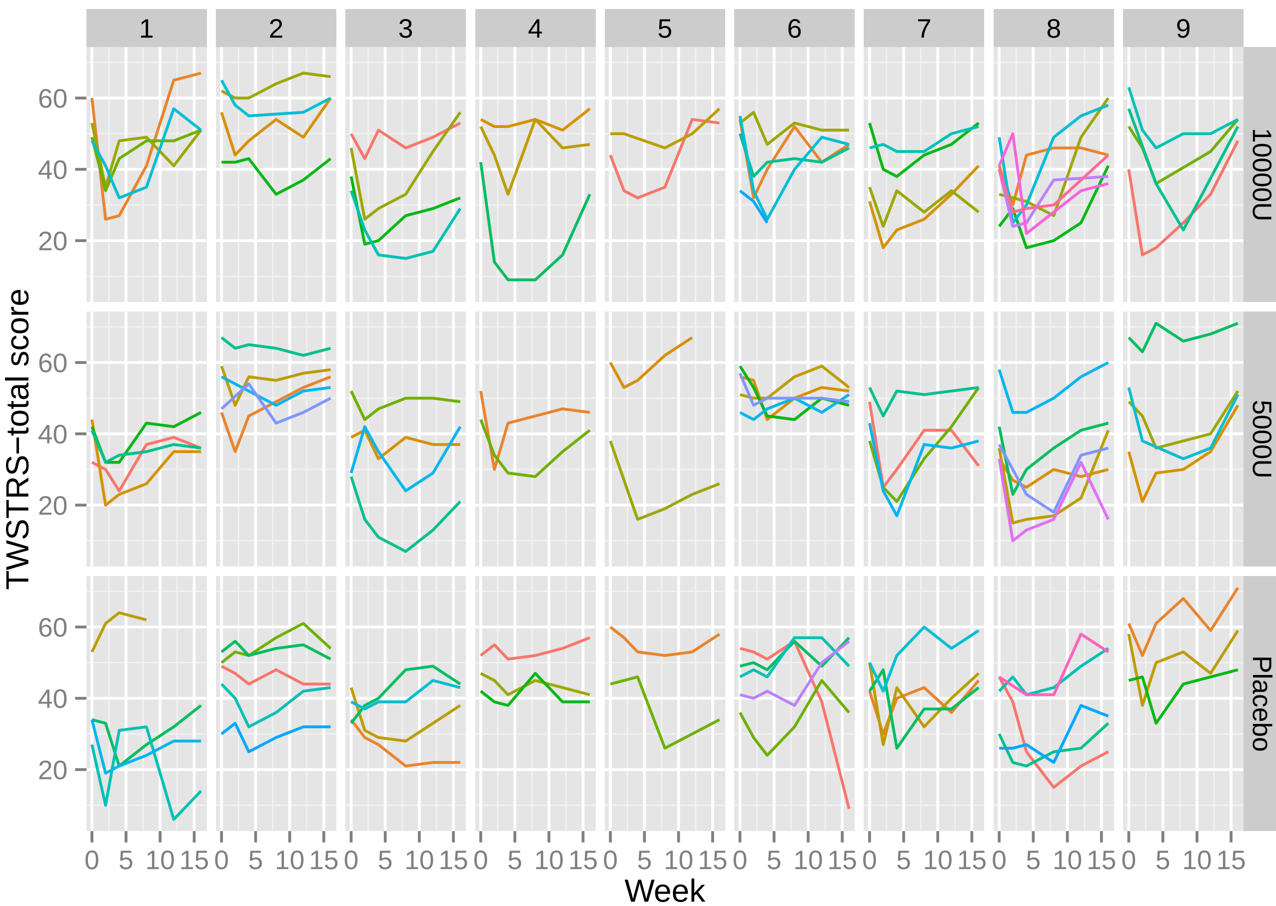


Figure 7.1: Time profiles for individual subjects, stratified by study site and dose

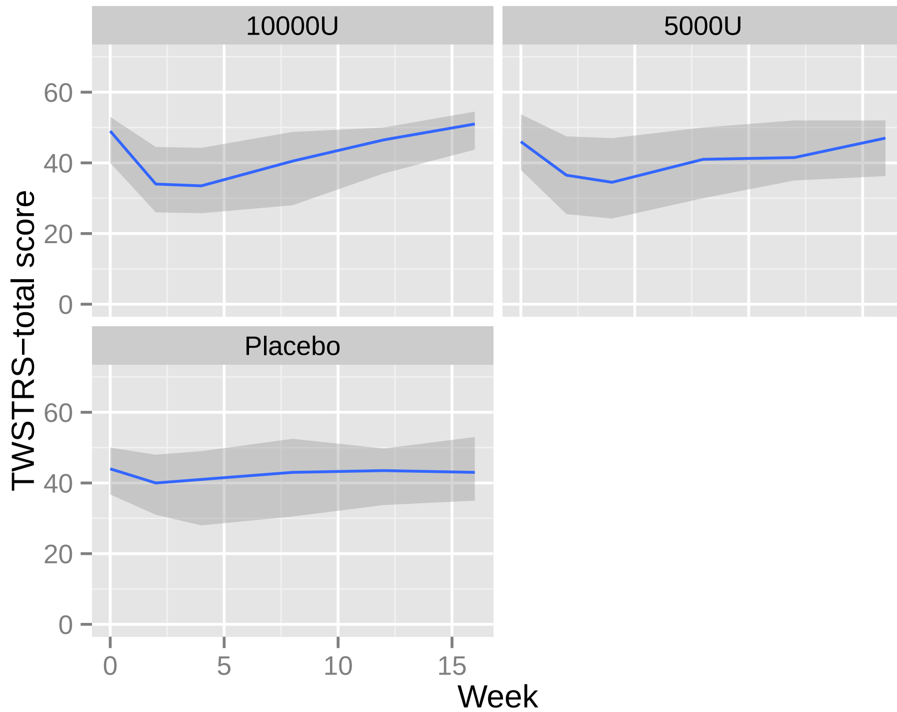


Figure 7.2: Quartiles of `TWSTRS` stratified by dose

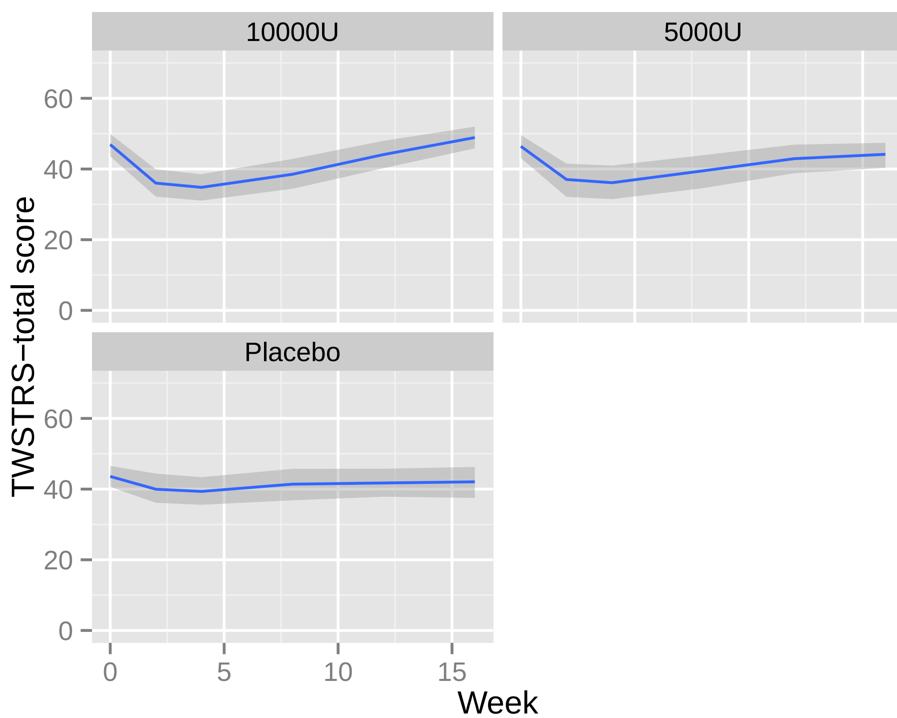


Figure 7.3: Mean responses and nonparametric bootstrap 0.95 confidence limits for population means, stratified by dose

Model with Y_{i0} as Baseline Covariate

```
baseline ← subset(data.frame(cdystonia,uid), week == 0, -week)
baseline ← upData(baseline, rename=c(twstrs='twstrs0'), print=FALSE)
followup ← subset(data.frame(cdystonia,uid), week > 0,
                  c(uid,week,twstrs))
rm(uid)
both     ← merge(baseline, followup, by='uid')

dd       ← datadist(both)
options(datadist='dd')
```

7.8.2 Using Generalized Least Squares

We stay with baseline adjustment and use a variety of correlation structures, with constant variance. Time is modeled as a restricted cubic spline with 3 knots, because there are only 3 unique interior values of `week`.

```
require(nlme)
```

```
cp ← list(corCAR1,corExp,corCompSymm,corLin,corGaus,corSpher)
z  ← vector('list',length(cp))
for(k in 1:length(cp)) {
  z[[k]] ← gls(twstrs ~ treat * rcs(week, 3) + rcs(twstrs0, 3) +
               rcs(age, 4) * sex, data=both,
               correlation=cp[[k]](form = ~week | uid))
}
```

```
anova(z[[1]],z[[2]],z[[3]],z[[4]],z[[5]],z[[6]])
```

```
       Model df      AIC      BIC    logLik
z[[1]]     1 20 3553.906 3638.357 -1756.953
z[[2]]     2 20 3553.906 3638.357 -1756.953
z[[3]]     3 20 3587.974 3672.426 -1773.987
z[[4]]     4 20 3575.079 3659.531 -1767.540
z[[5]]     5 20 3621.081 3705.532 -1790.540
z[[6]]     6 20 3570.958 3655.409 -1765.479
```

AIC computed above is set up so that smaller values are best. From this the continuous-time AR1 and exponential structures are tied for the best. For the remainder of the analysis use `corCAR1`, using `Gls`. 1

```
a ← Gls(twstrs ∼ treat * rcs(week, 3) + rcs(twstrs0, 3) +
        rcs(age, 4) * sex, data=both,
        correlation=corCAR1(form=∼week | uid))
```

```
print(a, latex=TRUE)
```

Generalized Least Squares Fit by REML

```
Gls(model = twstrs ~ treat * rcs(week, 3) + rcs(twstrs0, 3) +
    rcs(age, 4) * sex, data = both, correlation = corCAR1(form = ~week |
    uid))
```

Obs	522	Log-restricted-likelihood	-1756.95
Clusters	108	Model d.f.	17
g	11.334	σ	8.5917
		d.f.	504

	Coef	S.E.	t	Pr($>|t|$)
Intercept	-0.3093	11.8804	-0.03	0.9792
treat=5000U	0.4344	2.5962	0.17	0.8672
treat=Placebo	7.1433	2.6133	2.73	0.0065
week	0.2879	0.2973	0.97	0.3334
week'	0.7313	0.3078	2.38	0.0179
twstrs0	0.8071	0.1449	5.57	< 0.0001
twstrs0'	0.2129	0.1795	1.19	0.2360
age	-0.1178	0.2346	-0.50	0.6158
age'	0.6968	0.6484	1.07	0.2830
age''	-3.4018	2.5599	-1.33	0.1845
sex=M	24.2802	18.6208	1.30	0.1929
treat=5000U * week	0.0745	0.4221	0.18	0.8599
treat=Placebo * week	-0.1256	0.4243	-0.30	0.7674
treat=5000U * week'	-0.4389	0.4363	-1.01	0.3149
treat=Placebo * week'	-0.6459	0.4381	-1.47	0.1411
age * sex=M	-0.5846	0.4447	-1.31	0.1892
age' * sex=M	1.4652	1.2388	1.18	0.2375
age'' * sex=M	-4.0338	4.8123	-0.84	0.4023

```
Correlation Structure: Continuous AR(1)
 Formula: ~week | uid
 Parameter estimate(s):
      Phi
0.8666689
```

$\hat{\rho} = 0.8672$, the estimate of the correlation between two measurements taken one week apart on the same subject. The estimated correlation for measurements 10 weeks apart is $0.8672^{10} = 0.24$.

```
v ← Variogram(a, form=~ week | uid)
plot(v)  # Figure 7.4
```

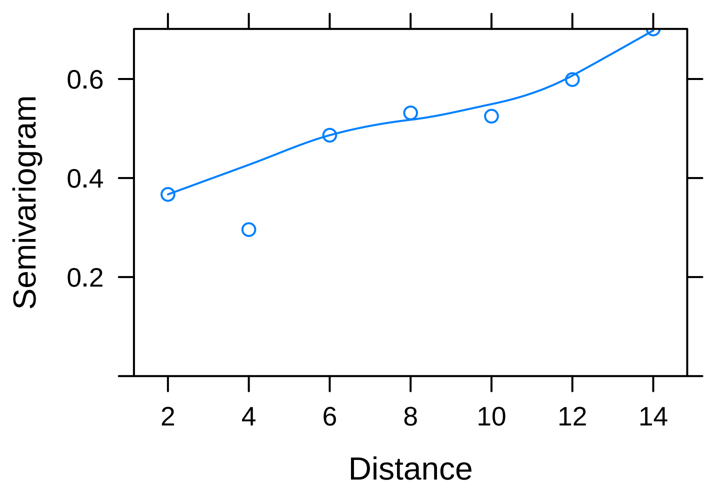


Figure 7.4: Variogram, with assumed correlation pattern superimposed

Check constant variance and normality assumptions:

```
both$resid ← resid(a); both$fitted ← fitted(a)
yl ← ylab('Residuals')
p1 ← ggplot(both, aes(x=fitted, y=resid)) + geom_point() +
```

```
      facet_grid(~ treat) + yl
p2 ← ggplot(both, aes(x=twstrs0, y=resid)) + geom_point() + yl
p3 ← ggplot(both, aes(x=week, y=resid)) + yl +
      stat_summary(fun.data="mean_sdl", geom='smooth') + ylim(−20,20)
p4 ← ggplot(both, aes(sample=resid)) + stat_qq() + yl
pmggplot(p1, p2, p3, p4)   # Figure 7.5
```

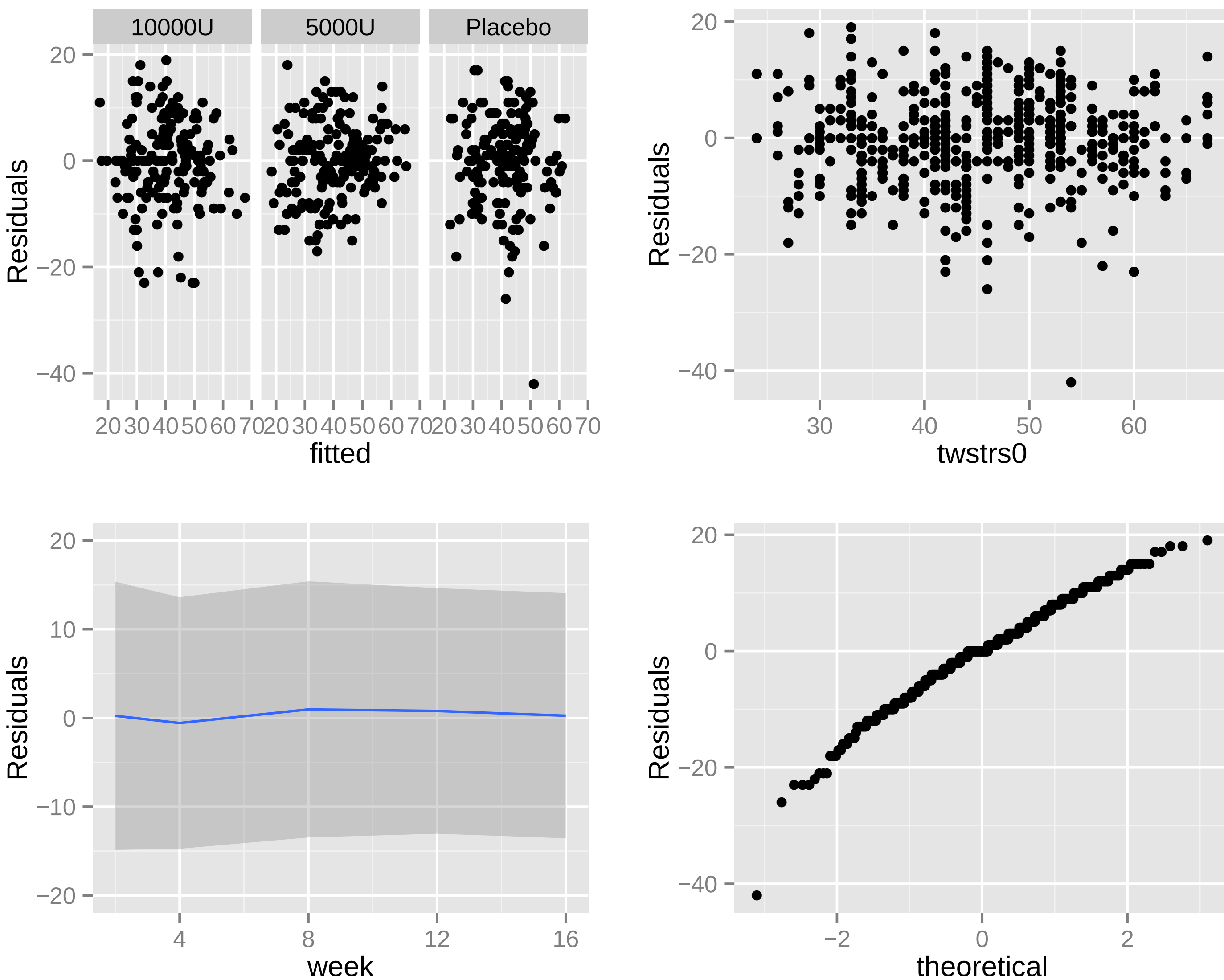


Figure 7.5: Three residual plots to check for absence of trends in central tendency and in variability. Upper right panel shows the baseline score on the x-axis. Bottom left panel shows the mean $\pm 2\times$SD. Bottom right panel is the QQ plot for checking normality of residuals from the GLS fit.

Now get hypothesis tests, estimates, and graphically interpret the model.

Table 7.2: Wald Statistics for `twstrs`

	χ^2	d.f.	P
treat (Factor+Higher Order Factors)	22.11	6	0.0012
All Interactions	14.94	4	0.0048
week (Factor+Higher Order Factors)	77.27	6	< 0.0001
All Interactions	14.94	4	0.0048
Nonlinear (Factor+Higher Order Factors)	6.61	3	0.0852
twstrs0	233.83	2	< 0.0001
Nonlinear	1.41	1	0.2354
age (Factor+Higher Order Factors)	9.68	6	0.1388
All Interactions	4.86	3	0.1826
Nonlinear (Factor+Higher Order Factors)	7.59	4	0.1077
sex (Factor+Higher Order Factors)	5.67	4	0.2252
All Interactions	4.86	3	0.1826
treat × week (Factor+Higher Order Factors)	14.94	4	0.0048
Nonlinear	2.27	2	0.3208
Nonlinear Interaction : f(A,B) vs. AB	2.27	2	0.3208
age × sex (Factor+Higher Order Factors)	4.86	3	0.1826
Nonlinear	3.76	2	0.1526
Nonlinear Interaction : f(A,B) vs. AB	3.76	2	0.1526
TOTAL NONLINEAR	15.03	8	0.0586
TOTAL INTERACTION	19.75	7	0.0061
TOTAL NONLINEAR + INTERACTION	28.54	11	0.0027
TOTAL	322.98	17	< 0.0001

```
latex(anova(a), file='', label='longit-anova')  # Table 7.2
```

```
plot(anova(a))     # Figure 7.6
```

```
ylm ← ylim(25, 60)
p1 ← ggplot(Predict(a, week, treat, conf.int=FALSE),
            adj.subtitle=FALSE, legend.position='top') + ylm
```

```
p2 ← ggplot(Predict(a, twstrs0), adj.subtitle=FALSE) + ylm
```

```
p3 ← ggplot(Predict(a, age, sex), adj.subtitle=FALSE,
            legend.position='top') + ylm
```

```
pmggplot(p1, p2, p3)    # Figure 7.7
```

```
latex(summary(a), file='', table.env=FALSE) # Shows for week 8
```

	Low	High	Δ	Effect	S.E.	Lower 0.95	Upper 0.95
week	4	12	8	6.69100	1.10570	4.5238	8.8582
twstrs0	39	53	14	13.55100	0.88618	11.8140	15.2880
age	46	65	19	2.50270	2.05140	-1.5179	6.5234
treat — 5000U:10000U	1	2		0.59167	1.99830	-3.3249	4.5083
treat — Placebo:10000U	1	3		5.49300	2.00430	1.5647	9.4212
sex — M:F	1	2		-1.08500	1.77860	-4.5711	2.4011

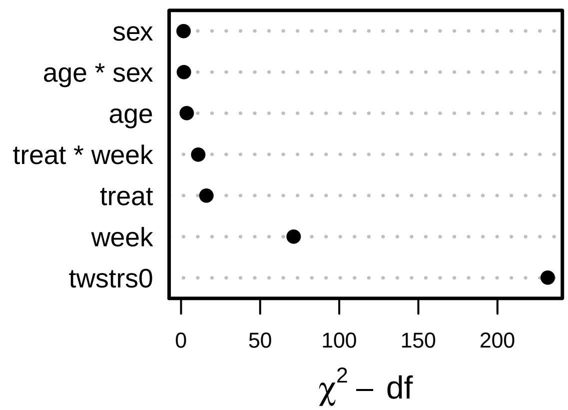


Figure 7.6: Results of `anova.rms` from generalized least squares fit with continuous time AR1 correlation structure

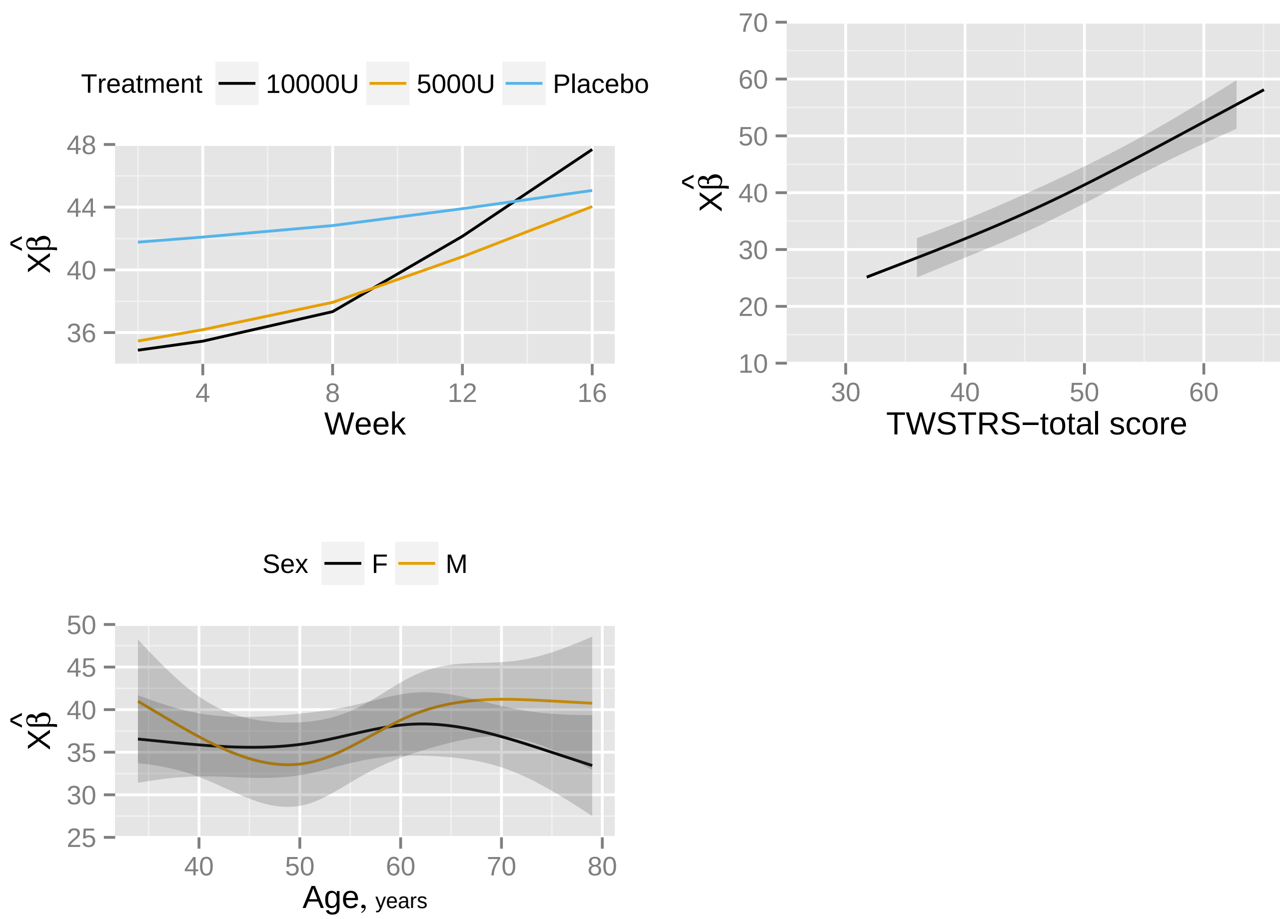


Figure 7.7: Estimated effects of time, baseline TWSTRS, age, and sex

```
# To get results for week 8 for a different reference group
# for treatment, use e.g. summary(a, week=4, treat='Placebo')

# Compare low dose with placebo, separately at each time
k1 ← contrast(a, list(week=c(2,4,8,12,16), treat='5000U'),
                 list(week=c(2,4,8,12,16), treat='Placebo'))
options(width=80)
print(k1, digits=4)
```

```
   week twstrs0 age sex Contrast  S.E.   Lower   Upper     Z Pr(>|z|)
1     2      46  56   F   -6.309 2.104 -10.432 -2.1859 -3.00   0.0027
2     4      46  56   F   -5.909 1.816  -9.468 -2.3490 -3.25   0.0011
3     8      46  56   F   -4.901 2.015  -8.850 -0.9527 -2.43   0.0150
4*   12      46  56   F   -3.066 1.748  -6.493  0.3607 -1.75   0.0795
5*   16      46  56   F   -1.024 2.100  -5.139  3.0924 -0.49   0.6260

Redundant contrasts are denoted by *

Confidence intervals are 0.95 individual intervals
```

```
# Compare high dose with placebo
k2 ← contrast(a, list(week=c(2,4,8,12,16), treat='10000U'),
                 list(week=c(2,4,8,12,16), treat='Placebo'))
print(k2, digits=4)
```

```
   week twstrs0 age sex Contrast  S.E.   Lower  Upper     Z Pr(>|z|)
1     2      46  56   F   -6.892 2.074 -10.957 -2.827 -3.32   0.0009
2     4      46  56   F   -6.641 1.793 -10.155 -3.127 -3.70   0.0002
3     8      46  56   F   -5.493 2.004  -9.421 -1.565 -2.74   0.0061
4*   12      46  56   F   -1.761 1.738  -5.168  1.645 -1.01   0.3109
5*   16      46  56   F    2.617 2.087  -1.474  6.707  1.25   0.2099

Redundant contrasts are denoted by *

Confidence intervals are 0.95 individual intervals
```

```
k1 ← as.data.frame(k1[c('week', 'Contrast', 'Lower', 'Upper')])
p1 ← ggplot(k1, aes(x=week, y=Contrast)) + geom_point() + geom_line() +
     geom_errorbar(aes(ymin=Lower, ymax=Upper), width=0) +
     ylab('Low Dose − Placebo')
k2 ← as.data.frame(k2[c('week', 'Contrast', 'Lower', 'Upper')])
p2 ← ggplot(k2, aes(x=week, y=Contrast)) + geom_point() + geom_line() +
     geom_errorbar(aes(ymin=Lower, ymax=Upper), width=0) +
     ylab('High Dose − Placebo')
pmggplot(p1, p2)   # Figure 7.8
```

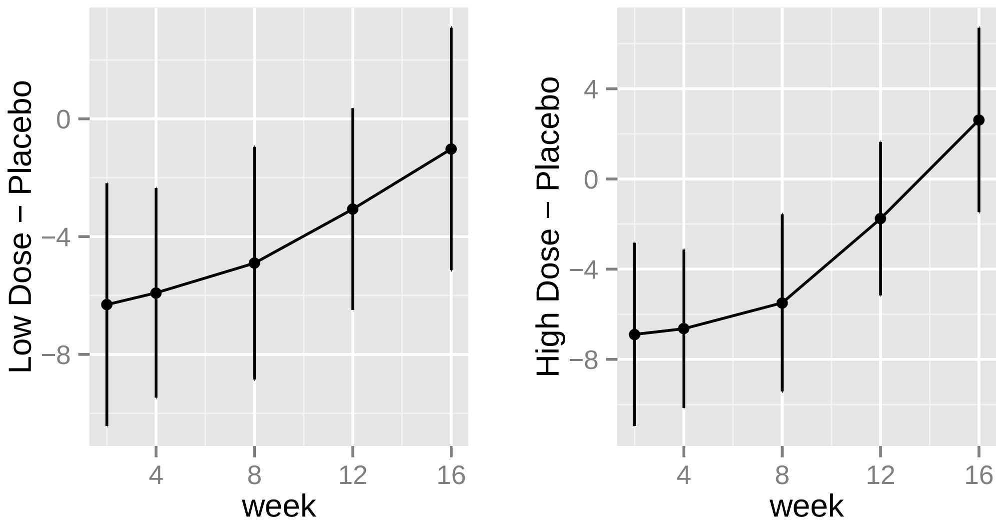


Figure 7.8: Contrasts and 0.95 confidence limits from GLS fit

Although multiple d.f. tests such as total treatment effects or treatment × time interaction tests are comprehensive, their increased degrees of freedom can dilute power. In a treatment comparison, treatment contrasts at the last time point (single d.f. tests) are often of major interest. Such contrasts are informed by all the measurements made by all subjects (up until dropout times) when a smooth time trend is assumed.

```
n ← nomogram(a, age=c(seq(20, 80, by=10), 85))
plot(n, cex.axis=.55, cex.var=.8, lmgp=.25)  # Figure 7.9
```

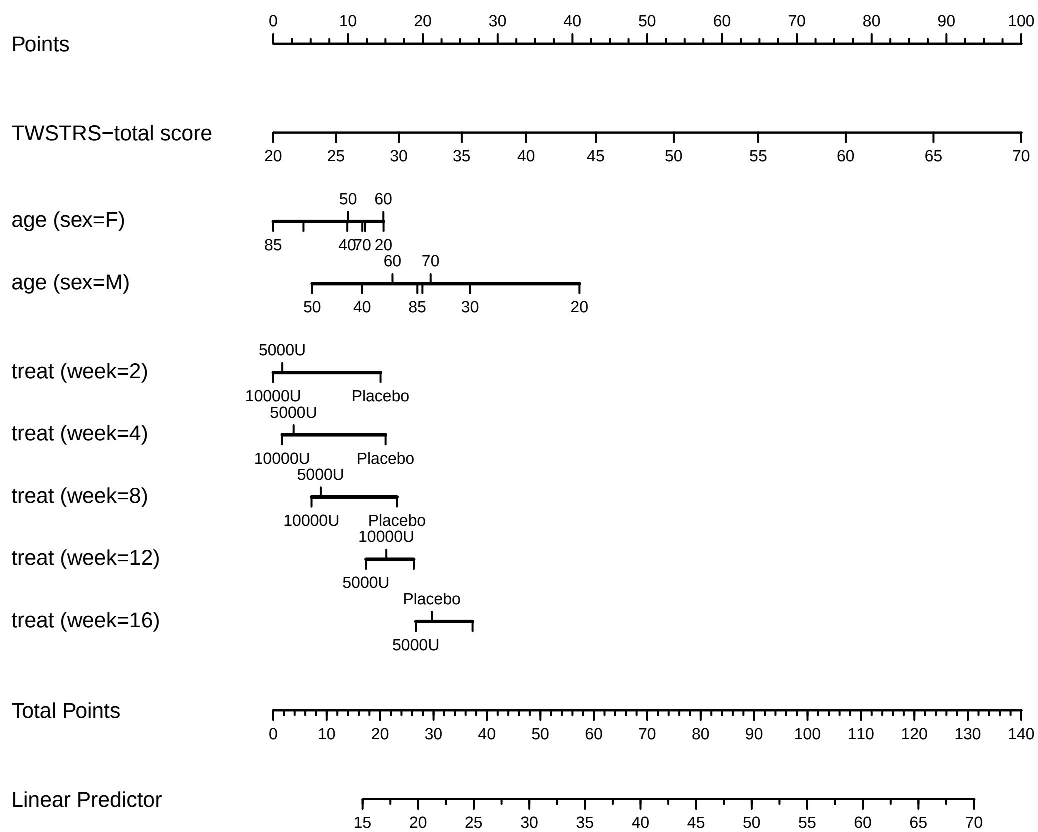


Figure 7.9: Nomogram from GLS fit. Second axis is the baseline score.

Chapter 8

Binary Logistic Regression

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

8.1 Model

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

- $O = \frac{P}{1-P}$

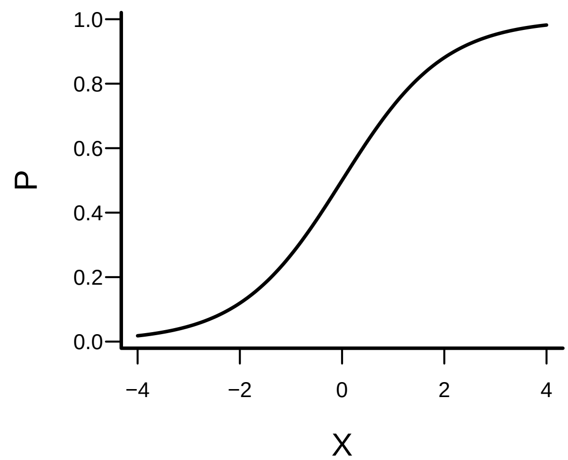


Figure 8.1: Logistic function

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

8.1.1 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 $d \to$ 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 pre-

dictor 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).

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

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

8.1.2 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

```
plot(0, 0, type="n", xlab="Risk for Subject Without Risk Factor",
     ylab="Increase in Risk",
     xlim=c(0,1), ylim=c(0,.6))   # Figure 8.2
i ← 0
or ← c(1.1,1.25,1.5,1.75,2,3,4,5,10)
for(h in or) {
  i ← i + 1
  p ← seq(.0001, .9999, length=200)
  logit ← log(p/(1 − p))  # same as qlogis(p)
  logit ← logit + log(h)  # modify by odds ratio
  p2 ← 1/(1 + exp(−logit))# same as plogis(logit)
  d ← p2 − p
  lines(p, d, lty=i)
  maxd ← max(d)
  smax ← p[d==maxd]
  text(smax, maxd + .02, format(h), cex=.6)
}
```

Let X_1 be a binary risk factor and let $A = \{X_2, \ldots, X_p\}$ be the other factors. Then the estimate of $\text{Prob}\{Y = 1|X_1 = 1, A\} - \text{Prob}\{Y = 1|X_1 = 0, A\}$ is

$$\frac{1}{1 + \exp -[\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p]} \\ - \frac{1}{1 + \exp -[\hat{\beta}_0 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p]}$$

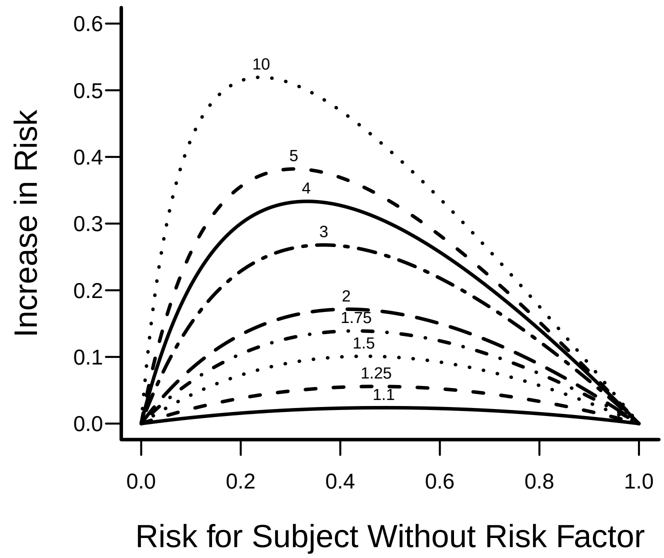


Figure 8.2: Absolute benefit as a function of risk of the event in a control subject and the relative effect (odds ratio) of the risk factor. The odds ratios are given for each curve.

$$= \frac{1}{1 + (\frac{1-\hat{R}}{\hat{R}})\exp(-\hat{\beta}_1)} - \hat{R},$$

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

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

8.1.3 Detailed Example

```
require(rms)
getHdata(sex.age.response)
d ← sex.age.response
dd ← datadist(d); options(datadist='dd')
f ← lrm(response ∼ sex + age, data=d)
fasr ← f   # Save for later
w ← function(...)
```

```
  with(d, {
    m ← sex=='male'
    f ← sex=='female'
    lpoints(age[f], response[f], pch=1)
    lpoints(age[m], response[m], pch=2)
    af ← cut2(age, c(45,55), levels.mean=TRUE)
    prop ← tapply(response, list(af, sex), mean,
                  na.rm=TRUE)
    agem ← as.numeric(row.names(prop))
    lpoints(agem, prop[,'female'],
            pch=4, cex=1.3, col='green')
    lpoints(agem, prop[,'male'],
            pch=5, cex=1.3, col='green')
    x ← rep(62, 4); y ← seq(.25, .1, length=4)
    lpoints(x, y, pch=c(1, 2, 4, 5),
            col=rep(c('blue','green'),each=2))
    ltext(x+5, y,
          c('F Observed','M Observed',
            'F Proportion','M Proportion'), cex=.8)
  } )   # Figure 8.3

plot(Predict(f, age=seq(34, 70, length=200), sex, fun=plogis),
     ylab='Pr[response]', ylim=c(-.02, 1.02), addpanel=w)
ltx ← function(fit) latex(fit, inline=TRUE, columns=54,
                          file='', after='$.', digits=3,
        size='Ssize', before='$X\\hat{\\beta}=')
ltx(f)
```

$X\hat{\beta} = -9.84 + 3.49[\text{male}] + 0.158\,\text{age}.$

```
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
```

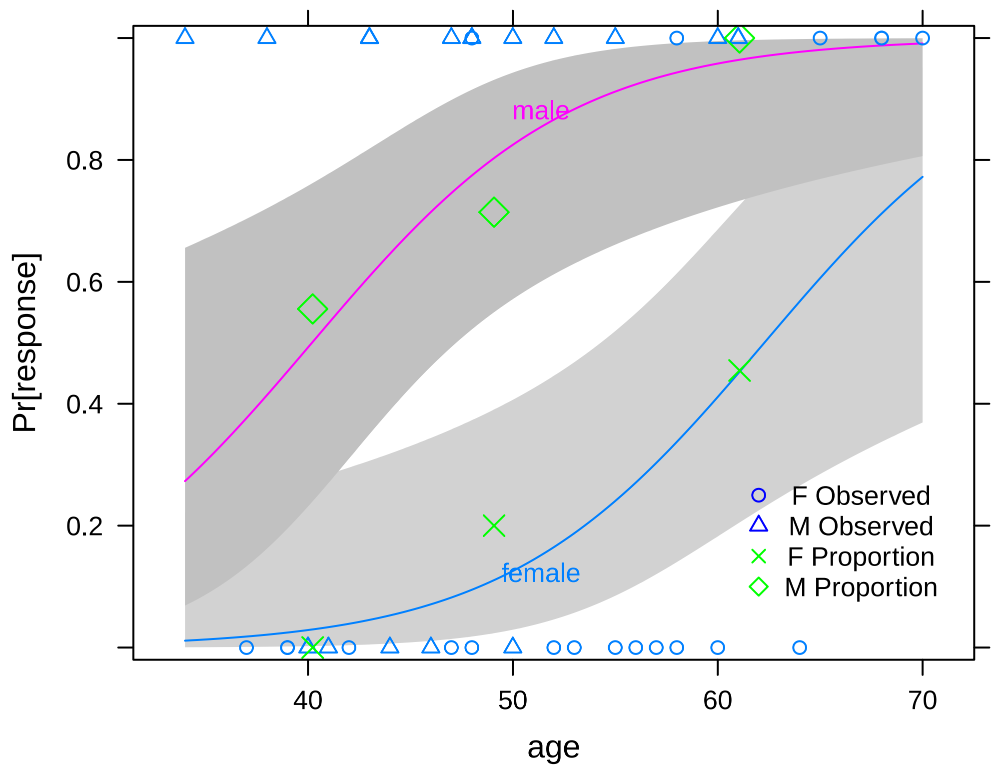


Figure 8.3: Data, subgroup proportions, and fitted logistic model, with 0.95 pointwise confidence bands

sex $\times$ response

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

Parameter	Estimate	Std Err	Wald χ^2	P
β_0	-0.847	0.488	3.015	
β_1	1.695	0.690	6.030	0.014

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.

```
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
```

Parameter	Estimate	Std Err	Wald χ^2	P
β_0	-2.734	1.838	2.213	
β_1	0.054	0.036	2.276	0.131

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) = 2.432

(Compare 2.432 with the Wald statistic 2.28.)

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

sex=F

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

sex=M

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

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

Parameter	Estimate	Std Err	Wald χ^2	P
β_0	-9.843	3.676	7.171	
β_1 (sex)	3.490	1.199	8.469	0.004
β_2 (age)	0.158	0.062	6.576	0.010

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) = 14.10
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 8.3. The fitted model is

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

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

8.1.4 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

8.2 Estimation

8.2.1 Maximum Likelihood Estimates

Like binomial case but Ps vary; $\hat{\beta}$ computed by trial and error using an iterative maximization technique

8.2.2 Estimation of Odds Ratios and Probabilities

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

8.2.3 Minimum Sample Size Requirement

- Simplest case: no covariates, only an intercept
- Consider margin of error of 0.1 in estimating $\theta = \text{Prob}[Y = 1]$ with 0.95 confidence
- Worst case: $\theta = \frac{1}{2}$
- Requires $n = 96$ observations[a]
- Single binary predictor with prevalence $\frac{1}{2}$: need $n = 96$ for each value of X
- Single continuous predictor X having a normal distribution with mean zero and standard deviation σ, with true $P = \frac{1}{1+\exp(-X)}$ so that the expected number of events is

[a] The general formula for the sample size required to achieve a margin of error of δ in estimating a true probability of θ at the 0.95 confidence level is $n = (\frac{1.96}{\delta})^2 \times \theta(1-\theta)$. Set $\theta = \frac{1}{2}$ for the worst case.

$\frac{n}{2}$. Compute mean of $\max_{X\in[-1.5,1.5]} |P - \hat{P}|$ over 1000 simulations for varying n and σ[b]

```
sigmas   ← c(.5, .75, 1, 1.25, 1.5, 1.75, 2, 2.5, 3, 4)
ns       ← seq(25, 300, by=25)
nsim     ← 1000
xs       ← seq(−1.5, 1.5, length=200)
pactual  ← plogis(xs)

dn ← list(sigma=format(sigmas), n=format(ns))
maxerr ← N1 ← array(NA, c(length(sigmas), length(ns)), dn)
require(rms)

i ← 0
for(s in sigmas) {
  i ← i + 1
  j ← 0
  for(n in ns) {
    j ← j + 1
    n1 ← maxe ← 0
    for(k in 1:nsim) {
      x ← rnorm(n, 0, s)
      P ← plogis(x)
      y ← ifelse(runif(n) ≤ P, 1, 0)
      n1 ← n1 + sum(y)
      beta ← lrm.fit(x, y)$coefficients
      phat ← plogis(beta[1] + beta[2] ∗ xs)
      maxe ← maxe + max(abs(phat − pactual))
    }
    n1 ← n1/nsim
    maxe ← maxe/nsim
    maxerr[i,j] ← maxe
    N1[i,j] ← n1
  }
}
xrange ← range(xs)
simerr ← llist(N1, maxerr, sigmas, ns, nsim, xrange)

maxe ← reShape(maxerr)
# Figure 8.4
xYplot(maxerr ∼ n, groups=sigma, data=maxe,
       ylab=expression(paste('Average Maximum  ',
            abs(hat(P) − P))),
```

[b] An average absolute error of 0.05 corresponds roughly to a 0.95 confidence interval margin of error of 0.1.

```
       type='l', lty=rep(1:2, 5), label.curve=FALSE,
       abline=list(h=c(.15, .1, .05), col=gray(.85)))
Key(.8, .68, other=list(cex=.7,
               title=expression(~~~~~~~~~~~sigma)))
```

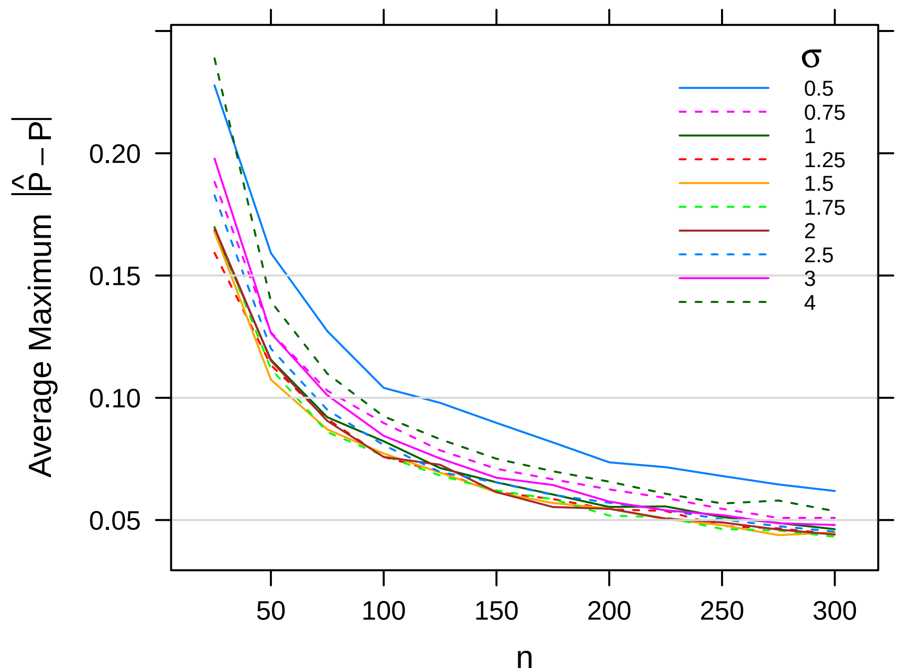


Figure 8.4: Simulated expected maximum error in estimating probabilities for $x \in [-1.5, 1.5]$ with a single normally distributed X with mean zero

8.3 Test Statistics

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

8.4 Residuals

Partial residuals (to check predictor transformations)

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

8.5 Assessment of Model Fit

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

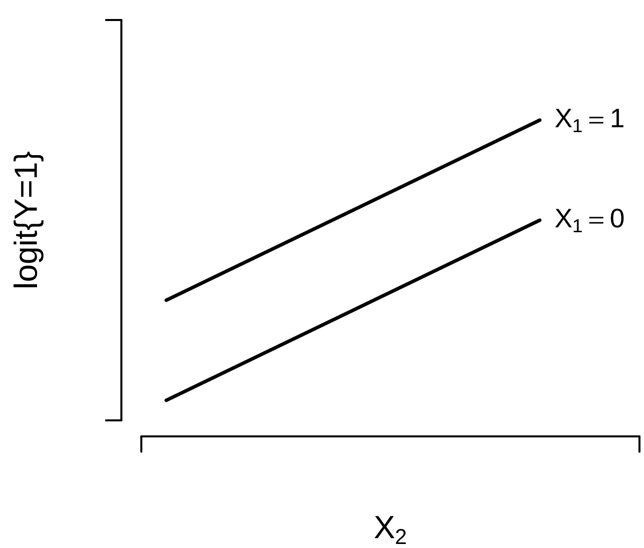


Figure 8.5: Logistic regression assumptions for one binary and one continuous predictor

```
getHdata(acath)
acath$sex ← factor(acath$sex, 0:1, c('male','female'))
dd ← datadist(acath); options(datadist='dd')
f ← lrm(sigdz ~ rcs(age, 4) * sex, data=acath)
```

```
w ← function(...)
  with(acath, {
    plsmo(age, sigdz, group=sex, fun=qlogis, lty='dotted',
          add=TRUE, grid=TRUE)
    af ← cut2(age, g=10, levels.mean=TRUE)
    prop ← qlogis(tapply(sigdz, list(af, sex), mean,
                         na.rm=TRUE))
    agem ← as.numeric(row.names(prop))
    lpoints(agem, prop[,'female'], pch=4, col='green')
    lpoints(agem, prop[,'male'],   pch=2, col='green')
  })   # Figure 8.6
plot(Predict(f, age, sex), ylim=c(-2,4), addpanel=w,
     label.curve=list(offset=unit(0.5, 'cm')))
```

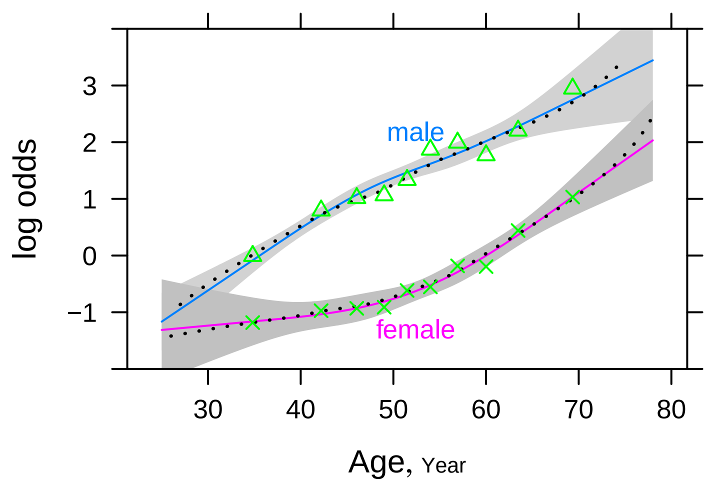


Figure 8.6: Logit proportions of significant coronary artery disease by sex and deciles of age for n=3504 patients, with spline fits (smooth curves). Spline fits are for $k = 4$ knots at age= 36, 48, 56, and 68 years, and interaction between age and sex is allowed. Shaded bands are pointwise 0.95 confidence limits for predicted log odds. Smooth nonparametric estimates are shown as dotted curves. Data courtesy of the Duke Cardiovascular Disease Databank.

- Can verify by plotting stratified proportions
- $\hat{P}$ = 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 of one or more predictors

$$\begin{aligned}\text{logit}\{Y=1|X\} &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_2' + \hat{\beta}_4 X_2'' \\ &= \hat{\beta}_0 + \hat{\beta}_1 X_1 + f(X_2),\end{aligned}$$

$$\begin{aligned}\text{logit}\{Y=1|X\} &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2' + \beta_4 X_2'' \\ &\quad + \beta_5 X_1 X_2 + \beta_6 X_1 X_2' + \beta_7 X_1 X_2''\end{aligned}$$

```
lr ← function(formula)
  {
    f ← lrm(formula, data=acath)
    stats ← f$stats[c('Model L.R.', 'd.f.')]
    cat('L.R. Chi-square:', round(stats[1],1),
        '  d.f.:', stats[2],'\n')
    f
  }
a ← lr(sigdz ∼ sex + age)
```

```
L.R. Chi-square: 766   d.f.: 2
```

```
b ← lr(sigdz ∼ sex * age)
```

```
L.R. Chi-square: 768.2   d.f.: 3
```

```r
c ← lr(sigdz ~ sex + rcs(age,4))
```

```
L.R. Chi-square: 769.4   d.f.: 4
```

```r
d ← lr(sigdz ~ sex * rcs(age,4))
```

```
L.R. Chi-square: 782.5   d.f.: 7
```

```r
lrtest(a, b)
```

```

Model 1: sigdz ~ sex + age
Model 2: sigdz ~ sex * age

L.R. Chisq       d.f.          P
 2.1964146  1.0000000  0.1383322
```

```r
lrtest(a, c)
```

```

Model 1: sigdz ~ sex + age
Model 2: sigdz ~ sex + rcs(age, 4)

L.R. Chisq       d.f.          P
 3.4502500  2.0000000  0.1781508
```

```r
lrtest(a, d)
```

```

Model 1: sigdz ~ sex + age
Model 2: sigdz ~ sex * rcs(age, 4)

  L.R. Chisq         d.f.            P
16.547036344  5.000000000  0.005444012
```

```r
lrtest(b, d)
```

```

Model 1: sigdz ~ sex * age
Model 2: sigdz ~ sex * rcs(age, 4)

  L.R. Chisq         d.f.            P
14.350621767  4.000000000  0.006256138
```

```r
lrtest(c, d)
```

```
Model 1: sigdz ~ sex + rcs(age, 4)
Model 2: sigdz ~ sex * rcs(age, 4)

  L.R. Chisq          d.f.            P
13.096786352  3.000000000  0.004431906
```

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

```
dz ← subset(acath, sigdz==1)
dd ← datadist(dz)
```

```
f ← lrm(tvdlm ~ rcs(cad.dur, 5), data=dz)
w ← function(...)
  with(dz, {
    plsmo(cad.dur, tvdlm, fun=qlogis, add=TRUE,
          grid=TRUE, lty='dotted')
    x ← cut2(cad.dur, g=15, levels.mean=TRUE)
    prop ← qlogis(tapply(tvdlm, x, mean, na.rm=TRUE))
    xm ← as.numeric(names(prop))
    lpoints(xm, prop, pch=2, col='green')
  } )   # Figure 8.7
plot(Predict(f, cad.dur), addpanel=w)
```

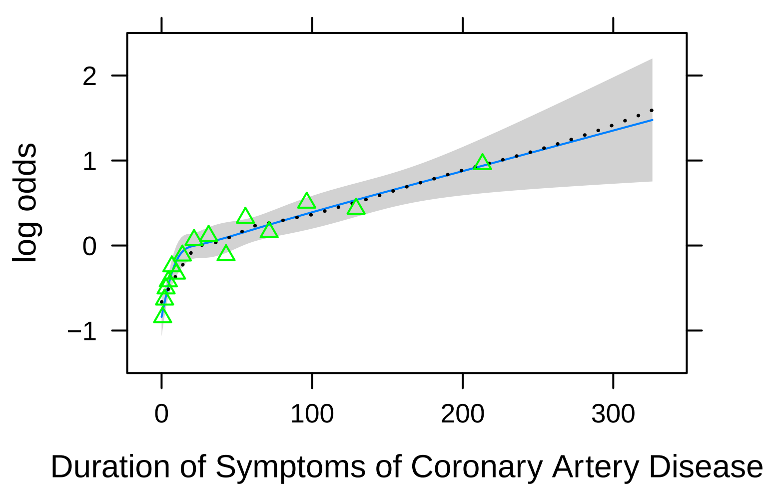


Figure 8.7: Estimated relationship between duration of symptoms and the log odds of severe coronary artery disease for $k = 5$. Knots are marked with arrows. Solid line is spline fit; dotted line is a nonparametric loess estimate.

```
f ← lrm(tvdlm ~ log10(cad.dur + 1), data=dz)
```

```
w ← function(...)
  with(dz, {
    x ← cut2(cad.dur, m=150, levels.mean=TRUE)
    prop ← tapply(tvdlm, x, mean, na.rm=TRUE)
    xm ← as.numeric(names(prop))
    lpoints(xm, prop, pch=2, col='green')
  } )
#   Figure 8.8
plot(Predict(f, cad.dur, fun=plogis), ylab='P',
     ylim=c(.2, .8), addpanel=w)
```

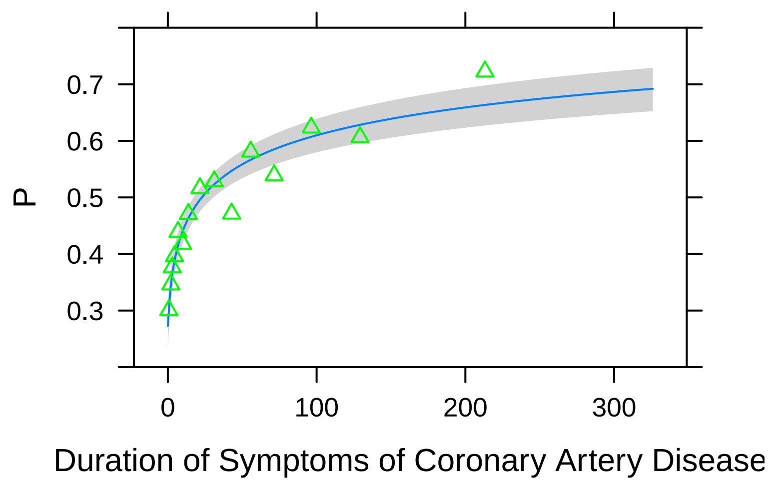


Figure 8.8: Fitted linear logistic model in $\log_{10}$(duration+1), with subgroup estimates using groups of 150 patients. Fitted equation is logit(`tvdlm`) $= -.9809 + .7122 \log_{10}(\text{months} + 1)$.

- Sample of 2258 pts[c]
- 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

[c] Many patients had missing cholesterol.

```
acath ← transform(acath,
                  cholesterol = choleste,
                  age.tertile = cut2(age,g=3),
                  sx = as.integer(acath$sex) - 1)
# sx for loess, need to code as numeric
dd ← datadist(acath); options(datadist='dd')

# First model stratifies age into tertiles to get more
# empirical estimates of age x cholesterol interaction

f ← lrm(sigdz ~ age.tertile*(sex + rcs(cholesterol,4)),
        data=acath)
print(f, latex=TRUE)
```

Logistic Regression Model

```
lrm(formula = sigdz ~ age.tertile * (sex + rcs(cholesterol, 4)),
    data = acath)
```

Frequencies of Missing Values Due to Each Variable

```
 sigdz age.tertile         sex cholesterol
     0           0           0        1246
```

	Model Likelihood Ratio Test	Discrimination Indexes	Rank Discrim. Indexes
Obs 2258	LR χ^2 533.52	R^2 0.291	C 0.780
0 768	d.f. 14	g 1.316	D_{xy} 0.560
1 1490	Pr$(> \chi^2) < 0.0001$	g_r 3.729	γ 0.562
$\max \lvert\frac{\partial \log L}{\partial \beta}\rvert$ 2×10^{-8}		g_p 0.252	τ_a 0.251
		Brier 0.173	

	Coef	S.E.	Wald Z	Pr$(>\lvert Z\rvert)$
Intercept	-0.4155	1.0987	-0.38	0.7053
age.tertile=[49,58)	0.8781	1.7337	0.51	0.6125
age.tertile=[58,82]	4.7861	1.8143	2.64	0.0083
sex=female	-1.6123	0.1751	-9.21	<0.0001
cholesterol	0.0029	0.0060	0.48	0.6347
cholesterol'	0.0384	0.0242	1.59	0.1126
cholesterol''	-0.1148	0.0768	-1.49	0.1350
age.tertile=[49,58) * sex=female	-0.7900	0.2537	-3.11	0.0018
age.tertile=[58,82] * sex=female	-0.4530	0.2978	-1.52	0.1283
age.tertile=[49,58) * cholesterol	0.0011	0.0095	0.11	0.9093
age.tertile=[58,82] * cholesterol	-0.0158	0.0099	-1.59	0.1111
age.tertile=[49,58) * cholesterol'	-0.0183	0.0365	-0.50	0.6162
age.tertile=[58,82] * cholesterol'	0.0127	0.0406	0.31	0.7550
age.tertile=[49,58) * cholesterol''	0.0582	0.1140	0.51	0.6095

Table 8.2: Crudely categorizing age into tertiles

	χ^2	d.f.	P
age.tertile (Factor+Higher Order Factors)	120.74	10	< 0.0001
All Interactions	21.87	8	0.0052
sex (Factor+Higher Order Factors)	329.54	3	< 0.0001
All Interactions	9.78	2	0.0075
cholesterol (Factor+Higher Order Factors)	93.75	9	< 0.0001
All Interactions	10.03	6	0.1235
Nonlinear (Factor+Higher Order Factors)	9.96	6	0.1263
age.tertile × sex (Factor+Higher Order Factors)	9.78	2	0.0075
age.tertile × cholesterol (Factor+Higher Order Factors)	10.03	6	0.1235
Nonlinear	2.62	4	0.6237
Nonlinear Interaction : f(A,B) vs. AB	2.62	4	0.6237
TOTAL NONLINEAR	9.96	6	0.1263
TOTAL INTERACTION	21.87	8	0.0052
TOTAL NONLINEAR + INTERACTION	29.67	10	0.0010
TOTAL	410.75	14	< 0.0001

	Coef	S.E.	Wald Z	Pr(> $\|Z\|$)
age.tertile=[58,82] * cholesterol"	-0.0092	0.1301	-0.07	0.9436

```
ltx(f)
```

$X\hat{\beta} = -0.415 + 0.878[\text{age.tertile} \in [49,58)] + 4.79[\text{age.tertile} \in [58,82]] - 1.61[\text{female}] + 0.00287\text{cholesterol} + 1.52\times10^{-6}(\text{cholesterol} - 160)_+^3 - 4.53\times10^{-6}(\text{cholesterol} - 208)_+^3 + 3.44\times10^{-6}(\text{cholesterol}-243)_+^3 - 4.28\times10^{-7}(\text{cholesterol}-319)_+^3 + [\text{female}][-0.79[\text{age.tertile} \in [49,58)] - 0.453[\text{age.tertile} \in [58,82]]] + [\text{age.tertile} \in [49,58)][0.00108\text{cholesterol} - 7.23\times10^{-7}(\text{cholesterol}-160)_+^3 + 2.3\times10^{-6}(\text{cholesterol}-208)_+^3 - 1.84\times10^{-6}(\text{cholesterol}-243)_+^3 + 2.69\times10^{-7}(\text{cholesterol}-319)_+^3] + [\text{age.tertile} \in [58,82]][-0.0158\text{cholesterol} + 5\times10^{-7}(\text{cholesterol} - 160)_+^3 - 3.64\times10^{-7}(\text{cholesterol} - 208)_+^3 - 5.15\times10^{-7}(\text{cholesterol} - 243)_+^3 + 3.78\times10^{-7}(\text{cholesterol} - 319)_+^3]$.

```
# Table 8.2:
latex(anova(f), file='', size='smaller',
      caption='Crudely categorizing age into tertiles',
      label='tab:anova-tertiles')
```

```
yl ← c(-1,5)
plot(Predict(f, cholesterol, age.tertile),
     adj.subtitle=FALSE, ylim=yl)   # Figure 8.9
```

- Now model age as continuous predictor
- Start with nonparametric surface using $Y = 0/1$

```
# Re-do model with continuous age
f ← loess(sigdz ~ age * (sx + cholesterol), data=acath,
```

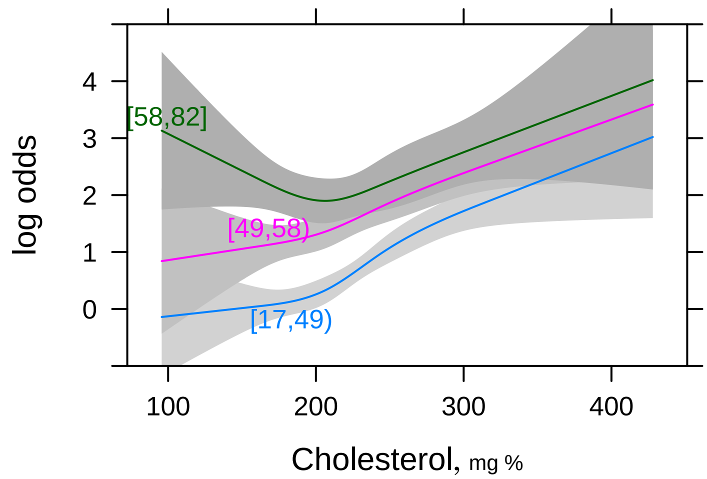


Figure 8.9: Log odds of significant coronary artery disease modeling age with two dummy variables

```
           parametric="sx", drop.square="sx")
ages  ← seq(25,   75, length=40)
chols ← seq(100, 400, length=40)
g ← expand.grid(cholesterol=chols, age=ages, sx=0)
# drop sex dimension of grid since held to 1 value
p ← drop(predict(f, g))
p[p < 0.001] ← 0.001
p[p > 0.999] ← 0.999
zl ← c(−3, 6)   # Figure 8.10
wireframe(qlogis(p) ∼ cholesterol∗age,
          xlab=list(rot=30), ylab=list(rot=−40),
          zlab=list(label='log odds', rot=90), zlim=zl,
          scales = list(arrows = FALSE), data=g)
```

- Next try parametric fit using linear spline in age, chol. (3 knots each), all product terms. For all the remaining 3-d plots we limit plotting to points that are supported by at least 5 subjects beyond those cholesterol/age combinations

```
f ← lrm(sigdz ∼ lsp(age,c(46,52,59)) ∗
        (sex + lsp(cholesterol,c(196,224,259))),
        data=acath)
```

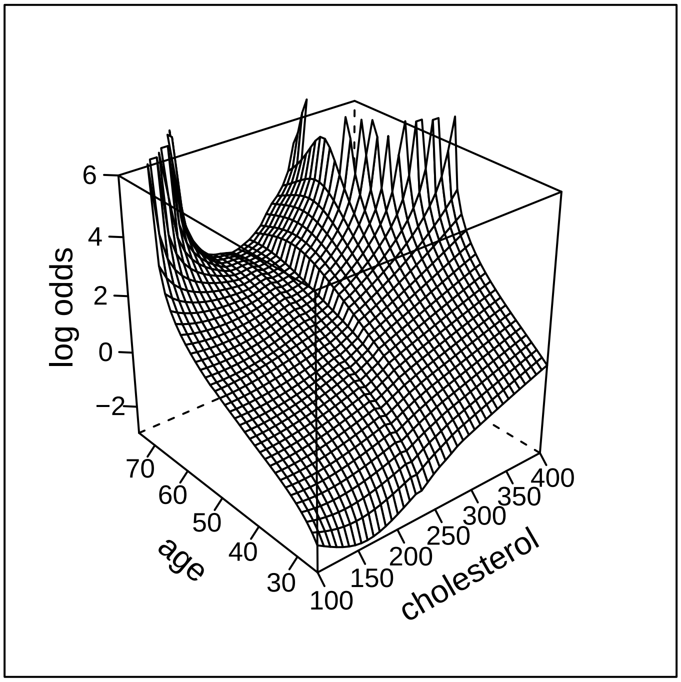


Figure 8.10: Local regression fit for the logit of the probability of significant coronary disease vs. age and cholesterol for males, based on the `loess` function.

```
ltx(f)
```

$X\hat{\beta} = -1.83 + 0.0232\ \text{age} + 0.0759(\text{age} - 46)_+ - 0.0025(\text{age} - 52)_+ + 2.27(\text{age} - 59)_+ + 3.02[\text{female}] - 0.0177\ \text{cholesterol} + 0.114(\text{cholesterol} - 196)_+ - 0.131(\text{cholesterol} - 224)_+ + 0.0651(\text{cholesterol} - 259)_+ + [\text{female}][-0.112\ \text{age} + 0.0852\ (\text{age} - 46)_+ - 0.0302\ (\text{age} - 52)_+ + 0.176(\text{age}-59)_+] + \text{age}[0.000577\text{cholesterol} - 0.00286(\text{cholesterol}-196)_+ + 0.00382(\text{cholesterol}-224)_+ - 0.00205(\text{cholesterol}-259)_+] + (\text{age}-46)_+[-0.000936\text{cholesterol} + 0.00643(\text{cholesterol}-196)_+ - 0.0115(\text{cholesterol}-224)_+ + 0.00756(\text{cholesterol}-259)_+] + (\text{age}-52)_+[0.000433\text{cholesterol} - 0.0037\ (\text{cholesterol} - 196)_+ + 0.00815\ (\text{cholesterol} - 224)_+ - 0.00715\ (\text{cholesterol} - 259)_+] + (\text{age} - 59)_+[-0.0124\ \text{cholesterol} + 0.015\ (\text{cholesterol} - 196)_+ - 0.0067\ (\text{cholesterol} - 224)_+ + 0.00752\ (\text{cholesterol} - 259)_+].$

```
latex(anova(f), caption='Linear spline surface', file='',
      size='smaller', label='tab:anova-lsp')   # Table 8.3
```

```
perim ← with(acath,
             perimeter(cholesterol, age, xinc=20, n=5))
zl ← c(-2, 4)   # Figure 8.11
bplot(Predict(f, cholesterol, age, np=40), perim=perim,
      lfun=wireframe, zlim=zl, adj.subtitle=FALSE)
```

- Next try smooth spline surface, include all cross-products

```
f ← lrm(sigdz ~ rcs(age,4)*(sex + rcs(cholesterol,4)),
```

Table 8.3: Linear spline surface

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	164.17	24	< 0.0001
All Interactions	42.28	20	0.0025
Nonlinear (Factor+Higher Order Factors)	25.21	18	0.1192
sex (Factor+Higher Order Factors)	343.80	5	< 0.0001
All Interactions	23.90	4	0.0001
cholesterol (Factor+Higher Order Factors)	100.13	20	< 0.0001
All Interactions	16.27	16	0.4341
Nonlinear (Factor+Higher Order Factors)	16.35	15	0.3595
age × sex (Factor+Higher Order Factors)	23.90	4	0.0001
Nonlinear	12.97	3	0.0047
Nonlinear Interaction : f(A,B) vs. AB	12.97	3	0.0047
age × cholesterol (Factor+Higher Order Factors)	16.27	16	0.4341
Nonlinear	11.45	15	0.7204
Nonlinear Interaction : f(A,B) vs. AB	11.45	15	0.7204
f(A,B) vs. Af(B) + Bg(A)	9.38	9	0.4033
Nonlinear Interaction in age vs. Af(B)	9.99	12	0.6167
Nonlinear Interaction in cholesterol vs. Bg(A)	10.75	12	0.5503
TOTAL NONLINEAR	33.22	24	0.0995
TOTAL INTERACTION	42.28	20	0.0025
TOTAL NONLINEAR + INTERACTION	49.03	26	0.0041
TOTAL	449.26	29	< 0.0001

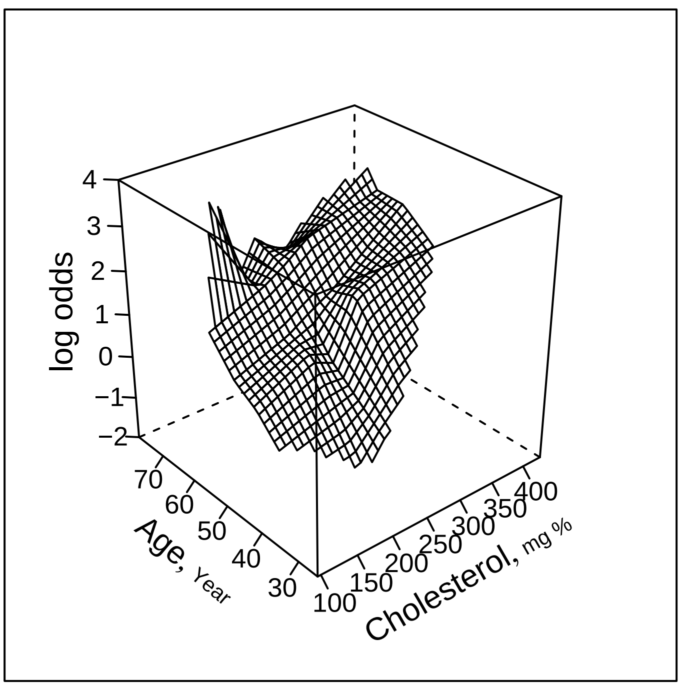


Figure 8.11: Linear spline surface for males, with knots for age at 46, 52, 59 and knots for cholesterol at 196, 224, and 259 (quartiles).

Table 8.4: Cubic spline surface

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	165.23	15	< 0.0001
All Interactions	37.32	12	0.0002
Nonlinear (Factor+Higher Order Factors)	21.01	10	0.0210
sex (Factor+Higher Order Factors)	343.67	4	< 0.0001
All Interactions	23.31	3	< 0.0001
cholesterol (Factor+Higher Order Factors)	97.50	12	< 0.0001
All Interactions	12.95	9	0.1649
Nonlinear (Factor+Higher Order Factors)	13.62	8	0.0923
age × sex (Factor+Higher Order Factors)	23.31	3	< 0.0001
Nonlinear	13.37	2	0.0013
Nonlinear Interaction : f(A,B) vs. AB	13.37	2	0.0013
age × cholesterol (Factor+Higher Order Factors)	12.95	9	0.1649
Nonlinear	7.27	8	0.5078
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
TOTAL NONLINEAR	29.22	14	0.0097
TOTAL INTERACTION	37.32	12	0.0002
TOTAL NONLINEAR + INTERACTION	45.41	16	0.0001
TOTAL	450.88	19	< 0.0001

```
          data=acath, tol=1e-11)
ltx(f)
```

$X\hat{\beta} = -6.41 + 0.166\text{age} - 0.00067(\text{age} - 36)_+^3 + 0.00543(\text{age} - 48)_+^3 - 0.00727(\text{age} - 56)_+^3 + 0.00251(\text{age} - 68)_+^3 + 2.87[\text{female}] + 0.00979\text{cholesterol} + 1.96\times10^{-6}(\text{cholesterol} - 160)_+^3 - 7.16\times10^{-6}(\text{cholesterol} - 208)_+^3 + 6.35\times10^{-6}(\text{cholesterol} - 243)_+^3 - 1.16\times10^{-6}(\text{cholesterol} - 319)_+^3 + [\text{female}][-0.109\text{age} + 7.52\times10^{-5}(\text{age} - 36)_+^3 + 0.00015(\text{age} - 48)_+^3 - 0.00045(\text{age} - 56)_+^3 + 0.000225(\text{age} - 68)_+^3] + \text{age}[-0.00028\text{cholesterol} + 2.68\times10^{-9}(\text{cholesterol} - 160)_+^3 + 3.03\times10^{-8}(\text{cholesterol} - 208)_+^3 - 4.99\times10^{-8}(\text{cholesterol} - 243)_+^3 + 1.69\times10^{-8}(\text{cholesterol} - 319)_+^3] + \text{age}'[0.00341\text{cholesterol} - 4.02\times10^{-7}(\text{cholesterol} - 160)_+^3 + 9.71\times10^{-7}(\text{cholesterol} - 208)_+^3 - 5.79\times10^{-7}(\text{cholesterol} - 243)_+^3 + 8.79\times10^{-9}(\text{cholesterol} - 319)_+^3] + \text{age}''[-0.029\text{cholesterol} + 3.04\times10^{-6}(\text{cholesterol} - 160)_+^3 - 7.34\times10^{-6}(\text{cholesterol} - 208)_+^3 + 4.36\times10^{-6}(\text{cholesterol} - 243)_+^3 - 5.82\times10^{-8}(\text{cholesterol} - 319)_+^3]$.

```
latex(anova(f), caption='Cubic spline surface', file='',
      size='smaller', label='tab:anova-rcs') #Table 8.4
```

```
# Figure 8.12:
bplot(Predict(f, cholesterol, age, np=40), perim=perim,
      lfun=wireframe, zlim=zl, adj.subtitle=FALSE)
```

- Now restrict surface by excluding doubly non-linear terms

```
f ← lrm(sigdz ~ sex*rcs(age,4) + rcs(cholesterol,4) +
          rcs(age,4) %ia% rcs(cholesterol,4), data=acath)
latex(anova(f), file='', size='smaller',
```

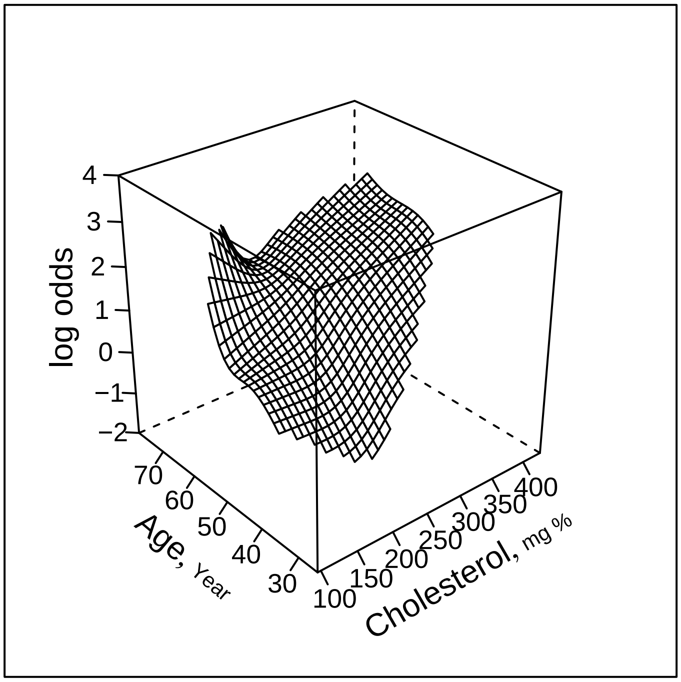


Figure 8.12: Restricted cubic spline surface in two variables, each with $k = 4$ knots

Table 8.5: Singly nonlinear cubic spline surface

	χ^2	d.f.	P
sex (Factor+Higher Order Factors)	343.42	4	< 0.0001
All Interactions	24.05	3	< 0.0001
age (Factor+Higher Order Factors)	169.35	11	< 0.0001
All Interactions	34.80	8	< 0.0001
Nonlinear (Factor+Higher Order Factors)	16.55	6	0.0111
cholesterol (Factor+Higher Order Factors)	93.62	8	< 0.0001
All Interactions	10.83	5	0.0548
Nonlinear (Factor+Higher Order Factors)	10.87	4	0.0281
age × cholesterol (Factor+Higher Order Factors)	10.83	5	0.0548
Nonlinear	3.12	4	0.5372
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	0.4400
sex × age (Factor+Higher Order Factors)	24.05	3	< 0.0001
Nonlinear	13.58	2	0.0011
Nonlinear Interaction : f(A,B) vs. AB	13.58	2	0.0011
TOTAL NONLINEAR	27.89	10	0.0019
TOTAL INTERACTION	34.80	8	< 0.0001
TOTAL NONLINEAR + INTERACTION	45.45	12	< 0.0001
TOTAL	453.10	15	< 0.0001

```
      caption='Singly nonlinear cubic spline surface',
      label='tab:anova-ria') #Table 8.5
```

```
# Figure 8.13:
bplot(Predict(f, cholesterol, age, np=40), perim=perim,
```

```
      lfun=wireframe, zlim=zl, adj.subtitle=FALSE)
ltx(f)
```

$X\hat{\beta} = -7.2 + 2.96[\text{female}] + 0.164\text{age} + 7.23\times10^{-5}(\text{age}-36)_+^3 - 0.000106(\text{age}-48)_+^3 - 1.63\times10^{-5}(\text{age}-56)_+^3 + 4.99\times10^{-5}(\text{age}-68)_+^3 + 0.0148\text{cholesterol} + 1.21\times10^{-6}(\text{cholesterol}-160)_+^3 - 5.5\times10^{-6}(\text{cholesterol}-208)_+^3 + 5.5\times10^{-6}(\text{cholesterol}-243)_+^3 - 1.21\times10^{-6}(\text{cholesterol}-319)_+^3 + \text{age}[-0.00029\text{cholesterol} + 9.28\times10^{-9}(\text{cholesterol}-160)_+^3 + 1.7\times10^{-8}(\text{cholesterol}-208)_+^3 - 4.43\times10^{-8}(\text{cholesterol}-243)_+^3 + 1.79\times10^{-8}(\text{cholesterol}-319)_+^3] + \text{cholesterol}[2.3\times10^{-7}(\text{age}-36)_+^3 + 4.21\times10^{-7}(\text{age}-48)_+^3 - 1.31\times10^{-6}(\text{age}-56)_+^3 + 6.64\times10^{-7}(\text{age}-68)_+^3] + [\text{female}][-0.111\text{age} + 8.03\times10^{-5}(\text{age}-36)_+^3 + 0.000135(\text{age}-48)_+^3 - 0.00044(\text{age}-56)_+^3 + 0.000224(\text{age}-68)_+^3]$.

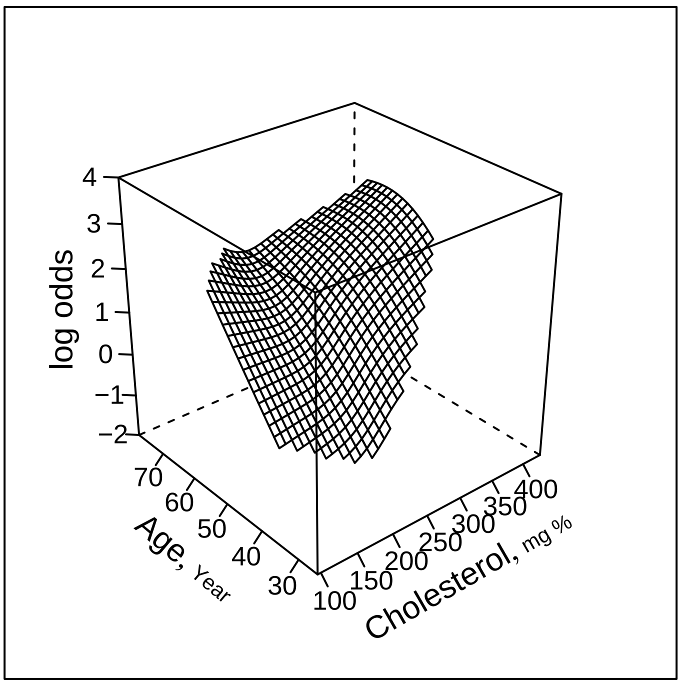


Figure 8.13: Restricted cubic spline fit with age × spline(cholesterol) and cholesterol × spline(age)

- Finally restrict the interaction to be a simple product

```
f ← lrm(sigdz ~ rcs(age,4)*sex + rcs(cholesterol,4) +
         age %ia% cholesterol, data=acath)
latex(anova(f), caption='Linear interaction surface', file='',
      size='smaller', label='tab:anova-lia') #Table 8.6
```

```
# Figure 8.14:
bplot(Predict(f, cholesterol, age, np=40), perim=perim,
      lfun=wireframe, zlim=zl, adj.subtitle=FALSE)
f.linia ← f  # save linear interaction fit for later
```

Table 8.6: Linear interaction surface

	χ^2	d.f.	P
age (Factor+Higher Order Factors)	167.83	7	< 0.0001
All Interactions	31.03	4	< 0.0001
Nonlinear (Factor+Higher Order Factors)	14.58	4	0.0057
sex (Factor+Higher Order Factors)	345.88	4	< 0.0001
All Interactions	22.30	3	0.0001
cholesterol (Factor+Higher Order Factors)	89.37	4	< 0.0001
All Interactions	7.99	1	0.0047
Nonlinear	10.65	2	0.0049
age × cholesterol (Factor+Higher Order Factors)	7.99	1	0.0047
age × sex (Factor+Higher Order Factors)	22.30	3	0.0001
Nonlinear	12.06	2	0.0024
Nonlinear Interaction : f(A,B) vs. AB	12.06	2	0.0024
TOTAL NONLINEAR	25.72	6	0.0003
TOTAL INTERACTION	31.03	4	< 0.0001
TOTAL NONLINEAR + INTERACTION	43.59	8	< 0.0001
TOTAL	452.75	11	< 0.0001

```
ltx(f)
```

$X\hat{\beta} = -7.36 + 0.182\text{age} - 5.18\times10^{-5}(\text{age}-36)_+^3 + 8.45\times10^{-5}(\text{age}-48)_+^3 - 2.91\times10^{-6}(\text{age}-56)_+^3 - 2.99\times10^{-5}(\text{age}-68)_+^3 + 2.8[\text{female}] + 0.0139\text{cholesterol} + 1.76\times10^{-6}(\text{cholesterol}-160)_+^3 - 4.88\times10^{-6}(\text{cholesterol}-208)_+^3 + 3.45\times10^{-6}(\text{cholesterol}-243)_+^3 - 3.26\times10^{-7}(\text{cholesterol}-319)_+^3 - 0.00034\,\text{age}\times\text{cholesterol} + [\text{female}][-0.107\text{age} + 7.71\times10^{-5}(\text{age}-36)_+^3 + 0.000115(\text{age}-48)_+^3 - 0.000398(\text{age}-56)_+^3 + 0.000205(\text{age}-68)_+^3]$.

The Wald test for age × cholesterol interaction yields $\chi^2 = 7.99$ with 1 d.f., p=.005.

- 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

```
# Make estimates of cholesterol effects for mean age in
# tertiles corresponding to initial analysis
mean.age ←
  with(acath,
       as.vector(tapply(age, age.tertile, mean, na.rm=TRUE)))
plot(Predict(f, cholesterol, age=round(mean.age,2),
             sex="male"),
```

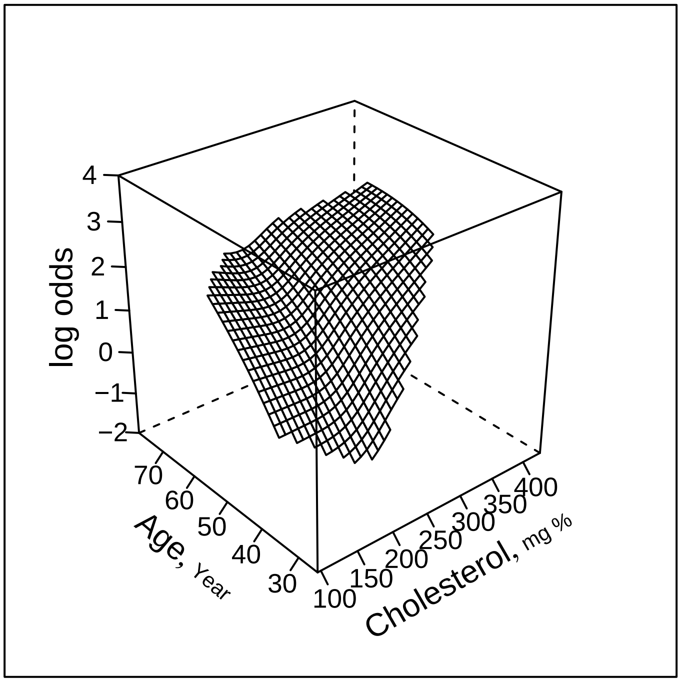


Figure 8.14: Spline fit with nonlinear effects of cholesterol and age and a simple product interaction

```
    adj.subtitle=FALSE, ylim=yl) #3 curves, Figure 8.15
```

- Using residuals for "duration of symptoms" example

```
f ← lrm(tvdlm ∼ cad.dur, data=dz, x=TRUE, y=TRUE)
resid(f, "partial", pl="loess", xlim=c(0,250), ylim=c(−3,3))
scat1d(dz$cad.dur)
log.cad.dur ← log10(dz$cad.dur + 1)
f ← lrm(tvdlm ∼ log.cad.dur, data=dz, x=TRUE, y=TRUE)
resid(f, "partial", pl="loess", ylim=c(−3,3))
scat1d(log.cad.dur)   # Figure 8.16
```

- Relative merits of strat., nonparametric, splines for checking fit

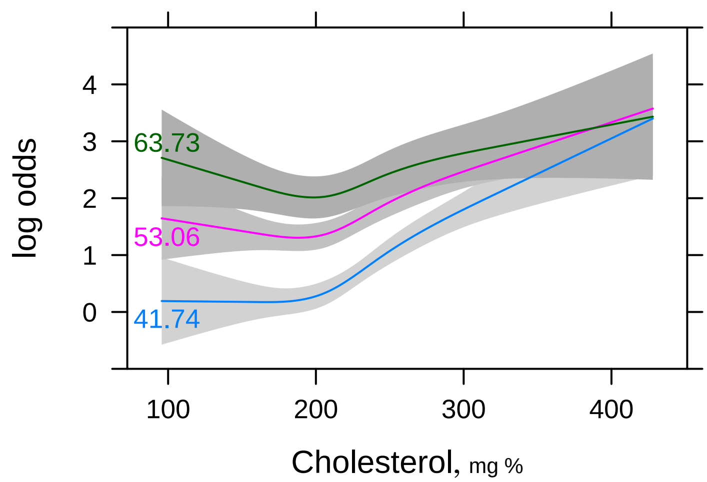


Figure 8.15: Predictions from linear interaction model with mean age in tertiles indicated.

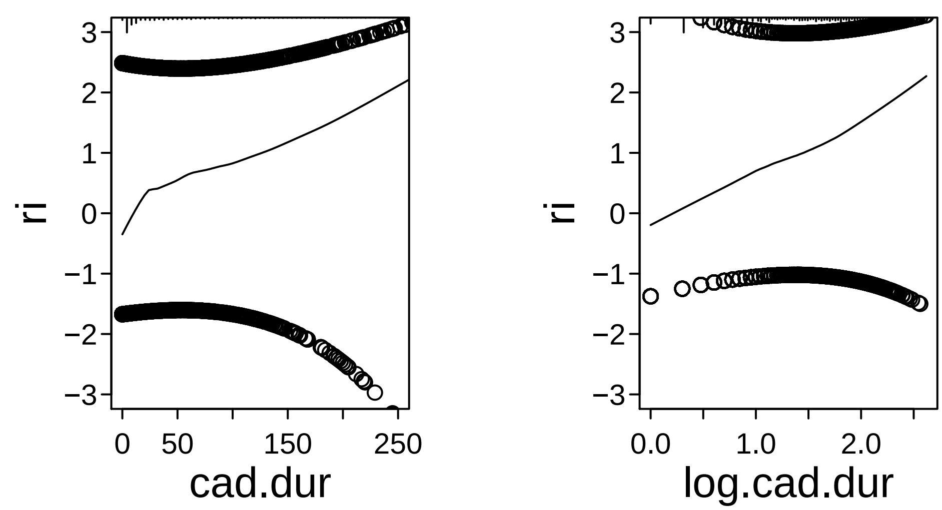


Figure 8.16: Partial residuals for duration and $\log_{10}$(duration+1). Data density shown at top of each plot.

Method	Choice Required	Assumes Additivity	Uses Ordering of X	Low Variance	Good Resolutio 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

8.6 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}
$$D_{xy} = 2(c - .5).$$
  - Good pure index of predictive discrimination for a single model
  - Not useful for comparing two models[38, 134][d]
- "Coefficient of discrimination"[164]: average $\hat{P}$ when $Y = 1$ minus average $\hat{P}$ when $Y = 0$
  - Has many advantages. Tjur shows how

[d] But see[132].

it ties in with sum of squares–based R^2 measures.

- "Percent classified correctly" has lots of problems
  - improper scoring rule; optimizing it will lead to incorrect model
  - arbitrary, insensitive, uses a strange loss (utility function)

8.7 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 equa-

tion

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

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

- Bootstrap validation of age-sex-response data, 150 samples
- 2 predictors forced into every model

```
d ← sex.age.response
dd ← datadist(d); options(datadist='dd')
f ← lrm(response ~ sex + age, data=d, x=TRUE, y=TRUE)
set.seed(3)  # for reproducibility
v1 ← validate(f, B=150)
```

```
latex(v1,
      caption='Bootstrap Validation, 2 Predictors Without Stepdown',
      digits=2, size='Ssize', file='')
```

Bootstrap Validation, 2 Predictors Without Stepdown

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.70	0.70	0.67	0.04	0.66	150
R^2	0.45	0.48	0.43	0.05	0.40	150
Intercept	0.00	0.00	0.01	−0.01	0.01	150
Slope	1.00	1.00	0.91	0.09	0.91	150
$E_{\max}$	0.00	0.00	0.02	0.02	0.02	150
D	0.39	0.44	0.36	0.07	0.32	150
U	−0.05	−0.05	0.04	−0.09	0.04	150
Q	0.44	0.49	0.32	0.16	0.28	150
B	0.16	0.15	0.18	−0.03	0.19	150
g	2.10	2.49	1.97	0.52	1.58	150
g_p	0.35	0.35	0.34	0.01	0.34	150

- Allow for step-down at each re-sample
- Use individual tests at $\alpha = 0.10$

- Both age and sex selected in 137 of 150, neither in 3 samples

```
v2 ← validate(f, B=150, bw=TRUE,
               rule='p', sls=.1, type='individual')
```

```
latex(v2,
      caption='Bootstrap Validation, 2 Predictors with Stepdown',
      digits=2, B=15, file='', size='Ssize')
```

Bootstrap Validation, 2 Predictors with Stepdown

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.70	0.70	0.64	0.07	0.63	150
R^2	0.45	0.49	0.41	0.09	0.37	150
Intercept	0.00	0.00	−0.04	0.04	−0.04	150
Slope	1.00	1.00	0.84	0.16	0.84	150
$E_{\max}$	0.00	0.00	0.05	0.05	0.05	150
D	0.39	0.45	0.34	0.11	0.28	150
U	−0.05	−0.05	0.06	−0.11	0.06	150
Q	0.44	0.50	0.28	0.22	0.22	150
B	0.16	0.14	0.18	−0.04	0.20	150
g	2.10	2.60	1.88	0.72	1.38	150
g_p	0.35	0.35	0.33	0.02	0.33	150

Factors Retained in Backwards Elimination
First 15 Resamples

sex	age
•	•
•	•
•	•
•	•
•	•
•	•
•	•
•	•
•	•
•	•
•	•
•	•
•	•
•	•
•	

Frequencies of Numbers of Factors Retained

0	1	2
3	10	137

- Try adding 5 noise candidate variables

```
set.seed(133)
n  ← nrow(d)
x1 ← runif(n)
x2 ← runif(n)
x3 ← runif(n)
x4 ← runif(n)
x5 ← runif(n)
f  ← lrm(response ∼ age + sex + x1 + x2 + x3 + x4 + x5,
         data=d, x=TRUE, y=TRUE)
v3 ← validate(f, B=150, bw=TRUE,
              rule='p', sls=.1, type='individual')
```

```
k ← attr(v3, 'kept')
# Compute number of x1-x5 selected
nx ← apply(k[,3:7], 1, sum)
# Get selections of age and sex
v ← colnames(k)
as ← apply(k[,1:2], 1,
           function(x) paste(v[1:2][x], collapse=', '))
table(paste(as, ' ', nx, 'Xs'))
```

```
           0 Xs              1 Xs       age   2 Xs
             50                 3                1
age, sex   0 Xs  age, sex   1 Xs  age, sex   2 Xs
             34                17               11
age, sex   3 Xs  age, sex   4 Xs       sex   0 Xs
              7                 1               12
     sex   1 Xs
              3
```

```
latex(v3,
 caption='Bootstrap Validation with 5 Noise Variables and Stepdown',
 digits=2, B=15, size='Ssize', file='')
```

Bootstrap Validation with 5 Noise Variables and Stepdown

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.70	0.47	0.38	0.09	0.60	139
R^2	0.45	0.34	0.23	0.11	0.34	139
Intercept	0.00	0.00	0.03	-0.03	0.03	139
Slope	1.00	1.00	0.78	0.22	0.78	139
$E_{\max}$	0.00	0.00	0.06	0.06	0.06	139
D	0.39	0.31	0.18	0.13	0.26	139
U	-0.05	-0.05	0.07	-0.12	0.07	139
Q	0.44	0.36	0.11	0.25	0.19	139
B	0.16	0.17	0.22	-0.04	0.20	139
g	2.10	1.81	1.06	0.75	1.36	139
g_p	0.35	0.23	0.19	0.04	0.31	139

Factors Retained in Backwards Elimination
First 15 Resamples

age	sex	x1	x2	x3	x4	x5
•	•		•	•	•	•
•	•	•				•
•	•					
•	•				•	•
•	•	•				
•	•					
•	•					
•	•		•			
•	•			•		

Frequencies of Numbers of Factors Retained

0	1	2	3	4	5	6
50	15	37	18	11	7	1

- Repeat but force age and sex to be in all models

```
v4 ← validate(f, B=150, bw=TRUE, rule='p', sls=.1,
              type='individual', force=1:2)
```

```
ap4 ← round(v4[,'index.orig'], 2)
bc4 ← round(v4[,'index.corrected'], 2)
```

```
latex(v4,
      caption='Bootstrap Validation with 5 Noise Variables and Stepdown, F
      digits=2, B=15, size='Ssize')
```

Bootstrap Validation with 5 Noise Variables and Stepdown, Forced Inclusion of age and sex

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.70	0.73	0.66	0.07	0.63	131
R^2	0.45	0.52	0.42	0.10	0.36	131
Intercept	0.00	0.00	-0.03	0.03	-0.03	131
Slope	1.00	1.00	0.80	0.20	0.80	131
$E_{\max}$	0.00	0.00	0.06	0.06	0.06	131
D	0.39	0.48	0.36	0.12	0.27	131
U	-0.05	-0.05	0.08	-0.13	0.08	131
Q	0.44	0.53	0.28	0.25	0.19	131
B	0.16	0.14	0.18	-0.04	0.20	131
g	2.10	2.75	1.93	0.82	1.28	131
g_p	0.35	0.36	0.34	0.03	0.32	131

Factors Retained in Backwards Elimination
First 15 Resamples

age	sex	x1	x2	x3	x4	x5
•	•					
•	•				•	•
•	•					
•	•					
•	•			•	•	•
•	•					
•	•	•				
•	•					
•	•					
•	•					
•	•			•	•	
•	•					
•	•					
•	•					
•	•					

Table 8.7: Effects Response : sigdz

	Low	High	Δ	Effect	S.E.	Lower 0.95	Upper 0.95
age	46	59	13	0.90629	0.18381	0.546030	1.26650
Odds Ratio	46	59	13	2.47510		1.726400	3.54860
cholesterol	196	259	63	0.75479	0.13642	0.487410	1.02220
Odds Ratio	196	259	63	2.12720		1.628100	2.77920
sex — female:male	1	2		-2.42970	0.14839	-2.720600	-2.13890
Odds Ratio	1	2		0.08806		0.065837	0.11778

Frequencies of Numbers of Factors Retained

2	3	4	5
95	24	9	3

8.8 Describing the Fitted Model

```
s ← summary(f.linia)    # Table 8.7
latex(s, file='', size='Ssize',
      label='tab:lrm-cholxage-confbar')
```

```
plot(s)    # Figure 8.17
```

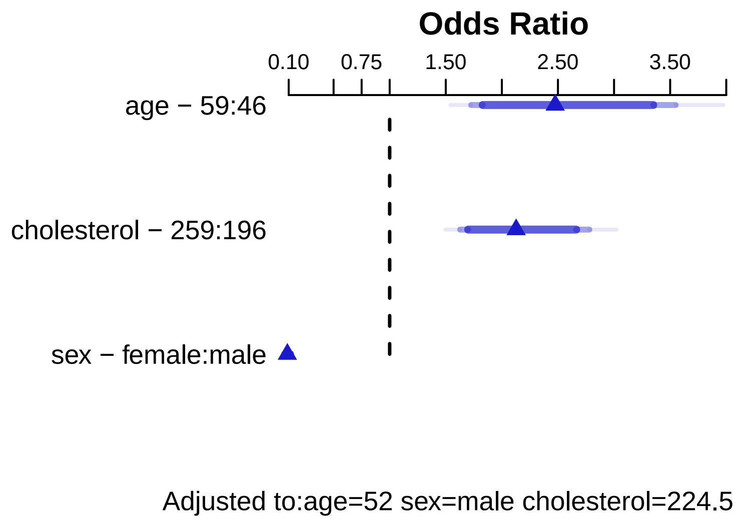


Figure 8.17: Odds ratios and confidence bars, using quartiles of age and cholesterol for assessing their effects on the odds of coronary disease.

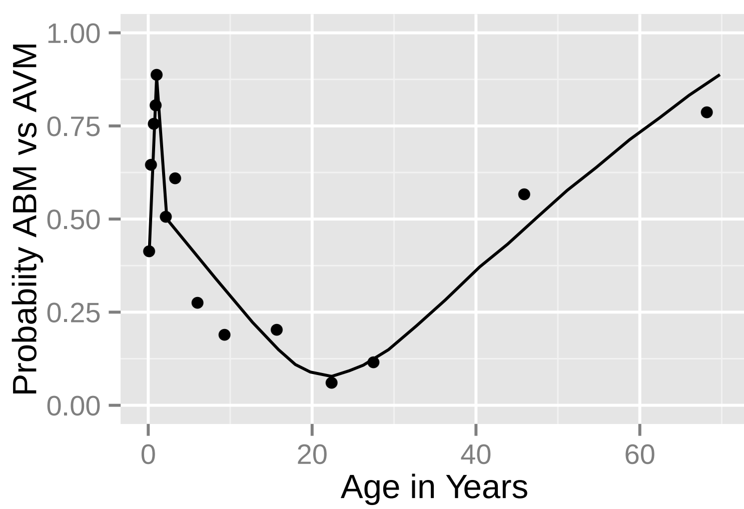


Figure 8.18: Linear spline fit for probability of bacterial vs. viral meningitis as a function of age at onset[153]. Points are simple proportions by age quantile groups.

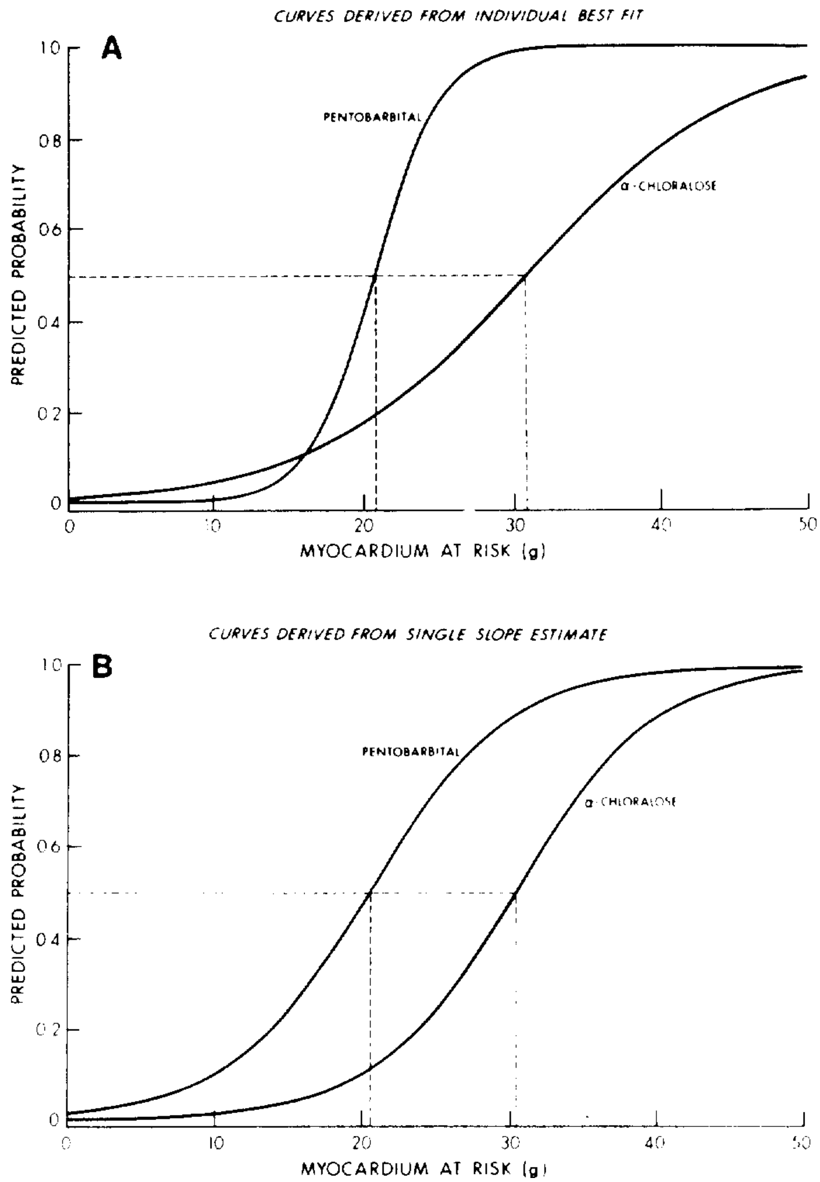


Figure 8.19: (A) Relationship between myocardium at risk and ventricular fibrillation, based on the individual best fit equations for animals anesthetized with pentobarbital and α-chloralose. The amount of myocardium at risk at which 0.5 of the animals are expected to fibrillate (MAR_{50}) is shown for 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^{182}. Reprinted by permission, NRC Research Press.

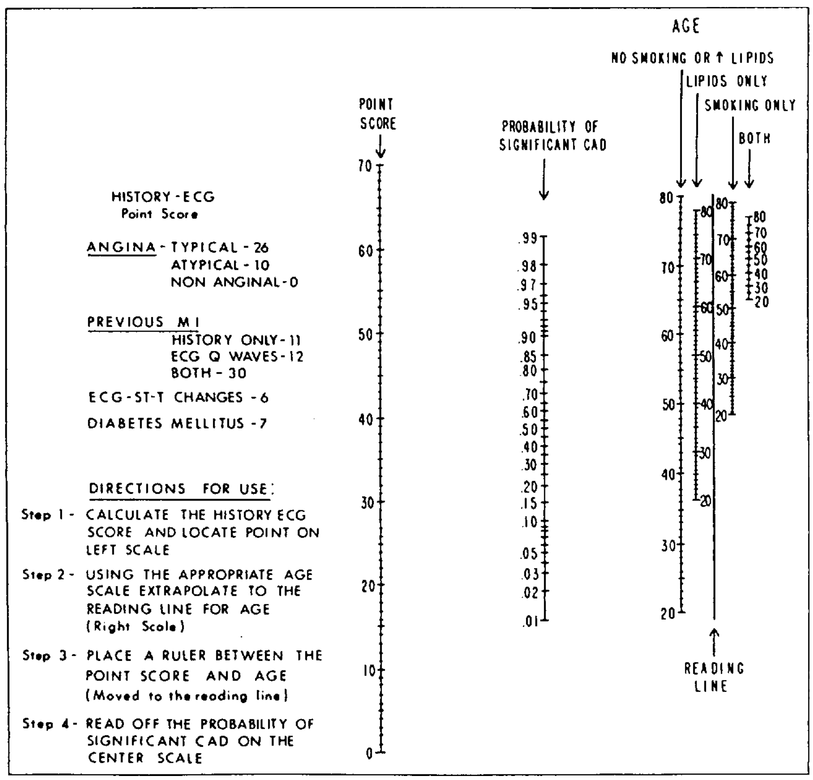


Figure 8.20: A nomogram for estimating the likelihood of significant coronary artery disease (CAD) in women. ECG = electrocardiographic; MI = myocardial infarction[138]. Reprinted from American Journal of Medicine, Vol 75, Pryor DB et al., “Estimating the likelihood of significant coronary artery disease”, p. 778, Copyright 1983, with permission from Excerpta Medica, Inc.

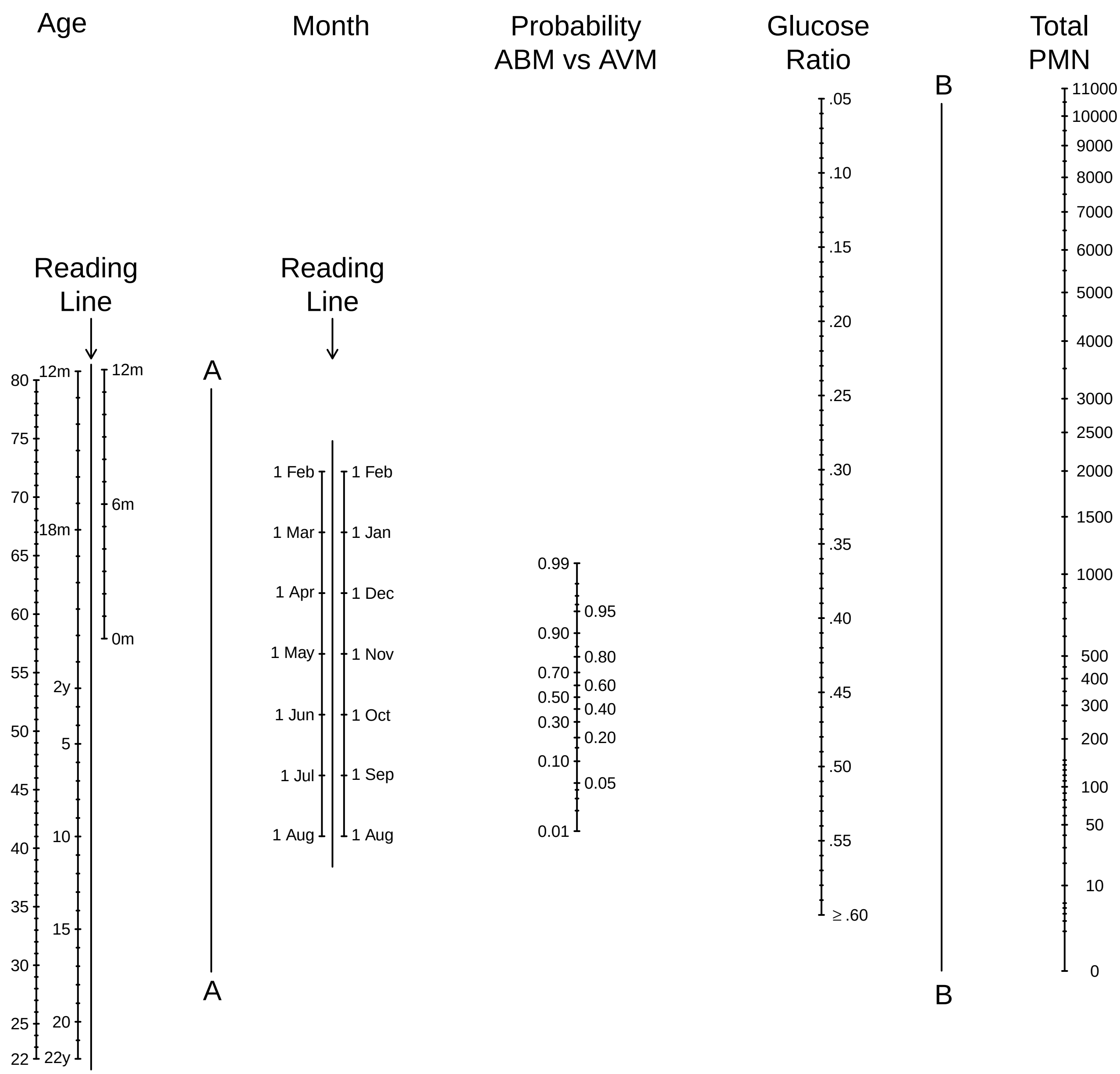


Figure 8.21: Nomogram for estimating probability of bacterial (ABM) vs. viral (AVM) meningitis. Step 1, place ruler on reading lines for patient's age and month of presentation and mark intersection with line A; step 2, place ruler on values for glucose ratio and total polymorphonuclear leukocyte (PMN) 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[153].

```
# Draw a nomogram that shows examples of confidence intervals
nom ← nomogram(f.linia, cholesterol=seq(150, 400, by=50),
               interact=list(age=seq(30, 70, by=10)),
               lp.at=seq(-2, 3.5, by=.5),
               conf.int=TRUE, conf.lp="all",
               fun=function(x)1/(1+exp(-x)),  # or plogis
               funlabel="Probability of CAD",
               fun.at=c(seq(.1, .9, by=.1), .95, .99)
               )  # Figure 8.22
plot(nom, col.grid = gray(c(0.8, 0.95)),
     varname.label=FALSE, ia.space=1, xfrac=.46, lmgp=.2)
```

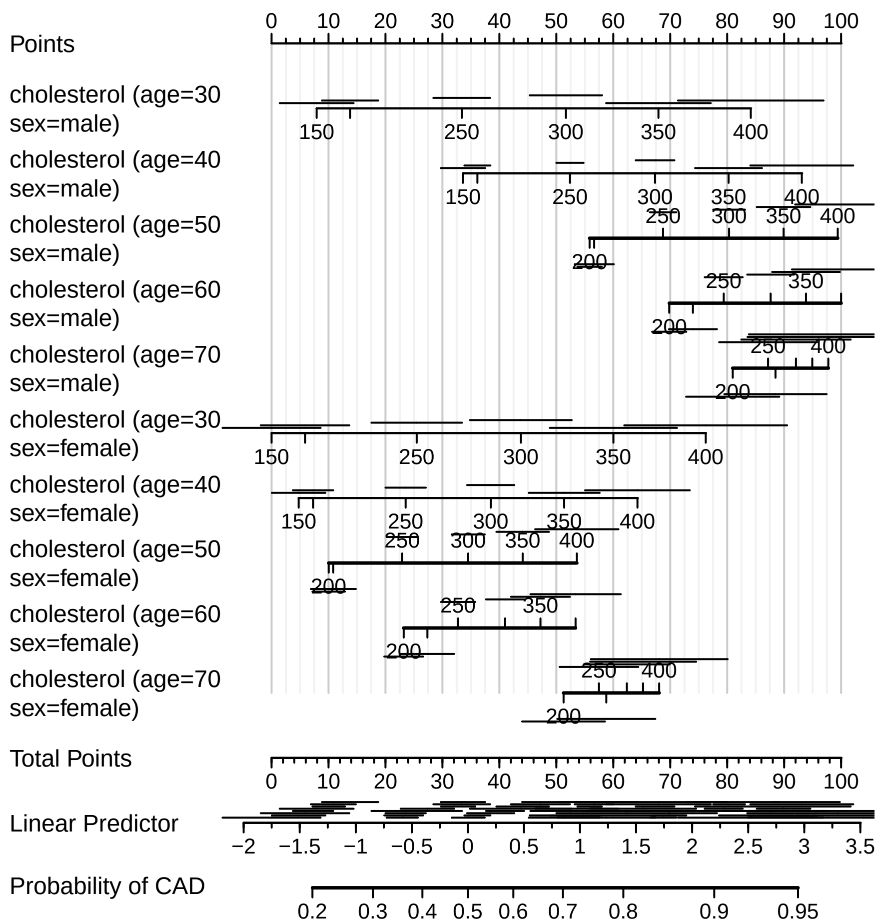


Figure 8.22: Nomogram relating age, sex, and cholesterol to the log odds and to the probability of significant coronary artery disease. Select one axis corresponding to sex and to age $\in \{30, 40, 50, 60, 70\}$. There was linear interaction between age and sex and between age and cholesterol. 0.70 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 9

Logistic Model Case Study: Survival of Titanic Passengers

Data source: *The Titanic Passenger List* edited by Michael A. Findlay, originally published in Eaton & Haas (1994) *Titanic: Triumph and Tragedy*, Patrick Stephens Ltd, and expanded with the help of the Internet community. The original `html` files were obtained from Philip Hind (1999) (`http://atschool.eduweb.co.uk/phind`). The dataset was compiled and interpreted by Thomas Cason. It is available in R, S-PLUS, and Excel formats from `biostat.mc.vanderbilt.edu/DataSets` under the name `titanic3`.

9.1 Descriptive Statistics

```
require(rms)
```

```
getHdata(titanic3)      # get dataset from web site
# List of names of variables to analyze
v ← c('pclass','survived','age','sex','sibsp','parch')
t3 ← titanic3[, v]
```

```
units(t3$age) ← 'years'
latex(describe(t3), file='')
```

t3
6 Variables 1309 Observations

pclass

n	missing	unique
1309	0	3

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

survived : Survived

n	missing	unique	Info	Sum	Mean
1309	0	2	0.71	500	0.382

age : Age [years]

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
1046	263	98	1	29.88	5	14	21	28	39	50	57

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

sex

n	missing	unique
1309	0	2

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

sibsp : Number of Siblings/Spouses Aboard

n	missing	unique	Info	Mean
1309	0	7	0.67	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	Info	Mean
1309	0	8	0.55	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(t3)
# describe distributions of variables to rms
options(datadist='dd')
s ← summary(survived ∼ age + sex + pclass +
                cut2(sibsp,0:3) + cut2(parch,0:3), data=t3)
plot(s, main='', subtitles=FALSE)    # Figure 9.1
```

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

```
tn ← transform(t3,
                   agec = ifelse(age < 21, 'child', 'adult'),
```

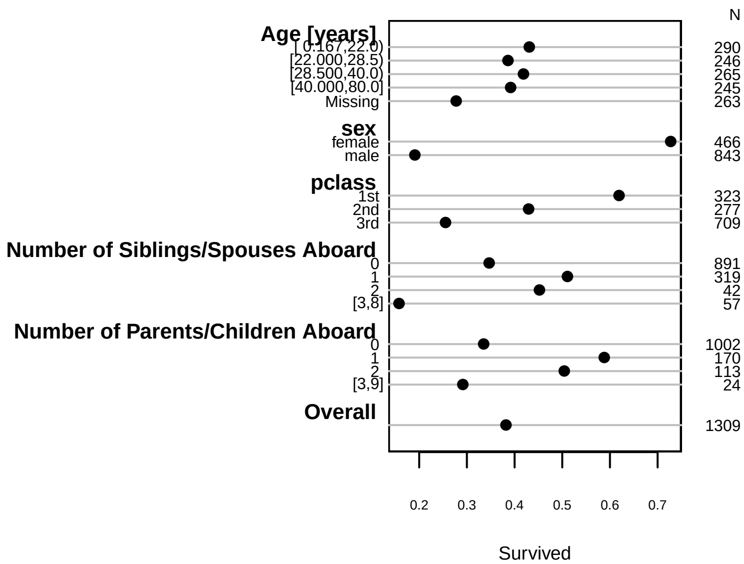


Figure 9.1: Univariable summaries of Titanic survival

```
                sibsp= ifelse(sibsp == 0, 'no sib/sp', 'sib/sp'),
                parch= ifelse(parch == 0, 'no par/child', 'par/child'))

g ← function(y) if(length(y) < 25) NA else mean(y)
s ← with(tn, summarize(survived,
           llist(agec, sex, pclass, sibsp, parch), g))
# llist, summarize in Hmisc package
# Figure 9.2:
ggplot(subset(s, agec != 'NA'), aes(x=survived, y=pclass, shape=sex)) +
  geom_point() + facet_grid(agec ~ sibsp * parch) +
  xlab('Proportion Surviving') + ylab('Passenger Class') +
  scale_x_continuous(breaks=c(0, .5, 1))
```

9.2 Exploring Trends with Nonparametric Regression

```
# Figure 9.3
b  ← scale_size_discrete(range=c(.1, .85))
yl ← ylab(NULL)
p1 ← ggplot(t3, aes(x=age, y=survived)) +
      histSpikeg(survived ~ age, lowess=TRUE, data=t3) +
      ylim(0,1) + yl
p2 ← ggplot(t3, aes(x=age, y=survived, color=sex)) +
      histSpikeg(survived ~ age + sex, lowess=TRUE,
```

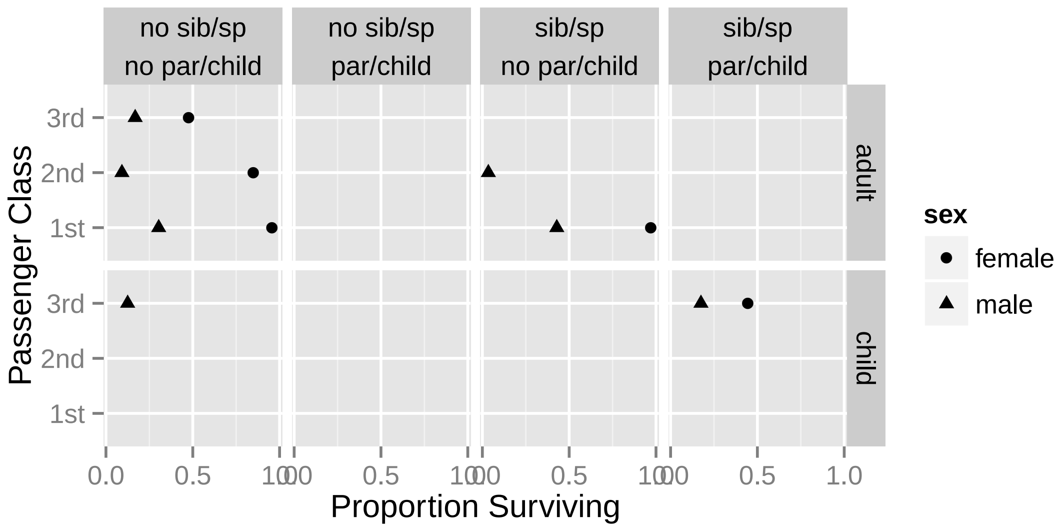


Figure 9.2: Multi-way summary of Titanic survival

```
                 data=t3) + ylim(0,1) + yl
p3 ← ggplot(t3, aes(x=age, y=survived, size=pclass)) +
      histSpikeg(survived ∼ age + pclass, lowess=TRUE,
                 data=t3) + b + ylim(0,1) + yl
p4 ← ggplot(t3, aes(x=age, y=survived, color=sex,
       size=pclass)) +
      histSpikeg(survived ∼ age + sex + pclass,
                 lowess=TRUE, data=t3) +
      b + ylim(0,1) + yl
gridExtra::grid.arrange(p1, p2, p3, p4, ncol=2)   # combine 4
```

```
# Figure 9.4
top ← theme(legend.position='top')
p1 ← ggplot(t3, aes(x=age, y=survived, color=cut2(sibsp,
       0:2))) + stat_plsmo() + b + ylim(0,1) + yl + top +
      scale_color_discrete(name='siblings/spouses')
p2 ← ggplot(t3, aes(x=age, y=survived, color=cut2(parch,
       0:2))) + stat_plsmo() + b + ylim(0,1) + yl + top +
      scale_color_discrete(name='parents/children')
gridExtra::grid.arrange(p1, p2, ncol=2)
```

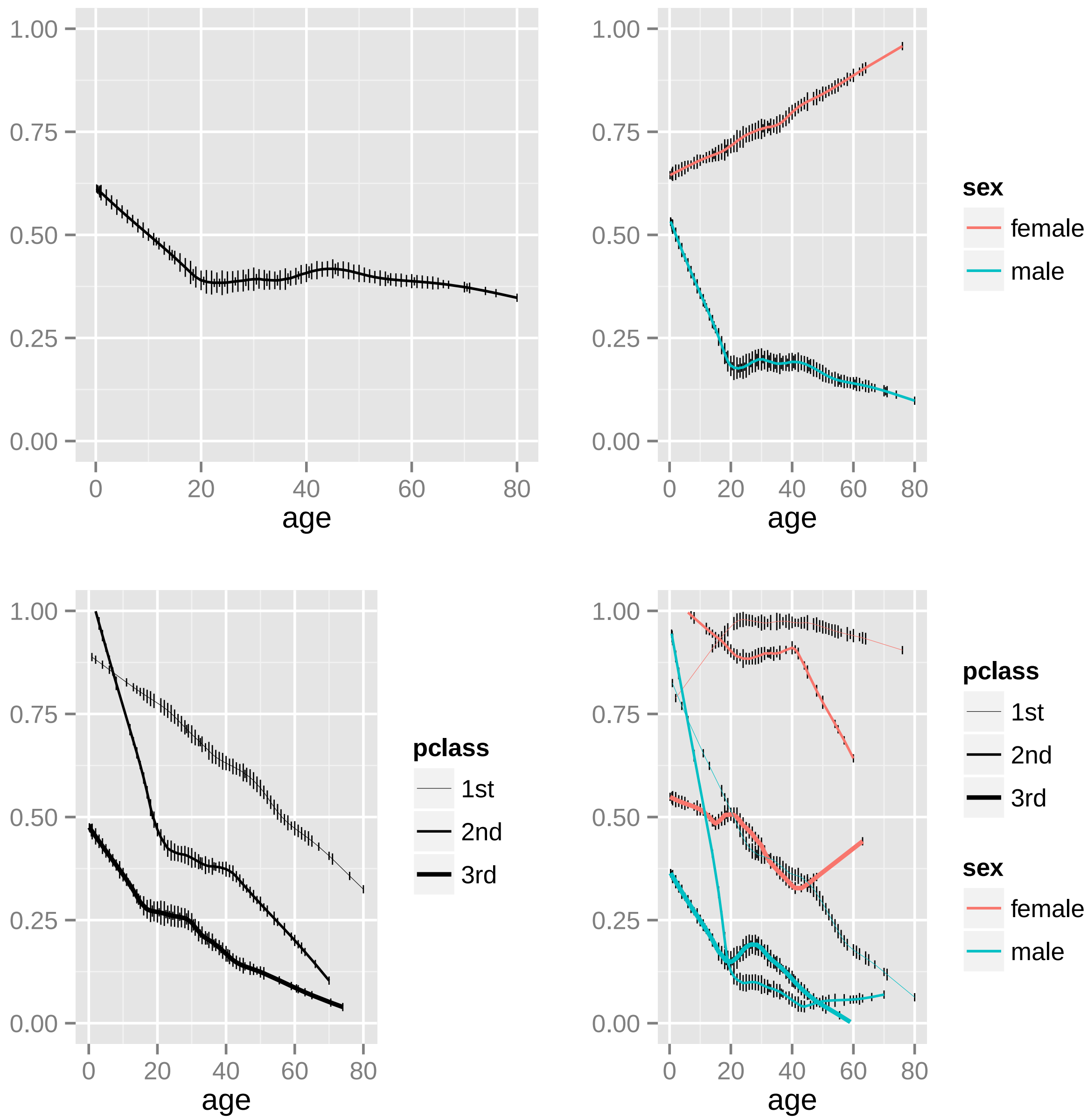


Figure 9.3: Nonparametric regression (`loess`) estimates of the relationship between age and the probability of surviving the Titanic, with tick marks depicting the age distribution. The top left panel shows unstratified estimates of the probability of survival. Other panels show nonparametric estimates by various stratifications.

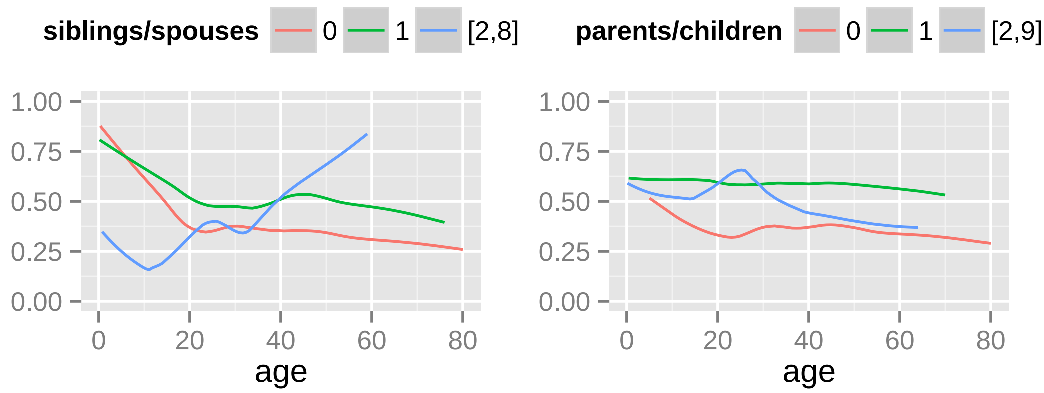


Figure 9.4: Relationship between age and survival stratified by the number of siblings or spouses on board (left panel) or by the number of parents or children of the passenger on board (right panel).

9.3 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), data=t3)   # Table 9.1
latex(anova(f1), file='', label='titanic−anova3', size='small')
```

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

```
f ← lrm(survived ∼ (sex + pclass + rcs(age,5))^2 +
        rcs(age,5)∗sibsp, data=t3)
print(f, latex=TRUE)
```

Logistic Regression Model

```
lrm(formula = survived ~ (sex + pclass + rcs(age, 5))^2 + rcs(age,
    5) * sibsp, data = t3)
```

Table 9.1: Wald Statistics for `survived`

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

Frequencies of Missing Values Due to Each Variable

```
survived      sex   pclass      age    sibsp
       0        0        0      263        0
```

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

	Coef	S.E.	Wald Z	Pr$(> \lvert Z\rvert)$
Intercept	3.3075	1.8427	1.79	0.0727
sex=male	-1.1478	1.0878	-1.06	0.2914
pclass=2nd	6.7309	3.9617	1.70	0.0893
pclass=3rd	-1.6437	1.8299	-0.90	0.3691
age	0.0886	0.1346	0.66	0.5102
age’	-0.7410	0.6513	-1.14	0.2552
age”	4.9264	4.0047	1.23	0.2186
age”’	-6.6129	5.4100	-1.22	0.2216
sibsp	-1.0446	0.3441	-3.04	0.0024
sex=male * pclass=2nd	-0.7682	0.7083	-1.08	0.2781
sex=male * pclass=3rd	2.1520	0.6214	3.46	0.0005
sex=male * age	-0.2191	0.0722	-3.04	0.0024
sex=male * age’	1.0842	0.3886	2.79	0.0053
sex=male * age”	-6.5578	2.6511	-2.47	0.0134
sex=male * age”’	8.3716	3.8532	2.17	0.0298
pclass=2nd * age	-0.5446	0.2653	-2.05	0.0401
pclass=3rd * age	-0.1634	0.1308	-1.25	0.2118
pclass=2nd * age’	1.9156	1.0189	1.88	0.0601
pclass=3rd * age’	0.8205	0.6091	1.35	0.1780
pclass=2nd * age”	-8.9545	5.5027	-1.63	0.1037
pclass=3rd * age”	-5.4276	3.6475	-1.49	0.1367
pclass=2nd * age”’	9.3926	6.9559	1.35	0.1769
pclass=3rd * age”’	7.5403	4.8519	1.55	0.1202
age * sibsp	0.0357	0.0340	1.05	0.2933
age’ * sibsp	-0.0467	0.2213	-0.21	0.8330
age” * sibsp	0.5574	1.6680	0.33	0.7382
age”’ * sibsp	-1.1937	2.5711	-0.46	0.6425

```
latex(anova(f), file='', label='titanic-anova2', size='small') #9.3
```

Table 9.3: Wald Statistics for `survived`

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

Show the many effects of predictors.

```
p ← Predict(f, age, sex, pclass, sibsp=0, fun=plogis)
ggplot(p)       # Fig. 9.5
```

```
ggplot(Predict(f, sibsp, age=c(10,15,20,50), conf.int=FALSE))
## Figure 9.6
```

Note that children having many siblings apparently had lower survival. Married adults had slightly higher survival than unmarried ones.

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

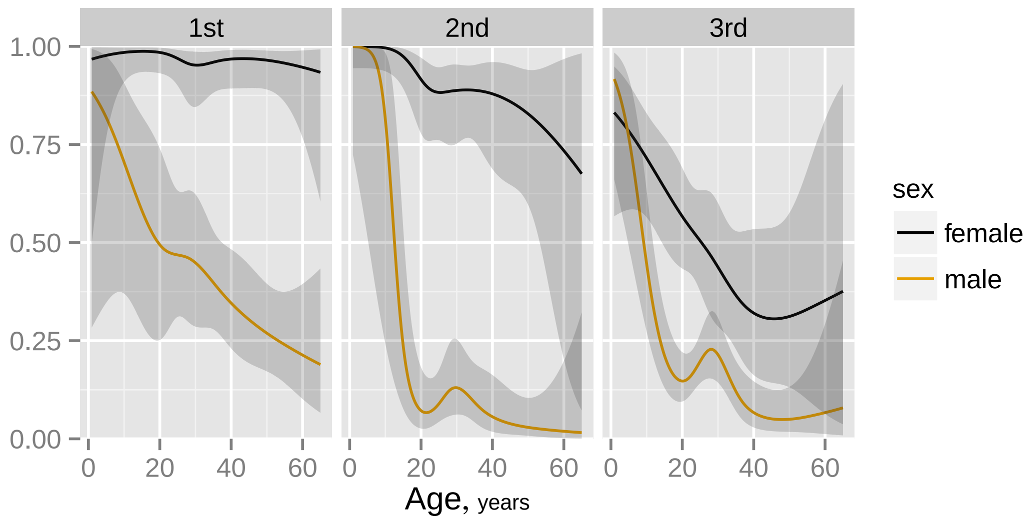


Figure 9.5: Effects of predictors on probability of survival of Titanic passengers, estimated for zero siblings or spouses

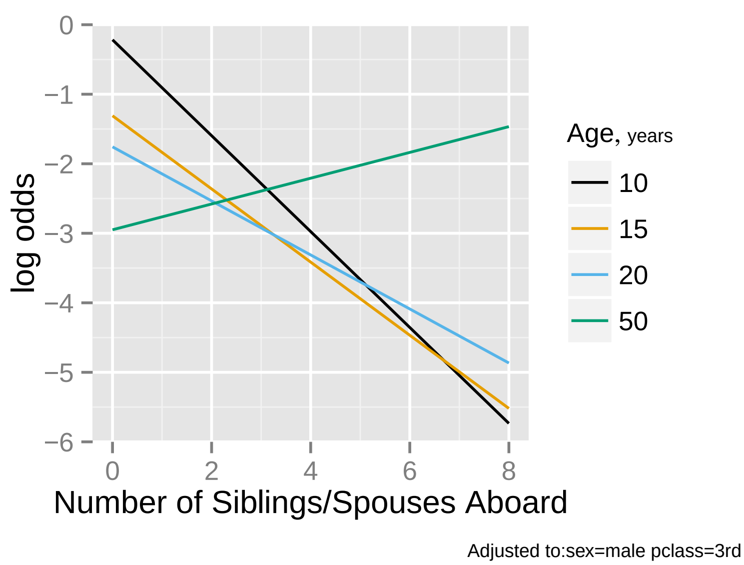


Figure 9.6: Effect of number of siblings and spouses on the log odds of surviving, for third class males

```
f ← update(f, x=TRUE, y=TRUE)
# x=TRUE, y=TRUE adds raw data to fit object so can bootstrap
set.seed(131)                    # so can replicate re-samples
latex(validate(f, B=200), digits=2, size='Ssize')
```

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.76	0.77	0.74	0.03	0.72	200
R^2	0.55	0.58	0.53	0.05	0.50	200
Intercept	0.00	0.00	−0.08	0.08	−0.08	200
Slope	1.00	1.00	0.87	0.13	0.87	200
$E_{\max}$	0.00	0.00	0.05	0.05	0.05	200
D	0.53	0.56	0.50	0.06	0.46	200
U	0.00	0.00	0.01	−0.01	0.01	200
Q	0.53	0.56	0.49	0.07	0.46	200
B	0.13	0.13	0.13	−0.01	0.14	200
g	2.43	2.75	2.37	0.37	2.05	200
g_p	0.37	0.37	0.35	0.02	0.35	200

```
cal ← calibrate(f, B=200)        # Figure 9.7
plot(cal, subtitles=FALSE)
```

```
n=1046   Mean absolute error=0.009   Mean squared error=0.00012
0.9 Quantile of absolute error=0.017
```

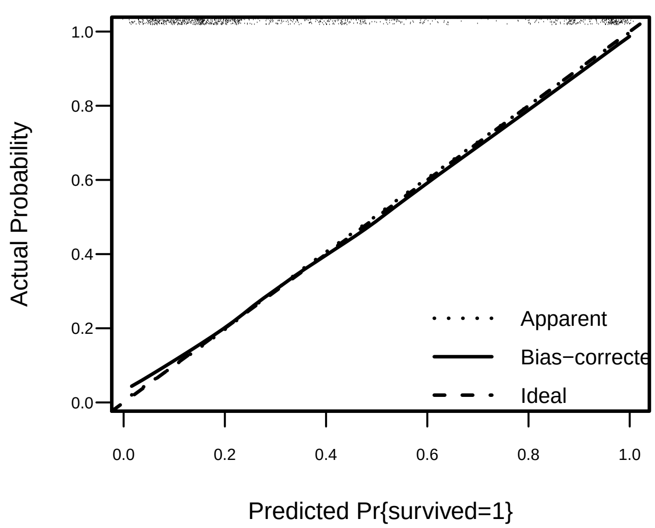


Figure 9.7: Bootstrap overfitting-corrected loess nonparametric calibration curve for casewise deletion model

But moderate problem with missing data

9.4 Examining Missing Data Patterns

```
na.patterns ← naclus(titanic3)
require(rpart)      # Recursive partitioning package
```

```
who.na ← rpart(is.na(age) ~ sex + pclass + survived +
               sibsp + parch, data=titanic3, minbucket=15)
naplot(na.patterns, 'na per var')
plot(who.na, margin=.1); text(who.na) # Figure 9.8
plot(na.patterns)
```

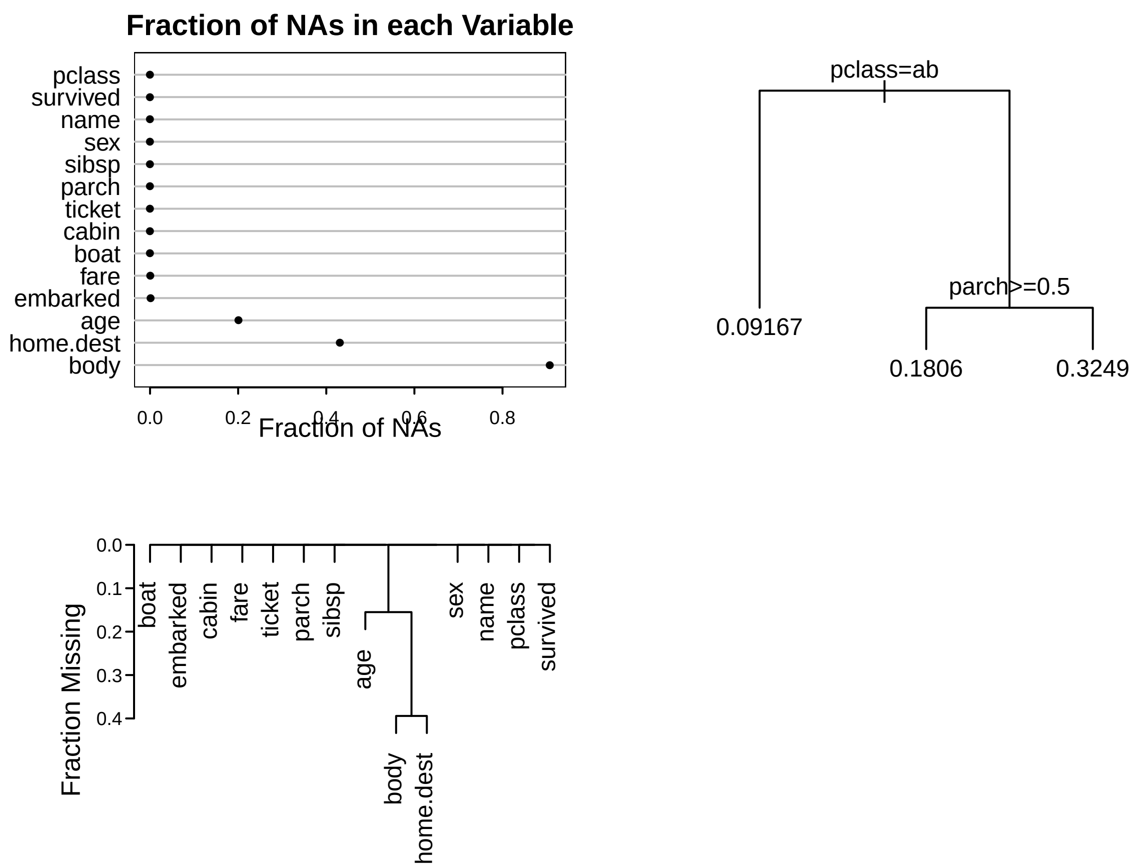


Figure 9.8: Patterns of missing data. Upper left panel shows the fraction of observations missing on each predictor. Lower panel depicts a hierarchical cluster analysis of missingness combinations. The similarity measure shown on the Y-axis is the fraction of observations for which both variables are missing. Right panel shows the result of recursive partitioning for predicting `is.na(age)`. The `rpart` function found only strong patterns according to passenger class.

```
plot(summary(is.na(age) ~ sex + pclass + survived +
             sibsp + parch, data=t3))  # Figure 9.9
```

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

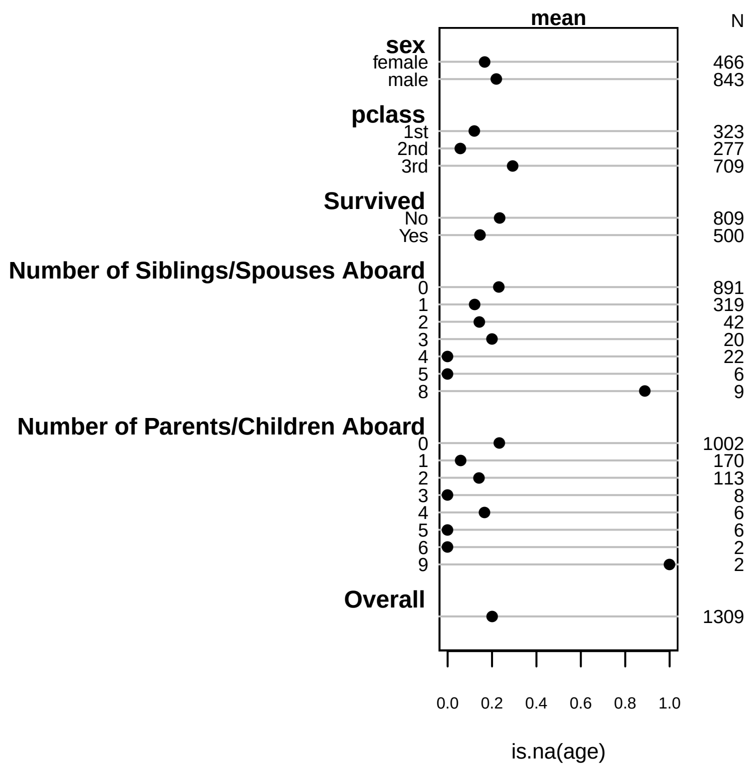


Figure 9.9: Univariable descriptions of proportion of passengers with missing age

```
        data=t3)
print(m, latex=TRUE, needspace='2in')
```

Logistic Regression Model

```
lrm(formula = is.na(age) ~ sex * pclass + survived + sibsp +
    parch, data = t3)
```

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

Table 9.5: 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

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

```
latex(anova(m), file='', label='titanic-anova.na')# Table 9.5
```

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

9.5 Single Conditional Mean Imputation

First try: conditional mean imputation
Default spline transformation for age caused distribution of imputed values to be much dif-

ferent from non-imputed ones; constrain to linear

```
xtrans ← transcan(~ I(age) + sex + pclass + sibsp + parch,
                  imputed=TRUE, pl=FALSE, pr=FALSE, data=t3)
```

```
summary(xtrans)
```

```
transcan(x = ~I(age) + sex + pclass + sibsp + parch, imputed = TRUE,
    pr = FALSE, pl = FALSE, data = t3)

Iterations: 5

R^2 achieved in predicting each variable:

   age     sex pclass  sibsp  parch
 0.264  0.076  0.242  0.249  0.291

Adjusted R^2:

   age     sex pclass  sibsp  parch
 0.260  0.073  0.239  0.245  0.288

Coefficients of canonical variates for predicting each (row) variable

       age   sex   pclass sibsp parch
age           0.92  6.05  -2.02 -2.65
sex     0.03       -0.56  -0.01 -0.75
pclass  0.08 -0.26         0.03  0.28
sibsp  -0.02  0.00  0.03         0.86
parch  -0.03 -0.30  0.23   0.75

Summary of imputed values

age
      n missing  unique    Info    Mean     .05     .10
    263       0      24    0.91   28.53   17.34   21.77
    .25     .50     .75     .90     .95
  26.17   28.10   28.10   42.77   42.77

lowest :  9.829 11.757 13.224 15.152 17.283
highest: 33.246 34.738 38.638 40.840 42.768

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 ← with(t3, impute(xtrans, age, data=t3))
i ← is.imputed(age.i)
with(t3, tapply(age.i[i], list(sex[i],pclass[i]), mean))
```

```
            1st      2nd      3rd
female 39.08396 31.31831 23.10548
male   42.76765 33.24650 26.87451
```

```
with(t3, tapply(age, list(sex,pclass), mean, na.rm=TRUE))
```

```
            1st      2nd      3rd
female 37.03759 27.49919 22.18531
male   41.02925 30.81540 25.96227
```

```
dd   ← datadist(dd, age.i)
f.si ← lrm(survived ~ (sex + pclass + rcs(age.i,5))^2 +
           rcs(age.i,5)*sibsp, data=t3)
print(f.si, coefs=FALSE, latex=TRUE)
```

Logistic Regression Model

```
lrm(formula = survived ~ (sex + pclass + rcs(age.i, 5))^2 + rcs(age.i,
    5) * sibsp, data = t3)
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	1309	LR χ^2	640.85	R^2	0.526	C	0.861
0	809	d.f.	26	g	2.223	D_{xy}	0.723
1	500	$\Pr(>\chi^2)$	<0.0001	g_r	9.233	γ	0.728
$\max\lvert\frac{\partial \log L}{\partial \beta}\rvert$	4×10^{-4}			g_p	0.346	τ_a	0.341
				Brier	0.133		

```
p1 ← Predict(f,    age,   pclass, sex, sibsp=0, fun=plogis)
p2 ← Predict(f.si, age.i, pclass, sex, sibsp=0, fun=plogis)
p  ← rbind('Casewise Deletion'=p1, 'Single Imputation'=p2,
           rename=c(age.i='age'))   # creates .set. variable
ggplot(p, groups='sex', ylab='Probability of Surviving')
# Figure 9.10
```

```
latex(anova(f.si), file='', label='titanic-anova.si')   # Table 9.6
```

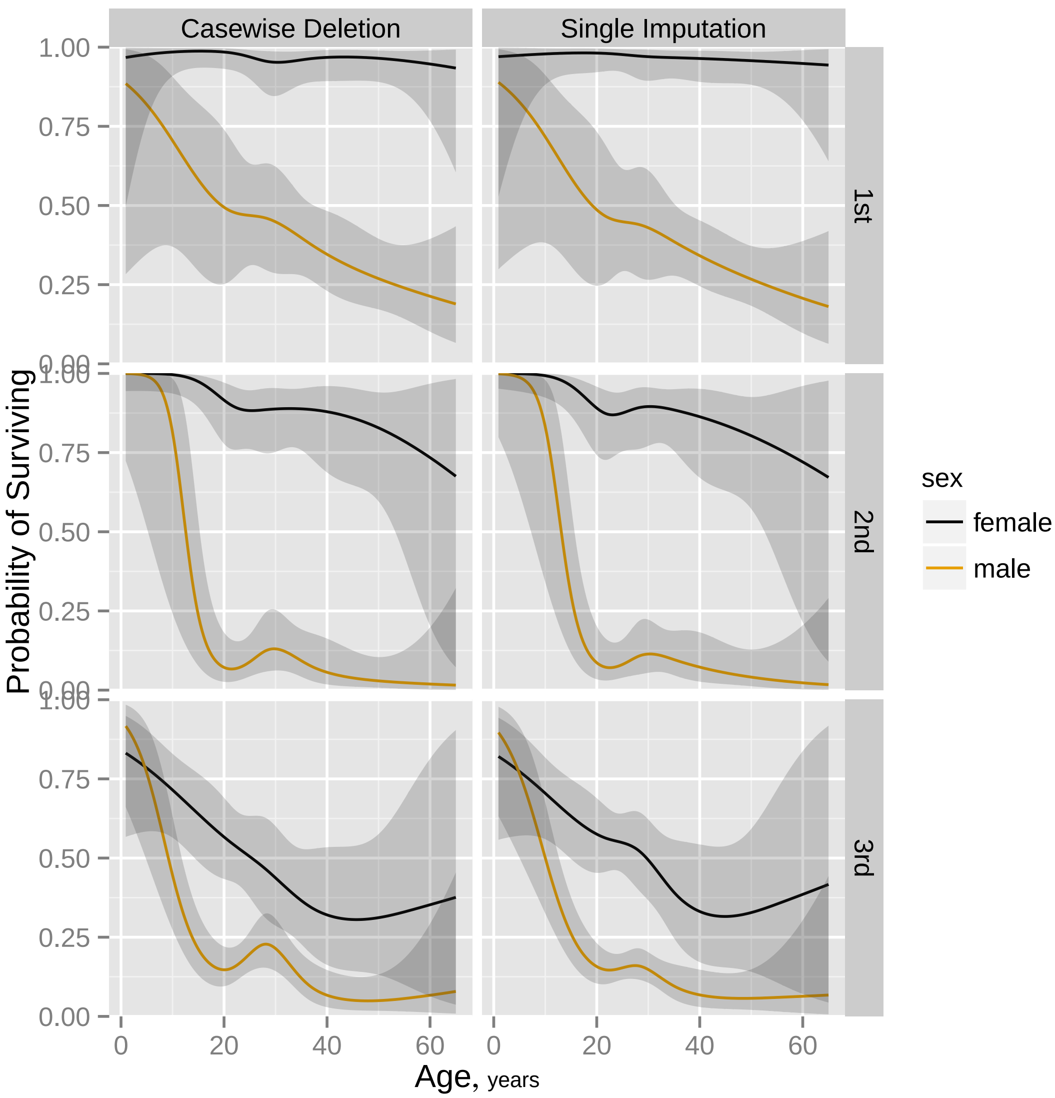


Figure 9.10: Predicted probability of survival for males from fit using casewise deletion (bottom) and single conditional mean imputation (top). `sibsp` is set to zero for these predicted values.

Table 9.6: Wald Statistics for `survived`

	χ^2	d.f.	P
sex (Factor+Higher Order Factors)	245.39	7	< 0.0001
All Interactions	52.85	6	< 0.0001
pclass (Factor+Higher Order Factors)	112.07	12	< 0.0001
All Interactions	36.79	10	0.0001
age.i (Factor+Higher Order Factors)	49.32	20	0.0003
All Interactions	25.62	16	0.0595
Nonlinear (Factor+Higher Order Factors)	19.71	15	0.1835
sibsp (Factor+Higher Order Factors)	22.02	5	0.0005
All Interactions	12.28	4	0.0154
sex × pclass (Factor+Higher Order Factors)	30.29	2	< 0.0001
sex × age.i (Factor+Higher Order Factors)	8.91	4	0.0633
Nonlinear	5.62	3	0.1319
Nonlinear Interaction : f(A,B) vs. AB	5.62	3	0.1319
pclass × age.i (Factor+Higher Order Factors)	6.05	8	0.6421
Nonlinear	5.44	6	0.4888
Nonlinear Interaction : f(A,B) vs. AB	5.44	6	0.4888
age.i × sibsp (Factor+Higher Order Factors)	12.28	4	0.0154
Nonlinear	2.05	3	0.5614
Nonlinear Interaction : f(A,B) vs. AB	2.05	3	0.5614
TOTAL NONLINEAR	19.71	15	0.1835
TOTAL INTERACTION	67.00	18	< 0.0001
TOTAL NONLINEAR + INTERACTION	69.53	21	< 0.0001
TOTAL	305.74	26	< 0.0001

9.6 Multiple Imputation

The following uses `aregImpute` with predictive mean matching. By default, `aregImpute` does not transform `age` when it is being predicted from the other variables. Four knots are used to transform `age` when used to impute other variables (not needed here as no other missings were present). Since the fraction of observations with missing age is $\frac{263}{1309} = 0.2$ we use 20 imputations.

```
set.seed(17)          # so can reproduce random aspects
mi ← aregImpute(~ age + sex + pclass +
                  sibsp + parch + survived,
                  data=t3, n.impute=20, nk=4, pr=FALSE)
```

```
mi
```

```

Multiple Imputation using Bootstrap and PMM

aregImpute(formula = ~age + sex + pclass + sibsp + parch + survived,
    data = t3, n.impute = 20, nk = 4, pr = FALSE)

n: 1309          p: 6    Imputations: 20          nk: 4

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

         type d.f.
age         s    1
sex         c    1
pclass      c    2
```

```
sibsp        s    2
parch        s    2
survived     l    1

Transformation of Target Variables Forced to be Linear

R-squares for Predicting Non-Missing Values for Each Variable
Using Last Imputations of Predictors
  age
0.295
```

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

```
    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
16    40   49   24   29 60.0   58   64   36   50    61
38    33   45   40   49 80.0    2   38   38   36    53
41    29   24   19   31 40.0   60   64   42   30    65
47    40   42   29   48 36.0   46   64   30   38    42
60    52   40   22   31 38.0   22   19   24   40    33
70    16   14   23   23 18.0   24   19   27   59    23
71    30   62   57   30 42.0   31   64   40   40    63
75    43   23   36   61 45.5   58   64   27   24    50
81    44   57   47   31 45.0   30   64   62   39    67
107   52   18   24   62 32.5   38   64   47   19    23
```

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

```
plot(mi)
Ecdf(t3$age, add=TRUE, col='gray', lwd=2,
     subtitles=FALSE) #Fig. 9.11
```

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,
```

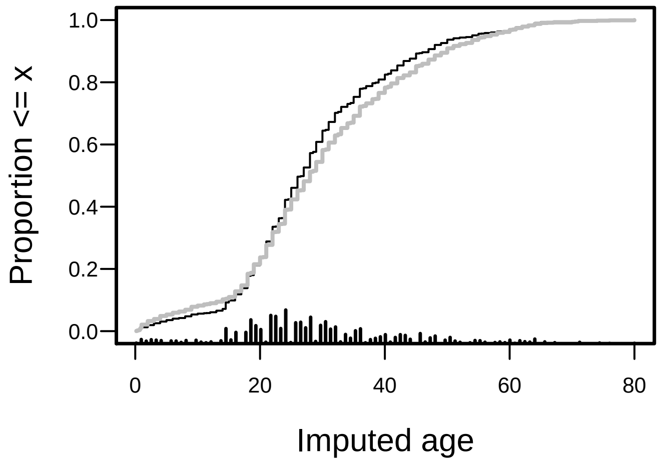


Figure 9.11: Distributions of imputed and actual ages for the Titanic dataset. Imputed values are in black and actual ages in gray.

Table 9.7: Wald Statistics for `survived`

	χ^2	d.f.	P
sex (Factor+Higher Order Factors)	240.42	7	< 0.0001
All Interactions	54.56	6	< 0.0001
pclass (Factor+Higher Order Factors)	114.21	12	< 0.0001
All Interactions	36.43	10	0.0001
age (Factor+Higher Order Factors)	50.37	20	0.0002
All Interactions	25.88	16	0.0557
Nonlinear (Factor+Higher Order Factors)	24.21	15	0.0616
sibsp (Factor+Higher Order Factors)	24.22	5	0.0002
All Interactions	12.86	4	0.0120
sex × pclass (Factor+Higher Order Factors)	30.99	2	< 0.0001
sex × age (Factor+Higher Order Factors)	11.38	4	0.0226
Nonlinear	8.15	3	0.0430
Nonlinear Interaction : f(A,B) vs. AB	8.15	3	0.0430
pclass × age (Factor+Higher Order Factors)	5.30	8	0.7246
Nonlinear	4.63	6	0.5918
Nonlinear Interaction : f(A,B) vs. AB	4.63	6	0.5918
age × sibsp (Factor+Higher Order Factors)	12.86	4	0.0120
Nonlinear	1.84	3	0.6058
Nonlinear Interaction : f(A,B) vs. AB	1.84	3	0.6058
TOTAL NONLINEAR	24.21	15	0.0616
TOTAL INTERACTION	67.12	18	< 0.0001
TOTAL NONLINEAR + INTERACTION	70.99	21	< 0.0001
TOTAL	298.78	26	< 0.0001

```
  lrm, mi, data=t3, pr=FALSE)
latex(anova(f.mi), file='', label='titanic-anova.mi', size='small')   # Table
```

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

Show estimated effects of age by classes.

```
p1 ← Predict(f.si,  age.i, pclass, sex, sibsp=0, fun=plogis)
p2 ← Predict(f.mi,  age,   pclass, sex, sibsp=0, fun=plogis)
p  ← rbind('Single Imputation'=p1, 'Multiple Imputation'=p2,
            rename=c(age.i='age'))
ggplot(p, groups='sex', ylab='Probability of Surviving')
# Figure 9.12
```

9.7 Summarizing the Fitted Model

Show odds ratios for changes in predictor values

```
# Get predicted values for certain types of passengers
s ← summary(f.mi, age=c(1,30), sibsp=0:1)
# override default ranges for 3 variables
plot(s, log=TRUE, main='')               # Figure 9.13
```

```
phat ← predict(f.mi,
                combos ←
         expand.grid(age=c(2,21,50),sex=levels(t3$sex),
                     pclass=levels(t3$pclass),
                     sibsp=0), type='fitted')
# Can also use Predict(f.mi, age=c(2,21,50), sex, pclass,
#                      sibsp=0, fun=plogis)$yhat
options(digits=1)
data.frame(combos, phat)
```

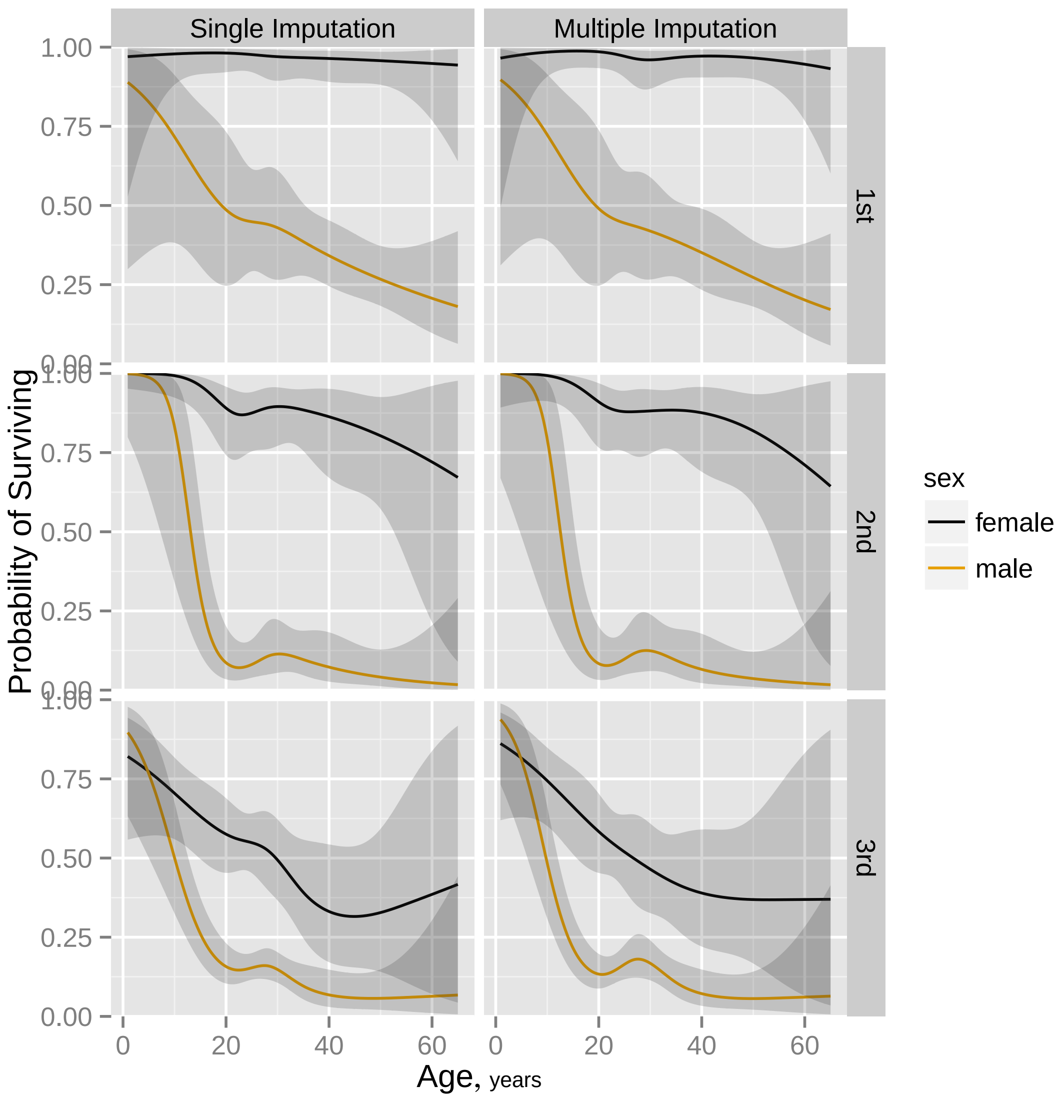


Figure 9.12: Predicted probability of survival for males from fit using single conditional mean imputation again (top) and multiple random draw imputation (bottom). Both sets of predictions are for `sibsp`=0.

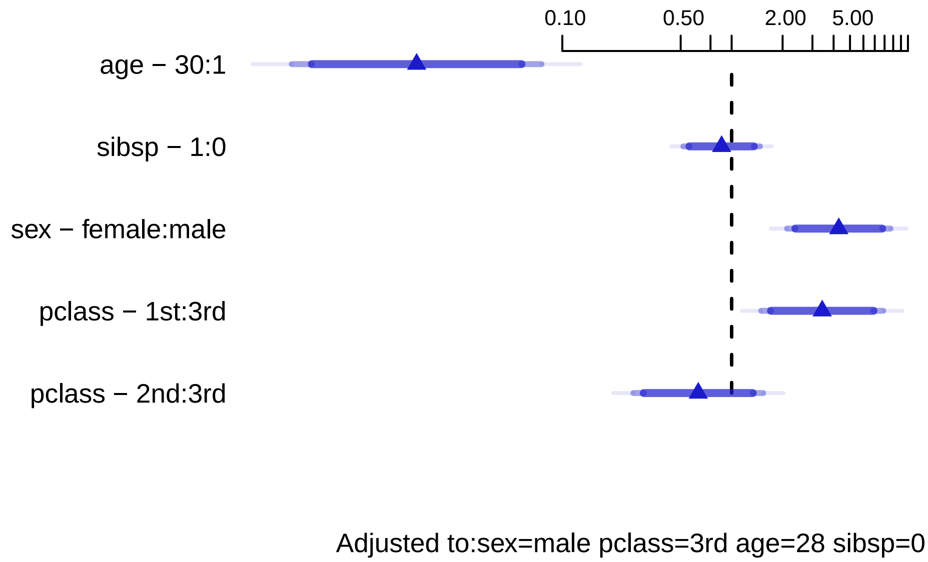


Figure 9.13: Odds ratios for some predictor settings

```
   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 0.88
5   21   male    1st     0 0.48
6   50   male    1st     0 0.27
7    2 female    2nd     0 1.00
8   21 female    2nd     0 0.90
9   50 female    2nd     0 0.82
10   2   male    2nd     0 1.00
11  21   male    2nd     0 0.08
12  50   male    2nd     0 0.04
13   2 female    3rd     0 0.85
14  21 female    3rd     0 0.57
15  50 female    3rd     0 0.37
16   2   male    3rd     0 0.91
17  21   male    3rd     0 0.13
18  50   male    3rd     0 0.06
```

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

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

```
pred.logit ← Function(f.mi)
# Note: if don't define sibsp to pred.logit, defaults to 0
# normally just type the function name to see its body
latex(pred.logit, file='', type='Sinput', size='small',
      width.cutoff=49)
```

```
pred.logit ← function (sex = "male", pclass = "3rd", age = 28,
    sibsp = 0)
{
    3.2427671 − 0.95431809 ∗ (sex == "male") + 5.4086505 ∗
        (pclass == "2nd") − 1.3378623 ∗ (pclass ==
        "3rd") + 0.091162649 ∗ age − 0.00031204327 ∗
        pmax(age − 6, 0)^3 + 0.0021750413 ∗ pmax(age −
        21, 0)^3 − 0.0027627032 ∗ pmax(age − 27, 0)^3 +
        0.0009805137 ∗ pmax(age − 36, 0)^3 − 8.0808484e−05 ∗
        pmax(age − 55.8, 0)^3 − 1.1567976 ∗ sibsp +
        (sex == "male") ∗ (−0.46061284 ∗ (pclass ==
            "2nd") + 2.0406523 ∗ (pclass == "3rd")) +
        (sex == "male") ∗ (−0.22469066 ∗ age + 0.00043708296 ∗
            pmax(age − 6, 0)^3 − 0.0026505136 ∗ pmax(age −
            21, 0)^3 + 0.0031201404 ∗ pmax(age − 27,
            0)^3 − 0.00097923749 ∗ pmax(age − 36,
            0)^3 + 7.2527708e−05 ∗ pmax(age − 55.8,
            0)^3) + (pclass == "2nd") ∗ (−0.46144083 ∗
        age + 0.00070194849 ∗ pmax(age − 6, 0)^3 −
        0.0034726662 ∗ pmax(age − 21, 0)^3 + 0.0035255387 ∗
        pmax(age − 27, 0)^3 − 0.0007900891 ∗ pmax(age −
        36, 0)^3 + 3.5268151e−05 ∗ pmax(age − 55.8,
        0)^3) + (pclass == "3rd") ∗ (−0.17513289 ∗
        age + 0.00035283358 ∗ pmax(age − 6, 0)^3 −
        0.0023049372 ∗ pmax(age − 21, 0)^3 + 0.0028978962 ∗
        pmax(age − 27, 0)^3 − 0.00105145 ∗ pmax(age −
        36, 0)^3 + 0.00010565735 ∗ pmax(age − 55.8,
        0)^3) + sibsp ∗ (0.040830773 ∗ age − 1.5627772e−05 ∗
        pmax(age − 6, 0)^3 + 0.00012790256 ∗ pmax(age −
        21, 0)^3 − 0.00025039385 ∗ pmax(age − 27,
        0)^3 + 0.00017871701 ∗ pmax(age − 36, 0)^3 −
        4.0597949e−05 ∗ pmax(age − 55.8, 0)^3)
}
```

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

```
[1] 0.914817 0.132640 0.056248
```

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

R Software Used

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

[a]Written by Atkinson & Therneau

Chapter 10

Ordinal Logistic Regression

10.1 Background

- Levels of Y are ordered; no spacing assumed
- If no model assumed, one can still assess association between X and Y
- Example: $Y = 0, 1, 2$ corresponds to no event, heart attack, death. Test of association between race (3 levels) and outcome (3 levels) can be obtained from a 2×2 d.f. χ^2 test for a contingency table
- If willing to assuming an ordering of Y *and* a model, can test for association using 2×1 d.f.

- Proportional odds model: generalization of Wilcoxon-Mann-Whitney-Kruskal-Wallis-Spearma
- Can have n categories for n observations!
- Continuation ratio model: discrete proportional hazards model

10.2 Ordinality Assumption

- Assume X is linearly related to some appropriate log odds
- Estimate mean $X|Y$ with and without assuming the model holds

10.3 Proportional Odds Model

10.3.1 Model

- Walker & Duncan[178] — most popular ordinal response model

- For convenience $Y = 0, 1, 2, \ldots, k$

  $$\Pr[Y \geq j|X] = \frac{1}{1 + \exp[-(\alpha_j + X\beta)]},$$

  where $j = 1, 2, \ldots, k$.
- α_j is the logit of Prob$[Y \geq j]$ when all Xs are zero
- Odds$[Y \geq j|X] = \exp(\alpha_j + X\beta)$
- Odds$[Y \geq j|X_m = a+1]$ / Odds$[Y \geq j|X_m = a] = e^{\beta_m}$
- Same odds ratio e^{β_k} for any $j = 1, 2, \ldots, k$
- Odds$[Y \geq j|X]$ / Odds$[Y \geq v|X] = \frac{e^{\alpha_j + X\beta}}{e^{\alpha_v + X\beta}} = e^{\alpha_j - \alpha_v}$
- Odds$[Y \geq j|X] = constant \times$ Odds$[Y \geq v|X]$
- Assumes OR for 1 unit increase in age is the same when considering the probability of death as when considering the probability of death or heart attack
- PO model only uses ranks of Y; same $\hat{\beta}$s if transform Y; is robust to outliers

10.3.2 Assumptions and Interpretation of Parameters

10.3.3 Estimation

10.3.4 Residuals

- Construct binary events $Y \geq j, j = 1, 2, \ldots, k$ and use corresponding predicted probabilities
$$\hat{P}_{ij} = \frac{1}{1 + \exp[-(\hat{\alpha}_j + X_i\hat{\beta})]},$$
- Score residual for subject i predictor m:
$$U_{im} = X_{im}([Y_i \geq j] - \hat{P}_{ij}),$$
- For each column of U plot mean $\bar{U}_{\cdot m}$ and C.L. against Y
- Partial residuals are more useful as they can also estimate covariable transformations[36, 105]:
$$r_{im} = \hat{\beta}_m X_{im} + \frac{Y_i - \hat{P}_i}{\hat{P}_i(1 - \hat{P}_i)},$$

where

$$\hat{P}_i = \frac{1}{1 + \exp[-(\alpha + X_i\hat{\beta})]}.$$

- Smooth r_{im} vs. X_{im} to estimate how X_m relates to the log relative odds that $Y = 1|X_m$
- For ordinal Y compute binary model partial res. for all cutoffs j:

$$r_{im} = \hat{\beta}_m X_{im} + \frac{[Y_i \geq j] - \hat{P}_{ij}}{\hat{P}_{ij}(1 - \hat{P}_{ij})},$$

Li and Shepherd[112] have a residual for ordinal models that serves for the entire range of Y without the need to consider cutoffs. Their residual is useful for checking functional form of predictors but not the proportional odds assumption.

10.3.5 Assessment of Model Fit

- Section 10.2

- Stratified proportions $Y \geq j, j = 1, 2, \ldots, k$, since $\text{logit}(Y \geq j|X) - \text{logit}(Y \geq i|X) = \alpha_j - \alpha_i$, for any constant X

```
require(Hmisc)
```

```
getHdata(support)
sfdm ← as.integer(support$sfdm2) − 1
sf ← function(y)
  c('Y≥1'=qlogis(mean(y ≥ 1)), 'Y≥2'=qlogis(mean(y ≥ 2)),
    'Y≥3'=qlogis(mean(y ≥ 3)))
s ← summary(sfdm ∼ adlsc + sex + age + meanbp, fun=sf, data=support)
plot(s, which=1:3, pch=1:3, xlab='logit', vnames='names', main='',
     width.factor=1.5)
```

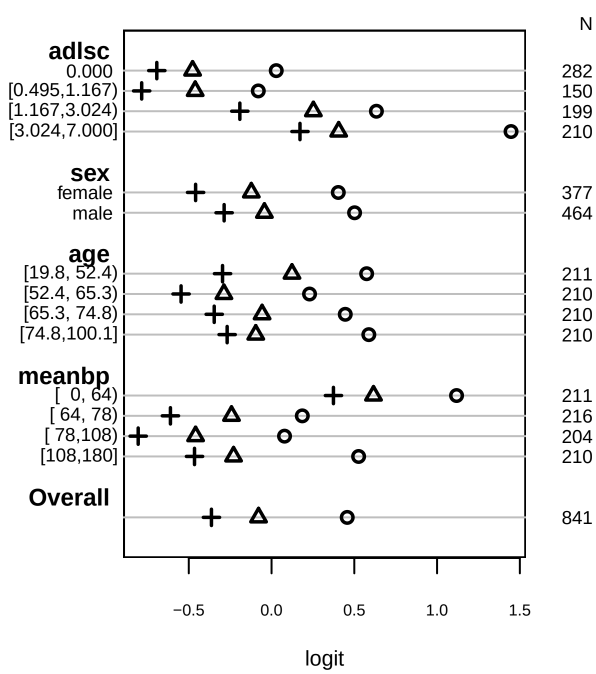


Figure 10.1: Checking PO assumption separately for a series of predictors. The circle, triangle, and plus sign correspond to $Y \geq 1, 2, 3$, respectively. PO is checked by examining the vertical constancy of distances between any two of these three symbols. Response variable is the severe functional disability scale `sfdm2` from the 1000-patient SUPPORT dataset, with the last two categories combined because of low frequency of coma/intubation.

When Y is continuous or almost continuous and X is discrete, the PO model assumes that the logit of the cumulative distribution function of Y is parallel across categories of X. The corresponding, more rigid, assumptions of the ordinary linear model (here, parametric ANOVA) are parallelism and linearity if the normal inverse cumulative distribution function across categories of X. As an example consider the web site's `diabetes` dataset, where we consider the distribution of log glycohemoglobin across subjects' body frames.

```
getHdata(diabetes)
a ← Ecdf(~ log(glyhb), group=frame, fun=qnorm, xlab='log(HbA1c)',
         label.curves=FALSE, data=diabetes,
         ylab=expression(paste(Phi^-1, (F[n](x)))))  # Figure 10.2
b ← Ecdf(~ log(glyhb), group=frame, fun=qlogis, xlab='log(HbA1c)',
         label.curves=list(keys='lines'), data=diabetes,
         ylab=expression(logit(F[n](x))))
print(a, more=TRUE, split=c(1,1,2,1))
print(b, split=c(2,1,2,1))
```

10.3.6 Quantifying Predictive Ability

10.3.7 Describing the Model

For PO models there are four and sometimes five types of relevant predictions:

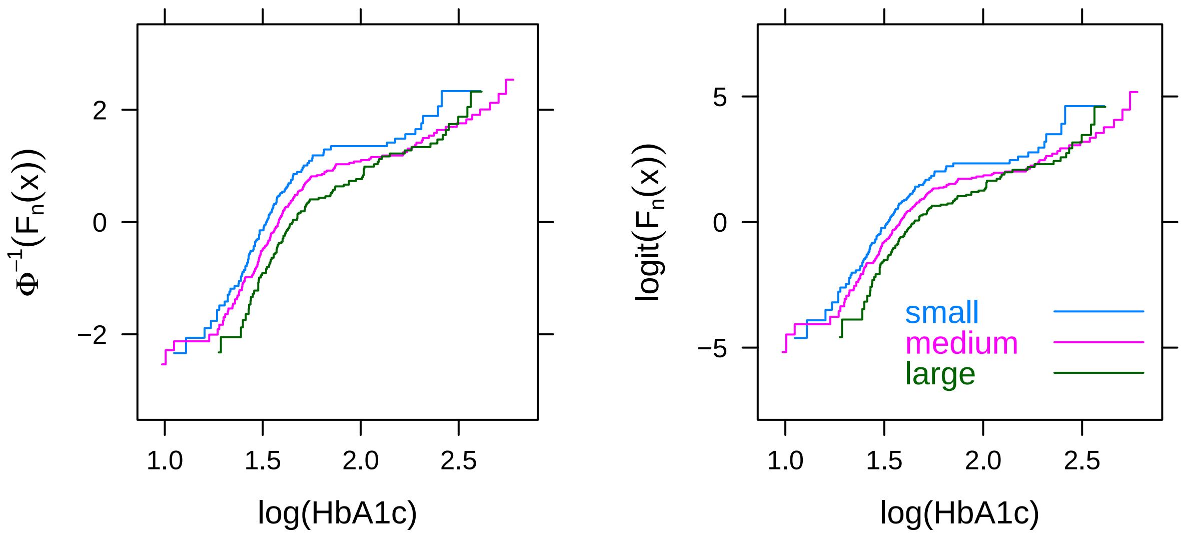


Figure 10.2: Transformed empirical cumulative distribution functions stratified by body frame in the `diabetes` dataset. Left panel: checking all assumptions of the parametric ANOVA. Right panel: checking all assumptions of the PO model (here, Kruskal–Wallis test).

1. logit$[Y \geq j|X]$, i.e., the linear predictor
2. Prob$[Y \geq j|X]$
3. Prob$[Y = j|X]$
4. Quantiles of $Y|X$ (e.g., the median[a])
5. $E(Y|X)$ if Y is interval scaled.

Graphics:

1. Partial effect plot (prob. scale or mean)
2. Odds ratio chart

[a] If Y does not have very many levels, the median will be a discontinuous function of X and may not be satisfactory.

3. Nomogram (possibly including the mean)

10.3.8 Validating the Fitted Model

10.3.9 R Functions

The `rms` package's `lrm` and `orm` functions fit the PO model directly, assuming that the levels of the response variable (e.g., the `levels` of a `factor` variable) are listed in the proper order. `predict` computes all types of estimates except for quantiles. `orm` allows for more link functions than the logistic and is intended to efficiently handle hundreds of intercepts as happens when Y is continuous.

The R functions `popower` and `posamsize` (in the `Hmisc` package) compute power and sample size estimates for ordinal responses using the proportional odds model.

The function `plot.xmean.ordinaly` in `rms` computes and graphs the quantities described in Section 10.2. It plots simple Y-stratified means overlaid with $\hat{E}(X|Y=j)$, with j on the x-axis. The $\hat{E}$s are computed for both PO and continuation ratio ordinal logistic models.

The `Hmisc` package's `summary.formula` function is also useful for assessing the PO assumption.

Generic `rms` functions such as `validate`, `calibrate`, and `nomogram` work with PO model fits from `lrm` as long as the analyst specifies which intercept(s) to use.

`rms` has a special function generator `Mean` for constructing an easy-to-use function for getting the predicted mean Y from a PO model. This is handy with `plot` and `nomogram`. If the fit has been run through `bootcov`, it is easy to use the `Predict` function to estimate bootstrap confidence limits for predicted means.

10.4 Continuation Ratio Model

10.4.1 Model

Unlike the PO model, which is based on *cumulative* probabilities, the continuation ratio (CR) model is based on *conditional* probabilities. The (forward) CR model[6,17,60] is stated as follows for $Y = 0, \ldots, k$:

$$\begin{aligned}
\Pr(Y = j|Y \geq j, X) &= \frac{1}{1 + \exp[-(\theta_j + X\gamma)]} \\
\mathrm{logit}(Y = 0|Y \geq 0, X) &= \mathrm{logit}(Y = 0|X) \\
&= \theta_0 + X\gamma \\
\mathrm{logit}(Y = 1|Y \geq 1, X) &= \theta_1 + X\gamma \\
&\ldots \\
\mathrm{logit}(Y = k-1|Y \geq k-1, X) &= \theta_{k-1} + X\gamma.
\end{aligned}$$

The CR model has been said to be likely to fit ordinal responses when subjects have to "pass through" one category to get to the next The

CR model is a discrete version of the Cox proportional hazards model. The discrete hazard function is defined as $\Pr(Y = j|Y \geq j)$.

Advantage of CR model: easy to allow unequal slopes across Y for selected X.

10.4.2 Assumptions and Interpretation of Parameters

10.4.3 Estimation

10.4.4 Residuals

To check CR model assumptions, binary logistic model partial residuals are again valuable. We separately fit a sequence of binary logistic models using a series of binary events and the corresponding applicable (increasingly small) subsets of subjects, and plot smoothed partial residuals against X for all of the binary events. Parallelism in these plots indicates that the CR model's constant γ assumptions are satisfied.

10.4.5 Assessment of Model Fit

10.4.6 Extended CR Model

10.4.7 Role of Penalization in Extended CR Model

10.4.8 Validating the Fitted Model

10.4.9 R Functions

The `cr.setup` function in `rms` returns a list of vectors useful in constructing a dataset used to trick a binary logistic function such as `lrm` into fitting CR models.

Chapter 11

Regression Models for Continuous Y and Case Study in Ordinal Regression

This chapter concerns univariate continuous Y. There are many multivariable models for predicting such response variables.

- linear models with assumed normal residuals, fitted with ordinary least squares
- generalized linear models and other parametric models based on special distributions such as the gamma
- generalized additive models (GAMs)
- generalization of GAMs to also nonparametrically transform Y
- quantile regression (see Section 11.3)

- other robust regression models that, like quantile regression, use an objective different from minimizing the sum of squared errors[171]
- semiparametric models based on the ranks of Y, such as the Cox proportional hazards model and the proportional odds ordinal logistic model
- cumulative probability models (often called *cumulative link models*) which are semiparametric models from a wider class of families than the logistic

Semiparametric models that treat Y as ordinal but not interval-scaled have many advantages including robustness and freedom of distributional assumptions for Y conditional on any given set of predictors.

Advantages are demonstrated in a case study of a cumulative probability ordinal model. Some of the results are compared to quantile regression and OLS. Many of the methods used in the case study also apply to ordinary linear models.

11.1 Dataset and Descriptive Statistics

- Diabetes Mellitus (DM) type II (adult onset diabetes) is strongly associated with obesity
- Primary laboratory test for diabetes: gylcosylated hemoglobin (HbA$_{1c}$), also called glycated hemoglobin, glycohemoglobin, or hemoglobin A_{1c}.
- HbA$_{1c}$ reflects average blood glucose for the preceding 60 to 90 days
- HbA$_{1c}$ > 7.0 usually taken as a positive diagnosis of diabetes
- Goal of analysis:
  - better understand effects of body size measurements on risk of DM
  - enhance screening for DM

- Best way to develop a model for DM screening is **not** to fit binary logistic model with HbA$_{1c} > 7$ as the response variable
  - All cutpoints are arbitrary; no justification for any putative cut
  - HbA$_{1c}$ 2=6.9, 7.1=10
  - Larger standard errors of $\hat{\beta}$, lower power, wider confidence bands
  - Better: predict continuous HbA$_{1c}$ using continuous response model, then convert to probability HbA$_{1c}$ exceeds any cutoff or estimate 0.9 quantile of HbA$_{1c}$
- Data: U.S. National Health and Nutrition Examination Survey (NHANES) from National Center for Health Statistics/CDC: `http://www.cdc.gov/nchs/nhanes.htm`[28]
- age ≥ 80 coded as 80 by CDC
- Subset with age ≥ 21, neither diagnosed nor treated for DM

```
require(rms)
```

```
getHdata(nhgh)
w ← subset(nhgh, age ≥ 21 & dx==0 & tx==0, select=−c(dx,tx))
latex(describe(w), file='')
```

w
18 Variables 4629 Observations

seqn : Respondent sequence number

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4629	0	4629	1	56902	52136	52633	54284	56930	59495	61079	61641

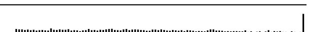
```
lowest : 51624 51629 51630 51645 51647
highest: 62152 62153 62155 62157 62158
```

sex

n	missing	unique
4629	0	2

```
male (2259, 49%), female (2370, 51%)
```

age : Age [years]



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4629	0	703	1	48.57	23.33	26.08	33.92	46.83	61.83	74.83	80.00

```
lowest : 21.00 21.08 21.17 21.25 21.33
highest: 79.67 79.75 79.83 79.92 80.00
```

re : Race/Ethnicity

n	missing	unique
4629	0	5

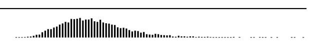
```
Mexican American (832, 18%), Other Hispanic (474, 10%)
Non-Hispanic White (2318, 50%), Non-Hispanic Black (756, 16%)
Other Race Including Multi-Racial (249, 5%)
```

income : Family Income

n	missing	unique
4389	240	14

```
[0,5000) (162, 4%), [5000,10000) (216, 5%), [10000,15000) (371, 8%)
[15000,20000) (300, 7%), [20000,25000) (374, 9%)
[25000,35000) (535, 12%), [35000,45000) (421, 10%)
[45000,55000) (346, 8%), [55000,65000) (257, 6%), [65000,75000) (188, 4%)
> 20000 (149, 3%), < 20000 (52, 1%), [75000,100000) (399, 9%)
>= 100000 (619, 14%)
```

wt : Weight [kg]



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4629	0	890	1	80.49	52.44	57.18	66.10	77.70	91.40	106.52	118.00

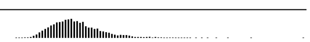
```
lowest :  33.2  36.1  37.9  38.5  38.7
highest: 184.3 186.9 195.3 196.6 203.0
```

ht : Standing Height [cm]

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4629	0	512	1	167.5	151.1	154.4	160.1	167.2	175.0	181.0	184.8

```
lowest : 123.3 135.4 137.5 139.4 139.8
highest: 199.2 199.3 199.6 201.7 202.7
```

bmi : Body Mass Index [kg/m^2]



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4629	0	1994	1	28.59	20.02	21.35	24.12	27.60	31.88	36.75	40.68

```
lowest : 13.18 14.59 15.02 15.40 15.49
highest: 61.20 62.81 65.62 71.30 84.87
```

leg : Upper Leg Length [cm]

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4474	155	216	1	38.39	32.0	33.5	36.0	38.4	41.0	43.3	44.6

```
lowest : 20.4 24.9 25.0 25.1 26.4, highest: 49.0 49.5 49.8 50.0 50.3
```

arml : Upper Arm Length [cm]

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4502	127	156	1	37.01	32.6	33.5	35.0	37.0	39.0	40.6	41.7

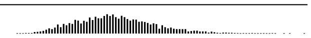
```
lowest : 24.8 27.0 27.5 29.2 29.5, highest: 45.2 45.5 45.6 46.0 47.0
```

armc : Arm Circumference [cm]

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4499	130	290	1	32.87	25.4	26.9	29.5	32.5	35.8	39.1	41.4

```
lowest : 17.9 19.0 19.3 19.5 19.9, highest: 54.2 54.9 55.3 56.0 61.0
```

waist : Waist Circumference [cm]



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4465	164	716	1	97.62	74.8	78.6	86.9	96.3	107.0	117.8	125.0

```
lowest :  59.7  60.0  61.5  62.0  62.4
highest: 160.0 160.6 162.2 162.7 168.7
```

tri : Triceps Skinfold [mm]

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4295	334	342	1	18.94	7.2	8.8	12.0	18.0	25.2	31.0	33.8

```
lowest :  2.6  3.1  3.2  3.3  3.4, highest: 39.6 39.8 40.0 40.2 40.6
```

sub : Subscapular Skinfold [mm]

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
3974	655	329	1	20.8	8.60	10.30	14.40	20.30	26.58	32.00	35.00

```
lowest :  3.8  4.2  4.6  4.8  4.9, highest: 40.0 40.1 40.2 40.3 40.4
```

gh : Glycohemoglobin [%]

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4629	0	63	0.99	5.533	4.8	5.0	5.2	5.5	5.8	6.0	6.3

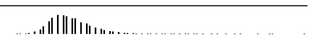
```
lowest :  4.0  4.1  4.2  4.3  4.4, highest: 11.9 12.0 12.1 12.3 14.5
```

albumin : Albumin [g/dL]

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4576	53	26	0.99	4.261	3.7	3.9	4.1	4.3	4.5	4.7	4.8

```
lowest : 2.6 2.7 3.0 3.1 3.2, highest: 4.9 5.0 5.1 5.2 5.3
```

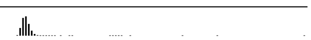
bun : Blood urea nitrogen [mg/dL]



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4576	53	50	0.99	13.03	7	8	10	12	15	19	22

```
lowest :  1  2  3  4  5, highest: 49 53 55 56 63
```

SCr : Creatinine [mg/dL]



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
4576	53	167	1	0.8887	0.58	0.62	0.72	0.84	0.99	1.14	1.25

```
lowest :  0.34  0.38  0.39  0.40  0.41
highest:  5.98  6.34  9.13 10.98 15.66
```

```
dd ← datadist(w); options(datadist='dd')
```

11.2 The Linear Model

The most popular multivariable model for analyzing a univariate continuous Y is the the linear model

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

where β is estimated using ordinary least squares, that is, by solving for $\hat{\beta}$ to minimize $\Sigma(Y_i - X\hat{\beta})^2$.

- To compute P-values and confidence limits using parametric methods (and for least squares estimates to coincide with maximum likelihood estimates) we would have to assume that $Y|X$ is normal with mean $X\beta$ and constant variance σ^2 [a]

11.2.1 Checking Assumptions of OLS and Other Models

- First see if `gh` would make a Gaussian residuals model fit

[a] The latter assumption may be dispensed with if we use a robust Huber–White or bootstrap covariance matrix estimate. Normality may sometimes be dispensed with by using bootstrap confidence intervals, but this would not fix inefficiency problems with OLS when residuals are non-normal.

- Use ordinary regression on 4 key variables to collapse into one variable (predicted mean from OLS model)
- Stratify predicted mean into 6 quantile groups
- Apply the normal inverse ECDF of `gh` to these strata and check for normality and constant σ^2
- ECDF is for $\text{Prob}[Y \leq y|X]$ but for ordinal modeling we want to state models in terms of $\text{Prob}[Y \geq y|X]$ so take 1 - ECDF before inverse transforming

```r
f ← ols(gh ~ rcs(age,5) + sex + re + rcs(bmi, 3), data=w)
pgh ← fitted(f)

p ← function(fun, row, col) {
  f ← substitute(fun); g ← function(F) eval(f)
  z ← Ecdf(~ gh, groups=cut2(pgh, g=6),
           fun=function(F) g(1 − F),
           ylab=as.expression(f), xlim=c(4.5, 7.75), data=w,
           label.curve=FALSE)
  print(z, split=c(col, row, 2, 2), more=row < 2 | col < 2)
}
p(log(F/(1−F)),   1, 1)
p(qnorm(F),       1, 2)
p(−log(−log(F)),  2, 1)
p(log(−log(1−F)), 2, 2)
# Get slopes of pgh for some cutoffs of Y
# Use glm complementary log-log link on Prob(Y < cutoff) to
# get log-log link on Prob(Y ≥ cutoff)
r ← NULL
for(link in c('logit','probit','cloglog'))
  for(k in c(5, 5.5, 6)) {
```

```
    co ← coef(glm(gh < k ~ pgh, data=w, family=binomial(link)))
    r ← rbind(r, data.frame(link=link, cutoff=k,
                            slope=round(co[2],2)))
}
print(r, row.names=FALSE)
```

```
    link cutoff slope
   logit    5.0 -3.39
   logit    5.5 -4.33
   logit    6.0 -5.62
  probit    5.0 -1.69
  probit    5.5 -2.61
  probit    6.0 -3.07
 cloglog    5.0 -3.18
 cloglog    5.5 -2.97
 cloglog    6.0 -2.51
```

- Upper right curves are not linear, implying that a normal conditional distribution cannot work for gh[b]
- There is non-parallelism for the logit model
- Other graphs will be used to guide selection of an ordinal model below

11.3 Quantile Regression

- Ruled out OLS and semiparametric proportional odds model

[b] They are not parallel either.

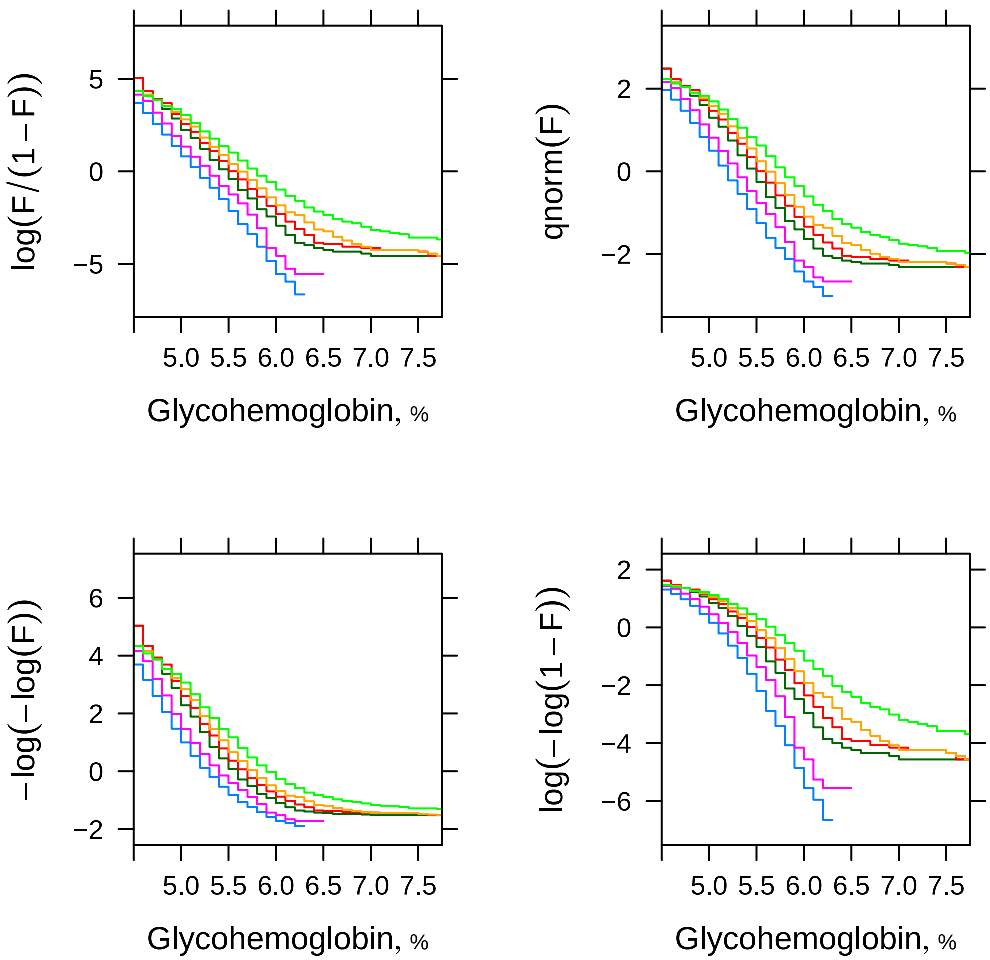


Figure 11.1: Examination of normality and constant variance assumption, and assumptions for various ordinal models

- Quantile regression[100, 102] is a different approach to modeling Y
- No distributional assumptions other than continuity of Y
- All the usual right hand side assumptions
- When there is a single predictor that is categorical, quantile regression coincides with ordinary sample quantiles stratified by that predictor
- Is transformation invariant - pre-transforming Y not important

Let $\rho_{\tau}(y) = y(\tau - [y < 0])$. The τ^{th} sample quantile is the minimizer q of $\Sigma_{i-1}^{n} \rho_{\tau}(y_i - q)$. For a conditional τ^{th} quantile of $Y|X$ the corresponding quantile regression estimator $\hat{\beta}_{\tau}$ minimizes $\Sigma_{i=1}^{n} \rho_{\tau}(Y_i - X\beta)$.

Quantile regression is not as efficient at estimating quantiles as is ordinary least squares at estimating the mean, if the latter's assumptions hold.

Koenker's `quantreg` package in R[101] implements quantile regression, and the `rms` package's `Rq` function provides a front-end that gives rise to various graphics and inference tools.

If we model the median `gh` as a function of covariates, only the $X\beta$ structure need be correct. Other quantiles (e.g., 90^{th} percentile) can be directly modeled but standard errors will be much larger as it is more difficult to precisely estimate outer quantiles.

11.4 Ordinal Regression Models for Continuous Y

- Advantages of semiparametric models (e.g., quantile regression and cumulative probability ordinal models
- For ordinal cumulative probability models, there is no distributional assumption for Y given a setting of X

- Assume only a connection between distributions of Y for different X
- Applying an increasing 1–1 transformation to Y results in no change to regression coefficient estimates[c]
- Regression coefficient estimates are completely robust to extreme Y values[d]
- Estimates of quantiles of Y are exactly transformation-preserving, e.g., estimate of median of $\log Y$ is exactly the log of the estimate of median Y

For a general continuous distribution function $F(y)$, an ordinal regression model based on cumulative probabilities may be stated as follows[e]. Let the ordered unique values of Y be denoted by $y_1, y_2, \ldots, y_k$ and let the intercepts associated with $y_1, \ldots, y_k$ be $\alpha_1, \alpha_2, \ldots, \alpha_k$, where $\alpha_1 = \infty$ because $\text{Prob}[Y \geq y_1] = 1$. Let $\alpha_y =$

[c] For symmetric distributions applying a decreasing transformation will negate the coefficients. For asymmetric distributions (e.g., Gumbel), reversing the order of Y will do more than change signs.

[d] Only an estimate of mean Y from these $\hat{\beta}$s is non-robust.

[e] It is more traditional to state the model in terms of $\text{Prob}[Y \leq y|X]$ but we use $\text{Prob}[Y \geq y|X]$ so that higher predicted values are associated with higher Y.

$\alpha_i, i : y_i = y$. Then

$$\text{Prob}[Y \geq y_i|X] = F(\alpha_i + X\beta) = F(\alpha_{y_i} + X\beta)$$

For the OLS fully parametric case, the model may be restated

$$\begin{aligned}\text{Prob}[Y \geq y|X] &= \text{Prob}[\frac{Y - X\beta}{\sigma} \geq \frac{y - X\beta}{\sigma}] \\ &= 1 - \Phi(\frac{y - X\beta}{\sigma}) = \Phi(\frac{-y}{\sigma} + \frac{X\beta}{\sigma})\end{aligned}$$

so that to within an additive constant[f] $\alpha_y = \frac{-y}{\sigma}$ (intercepts α are linear in y whereas they are arbitrarily descending in the ordinal model), and σ is absorbed in β to put the OLS model into the new notation.

The general ordinal regression model assumes that for fixed X_1, X_2,

$$\begin{aligned}F^{-1}(\text{Prob}[Y \geq y|X_2]) - F^{-1}(\text{Prob}[Y \geq y|X_1]) \\ = (X_2 - X_1)\beta\end{aligned}$$

[f] $\hat{\alpha_y}$ are unchanged if a constant is added to all y.

Table 11.1: Distribution families used in ordinal cumulative probability models. Φ denotes the Gaussian cumulative distribution function. For the Connection column, $P_1 = \text{Prob}[Y \geq y|X_1], P_2 = \text{Prob}[Y \geq y|X_2], \Delta = (X_2 - X_1)\beta$. The connection specifies the only distributional assumption if the model is fitted semiparametrically, i.e, contains an intercept for every unique Y value less one. For parametric models, P_1 must be specified absolutely instead of just requiring a relationship between P_1 and P_2. For example, the traditional Gaussian parametric model specifies that $\text{Prob}[Y \geq y|X] = 1 - \Phi(\frac{y-X\beta}{\sigma}) = \Phi(\frac{-y+X\beta}{\sigma})$.

Distribution	F	Inverse (Link Function)	Link Name	Connection
Logistic	$[1+\exp(-y)]^{-1}$	$\log(\frac{y}{1-y})$	logit	$\frac{P_2}{1-P_2} = \frac{P_1}{1-P_1}\exp(\Delta)$
Gaussian	$\Phi(y)$	$\Phi^{-1}(y)$	probit	$P_2 = \Phi(\Phi^{-1}(P_1)+\Delta)$
Gumbel maximum value	$\exp(-\exp(-y))$	$\log(-\log(y))$	$\log-\log$	$P_2 = P_1^{\exp(\Delta)}$
Gumbel minimum value	$1-\exp(-\exp(y))$	$\log(-\log(1-y))$	complementary $\log-\log$	$1-P_2 = (1-P_1)^{\exp(\Delta)}$
Cauchy	$\frac{1}{\pi}\tan^{-1}(y)+\frac{1}{2}$	$\tan[\pi(y-\frac{1}{2})]$	cauchit	

independent of the αs (parallelism assumption). If $F = [1+\exp(-y)]^{-1}$, this is the proportional odds assumption.

Common choices of F, implemented in the `rms` `orm` function, are shown in Table 11.1.

The Gumbel maximum value distribution is also called the extreme value type I distribution. This distribution ($\log - \log$ link) also represents a continuous time proportional hazards model. The hazard ratio when X changes from X_1 to X_2 is $\exp(-(X_2 - X_1)\beta)$.

The mean of $Y|X$ is easily estimated by computing

$$\sum_{i=1}^{n} y_i \hat{\mathrm{Prob}}[Y = y_i|X]$$

and the q^{th} quantile of $Y|X$ is y such that $F^{-1}(1-q) - X\hat{\beta} = \hat{\alpha}_y$.[g]

The `orm` function in the `rms` package takes advantage of the information matrix being of a sparse tri-band diagonal form for the intercept parameters. This makes the computations efficient even for hundreds of intercepts (i.e., unique values of Y). `orm` is made to handle continuous Y.

Ordinal regression has nice properties in addition to those listed above, allowing for

- estimation of quantiles as efficiently as quantile regression if the parallel slopes assumptions hold
- efficient estimation of mean Y

[g] The intercepts have to be shifted to the left one position in solving this equation because the quantile is such that $\mathrm{Prob}[Y \leq y] = q$ whereas the model is stated in terms of $\mathrm{Prob}[Y \geq y]$.

- direct estimation of $\text{Prob}[Y \geq y|X]$
- arbitrary clumping of values of Y, while still estimating β and mean Y efficiently[h]
- solutions for $\hat{\beta}$ using ordinary Newton-Raphson or other popular optimization techniques
- being based on a standard likelihood function, penalized estimation can be straightforward
- Wald, score, and likelihood ratio χ^2 tests that are more powerful than tests from quantile regression

To summarize how assumptions of parametric models compare to assumptions of semi-parametric models, consider the ordinary linear model or its special case the equal variance two-sample t-test, vs. the probit or logit (proportional odds) ordinal model or their special cases the Van der Waerden (normal-scores) two-sample test or the Wilcoxon test. All the

[h] But it is not sensible to estimate quantiles of Y when there are heavy ties in Y in the area containing the quantile.

assumptions of the linear model other than independence of residuals are captured in the following (written in traditional $Y \leq y$ form):

$$F(y|X) = \mathrm{Prob}[Y \leq y|X] = \Phi(\frac{y - X\beta}{\sigma})$$
$$\Phi^{-1}(F(y|X)) = \frac{y - X\beta}{\sigma}$$

On the other hand, ordinal models assume

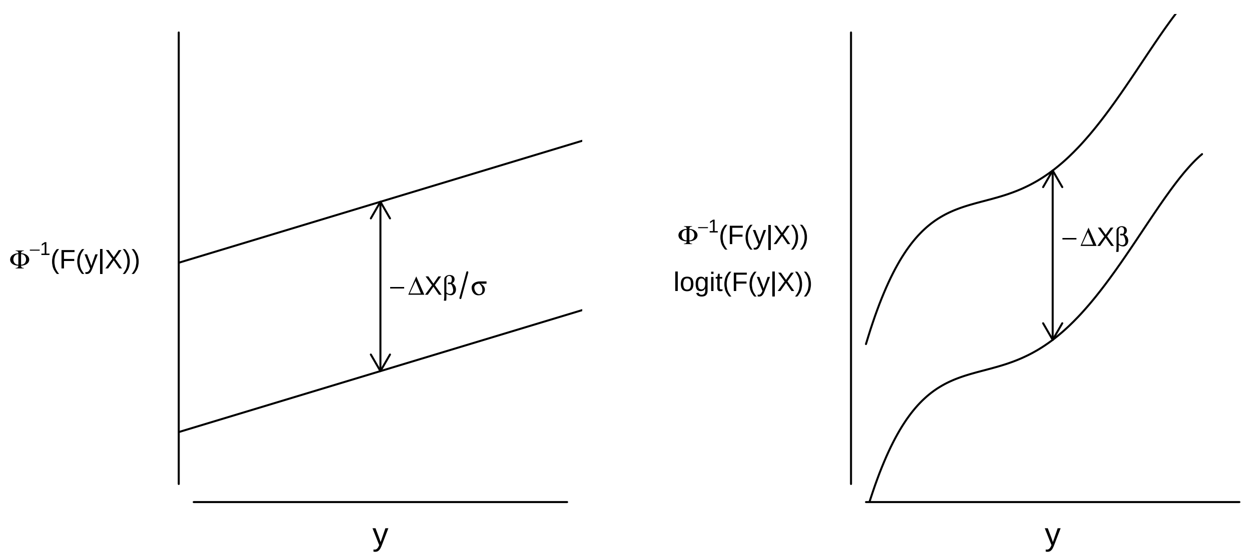


Figure 11.2: Assumptions of the linear model (left panel) and semiparametric ordinal probit or logit (proportional odds) models (right panel). Ordinal models do not assume any shape for the distribution of Y for a given X; they only assume parallelism.

the following:

$$\mathrm{Prob}[Y \leq y|X] = F(g(y) - X\beta),$$

where g is unknown and may be discontinuous.

From this point we revert back to $Y \geq y$ notation so that Y increases as $X\beta$ increases.

11.5 Ordinal Regression Applied to HbA_{1c}

- In Figure 11.1, logit inverse curves are not parallel so proportional odds assumption does not hold
- log-log link yields highest degree of parallelism and most constant regression coeffients across cutoffs of `gh` so use this link in an ordinal regression model (linearity of the curves is not required)

11.5.1 Checking Fit for Various Models Using Age

Another way to examine model fit is to flexibly fit the single most important predictor (age) using a variety of methods, and comparing predictions to sample quantiles and means based on overlapping subsets on age, each subset being subjects having age < 5 years away from the point being predicted by the models. Here we predict the 0.5, 0.75, and 0.9 quantiles and the mean. For quantiles we can compare to quantile regression(discussed below) and for means we compare to OLS.

```
ag ← 25:75
lag ← length(ag)
q2 ← q3 ← p90 ← means ← numeric(lag)
for(i in 1:lag) {
  s ← which(abs(w$age − ag[i]) < 5)
  y ← w$gh[s]
  a ← quantile(y, probs=c(.5, .75, .9))
  q2[i]    ← a[1]
  q3[i]    ← a[2]
  p90[i]   ← a[3]
  means[i] ← mean(y)
}
fams ← c('logistic', 'probit', 'loglog', 'cloglog')
fe   ← function(pred, target) mean(abs(pred$yhat − target))
mod  ← gh ~ rcs(age,6)
P    ← Er ← list()
for(est in c('q2', 'q3', 'p90', 'mean')) {
  meth ← if(est == 'mean') 'ols' else 'QR'
  p ← list()
  er ← rep(NA, 5)
  names(er) ← c(fams, meth)
  for(family in fams) {
    h   ← orm(mod, family=family, data=w)
    fun ← if(est == 'mean') Mean(h)
    else {
      qu ← Quantile(h)
      switch(est, q2  = function(x) qu(.5,  x),
```

```
                  q3  = function(x) qu(.75, x),
                  p90 = function(x) qu(.9,  x))
    }
    p[[family]] ← z ← Predict(h, age=ag, fun=fun, conf.int=FALSE)
    er[family] ← fe(z, switch(est, mean=means, q2=q2, q3=q3, p90=p90))
  }
  h ← switch(est,
              mean= ols(mod, data=w),
              q2  = Rq (mod, data=w),
              q3  = Rq (mod, tau=0.75, data=w),
              p90 = Rq (mod, tau=0.90, data=w))
  p[[meth]] ← z ← Predict(h, age=ag, conf.int=FALSE)
  er[meth] ← fe(z, switch(est, mean=means, q2=q2, q3=q3, p90=p90))

  Er[[est]] ← er
  pr ← do.call('rbind', p)
  pr$est ← est
  P ← rbind.data.frame(P, pr)
}

xyplot(yhat ~ age | est, groups=.set., data=P, type='l', # Figure 11.3
       auto.key=list(x=.75, y=.2, points=FALSE, lines=TRUE),
       panel=function(..., subscripts) {
         panel.xyplot(..., subscripts=subscripts)
         est ← P$est[subscripts[1]]
         lpoints(ag, switch(est, mean=means, q2=q2, q3=q3, p90=p90),
                 col=gray(.7))
         er ← format(round(Er[[est]],3), nsmall=3)
         ltext(26, 6.15, paste(names(er), collapse='\n'),
               cex=.7, adj=0)
         ltext(40, 6.15, paste(er, collapse='\n'),
               cex=.7, adj=1)})
```

It can be seen in Figure 11.3 that models dedicated to a specific task (quantile reqression for quantiles and OLS for means) were best for those tasks. Although the log-log ordinal cumulative probability model did not estimate the median as accurately as some other methods,

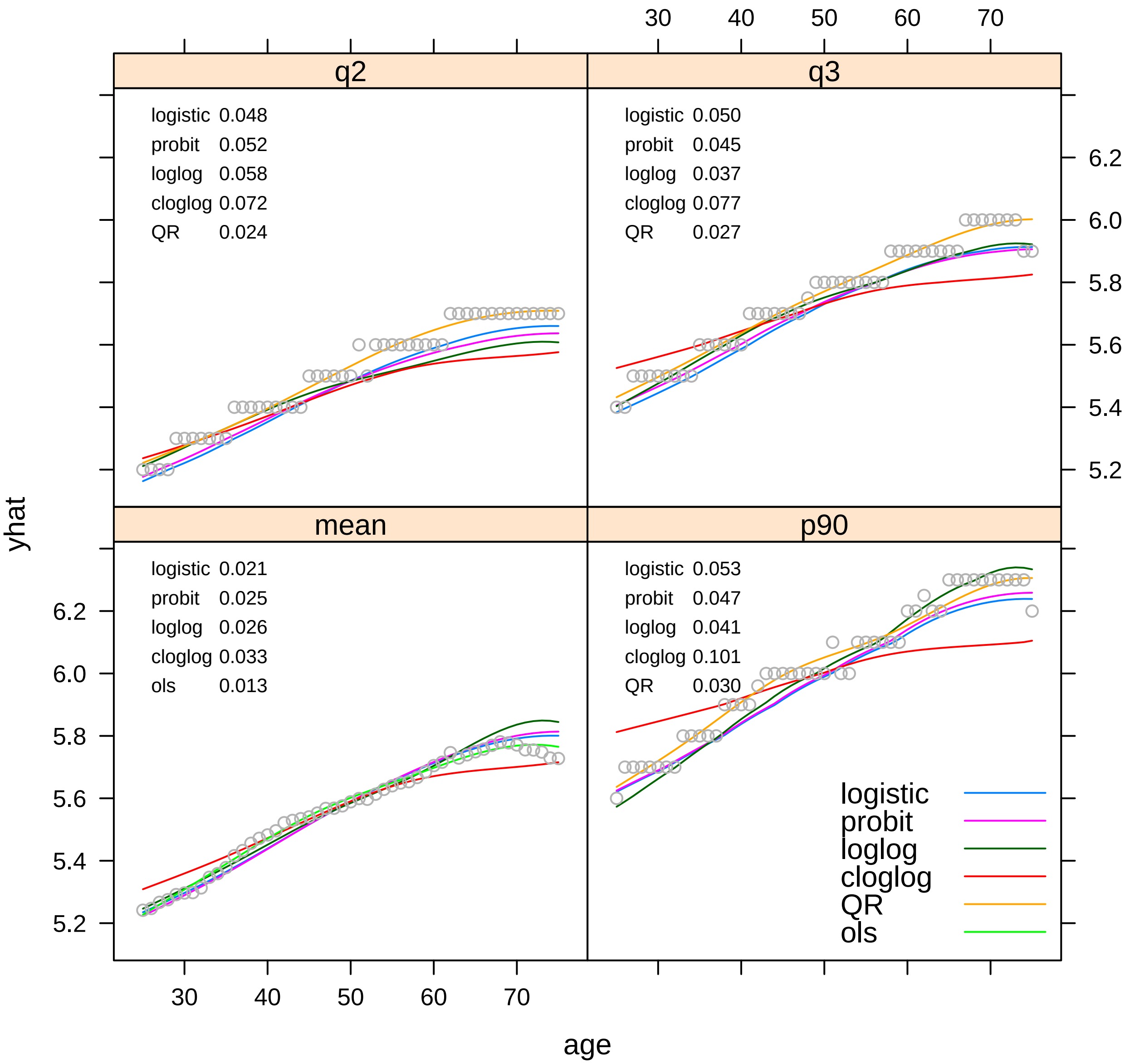


Figure 11.3: Three estimated quantiles and estimated mean using 6 methods, compared against caliper-matched sample quantiles/means (circles). Numbers are mean absolute differences between predicted and sample quantities using overlapping intervals of age and caliper matching. QR:quantile regression.

it does well for the 0.75 and 0.9 quantiles and is the best compromise overall because of its ability to also directly predict the mean as well as quantitles such as $\mathrm{Prob}[\mathrm{HbA}_{1c} > 7|X]$.

For here on we focus on the log-log ordinal model.

Going back to the bottom left of figure 11.1, let's look at quantile groups of predicted HbA_{1c} by OLS and plot predicted distributions of actual HbA_{1c} against empirical distributions.

```
w$pghg ← cut2(pgh, g=6)
f  ← orm(gh ~ pghg, data=w)
lp ← predict(f, newdata=data.frame(pghg=levels(w$pghg)))
ep ← ExProb(f)  # Exceedance prob. functn. generator in rms
z  ← ep(lp)
j  ← order(w$pghg)  # puts in order of lp (levels of pghg)
plot(z, xlim=c(4, 7.5), data=w[j,c('pghg', 'gh')]) # Fig. 11.4
```

Agreement between predicted and observed exceedance probability distributions is excellent in Figure 11.4.

To return to the initial look at a linear model with assumed Gaussian residuals, fit a probit ordinal model and compare the estimated in-

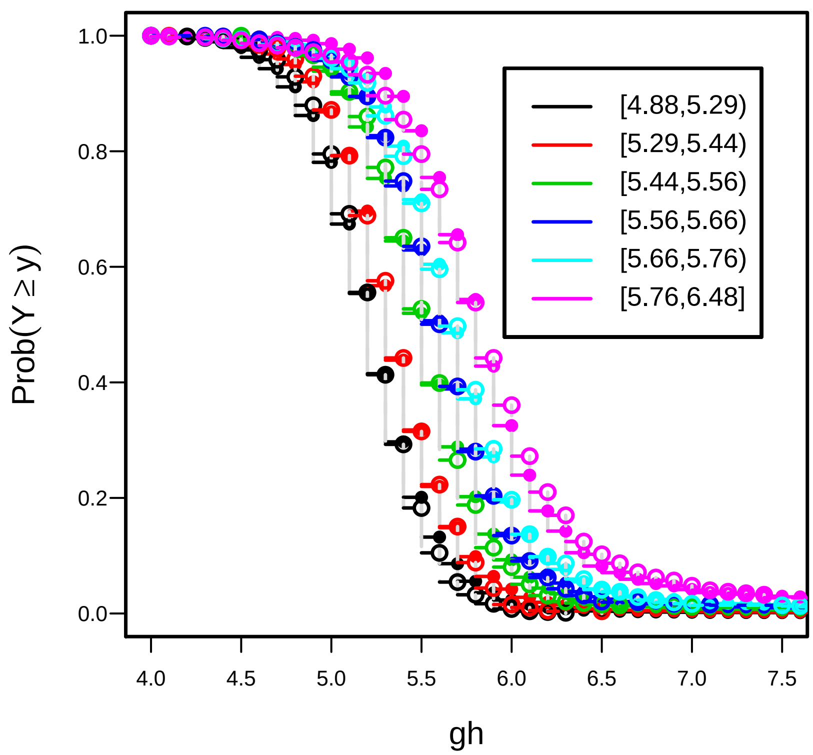


Figure 11.4: Observed (dashed lines, open circles) and predicted (solid lines, closed circles) exceedance probability distributions from a model using 6-tiles of OLS-predicted HbA_{1c}. Key shows quantile group intervals of predicted mean HbA_{1c}.

tercepts to the linear relationship with gh that is assumed by the normal distribution.

```
f ← orm(gh ~ rcs(age,6), family=probit, data=w)
g ← ols(gh ~ rcs(age,6), data=w)
s ← g$stats['Sigma']
yu ← f$yunique[−1]
r ← quantile(w$gh, c(.005, .995))
alphas ← coef(f)[1:num.intercepts(f)]
plot(−yu / s, alphas, type='l', xlim=rev(− r / s), # Fig. 11.5
     xlab=expression(−y/hat(sigma)), ylab=expression(alpha[y]))
```

Figure 11.5 depicts a significant departure from that implied by Gaussian residuals.

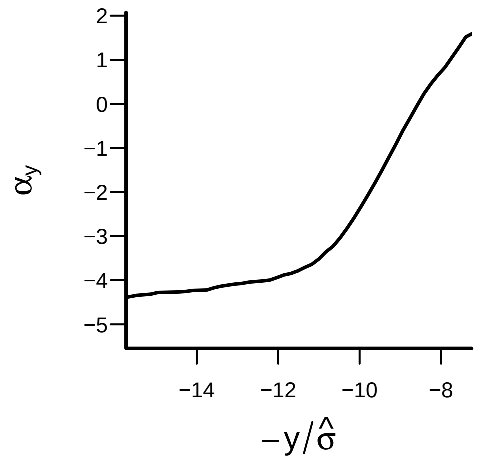


Figure 11.5: Estimated intercepts from probit model

11.5.2 Examination of BMI

Using the log-log model, we first check the adequacy of BMI as a summary of height and weight for estimating median `gh`.

- Adjust for age (without assuming linearity) in every case
- Look at ratio of coefficients of log height and log weight
- Use AIC to judge whether BMI is an adequate summary of height and weight

```
f ← orm(gh ∼ rcs(age,5) + log(ht) + log(wt),
        family=loglog, data=w)
print(f, latex=TRUE)
```

-log-log Ordinal Regression Model

```
orm(formula = gh ~ rcs(age, 5) + log(ht) + log(wt), data = w,
    family = loglog)
```

		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	4629	LR χ^2	1126.94	R^2	0.217	ρ	0.486
Unique Y	63	d.f.	6	g	0.627		
$Y_{0.5}$	5.5	Pr($>\chi^2$)	<0.0001	g_r	1.872		
$\max\|\frac{\partial \log L}{\partial \beta}\|$	1×10^{-6}	Score χ^2	1262.81	$\|\Pr(Y \geq Y_{0.5}) - \frac{1}{2}\|$	0.153		
		Pr($>\chi^2$)	<0.0001				

	Coef	S.E.	Wald Z	Pr($>\|Z\|$)
age	0.0398	0.0055	7.29	<0.0001
age'	-0.0158	0.0275	-0.57	0.5657
age''	-0.0072	0.0866	-0.08	0.9333
age'''	0.0309	0.1135	0.27	0.7853
ht	-3.0680	0.2789	-11.00	<0.0001
wt	1.2748	0.0704	18.10	<0.0001

```
aic ← NULL
for(mod in list(gh ∼ rcs(age,5) + rcs(log(bmi),5),
                gh ∼ rcs(age,5) + rcs(log(ht),5) + rcs(log(wt),5),
                gh ∼ rcs(age,5) + rcs(log(ht),4) * rcs(log(wt),4)))
  aic ← c(aic, AIC(orm(mod, family=loglog, data=w)))
print(aic)
```

```
[1] 25910.77 25910.17 25906.03
```

The ratio of the coefficient of log height to the coefficient of log weight is -2.4, which is between what BMI uses and the more dimensionally reasonable weight / height^3. By AIC, a spline interaction surface between height and weight does slightly better than BMI in predicting HbA$_{1c}$, but a nonlinear function of BMI is barely worse. It will require other body size measures to displace BMI as a predictor.

As an aside, compare this model fit to that from the Cox proportional hazards model. The Cox model uses a conditioning argument to obtain a partial likelihood free of the intercepts α (and requires a second step to estimate these log discrete hazard components) whereas we are using a full marginal likelihood of the ranks of Y[95].

```
print(cph(Surv(gh) ~ rcs(age,5) + log(ht) + log(wt), data=w),
      latex=TRUE)
```

Cox Proportional Hazards Model

```
cph(formula = Surv(gh) ~ rcs(age, 5) + log(ht) + log(wt), data = w)
```

		Model Tests		Discrimination Indexes	
Obs	4629	LR χ^2	1120.20	R^2	0.215
Events	4629	d.f.	6	D_{xy}	0.359
Center	8.3792	Pr$(>\chi^2)$	0.0000	g	0.622
		Score χ^2	1258.07	g_r	1.863
		Pr$(>\chi^2)$	0.0000		

	Coef	S.E.	Wald Z	Pr$(>\lvert Z\rvert)$
age	-0.0392	0.0054	-7.24	<0.0001
age'	0.0148	0.0274	0.54	0.5888
age''	0.0093	0.0862	0.11	0.9144
age'''	-0.0321	0.1131	-0.28	0.7767
ht	3.0477	0.2779	10.97	<0.0001
wt	-1.2653	0.0701	-18.04	<0.0001

Back up and look at all body size measures, and examine their redundancies.

```
v ← varclus(~ wt + ht + bmi + leg + arml + armc + waist +
              tri + sub + age + sex + re, data=w)
plot(v)
# Omit wt so it won't be removed before bmi
redun(~ ht + bmi + leg + arml + armc + waist + tri + sub,
      data=w, r2=.75)
```

```
Redundancy Analysis

redun(formula = ~ht + bmi + leg + arml + armc + waist + tri +
    sub, data = w, r2 = 0.75)

n: 3853          p: 8    nk: 3

Number of NAs:   776
Frequencies of Missing Values Due to Each Variable
   ht    bmi    leg   arml   armc  waist    tri    sub
    0      0    155    127    130    164    334    655


Transformation of target variables forced to be linear

R2 cutoff: 0.75          Type: ordinary

R2 with which each variable can be predicted from all other variables:

   ht    bmi    leg   arml   armc  waist    tri    sub
0.829 0.924 0.682 0.748 0.843 0.864 0.531 0.594

Rendundant variables:

bmi ht

Predicted from variables:

leg arml armc waist tri sub

  Variable Deleted    R2 R2 after later deletions
1              bmi 0.924                      0.909
2               ht 0.792
```

Six size measures adequately capture the entire set. Height and BMI are removed. An ad-

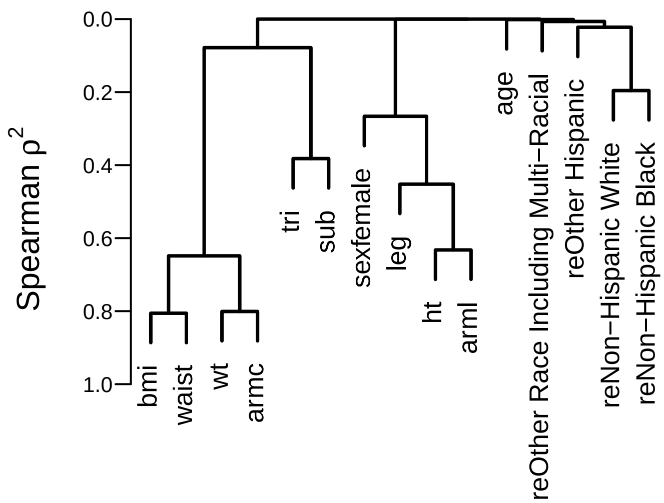


Figure 11.6: Variable clustering for all potential predictors

vantage of removing height is that it is age-dependent due to vertebral compression in the elderly:

```
f ← orm(ht ~ rcs(age,4)*sex, data=w)   # Prop. odds model
qu ← Quantile(f); med ← function(x) qu(.5, x)
ggplot(Predict(f, age, sex, fun=med, conf.int=FALSE),
       ylab='Predicted Median Height, cm')
```

Next allocate d.f. according to generalized Spearman ρ^2[i].

```
s ← spearman2(gh ~ age + sex + re + wt + leg + arml + armc +
              waist + tri + sub, data=w, p=2)
plot(s)
```

Parameters will be allocated in descending order of ρ^2. But note that subscapular skinfold has a large number of `NA`s and other predictors

[i]Competition between collinear size measures hurts interpretation of partial tests of association in a saturated additive model.

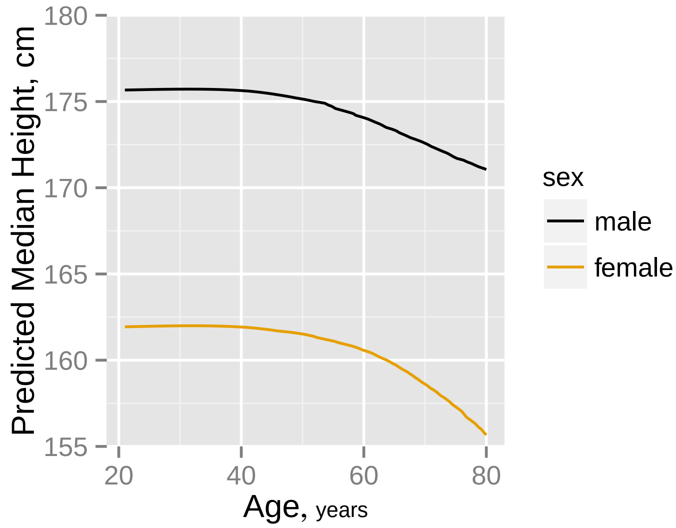


Figure 11.7: Estimated median height as a smooth function of age, allowing age to interact with sex, from a proportional odds model

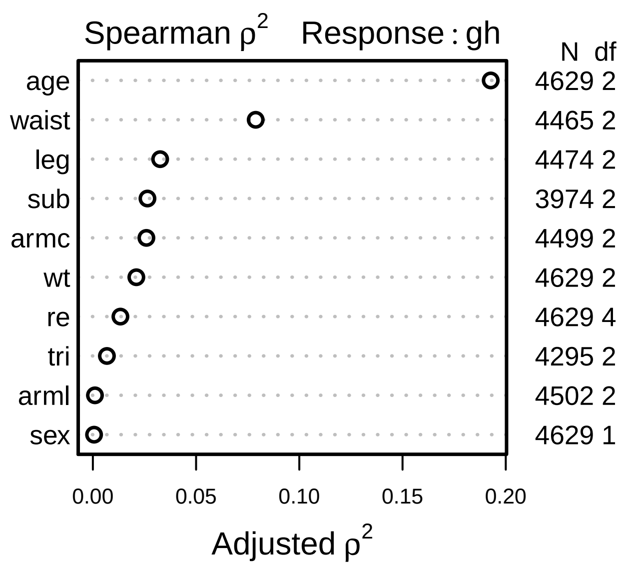


Figure 11.8: Generalized squared rank correlations

also have NAs. Suboptimal casewise deletion will be used until the final model is fitted.

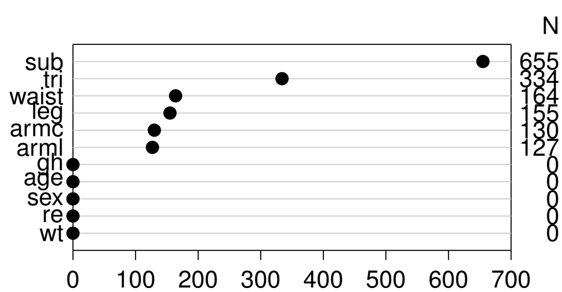
Because there are many competing body measures, we use backwards stepdown to arrive at a set of predictors. The bootstrap will be used to penalize predictive ability for variable selection. First the full model is fit using casewise deletion, then we do a composite test to assess whether any of the frequently–missing predictors is important.

```
f ← orm(gh ∼ rcs(age,5) + sex + re + rcs(wt,3) + rcs(leg,3) + arml +
         rcs(armc,3) + rcs(waist,4) + tri + rcs(sub,3),
         family='loglog', data=w, x=TRUE, y=TRUE)
print(f, latex=TRUE, coefs=FALSE)
```

-log-log Ordinal Regression Model

```
orm(formula = gh ~ rcs(age, 5) + sex + re + rcs(wt, 3) + rcs(leg,
    3) + arml + rcs(armc, 3) + rcs(waist, 4) + tri + rcs(sub,
    3), data = w, x = TRUE, y = TRUE, family = "loglog")
```

Frequencies of Missing Values Due to Each Variable



		Model Likelihood Ratio Test		Discrimination Indexes		Rank Discrim. Indexes	
Obs	3853	LR χ^2	1180.13	R^2	0.265	ρ	0.520
Unique Y	60	d.f.	22	g	0.732		
$Y_{0.5}$	5.5	Pr$(>\chi^2)$	<0.0001	g_r	2.080		
$\max\lvert\frac{\partial \log L}{\partial \beta}\rvert$	3×10^{-5}	Score χ^2	1298.88	$\lvert\Pr(Y\geq Y_{0.5})-\frac{1}{2}\rvert$	0.172		
		Pr$(>\chi^2)$	<0.0001				

```
## Composite test:
lan ← function(a) latex(a, table.env=FALSE, file='')
lan(anova(f, leg, arml, armc, waist, tri, sub))
```

	χ^2	d.f.	P
leg	8.30	2	0.0158
Nonlinear	3.32	1	0.0685
arml	0.16	1	0.6924
armc	6.66	2	0.0358
Nonlinear	3.29	1	0.0695
waist	29.40	3	<0.0001
Nonlinear	4.29	2	0.1171
tri	16.62	1	<0.0001
sub	40.75	2	<0.0001
Nonlinear	4.50	1	0.0340
TOTAL NONLINEAR	14.95	5	0.0106
TOTAL	128.29	11	<0.0001

The model yields Spearman $\rho = 0.52$, the rank correlation between predicted and observed HbA$_{1c}$.

Show predicted mean and median HbA$_{1c}$ as a function of age, adjusting other variables to median/mode. Compare the estimate of the median with that from quantile regression (discussed below).

```
M    ← Mean(f)
qu   ← Quantile(f)
med  ← function(x) qu(.5, x)
p90  ← function(x) qu(.9, x)
```

```
fq     ← Rq(formula(f), data=w)
fq90   ← Rq(formula(f), data=w, tau=.9)
```

```
pmean  ← Predict(f,    age, fun=M,    conf.int=FALSE)
pmed   ← Predict(f,    age, fun=med,  conf.int=FALSE)
p90    ← Predict(f,    age, fun=p90,  conf.int=FALSE)
pmedqr ← Predict(fq,   age, conf.int=FALSE)
p90qr  ← Predict(fq90, age, conf.int=FALSE)
z ← rbind('orm mean'=pmean, 'orm median'=pmed, 'orm P90'=p90,
          'QR median'=pmedqr, 'QR P90'=p90qr)
ggplot(z, groups='.set.',
       adj.subtitle=FALSE, legend.label=FALSE)
```

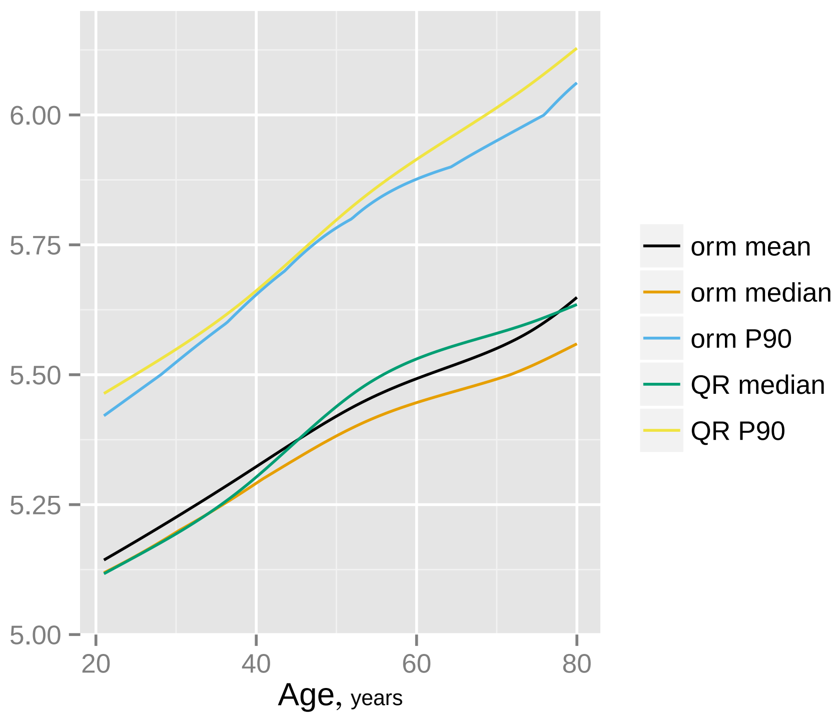


Figure 11.9: Estimated mean and 0.5 and 0.9 quantiles from the log-log ordinal model using casewise deletion, along with predictions of 0.5 and 0.9 quantiles from quantile regression (QR). Age is varied and other predictors are held constant to medians/modes.

```
print(fastbw(f, rule='p'), estimates=FALSE)
```

```
 Deleted Chi-Sq d.f. P      Residual d.f. P      AIC
 arml    0.16   1    0.6924 0.16     1    0.6924 -1.84
 sex     0.45   1    0.5019 0.61     2    0.7381 -3.39
 wt      5.72   2    0.0572 6.33     4    0.1759 -1.67
 armc    3.32   2    0.1897 9.65     6    0.1400 -2.35

Factors in Final Model
```

```
[1] age   re    leg   waist tri   sub
```

```
set.seed(13)  # so can reproduce results
v ← validate(f, B=100, bw=TRUE, estimates=FALSE, rule='p')
```

```
                Backwards Step-down - Original Model

 Deleted Chi-Sq d.f. P      Residual d.f. P      AIC
 arml    0.16   1    0.6924 0.16     1    0.6924 -1.84
 sex     0.45   1    0.5019 0.61     2    0.7381 -3.39
 wt      5.72   2    0.0572 6.33     4    0.1759 -1.67
 armc    3.32   2    0.1897 9.65     6    0.1400 -2.35

Factors in Final Model

[1] age   re    leg   waist tri   sub
```

```
# Show number of variables selected in first 30 boots
latex(v, B=30, file='', size='small')
```

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
ρ	0.5225	0.5290	0.5208	0.0083	0.5142	100
R^2	0.2712	0.2788	0.2692	0.0095	0.2617	100
Slope	1.0000	1.0000	0.9761	0.0239	0.9761	100
g	1.2276	1.2505	1.2207	0.0298	1.1978	100
$\overline{\lvert \Pr(Y \geq Y_{0.5}) - \frac{1}{2} \rvert}$	0.2007	0.2050	0.1987	0.0064	0.1943	100

Factors Retained in Backwards Elimination
First 30 Resamples

age	sex	re	wt	leg	arml	armc	waist	tri	sub
•	•	•	•	•		•	•		•
•		•		•		•	•	•	•
•		•		•		•	•	•	•
•	•	•	•	•		•	•		•
•	•	•	•	•		•	•		•
•		•		•		•	•	•	•
•	•	•	•			•	•		•
•	•	•	•	•		•	•		
•		•	•	•		•	•	•	•
•	•	•	•	•		•	•		•
•		•	•			•	•	•	•
•		•		•	•	•	•	•	•
•		•	•	•		•	•	•	•
•		•	•	•		•	•	•	•
•		•		•		•	•	•	•
•		•	•	•		•	•	•	•
•	•	•	•	•		•	•	•	•
•		•		•			•	•	•
•		•	•	•		•	•	•	•
•		•	•	•		•	•	•	•
•		•	•	•		•	•	•	•
•		•	•			•	•	•	•
•		•	•	•		•	•	•	•
•	•	•		•			•		•
•		•		•		•		•	•
•		•	•	•		•	•	•	•
•		•	•			•	•	•	•
•		•	•	•		•	•	•	•
•		•	•	•		•	•	•	•
•		•		•		•		•	•

Frequencies of Numbers of Factors Retained

5	6	7	8	9	10
1	19	29	46	4	1

Next fit the reduced model. Use multiple imputation to impute missing predictors.

```
a ← aregImpute(∼ gh + wt + ht + bmi + leg + arml + armc + waist +
                 tri + sub + age +re, data=w, n.impute=5, pr=FALSE)
g ← fit.mult.impute(gh ∼ rcs(age,5) + re + rcs(leg,3) +
                      rcs(waist,4) + tri + rcs(sub,4),
                      orm, a, family=loglog, data=w, pr=FALSE)
```

```
print(g, latex=TRUE, needspace='1.5in')
```

-log-log Ordinal Regression Model

```
fit.mult.impute(formula = gh ~ rcs(age, 5) + re + rcs(leg, 3) +
    rcs(waist, 4) + tri + rcs(sub, 4), fitter = orm, xtrans = a,
    data = w, pr = FALSE, family = loglog)
```

	Model Likelihood Ratio Test	Discrimination Indexes	Rank Discrim. Indexes
Obs 4629	LR χ^2 1448.42	R^2 0.269	ρ 0.513
Unique Y 63	d.f. 17	g 0.743	
$Y_{0.5}$ 5.5	Pr($>\chi^2$) <0.0001	g_r 2.102	
$\max\lvert\frac{\partial \log L}{\partial \beta}\rvert$ 1×10^{-5}	Score χ^2 1569.21	$\lvert\Pr(Y\geq Y_{0.5})-\frac{1}{2}\rvert$ 0.173	
	Pr($>\chi^2$) <0.0001		

	Coef	S.E.	Wald Z	Pr($>\lvert Z\rvert$)
age	0.0404	0.0055	7.29	<0.0001
age'	-0.0228	0.0279	-0.82	0.4137
age''	0.0126	0.0876	0.14	0.8857
age'''	0.0424	0.1148	0.37	0.7116
re=Other Hispanic	-0.0766	0.0597	-1.28	0.1992
re=Non-Hispanic White	-0.4121	0.0449	-9.17	<0.0001
re=Non-Hispanic Black	0.0645	0.0566	1.14	0.2543
re=Other Race Including Multi-Racial	-0.0555	0.0750	-0.74	0.4593
leg	-0.0339	0.0091	-3.73	0.0002
leg'	0.0153	0.0105	1.46	0.1434
waist	0.0073	0.0050	1.47	0.1428
waist'	0.0304	0.0158	1.93	0.0536
waist''	-0.0910	0.0508	-1.79	0.0732
tri	-0.0163	0.0026	-6.28	<0.0001
sub	-0.0027	0.0097	-0.28	0.7817
sub'	0.0674	0.0289	2.33	0.0198
sub''	-0.1895	0.0922	-2.06	0.0398

```
an ← anova(g)
lan(an)
```

	χ^2	d.f.	P
age	692.50	4	< 0.0001
Nonlinear	28.47	3	< 0.0001
re	168.91	4	< 0.0001
leg	24.37	2	< 0.0001
Nonlinear	2.14	1	0.1434
waist	128.31	3	< 0.0001
Nonlinear	4.05	2	0.1318
tri	39.44	1	< 0.0001
sub	39.30	3	< 0.0001
Nonlinear	6.63	2	0.0363
TOTAL NONLINEAR	46.80	8	< 0.0001
TOTAL	1464.24	17	< 0.0001

```
b  ← anova(g, leg, waist, tri, sub)
# Add new lines to the plot with combined effect of 4 size var.
s ← rbind(an, size=b['TOTAL', ])
class(s) ← 'anova.rms'
plot(s)
```

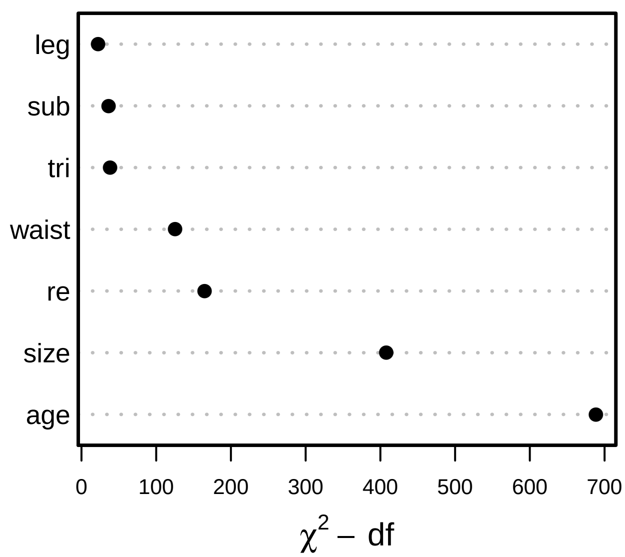


Figure 11.10: ANOVA for reduced model after multiple imputation, with addition of a combined effect for four size variables

```
ggplot(Predict(g), abbrev=TRUE, ylab=NULL)    # Figure 11.11
```

Compare the estimated age partial effects and confidence intervals with those from a model using casewise deletion, and with bootstrap

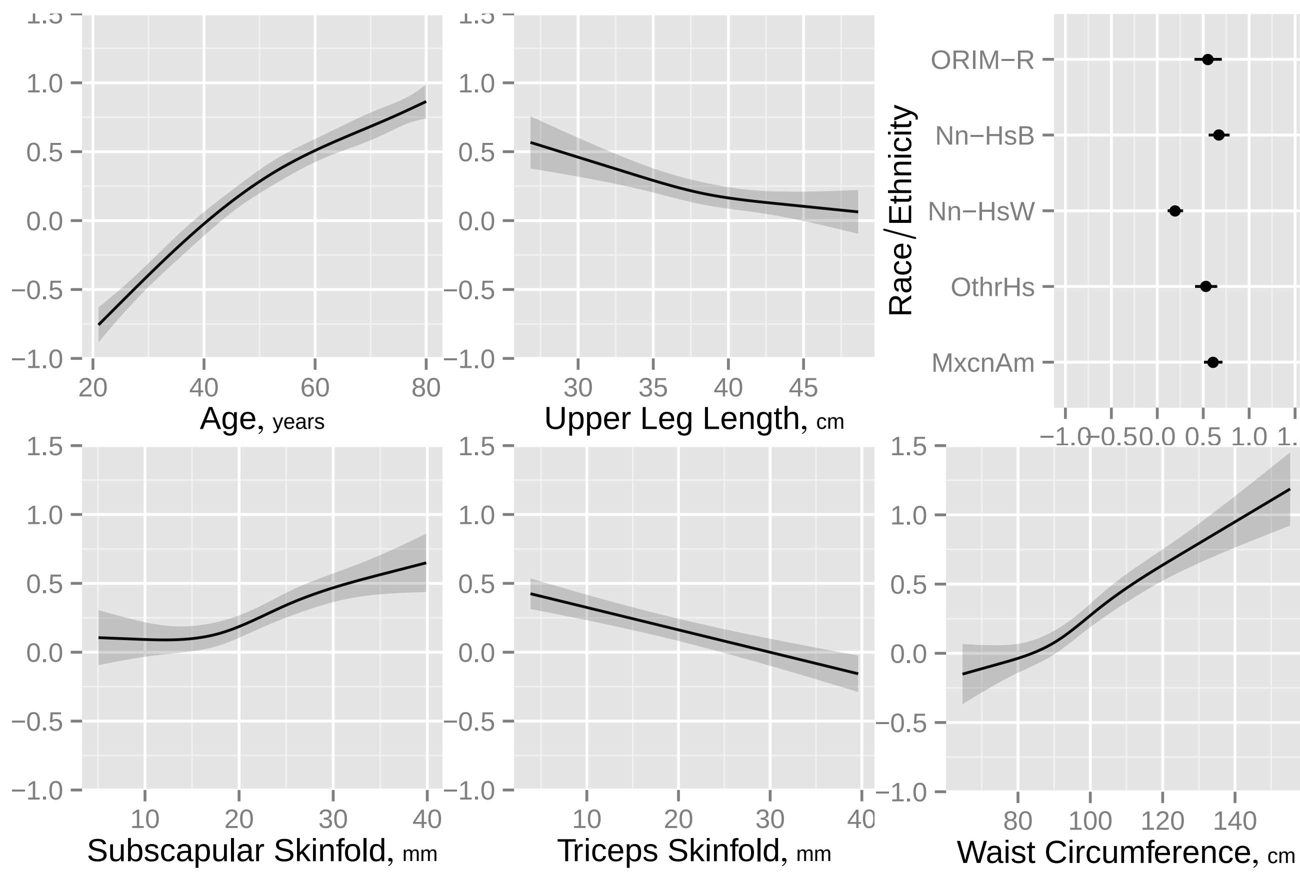


Figure 11.11: Partial effects (log hazard or log-log cumulative probability scale) of all predictors in reduced model, after multiple imputation

nonparametric confidence intervals (also with casewise deletion).

```
gc ← orm(gh ~ rcs(age,5) + re + rcs(leg,3) +
         rcs(waist,4) + tri + rcs(sub,4),
         family=loglog, data=w, x=TRUE, y=TRUE)
gb ← bootcov(gc, B=300)
```

```
bootclb ← Predict(gb, age, boot.type='basic')
```

```
bootclp ← Predict(gb, age, boot.type='percentile')
multimp ← Predict(g, age)
plot(Predict(gc, age), addpanel=function(...) {
  with(bootclb, {llines(age, lower, col='blue')
                llines(age, upper, col='blue')})
  with(bootclp, {llines(age, lower, col='blue', lty=2)
                llines(age, upper, col='blue', lty=2)})
  with(multimp, {llines(age, lower, col='red')
                llines(age, upper, col='red')
                llines(age, yhat, col='red')} ) },
     col.fill=gray(.9), adj.subtitle=FALSE)   # Figure 11.12
```

```
M   ← Mean(g)
qu  ← Quantile(g)
med ← function(lp) qu(.5, lp)
q90 ← function(lp) qu(.9, lp)
lp  ← predict(g)
lpr ← quantile(predict(g), c(.002, .998), na.rm=TRUE)
lps ← seq(lpr[1], lpr[2], length=200)
pmn ← M(lps)
pme ← med(lps)
p90 ← q90(lps)
plot(pmn, pme,   # Figure 11.13
     xlab=expression(paste('Predicted Mean ', HbA["1c"])),
     ylab='Median and 0.9 Quantile', type='l',
     xlim=c(4.75, 8.0), ylim=c(4.75, 8.0), bty='n')
box(col=gray(.8))
lines(pmn, p90, col='blue')
abline(a=0, b=1, col=gray(.8))
text(6.5, 5.5, 'Median')
text(5.5, 6.3, '0.9', col='blue')
nint ← 350
scat1d(M(lp),   nint=nint)
scat1d(med(lp), side=2, nint=nint)
scat1d(q90(lp), side=4, col='blue', nint=nint)
```

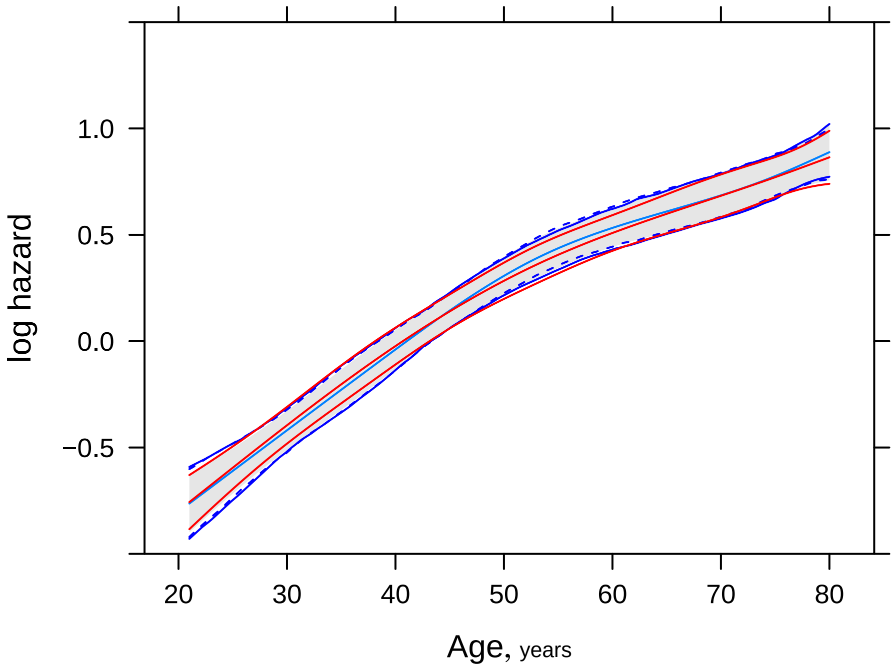


Figure 11.12: Partial effect for age from multiple imputation (center red line) and casewise deletion (center blue line) with symmetric Wald 0.95 confidence bands using casewise deletion (gray shaded area), basic bootstrap confidence bands using casewise deletion (blue lines), percentile bootstrap confidence bands using casewise deletion (dashed blue lines), and symmetric Wald confidence bands accounting for multiple imputation (red lines).

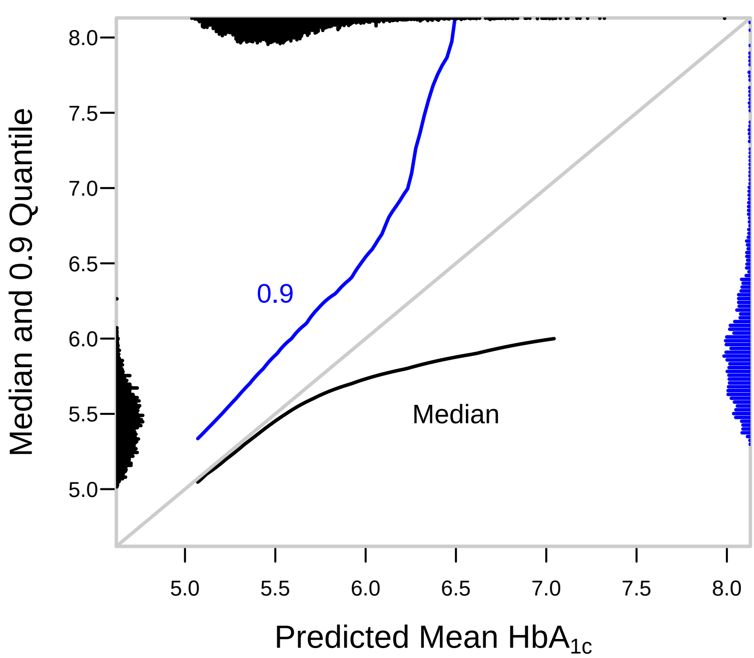


Figure 11.13: Predicted mean HbA_{1c}vs. predicted median and 0.9 quantile along with their marginal distributions

```
g       ← Newlevels(g, list(re=abbreviate(levels(w$re))))
exprob  ← ExProb(g)
nom ←
  nomogram(g, fun=list(Mean=M,
                'Median Glycohemoglobin' = med,
                '0.9 Quantile'           = q90,
                'Prob(HbA1c ≥ 6.5)'=
                     function(x) exprob(x, y=6.5),
                'Prob(HbA1c ≥ 7.0)'=
                     function(x) exprob(x, y=7),
                'Prob(HbA1c ≥ 7.5)'=
                     function(x) exprob(x, y=7.5)),
           fun.at=list(seq(5, 8, by=.5),
             c(5,5.25,5.5,5.75,6,6.25),
             c(5.5,6,6.5,7,8,10,12,14),
             c(.01,.05,.1,.2,.3,.4),
             c(.01,.05,.1,.2,.3,.4),
             c(.01,.05,.1,.2,.3,.4)))
plot(nom, lmgp=.28)    # Figure 11.14
```

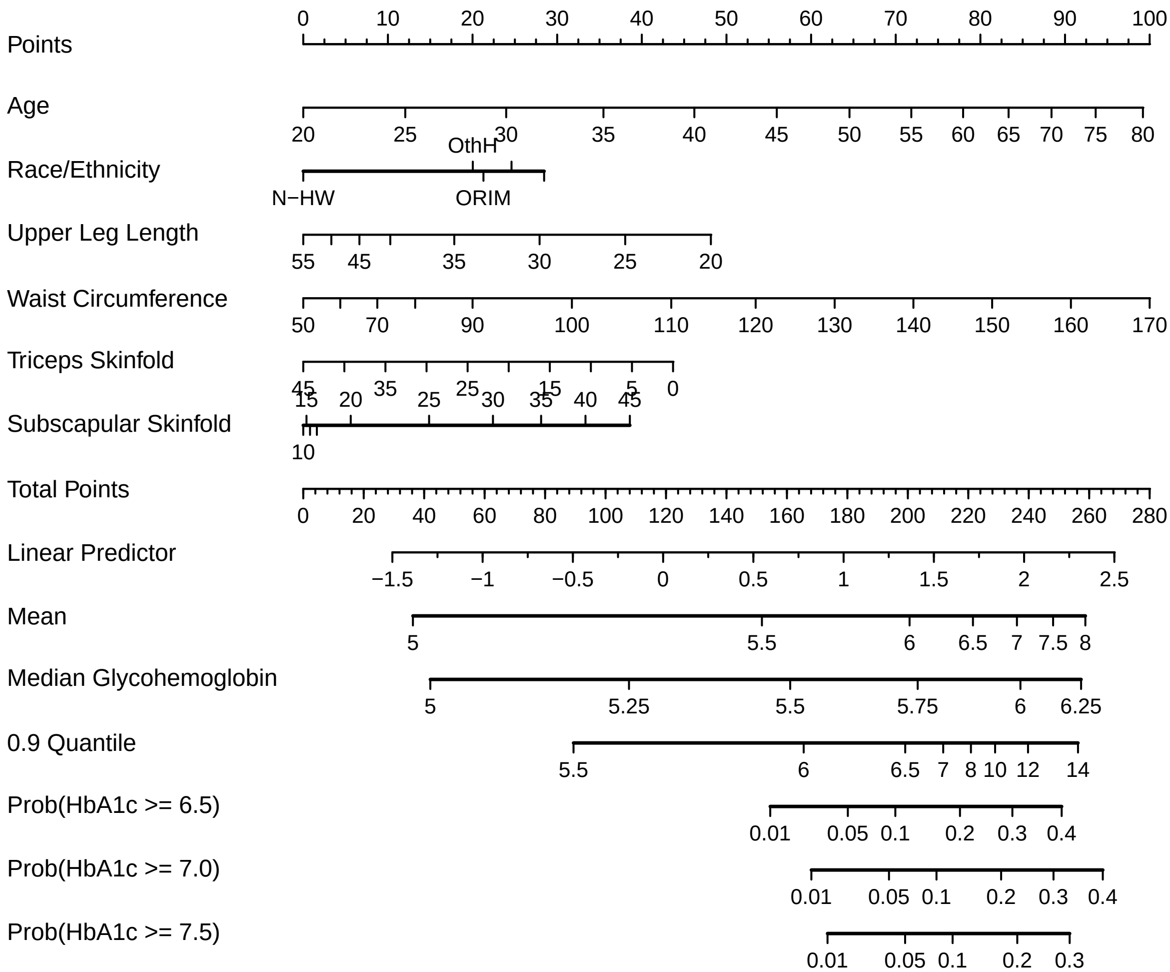


Figure 11.14: Nomogram for predicting median, mean, and 0.9 quantile of glycohemoglobin, along with the estimated probability that $\mathrm{HbA}_{1c} \geq 6.5, 7$, or 7.5, all from the log-log ordinal model

Chapter 12

Case Study in Parametric Survival Modeling and Model Approximation

Data source: Random sample of 1000 patients from Phases I & II of SUPPORT (Study to Understand Prognoses Preferences Outcomes and Risks of Treatment, funded by the Robert Wood Johnson Foundation). See[98]. The dataset is available from `http://biostat.mc.vanderbilt.edu/DataSets`.

- Analyze acute disease subset of SUPPORT (acute respiratory failure, multiple organ system failure, coma) — the shape of the survival curves is different between acute and chronic disease categories

- Patients had to survive until day 3 of the study to qualify
- Baseline physiologic variables measured during day 3

12.1 Descriptive Statistics

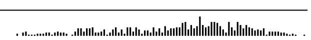
Create a variable `acute` to flag categories of interest; print univariable descriptive statistics.

```
require(rms)
```

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

support[acute,]
35 Variables 537 Observations

age : Age



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	529	1	60.7	28.49	35.22	47.93	63.67	74.49	81.54	85.56

```
lowest : 18.04 18.41 19.76 20.30 20.31
highest: 91.62 91.82 91.93 92.74 95.51
```

death : Death at any time up to NDI date:31DEC94

n	missing	unique	Info	Sum	Mean
537	0	2	0.67	356	0.6629

sex

n	missing	unique
537	0	2

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

hospdead : Death in Hospital

n	missing	unique	Info	Sum	Mean
537	0	2	0.7	201	0.3743

slos : Days from Study Entry to Discharge

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	85	1	23.44	4.0	5.0	9.0	15.0	27.0	47.4	68.2

```
lowest :   3   4   5   6   7, highest: 145 164 202 236 241
```

d.time : Days of Follow-Up

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	340	1	446.1	4	6	16	182	724	1421	1742

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

dzgroup

n	missing	unique
537	0	3

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

dzclass

n	missing	unique
537	0	2

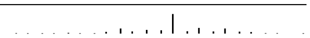
```
ARF/MOSF (477, 89%), Coma (60, 11%)
```

num.co : number of comorbidities

n	missing	unique	Info	Mean
537	0	7	0.93	1.525

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

edu : Years of Education



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
411	126	22	0.96	12.03	7	8	10	12	14	16	17

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

income

n	missing	unique
335	202	4

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

scoma : SUPPORT Coma Score based on Glasgow D3

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	11	0.82	19.24	0	0	0	0	37	55	100

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

charges : Hospital Charges

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
517	20	516	1	86652	11075	15180	27389	51079	100904	205562	283411

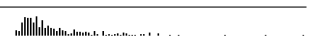
```
lowest :   3448   4432   4574   5555   5849
highest: 504660 538323 543761 706577 740010
```

totcst : Total RCC cost

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
471	66	471	1	46360	6359	8449	15412	29308	57028	108927	141569

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

totmcst : Total micro-cost



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
331	206	328	1	39022	6131	8283	14415	26323	54102	87495	111920

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

avtisst : Average TISS, Days 3-25

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

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

---

race

n	missing	unique
535	2	5

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

---

meanbp : Mean Arterial Blood Pressure Day 3

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	109	1	83.28	41.8	49.0	59.0	73.0	111.0	124.4	135.0

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

---

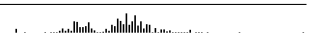
wblc : White Blood Cell Count Day 3

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
532	5	241	1	14.1	0.8999	4.5000	7.9749	12.3984	18.1992	25.1891	30.1873

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

---

hrt : Heart Rate Day 3



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	111	1	105	51	60	75	111	126	140	155

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

---

resp : Respiration Rate Day 3

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	45	1	23.72	8	10	12	24	32	39	40

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

---

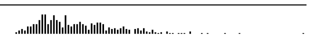
temp : Temperature (celcius) Day 3

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	61	1	37.52	35.50	35.80	36.40	37.80	38.50	39.09	39.50

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

---

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



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
500	37	357	1	227.2	86.99	105.08	137.88	202.56	290.00	390.49	433.31

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

---

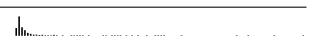
alb : Serum Albumin Day 3

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
346	191	34	1	2.668	1.700	1.900	2.225	2.600	3.100	3.400	3.800

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

---

bili : Bilirubin Day 3

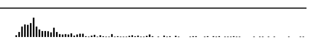


n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
386	151	88	1	2.678	0.3000	0.4000	0.6000	0.8999	2.0000	6.5996	13.1743

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

---

crea : Serum creatinine Day 3



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	84	1	2.232	0.6000	0.7000	0.8999	1.3999	2.5996	5.2395	7.3197

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

---

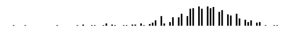
sod : Serum sodium Day 3

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	38	1	138.1	129	131	134	137	142	147	150

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

---

ph : Serum pH (arterial) Day 3

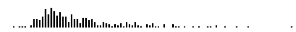


n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
500	37	49	1	7.416	7.270	7.319	7.380	7.420	7.470	7.510	7.529

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

---

glucose : Glucose Day 3



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
297	240	179	1	167.7	76.0	89.0	106.0	141.0	200.0	292.4	347.2

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

---

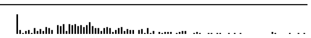
bun : BUN Day 3

n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
304	233	100	1	38.91	8.00	11.00	16.75	30.00	56.00	79.70	100.70

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

---

urine : Urine Output Day 3



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
303	234	262	1	2095	20.3	364.0	1156.5	1870.0	2795.0	4008.6	4817.5

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

---

adlp : ADL Patient Day 3

n	missing	unique	Info	Mean
104	433	8	0.87	1.577

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

---

adls : ADL Surrogate Day 3

n	missing	unique	Info	Mean
392	145	8	0.89	1.86

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

---

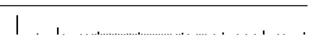
sfdm2

n	missing	unique
468	69	5

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

---

adlsc : Imputed ADL Calibrated to Surrogate



n	missing	unique	Info	Mean	.05	.10	.25	.50	.75	.90	.95
537	0	144	0.96	2.119	0.000	0.000	0.000	1.839	3.375	6.000	6.000

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

```
# Show patterns of missing data
plot(naclus(support[acute,]))          # Figure 12.1
```

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

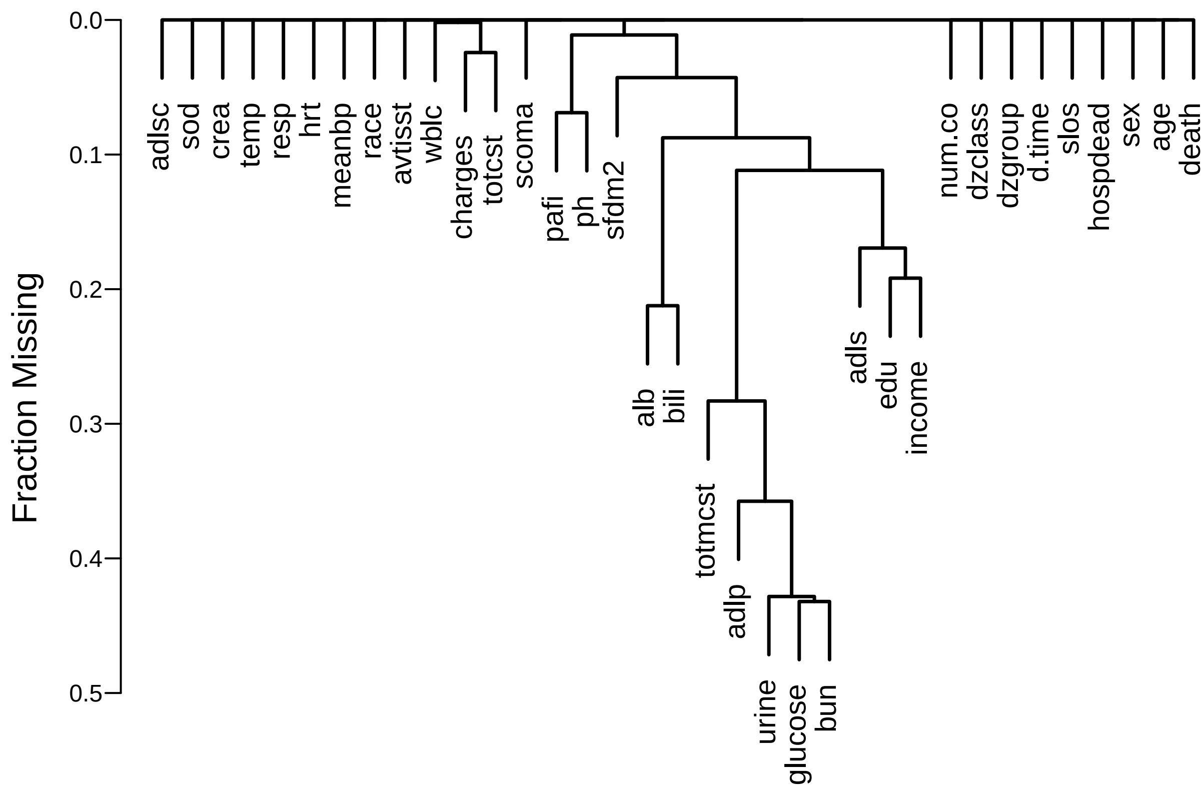


Figure 12.1: Cluster analysis showing which predictors tend to be missing on the same patients

```
ac ← support[acute,]
ac$dzgroup ← ac$dzgroup[drop=TRUE]    # Remove unused levels
attach(ac)
vc ← varclus(~ age+sex+dzgroup+num.co+edu+income+scoma+race+
             meanbp+wblc+hrt+resp+temp+pafi+alb+bili+crea+sod+
             ph+glucose+bun+urine+adlsc, sim='hoeffding')
plot(vc)                                 # Figure 12.2
```

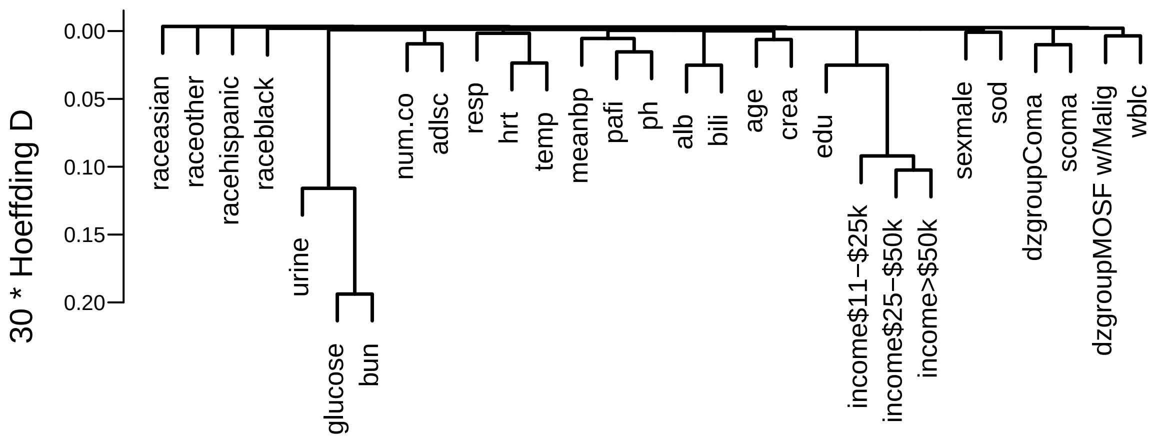


Figure 12.2: Hierarchical clustering of potential predictors using Hoeffding D as a similarity measure. Categorical predictors are automatically expanded into dummy variables.

12.2 Checking Adequacy of Log-Normal Accelerated Failure Time Model

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

# Generate right-censored survival time variable
years ← d.time/365.25
units(years) ← 'Year'
S ← Surv(years, death)

# Show normal inverse Kaplan-Meier estimates
# stratified by dzgroup
survplot(npsurv(S ~ dzgroup), conf='none',
         fun=qnorm, logt=TRUE)    # Figure 12.3
```

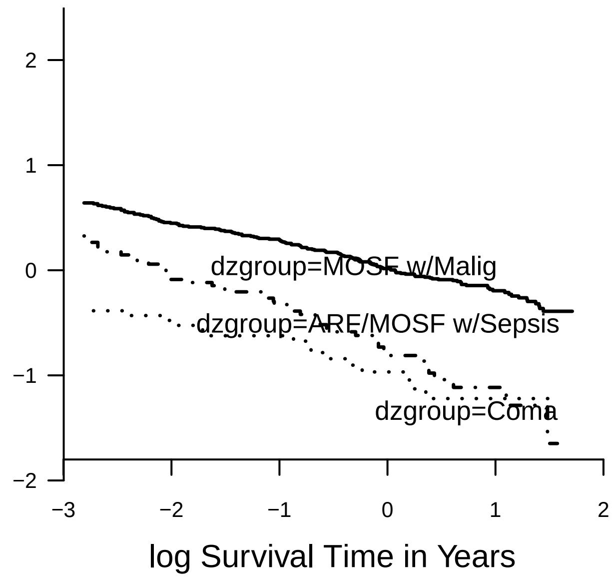


Figure 12.3: $\Phi^{-1}(S_{KM}(t))$ stratified by `dzgroup`. Linearity and semi-parallelism indicate a reasonable fit to the log-normal accelerated failure time model with respect to one predictor.

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

```
f ← psm(S ~ dzgroup + rcs(age,5) + rcs(meanbp,5),
              dist='lognormal', y=TRUE) # dist='gaussian' for S+
r ← resid(f)

survplot(r, dzgroup, label.curve=FALSE)
survplot(r, age,     label.curve=FALSE)
survplot(r, meanbp,  label.curve=FALSE)
random.number ← runif(length(age))
survplot(r, random.number, label.curve=FALSE)   # Figure 12.4
```

The fit for `dzgroup` is not great but overall fit is good.

Remove from consideration predictors that are missing in > 0.2 of the patients. Many of these

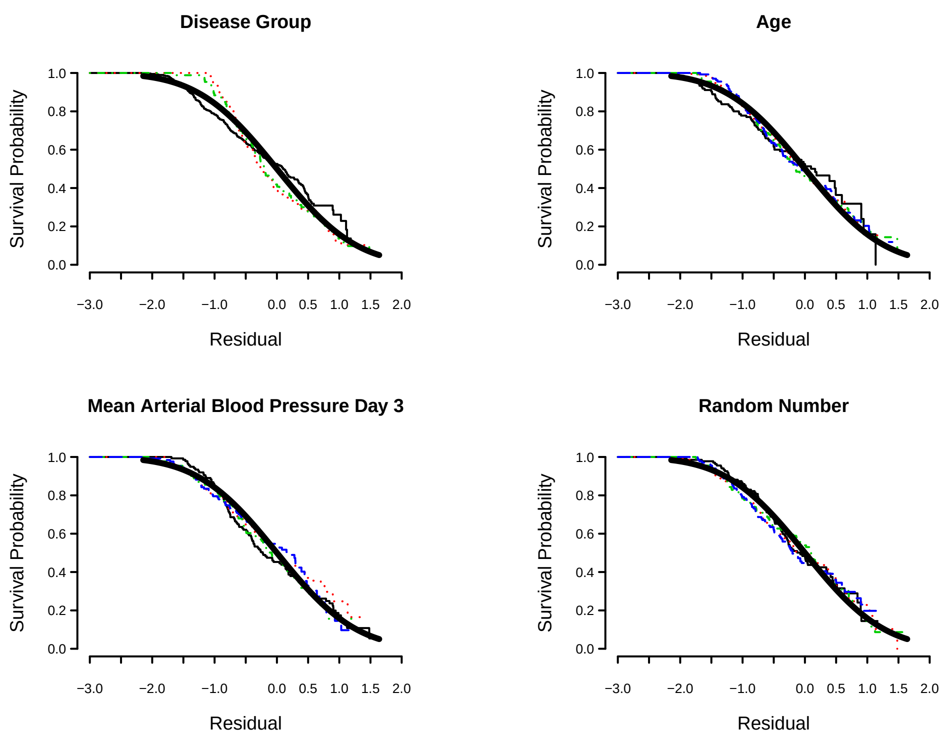


Figure 12.4: Kaplan-Meier estimates of distributions of normalized, right-censored residuals from the fitted log-normal survival model. Residuals are stratified by important variables in the model (by quartiles of continuous variables), plus a random variable to depict the natural variability (in the lower right plot). Theoretical standard Gaussian distributions of residuals are shown with a thick solid line. The upper left plot is with respect to disease group.

were only collected for the second phase of SUPPORT.

Of those variables to be included in the model, find which ones have enough potential predictive power to justify allowing for nonlinear relationships or multiple categories, which spend more d.f. For each variable compute Spearman ρ^2 based on multiple linear regression of rank(x), rank(x)^2 and the survival time, truncating survival time at the shortest follow-up for survivors (356 days). This rids the data of censoring but creates many ties at 356 days.

```
shortest.follow.up ← min(d.time[death==0], na.rm=TRUE)
d.timet ← pmin(d.time, shortest.follow.up)

w ← spearman2(d.timet ~ age + num.co + scoma + meanbp +
              hrt + resp + temp + crea + sod + adlsc +
              wblc + pafi + ph + dzgroup + race, p=2)
plot(w, main='')        # Figure 12.5
```

A better approach is to use the complete information in the failure and censoring times by computing Somers' D_{xy} rank correlation allowing for censoring.

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

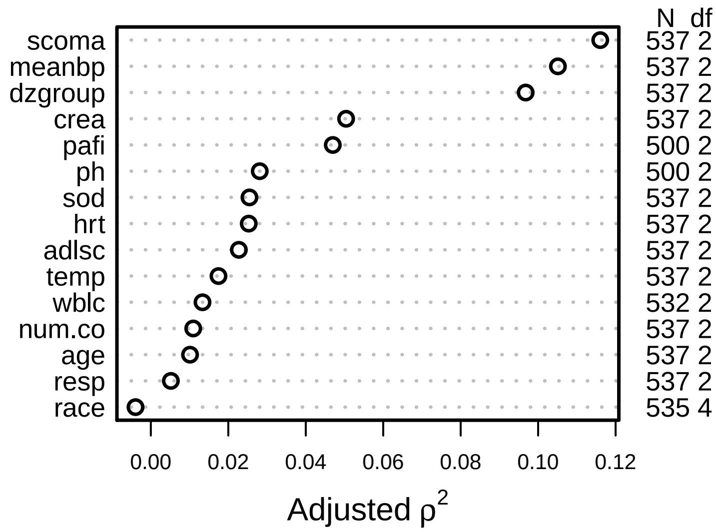


Figure 12.5: Generalized Spearman ρ^2 rank correlation between predictors and truncated survival time

```
                    dzgroup + race)
plot(w, main='')             # Figure 12.6
```

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

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

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

Do a formal redundancy analysis using more than pairwise associations, and allow for non-

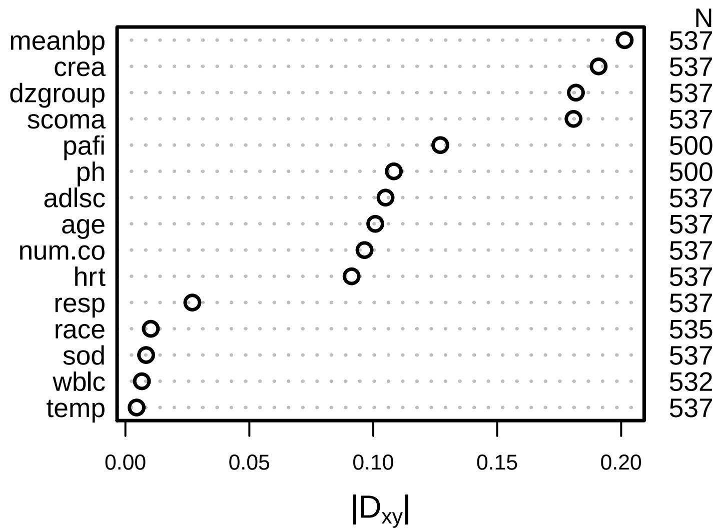


Figure 12.6: Somers' D_{xy} rank correlation between predictors and original survival time. For `dzgroup` or `race`, the correlation coefficient is the maximum correlation from using a dummy variable to represent the most frequent or one to represent the second most frequent category.',scap='Somers' D_{xy} rank correlation between predictors and original survival time

monotonic transformations in predicting each predictor from all other predictors. This analysis requires missing values to be imputed so as to not greatly reduce the sample size.

```
redun(~ crea + age + sex + dzgroup + num.co + scoma + adlsc + race2 +
      meanbp + hrt + resp + temp + sod + wblc.i + pafi.i + ph.i, nk=4)
```

```

Redundancy Analysis

redun(formula = ~crea + age + sex + dzgroup + num.co + scoma +
    adlsc + race2 + meanbp + hrt + resp + temp + sod + wblc.i +
    pafi.i + ph.i, nk = 4)

n: 537  p: 16   nk: 4

Number of NAs:   0

Transformation of target variables forced to be linear

R2 cutoff: 0.9  Type: ordinary

```

```
R^2 with which each variable can be predicted from all other variables:

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

No redundant variables
```

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

- Allow all continuous variables to have a the maximum number of knots entertained, in a log-normal survival model
- Must use imputation to avoid losing data
- Fit a "saturated" main effects model
- Makes full use of censored data
- Had to limit to 4 knots, force `scoma` to be linear, and omit `ph.i` to avoid singularity

```
k ← 4
f ← psm(S ~ rcs(age,k)+sex+dzgroup+pol(num.co,2)+scoma+
          pol(adlsc,2)+race+rcs(meanbp,k)+rcs(hrt,k)+rcs(resp,k)+
          rcs(temp,k)+rcs(crea,3)+rcs(sod,k)+rcs(wblc.i,k)+
          rcs(pafi.i,k), dist='lognormal')
plot(anova(f))    # Figure 12.7
```

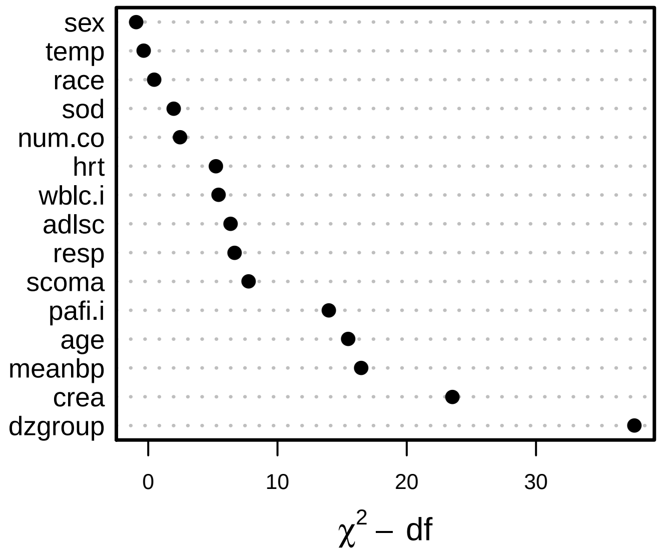


Figure 12.7: Partial χ^2 statistics for association of each predictor with response from saturated main effects model, penalized for d.f.

- Figure 12.7 properly blinds the analyst to the form of effects (tests of linearity).
- Fit a log-normal survival model with number of parameters corresponding to nonlinear effects determined from Figure 12.7. For the most promising predictors, five knots can be allocated, as there are fewer singularity problems once less promising predictors are simplified.

```
f ← psm(S ~ rcs(age,5)+sex+dzgroup+num.co+
              scoma+pol(adlsc,2)+race2+rcs(meanbp,5)+
              rcs(hrt,3)+rcs(resp,3)+temp+
              rcs(crea,4)+sod+rcs(wblc.i,3)+rcs(pafi.i,4),
          dist='lognormal')  # 'gaussian' for S+
print(f, latex=TRUE)
```

Parametric Survival Model: Log Normal Distribution

```
psm(formula = S ~ rcs(age, 5) + sex + dzgroup + num.co + scoma +
    pol(adlsc, 2) + race2 + rcs(meanbp, 5) + rcs(hrt, 3) + rcs(resp,
    3) + temp + rcs(crea, 4) + sod + rcs(wblc.i, 3) + rcs(pafi.i,
    4), dist = "lognormal")
```

	Model Likelihood Ratio Test	Discrimination Indexes
Obs 537	LR χ^2 236.83	R^2 0.594
Events 356	d.f. 30	D_{xy} 0.485
σ 2.230782	Pr$(> \chi^2) < 0.0001$	g 0.033
		g_r 1.959

	Coef	S.E.	Wald Z	Pr$(>\lvert Z\rvert)$
(Intercept)	-5.6883	3.7851	-1.50	0.1329
age	-0.0148	0.0309	-0.48	0.6322
age'	-0.0412	0.1078	-0.38	0.7024
age"	0.1670	0.5594	0.30	0.7653
age"'	-0.2099	1.3707	-0.15	0.8783
sex=male	-0.0737	0.2181	-0.34	0.7354
dzgroup=Coma	-2.0676	0.4062	-5.09	< 0.0001
dzgroup=MOSF w/Malig	-1.4664	0.3112	-4.71	< 0.0001
num.co	-0.1917	0.0858	-2.23	0.0255
scoma	-0.0142	0.0044	-3.25	0.0011
adlsc	-0.3735	0.1520	-2.46	0.0140
adlsc^2	0.0442	0.0243	1.82	0.0691
race2=other	0.2979	0.2658	1.12	0.2624
meanbp	0.0702	0.0210	3.34	0.0008
meanbp'	-0.3080	0.2261	-1.36	0.1732
meanbp"	0.8438	0.8556	0.99	0.3241
meanbp"'	-0.5715	0.7707	-0.74	0.4584
hrt	-0.0171	0.0069	-2.46	0.0140
hrt'	0.0064	0.0063	1.02	0.3090
resp	0.0454	0.0230	1.97	0.0483
resp'	-0.0851	0.0291	-2.93	0.0034
temp	0.0523	0.0834	0.63	0.5308
crea	-0.4585	0.6727	-0.68	0.4955
crea'	-11.5176	19.0027	-0.61	0.5444
crea"	21.9840	31.0113	0.71	0.4784
sod	0.0044	0.0157	0.28	0.7792
wblc.i	0.0746	0.0331	2.25	0.0242
wblc.i'	-0.0880	0.0377	-2.34	0.0195
pafi.i	0.0169	0.0055	3.07	0.0021
pafi.i'	-0.0569	0.0239	-2.38	0.0173
pafi.i"	0.1088	0.0482	2.26	0.0239

	Coef	S.E.	Wald Z	Pr$(>|Z|)$
Log(scale)	0.8024	0.0401	19.99	<0.0001

12.3 Summarizing the Fitted Model

- Plot the shape of the effect of each predictor on log survival time.
- All effects centered: can be placed on common scale
- Wald χ^2 statistics, penalized for d.f., plotted in descending order

```
ggplot(Predict(f, ref.zero=TRUE), vnames='names',
       sepdiscrete='vertical')  # Figure 12.8
```

```
latex(anova(f), file='', label='support-anovat')   # Table 12.2
```

```
plot(anova(f))              # Figure 12.9
```

```
options(digits=3)
plot(summary(f), log=TRUE, main='')   # Figure 12.10
```

12.4 Internal Validation of the Fitted Model Using the Bootstrap

Validate indexes describing the fitted model.

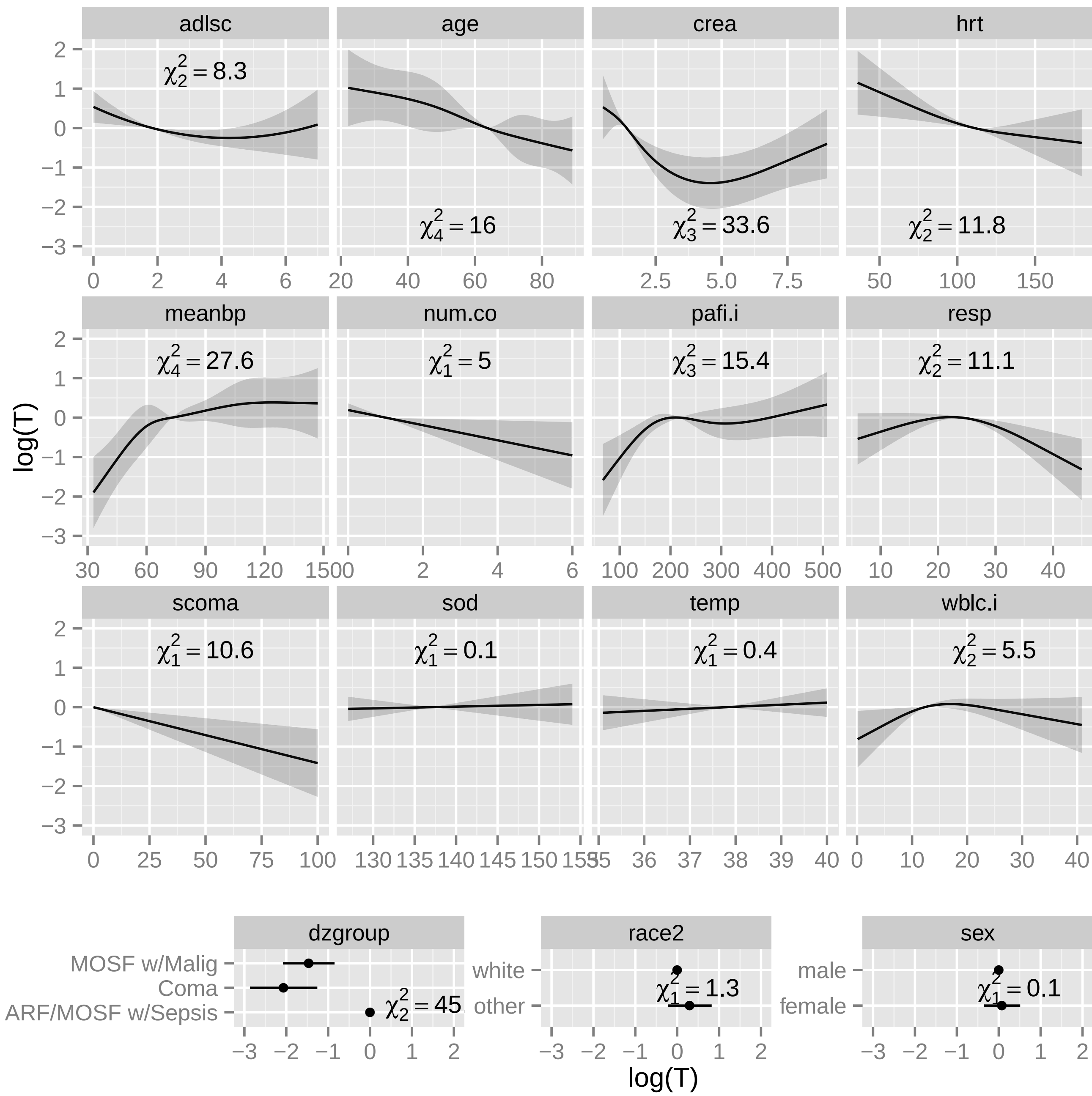


Figure 12.8: Effect of each predictor on log survival time. Predicted values have been centered so that predictions at predictor reference values are zero. Pointwise 0.95 confidence bands are also shown. As all Y-axes have the same scale, it is easy to see which predictors are strongest.

Table 12.2: Wald Statistics for S

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

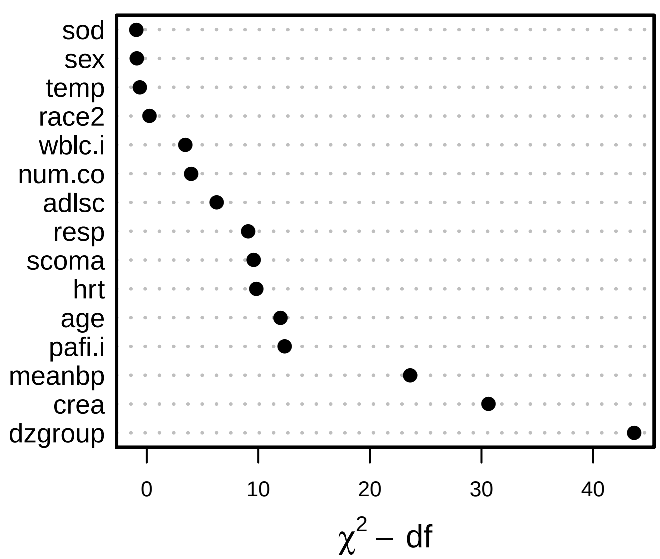


Figure 12.9: Contribution of variables in predicting survival time in log-normal model

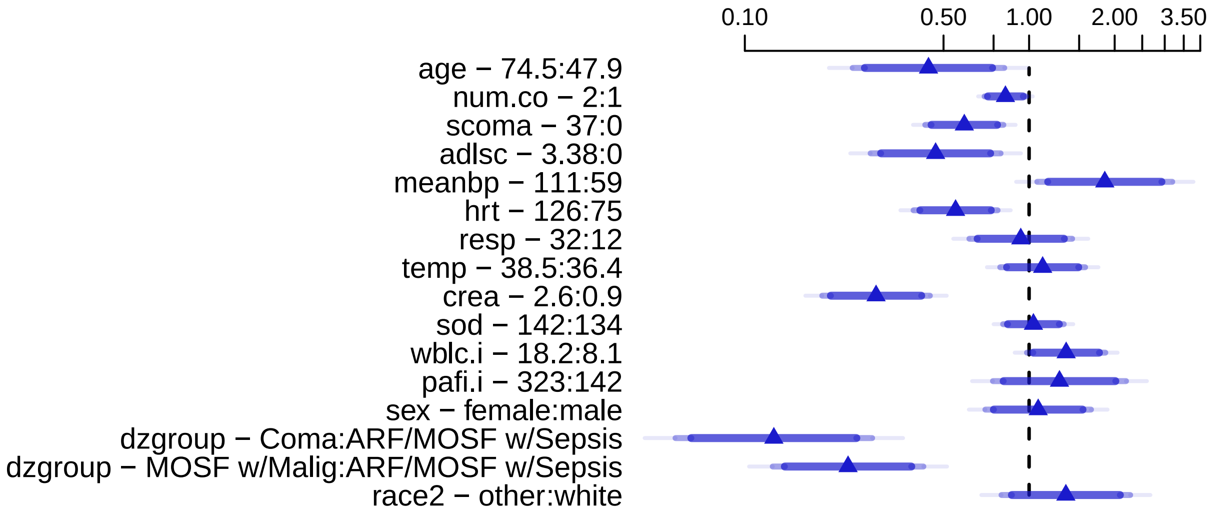


Figure 12.10: Estimated survival time ratios for default settings of predictors. For example, when age changes from its lower quartile to the upper quartile (47.9y to 74.5y), median survival time decreases by more than half. Different shaded areas of bars indicate different confidence levels (0.9, 0.95, 0.99).

```
# First add data to model fit so bootstrap can re-sample
#  from the data
g ← update(f, x=TRUE, y=TRUE)
set.seed(717)
latex(validate(g, B=120, dxy=TRUE), digits=2, size='Ssize')
```

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

- From D_{xy} and R^2 there is a moderate amount of overfitting.
- Slope shrinkage factor (0.90) is not trouble-

some

- Almost unbiased estimate of future predictive discrimination on similar patients is the corrected D_{xy} of 0.43.

Validate predicted 1-year survival probabilities. Use a smooth approach that does not require binning[103] and use less precise Kaplan-Meier estimates obtained by stratifying patients by the predicted probability, with at least 60 patients per group.

```
set.seed(717)
cal ← calibrate(g, u=1, B=120)
plot(cal, subtitles=FALSE)
cal ← calibrate(g, cmethod='KM', u=1, m=60, B=120, pr=FALSE)
plot(cal, add=TRUE)     # Figure 12.11
```

12.5 Approximating the Full Model

The fitted log-normal model is perhaps too complex for routine use and for routine data collection. Let us develop a simplified model that can predict the predicted values of the full model

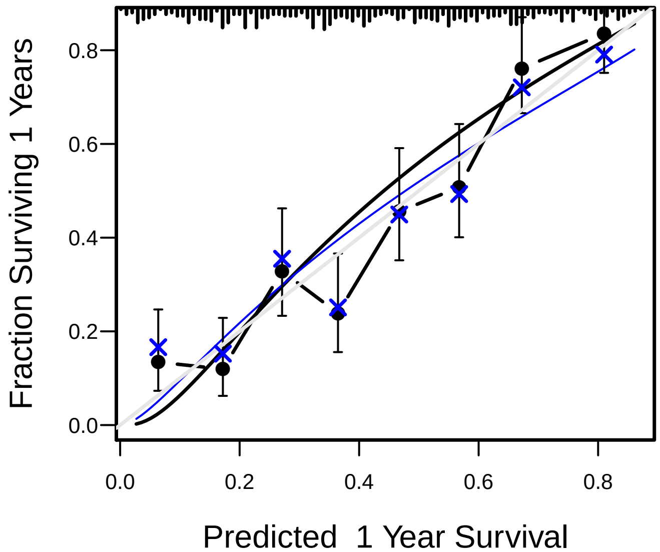


Figure 12.11: Bootstrap validation of calibration curve. Dots represent apparent calibration accuracy; × are bootstrap estimates corrected for overfitting, based on binning predicted survival probabilities and and computing Kaplan-Meier estimates. Black curve is the estimated observed relationship using `hare` and the blue curve is the overfitting-corrected `hare` estimate. The gray-scale line depicts the ideal relationship.

with high accuracy ($R^2 = 0.96$). The simplification is done using a fast backward stepdown against the full model predicted values.

```
Z ← predict(f)     # X*beta hat
a ← ols(Z ~ rcs(age,5)+sex+dzgroup+num.co+
            scoma+pol(adlsc,2)+race2+
            rcs(meanbp,5)+rcs(hrt,3)+rcs(resp,3)+
            temp+rcs(crea,4)+sod+rcs(wblc.i,3)+
            rcs(pafi.i,4), sigma=1)
# sigma=1 is used to prevent sigma hat from being zero when
# R2=1.0 since we start out by approximating Z with all
#  component variables
fastbw(a, aics=10000)    # fast backward stepdown
```

```
Deleted  Chi-Sq d.f. P      Residual d.f. P      AIC    R2
sod       0.43  1    0.512   0.43    1    0.5117  -1.57 1.000
sex       0.57  1    0.451   1.00    2    0.6073  -3.00 0.999
temp      2.20  1    0.138   3.20    3    0.3621  -2.80 0.998
race2     6.81  1    0.009  10.01    4    0.0402   2.01 0.994
wblc.i   29.52  2    0.000  39.53    6    0.0000  27.53 0.976
num.co   30.84  1    0.000  70.36    7    0.0000  56.36 0.957
```

```
 resp      54.18 2     0.000  124.55   9    0.0000  106.55 0.924
 adlsc     52.46 2     0.000  177.00  11    0.0000  155.00 0.892
 pafi.i    66.78 3     0.000  243.79  14    0.0000  215.79 0.851
 scoma     78.07 1     0.000  321.86  15    0.0000  291.86 0.803
 hrt       83.17 2     0.000  405.02  17    0.0000  371.02 0.752
 age       68.08 4     0.000  473.10  21    0.0000  431.10 0.710
 crea     314.47 3     0.000  787.57  24    0.0000  739.57 0.517
 meanbp   403.04 4     0.000 1190.61  28    0.0000 1134.61 0.270
 dzgroup  441.28 2     0.000 1631.89  30    0.0000 1571.89 0.000

Approximate Estimates after Deleting Factors

        Coef     S.E. Wald Z P
[1,] -0.5928 0.04315 -13.74 0

Factors in Final Model

None
```

```
f.approx ← ols(Z ~ dzgroup + rcs(meanbp,5) + rcs(crea,4) + rcs(age,5) +
               rcs(hrt,3) + scoma + rcs(pafi.i,4) + pol(adlsc,2)+
               rcs(resp,3), x=TRUE)
f.approx$stats
```

```
        n  Model L.R.       d.f.         R2           g
  537.000   1688.225     23.000      0.957       1.915
    Sigma
    0.370
```

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

```
V ← vcov(f,regcoef.only=TRUE)         # var(full model)
X ← cbind(Intercept=1, g$x)           # full model design
x ← cbind(Intercept=1, f.approx$x)    # approx. model design
w ← solve(t(x) %*% x, t(x)) %*% X     # contrast matrix
v ← w %*% V %*% t(w)
```

Table 12.3: Wald Statistics for Z

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

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

```
f.sub ← psm(S ∼ dzgroup + rcs(meanbp,5) + rcs(crea,4) + rcs(age,5) +
            rcs(hrt,3) + scoma + rcs(pafi.i,4) + pol(adlsc,2)+
            rcs(resp,3), dist='lognormal')  # 'gaussian' for S+

r ← diag(v)/diag(vcov(f.sub,regcoef.only=TRUE))
r[c(which.min(r), which.max(r))]
```

```
 hrt'   age
0.976 0.982
```

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

Equation for simplified model:

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

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

$$
\begin{aligned}
X\hat{\beta} = &\\
&-2.51\\
&-1.94[\text{Coma}] - 1.75[\text{MOSF w/Malig}]\\
&+0.068\text{meanbp} - 3.08\times10^{-5}(\text{meanbp} - 41.8)_+^3 + 7.9\times10^{-5}(\text{meanbp} - 61)_+^3\\
&-4.91\times10^{-5}(\text{meanbp} - 73)_+^3 + 2.61\times10^{-6}(\text{meanbp} - 109)_+^3 - 1.7\times10^{-6}(\text{meanbp} - 135)_+^3\\
&-0.553\text{crea} - 0.229(\text{crea} - 0.6)_+^3 + 0.45(\text{crea} - 1.1)_+^3 - 0.233(\text{crea} - 1.94)_+^3\\
&+0.0131(\text{crea} - 7.32)_+^3\\
&-0.0165\text{age} - 1.13\times10^{-5}(\text{age} - 28.5)_+^3 + 4.05\times10^{-5}(\text{age} - 49.5)_+^3\\
&-2.15\times10^{-5}(\text{age} - 63.7)_+^3 - 2.68\times10^{-5}(\text{age} - 72.7)_+^3 + 1.9\times10^{-5}(\text{age} - 85.6)_+^3\\
&-0.0136\text{hrt} + 6.09\times10^{-7}(\text{hrt} - 60)_+^3 - 1.68\times10^{-6}(\text{hrt} - 111)_+^3 + 1.07\times10^{-6}(\text{hrt} - 140)_+^3\\
&-0.0135\ \text{scoma}\\
&+0.0161\text{pafi.i} - 4.77\times10^{-7}(\text{pafi.i} - 88)_+^3 + 9.11\times10^{-7}(\text{pafi.i} - 167)_+^3\\
&-5.02\times10^{-7}(\text{pafi.i} - 276)_+^3 + 6.76\times10^{-8}(\text{pafi.i} - 426)_+^3 - 0.369\ \text{adlsc} + 0.0409\ \text{adlsc}^2\\
&+0.0394\text{resp} - 9.11\times10^{-5}(\text{resp} - 10)_+^3 + 0.000176(\text{resp} - 24)_+^3 - 8.5\times10^{-5}(\text{resp} - 39)_+^3
\end{aligned}
$$

and $[c] = 1$ if subject is in group c, 0 otherwise; $(x)_+ = x$ if $x > 0$, 0 otherwise.

Nomogram for predicting median and mean survival time, based on approximate model:

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

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

```
latex(quantile.surv, file='', type='Sinput')
```

```
quantile.surv ← function (q = 0.5, lp = NULL, parms = 0.802352037606488)
{
```

```
    names(parms) ← NULL
    f ← function(lp, q, parms) lp + exp(parms) * qnorm(q)
    names(q) ← format(q)
    drop(exp(outer(lp, q, FUN = f, parms = parms)))
}
```

```
median.surv   ← function(x) quantile.surv(lp=x)
```

```
# Improve variable labels for the nomogram
f.approx ← Newlabels(f.approx, c('Disease Group','Mean Arterial BP',
          'Creatinine','Age','Heart Rate','SUPPORT Coma Score',
          'PaO2/(.01*FiO2)','ADL','Resp. Rate'))
nom ←
  nomogram(f.approx,
           pafi.i=c(0, 50, 100, 200, 300, 500, 600, 700, 800, 900),
           fun=list('Median Survival Time'=median.surv,
                    'Mean Survival Time'  =expected.surv),
           fun.at=c(.1,.25,.5,1,2,5,10,20,40))
plot(nom, cex.var=1, cex.axis=.75, lmgp=.25)
# Figure 12.12
```

S Packages and Functions Used

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

Note: All packages are available from CRAN

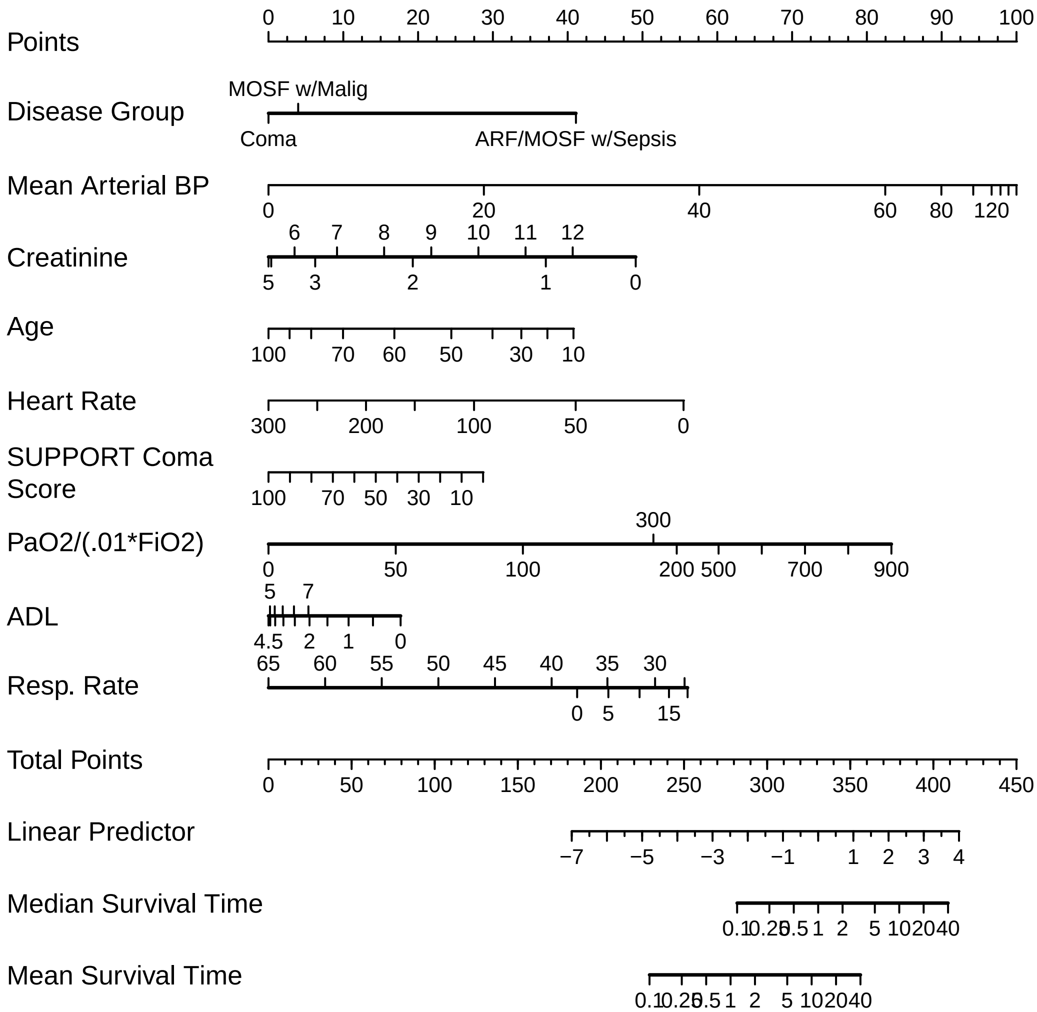


Figure 12.12: Nomogram for predicting median and mean survival time, based on approximation of full model

Chapter 13

Case Study in Cox Regression

13.1 Choosing the Number of Parameters and Fitting the Model

- Clinical trial of estrogen for prostate cancer
- Response is time to death, all causes
- Base analysis on Cox proportional hazards model[41]
- $S(t|X)$ = probability of surviving at least to time t given set of predictor values X
- $S(t|X) = S_0(t)^{\exp(X\beta)}$
- Censor time to death at time of last follow-up for patients still alive at end of study (treat survival time for pt. censored at 24m as 24m+)

- Use simple, partial approaches to data reduction
- Use `transcan` for single imputation
- Again combine last 2 categories for `ekg,pf`
- See if we can use a full additive model (4 knots for continuous X)

Predictor	Name	d.f.	Original Levels
Dose of estrogen	`rx`	3	placebo, 0.2, 1.0, 5.0 mg estrogen
Age in years	`age`	3	
Weight index: wt(kg)-ht(cm)+200	`wt`	3	
Performance rating	`pf`	2	normal, in bed $<$ 50% of time, in bed $>$ 50%, in bed always
History of cardiovascular disease	`hx`	1	present/absent
Systolic blood pressure/10	`sbp`	3	
Diastolic blood pressure/10	`dbp`	3	
Electrocardiogram code	`ekg`	5	normal, benign, rhythm disturb., block, strain, old myocardial infarction, new MI
Serum hemoglobin (g/100ml)	`hg`	3	
Tumor size (cm^2)	`sz`	3	
Stage/histologic grade combination	`sg`	3	
Serum prostatic acid phosphatase	`ap`	3	
Bone metastasis	`bm`	1	present/absent

- Total of 36 candidate d.f.
- Impute missings and estimate shrinkage

```
require(rms)
```

```
getHdata(prostate)
levels(prostate$ekg)[levels(prostate$ekg) %in%
                     c('old MI','recent MI')] ← 'MI'
# combines last 2 levels and uses a new name, MI

prostate$pf.coded ← as.integer(prostate$pf)
# save original pf, re-code to 1-4
levels(prostate$pf)  ← c(levels(prostate$pf)[1:3],
                        levels(prostate$pf)[3])
# combine last 2 levels

w ← transcan(~ sz + sg + ap + sbp + dbp + age +
             wt + hg + ekg + pf + bm + hx,
             imputed=TRUE, data=prostate, pl=FALSE, pr=FALSE)

attach(prostate)
sz  ← impute(w, sz, data=prostate)
sg  ← impute(w, sg, data=prostate)
age ← impute(w, age,data=prostate)
wt  ← impute(w, wt, data=prostate)
ekg ← impute(w, ekg,data=prostate)

dd ← datadist(prostate)
options(datadist='dd')

units(dtime) ← 'Month'
S ← Surv(dtime, status!='alive')

f ← cph(S ~ rx + rcs(age,4) + rcs(wt,4) + pf + hx +
        rcs(sbp,4) + rcs(dbp,4) + ekg + rcs(hg,4) +
        rcs(sg,4) + rcs(sz,4) + rcs(log(ap),4) + bm)
```

```
print(f, latex=TRUE)
```

Cox Proportional Hazards Model

```
cph(formula = S ~ rx + rcs(age, 4) + rcs(wt, 4) + pf + hx + rcs(sbp,
    4) + rcs(dbp, 4) + ekg + rcs(hg, 4) + rcs(sg, 4) + rcs(sz,
    4) + rcs(log(ap), 4) + bm)
```

		Model Tests		Discrimination Indexes	
Obs	502	LR χ^2	135.44	R^2	0.237
Events	354	d.f.	36	D_{xy}	0.332
Center	-2.9844	Pr$(>\chi^2)$	0.0000	g	0.783
		Score χ^2	143.22	g_r	2.189
		Pr$(>\chi^2)$	0.0000		

	Coef	S.E.	Wald Z	Pr$(>\|Z\|)$
rx=0.2 mg estrogen	0.0106	0.1551	0.07	0.9454
rx=1.0 mg estrogen	-0.3607	0.1703	-2.12	0.0342
rx=5.0 mg estrogen	-0.0479	0.1614	-0.30	0.7665
age	0.0031	0.0244	0.13	0.9004
age'	-0.0009	0.0397	-0.02	0.9827
age"	0.5690	0.5075	1.12	0.2622
wt	-0.0063	0.0165	-0.38	0.7029
wt'	-0.0479	0.0528	-0.91	0.3646
wt"	0.2592	0.2074	1.25	0.2114
pf=in bed $<$ 50% daytime	0.3640	0.2082	1.75	0.0804
pf=in bed $>$ 50% daytime	0.4971	0.3283	1.51	0.1299
hx	0.4650	0.1207	3.85	0.0001
sbp	-0.1033	0.1069	-0.97	0.3339
sbp'	0.2570	0.3962	0.65	0.5166
sbp"	-0.5495	1.0059	-0.55	0.5849
dbp	-0.0628	0.1284	-0.49	0.6249
dbp'	0.4371	0.2552	1.71	0.0867
dbp"	-3.2265	1.4725	-2.19	0.0284
ekg=benign	0.0809	0.2789	0.29	0.7718
ekg=rhythmic disturb & electrolyte ch	0.4079	0.1948	2.09	0.0363
ekg=heart block or conduction def	0.0686	0.2763	0.25	0.8039
ekg=heart strain	0.4019	0.1452	2.77	0.0056
ekg=MI	0.0768	0.1781	0.43	0.6662
hg	-0.1597	0.0815	-1.96	0.0500
hg'	0.0320	0.2183	0.15	0.8836
hg"	0.9678	1.2979	0.75	0.4559
sg	-0.0297	0.1125	-0.26	0.7920

	Coef	S.E.	Wald Z	Pr($>|Z|$)
sg'	0.5856	0.6823	0.86	0.3907
sg"	-1.0621	1.2270	-0.87	0.3867
sz	0.0507	0.0345	1.47	0.1424
sz'	-0.3800	0.3260	-1.17	0.2437
sz"	0.6267	0.5258	1.19	0.2333
ap	-0.4322	0.2147	-2.01	0.0441
ap'	4.9772	2.6931	1.85	0.0646
ap"	-10.0357	5.6126	-1.79	0.0738
bm	0.0781	0.1898	0.41	0.6808

- Global LR χ^2 is 135 and very significant $\rightarrow$ modeling warranted
- AIC on χ^2 scale = $135 - 2 \times 36 = 63$
- Rough shrinkage: 0.73 ($\frac{135.44-36}{135.44}$)
- Informal data reduction (increase for `ap`)

Variables	Reductions	d.f. Saved
`wt`	Assume variable not important enough for 4 knots; use 3 knots	1
`pf`	Assume linearity	1
`hx,ekg`	Make new 0,1,2 variable and assume linearity: 2=`hx` and `ekg` not normal or benign, 1=either, 0=none	5
`sbp,dbp`	Combine into mean arterial bp and use 3 knots: map=$\frac{2}{3}$ `dbp` $+\frac{1}{3}$ `sbp`	4
`sg`	Use 3 knots	1
`sz`	Use 3 knots	1
`ap`	Look at shape of effect of `ap` in detail, and take log before expanding as spline to achieve numerical stability: add 2 knots	-2

```
heart ← hx + ekg %nin% c('normal','benign')
label(heart) ← 'Heart Disease Code'
map    ← (2*dbp + sbp)/3
label(map) ← 'Mean Arterial Pressure/10'
dd ← datadist(dd, heart, map)

f ← cph(S ~ rx + rcs(age,4) + rcs(wt,3) + pf.coded +
         heart + rcs(map,3) + rcs(hg,4) +
         rcs(sg,3) + rcs(sz,3) + rcs(log(ap),6) + bm,
         x=TRUE, y=TRUE, surv=TRUE, time.inc=5*12)
```

```
print(f, latex=TRUE, coefs=FALSE)
```

Cox Proportional Hazards Model

```
cph(formula = S ~ rx + rcs(age, 4) + rcs(wt, 3) + pf.coded +
    heart + rcs(map, 3) + rcs(hg, 4) + rcs(sg, 3) + rcs(sz, 3) +
    rcs(log(ap), 6) + bm, x = TRUE, y = TRUE, surv = TRUE, time.inc = 5 *
    12)
```

		Model Tests		Discrimination Indexes	
Obs	502	LR χ^2	114.93	R^2	0.205
Events	354	d.f.	25	D_{xy}	0.343
Center	-2.9465	Pr$(>\chi^2)$	0.0000	g	0.794
		Score χ^2	134.11	g_r	2.212
		Pr$(>\chi^2)$	0.0000		

```
# x, y for predict, validate, calibrate;
# surv, time.inc for calibrate
latex(anova(f), file='')
```

- Savings of 11 d.f.
- AIC=65, shrinkage 0.78

13.2 Checking Proportional Hazards

- This is our tentative model

Table 13.2: Wald Statistics for S

	χ^2	d.f.	P
rx	5.17	3	0.1601
age	25.41	3	< 0.0001
Nonlinear	15.75	2	0.0004
wt	10.03	2	0.0066
Nonlinear	3.03	1	0.0815
pf.coded	12.08	1	0.0005
heart	21.90	1	< 0.0001
map	0.14	2	0.9325
Nonlinear	0.07	1	0.7918
hg	20.50	3	0.0001
Nonlinear	9.58	2	0.0083
sg	2.22	2	0.3295
Nonlinear	0.01	1	0.9278
sz	24.05	2	< 0.0001
Nonlinear	2.71	1	0.0998
ap	16.66	5	0.0052
Nonlinear	16.50	4	0.0024
bm	0.03	1	0.8742
TOTAL NONLINEAR	48.48	12	< 0.0001
TOTAL	190.74	25	< 0.0001

- Examine distributional assumptions using scaled Schoenfeld residuals
- Complication arising from predictors using multiple d.f.
- Transform to 1 d.f. empirically using $X\hat{\beta}$
- Following analysis approx. since internal coefficients estimated

```
z ← predict(f, type='terms')
# required x=T above to store design matrix
f.short ← cph(S ∼ z, x=TRUE, y=TRUE)
# store raw x, y so can get residuals
```

- Fit `f.short` has same LR χ^2 of 126 as the fit `f`, but with falsely low d.f.

- All $\beta = 1$

```
phtest ← cox.zph(f.short, transform='identity')
phtest
```

```
                rho   chisq      p
rx          0.09875 3.68894 0.0548
age        -0.03940 0.53133 0.4660
wt          0.03047 0.32617 0.5679
pf.coded   -0.03772 0.51857 0.4715
heart       0.01590 0.10930 0.7409
map        -0.06092 1.09209 0.2960
hg         -0.01838 0.11797 0.7312
sg         -0.03814 0.49720 0.4807
sz         -0.00516 0.00888 0.9249
ap         -0.00424 0.00651 0.9357
bm          0.04319 0.72397 0.3948
GLOBAL           NA 6.92647 0.8050
```

```
plot(phtest, var='rx')
```

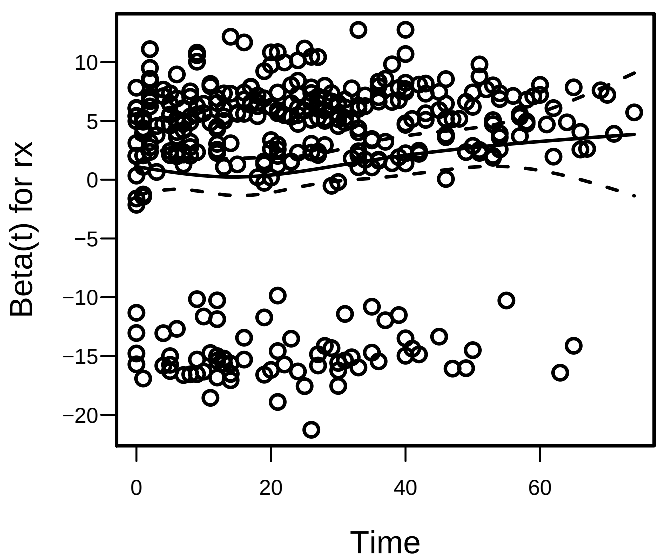


Figure 13.1: Raw and spline-smoothed scaled Schoenfeld residuals for dose of estrogen, nonlinearly coded from the Cox model fit, with $\pm$ 2 standard errors.

- Only the drug effect significantly changes over time

Table 13.3: Wald Statistics for `S`

	χ^2	d.f.	P
z.dose (Factor+Higher Order Factors)	16.94	11	0.1097
All Interactions	10.65	10	0.3856
z.other (Factor+Higher Order Factors)	132.33	20	< 0.0001
All Interactions	10.65	10	0.3856
z.dose $\times$ z.other (Factor+Higher Order Factors)	10.65	10	0.3856
TOTAL	135.35	21	< 0.0001

- Global test of PH $P = 0.84$

13.3 Testing Interactions

- Will ignore non-PH for dose even though it makes sense
- More accurate predictions could be obtained using stratification or time dep. cov.
- Test all interactions with dose
  Reduce to 1 d.f. as before

```
z.dose ← z[,"rx"]   # same as saying z[,1] - get first column
z.other ← z[,−1]    # all but the first column of z
f.ia ← cph(S ~ z.dose * z.other)
latex(anova(f.ia), file='')
```

13.4 Describing Predictor Effects

- Plot relationship between each predictor and $\log \lambda$

```
ggplot(Predict(f), sepdiscrete='vertical', nlevels=4,
       vnames='names')   # Figure 13.2
```

13.5 Validating the Model

- Validate for D_{xy} and slope shrinkage

```
set.seed(1)   # so can reproduce results
v ← validate(f, B=200, dxy=TRUE)
```

Divergence or singularity in 195 samples

```
latex(v, file='')
```

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.3427	0.3332	0.3004	0.0328	0.3099	5
R^2	0.2047	0.2340	0.1813	0.0527	0.1520	5
Slope	1.0000	1.0000	0.8353	0.1647	0.8353	5
D	0.0283	0.0330	0.0247	0.0083	0.0200	5
U	−0.0005	−0.0005	0.0015	−0.0020	0.0015	5
Q	0.0288	0.0335	0.0233	0.0103	0.0185	5
g	0.7940	0.7786	0.6454	0.1332	0.6608	5

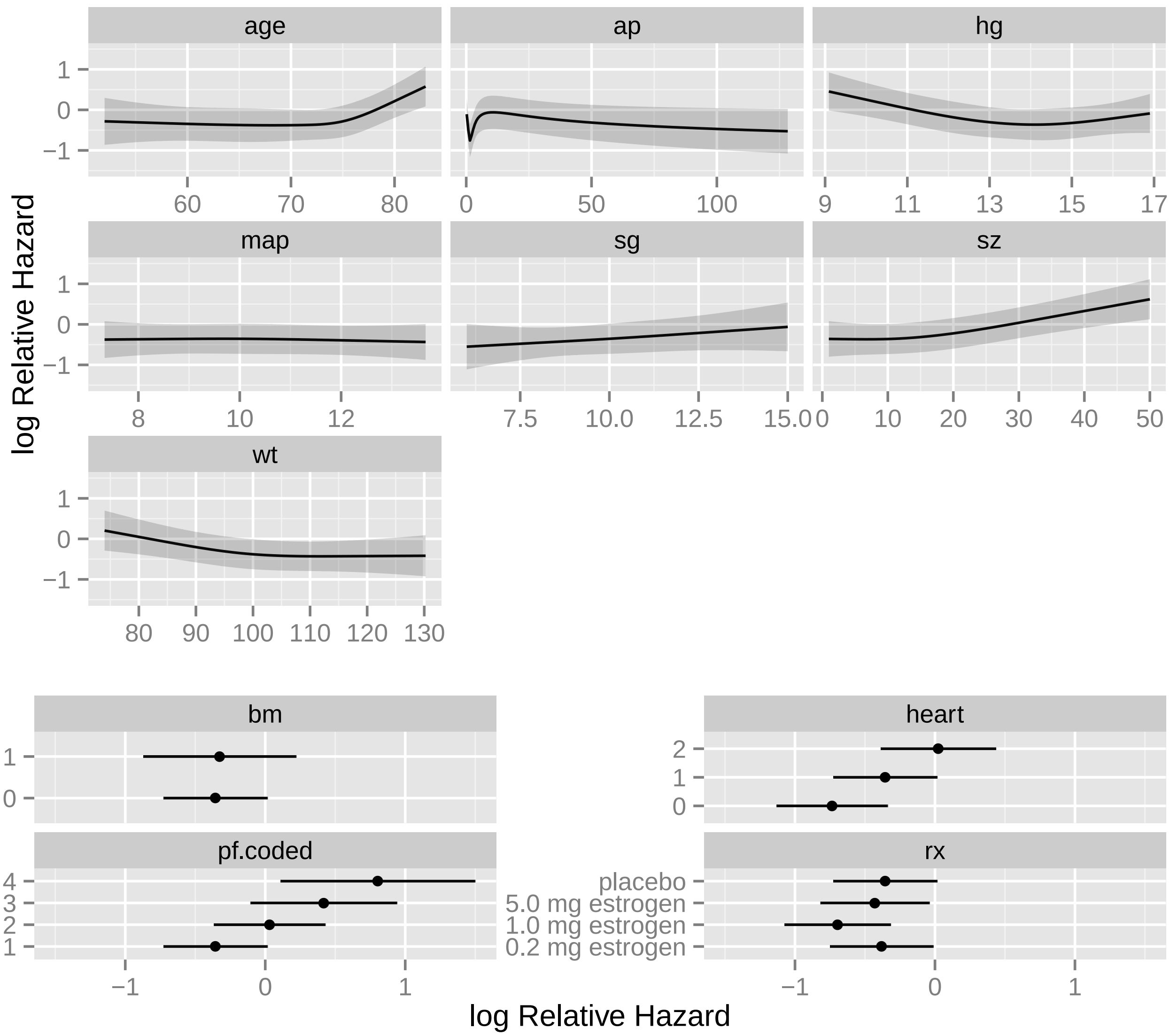


Figure 13.2: Shape of each predictor on log hazard of death. Y-axis shows $X\hat{\beta}$, but the predictors not plotted are set to reference values. Note the highly non-monotonic relationship with `ap`, and the increased slope after age 70 which has been found in outcome models for various diseases.

- Shrinkage surprisingly close to heuristic estimate of 0.78
- Now validate 5-year survival probability estimates

```
cal ← calibrate(f, B=200, u=5*12, maxdim=4)
```

```
Using Cox survival estimates at 60 Months
```

```
plot(cal)
```

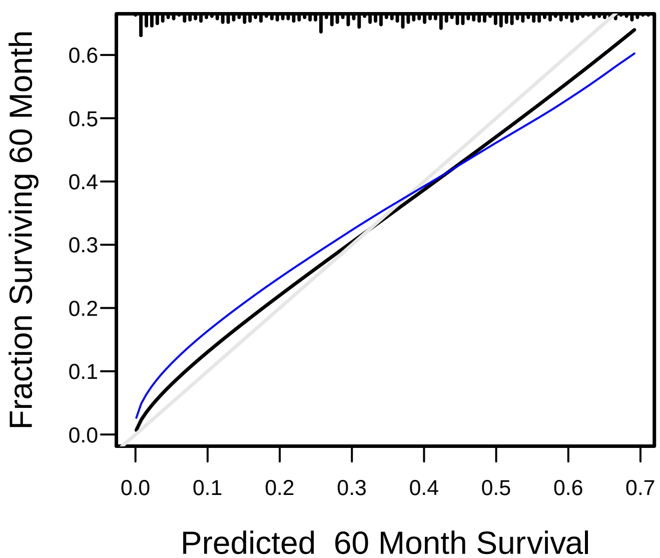


Figure 13.3: Bootstrap estimate of calibration accuracy for 5-year estimates from the final Cox model, using adaptive linear spline hazard regression. Line nearer the ideal line corresponds to apparent predictive accuracy. The blue curve corresponds to bootstrap-corrected estimates.

13.6 Presenting the Model

- Display hazard ratios, overriding default for `ap`

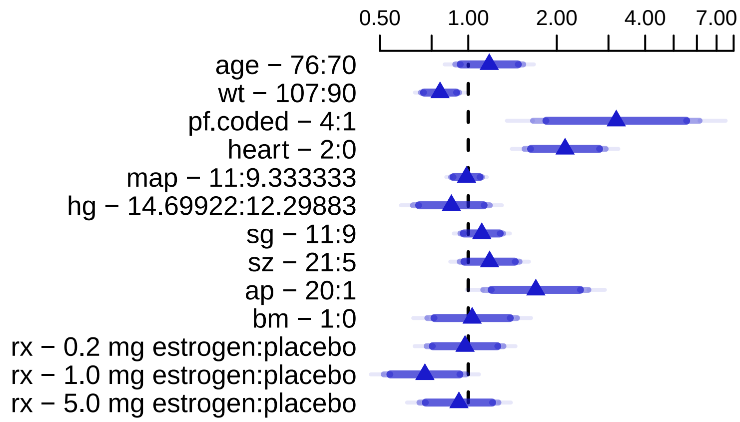


Figure 13.4: Hazard ratios and multi-level confidence bars for effects of predictors in model, using default ranges except for `ap`

```
plot(summary(f, ap=c(1,20)), log=TRUE, main='')
```

- Draw nomogram, with predictions stated 4 ways

```
surv  ← Survival(f)
surv3 ← function(x) surv(3*12,lp=x)
surv5 ← function(x) surv(5*12,lp=x)
quan  ← Quantile(f)
med   ← function(x) quan(lp=x)/12
ss    ← c(.05,.1,.2,.3,.4,.5,.6,.7,.8,.9,.95)

nom ← nomogram(f, ap=c(.1,.5,1,2,3,4,5,10,20,30,40),
               fun=list(surv3, surv5, med),
               funlabel=c('3-year Survival','5-year Survival',
                 'Median Survival Time (years)'),
               fun.at=list(ss, ss, c(.5,1:6)))
plot(nom, xfrac=.65, lmgp=.35)
```

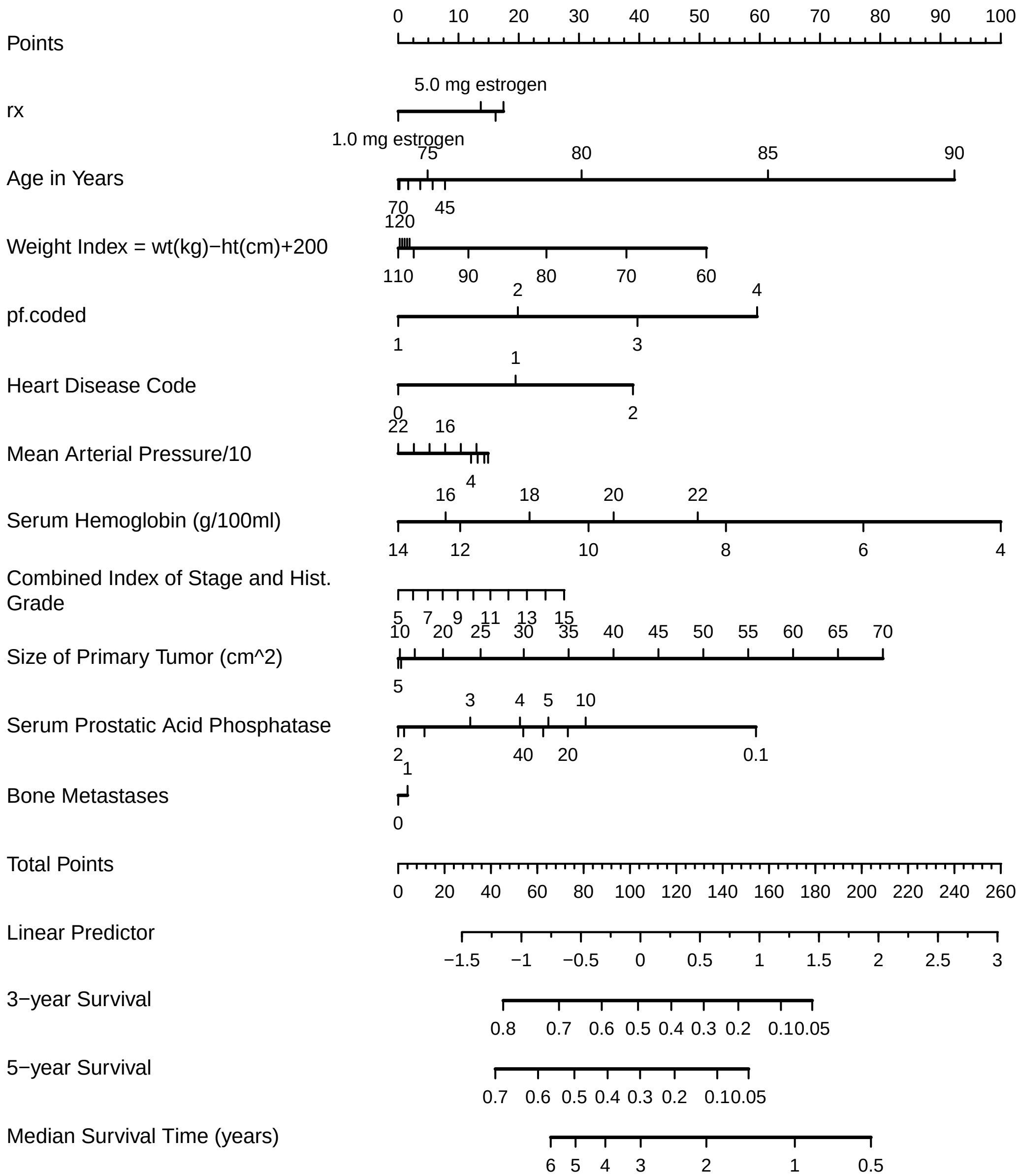


Figure 13.5: Nomogram for predicting death in prostate cancer trial

Bibliography

[1] P. D. Allison. *Missing Data*. Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Sage, Thousand Oaks CA, 2001. pages 66, 69

[2] D. G. Altman. Categorising continuous covariates (letter to the editor). *Brit J Cancer*, 64:975, 1991. pages 29

[3] D. G. Altman. Suboptimal analysis using 'optimal' cutpoints. *Brit J Cancer*, 78:556–557, 1998. pages 29

[4] D. G. Altman and P. K. Andersen. Bootstrap investigation of the stability of a Cox regression model. *Stat Med*, 8:771–783, 1989. pages 99

[5] D. G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher. Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors. *J Nat Cancer Inst*, 86:829–835, 1994. pages 29, 32

[6] B. G. Armstrong and M. Sloan. Ordinal regression models for epidemiologic data. *Am J Epi*, 129:191–204, 1989. See letter to editor by Peterson. pages 287

[7] A. C. Atkinson. A note on the generalized information criterion for choice of a model. *Biometrika*, 67:413–418, 1980. pages 44, 98

[8] P. C. Austin. Bootstrap model selection had similar performance for selecting authentic and noise variables compared to backward variable elimination: a simulation study. *J Clin Epi*, 61:1009–1017, 2008. pages 99

[9] P. C. Austin, J. V. Tu, and D. S. Lee. Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure. *J Clin Epi*, 63:1145–1155, 2010. pages 50

[10] S. A. Barnes, S. R. Lindborg, and J. W. Seaman. Multiple imputation techniques in small sample clinical trials. *Stat Med*, 25:233–245, 2006. pages 80

[11] F. Barzi and M. Woodward. Imputations of missing values in practice: Results from imputations of serum cholesterol in 28 cohort studies. *Am J Epi*, 160:34–45, 2004. pages 72, 80

[12] H. Belcher. The concept of residual confounding in regression models and some applications. *Stat Med*, 11:1747–1758, 1992. pages 29

[13] D. A. Belsley. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. Wiley, New York, 1991. pages 106

[14] D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York, 1980. pages 123, 124

[15] 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. pages 101

[16] K. Berhane, M. Hauptmann, and B. Langholz. Using tensor product splines in modeling exposure–time–response relationships: Application to the Colorado Plateau Uranium Miners cohort. *Stat Med*, 27:5484–5496, 2008. pages 63

[17] D. M. Berridge and J. Whitehead. Analysis of failure time data with ordinal categories of response. *Stat Med*, 10:1703–1710, 1991. pages 287

[18] M. Blettner and W. Sauerbrei. Influence of model-building strategies on the results of a case-control study. *Stat Med*, 12:1325–1338, 1993. pages 156

[19] J. G. Booth and S. Sarkar. Monte Carlo approximation of bootstrap variances. *Am Statistician*, 52:354–357, 1998. pages 144

[20] R. Bordley. Statistical decisionmaking without math. *Chance*, 20(3):39–44, 2007. pages 9

[21] L. Breiman. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *J Am Stat Assoc*, 87:738–754, 1992. pages 98, 99, 150

[22] L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation (with discussion). *J Am Stat Assoc*, 80:580–619, 1985. pages 116

[23] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1984. pages 48

[24] W. M. Briggs and R. Zaretzki. The skill plot: A graphical technique for evaluating continuous diagnostic tests (with discussion). *Biometrics*, 64:250–261, 2008. pages 9

[25] 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. pages 156

[26] P. Buettner, C. Garbe, and I. Guggenmoos-Holzmann. Problems in defining cutoff points of continuous prognostic factors: Example of tumor thickness in primary cutaneous melanoma. *J Clin Epi*, 50:1201–1210, 1997. pages 29

[27] S. Buuren. *Flexible imputation of missing data*. Chapman & Hall/CRC, Boca Raton, FL, 2012. pages 66, 79, 84

[28] Centers for Disease Control and Prevention CDC. National Center for Health Statistics NCHS. National Health and Nutrition Examination Survey, 2010. pages 293

[29] J. M. Chambers and T. J. Hastie, editors. *Statistical Models in S*. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1992. pages 63

[30] C. Chatfield. Avoiding statistical pitfalls (with discussion). *Statistical Sci*, 6:240–268, 1991. pages 124

[31] C. Chatfield. Model uncertainty, data mining and statistical inference (with discussion). *J Roy Stat Soc A*, 158:419–466, 1995. pages 93, 156

[32] S. Chatterjee and A. S. Hadi. *Regression Analysis by Example*. Wiley, New York, fifth edition, 2012. pages 104

[33] M. Chavent, V. Kuentz-Simonet, B. Liquet, and J. Saracco. ClustOfVar: An R package for the clustering of variables. *J Stat Software*, 50(13):1–16, Sept. 2012. pages 109

[34] A. Ciampi, J. Thiffault, J. P. Nakache, and B. Asselain. Stratification by stepwise regression, correspondence analysis and recursive partition. *Comp Stat Data Analysis*, 1986:185–204, 1986. pages 110

[35] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J Am Stat Assoc*, 74:829–836, 1979. pages 45

[36] D. Collett. *Modelling Binary Data*. Chapman and Hall, London, second edition, 2002. pages 280

[37] E. F. Cook and L. Goldman. Asymmetric stratification: An outline for an efficient method for controlling confounding in cohort studies. *Am J Epi*, 127:626–639, 1988. pages 50

[38] N. R. Cook. Use and misues of the receiver operating characteristic curve in risk prediction. *Circulation*, 115:928–935, 2007. pages 238

[39] J. B. Copas. Regression, prediction and shrinkage (with discussion). *J Roy Stat Soc B*, 45:311–354, 1983. pages 103, 104

[40] J. B. Copas. Cross-validation shrinkage of regression predictors. *J Roy Stat Soc B*, 49:175–183, 1987. pages 154

[41] D. R. Cox. Regression models and life-tables (with discussion). *J Roy Stat Soc B*, 34:187–220, 1972. pages 65, 358

[42] S. L. Crawford, S. L. Tennstedt, and J. B. McKinlay. A comparison of analytic methods for non-random missingness of outcome data. *J Clin Epi*, 48:209–219, 1995. pages 68, 129

[43] N. J. Crichton and J. P. Hinde. Correspondence analysis as a screening method for indicants for clinical diagnosis. *Stat Med*, 8:1351–1362, 1989. pages 110

[44] R. B. D'Agostino, A. J. Belanger, E. W. Markson, M. Kelly-Hayes, and P. A. Wolf. Development of health risk appraisal functions in the presence of multiple indicators: The Framingham Study nursing home institutionalization model. *Stat Med*, 14:1757–1770, 1995. pages 106, 109

[45] 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. H. Moolgavkar and R. L. Prentice, editors, *Modern Statistical Methods in Chronic Disease Epi*, pages 140–147. Wiley, New York, 1986. pages 105

[46] C. S. Davis. *Statistical Methods for the Analysis of Repeated Measurements*. Springer, New York, 2002. pages 175, 190

[47] S. Derksen and H. J. Keselman. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British J Math Stat Psych*, 45:265–282, 1992. pages 93

[48] 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. pages 39

[49] P. J. Diggle, P. Heagerty, K.-Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Clarendon Press, Oxford UK, second edition, 2002. pages 175, 185

[50] Donders, G. J. M. G. van der Heijden, T. Stijnen, and K. G. M. Moons. Review: A gentle introduction to imputation of missing values. *J Clin Epi*, 59:1087–1091, 2006. pages 66, 69, 70

[51] W. D. Dupont. *Statistical Modeling for Biomedical Researchers*. Cambridge University Press, Cambridge, UK, second edition, 2008. pages 383

[52] S. Durrleman and R. Simon. Flexible regression models with cubic splines. *Stat Med*, 8:551–561, 1989. pages 43

[53] B. Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J Am Stat Assoc*, 78:316–331, 1983. pages 151, 154

[54] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993. pages 154

[55] B. Efron and R. Tibshirani. Improvements on cross-validation: The .632+ bootstrap method. *J Am Stat Assoc*, 92:548–560, 1997. pages 154

[56] J. Fan and R. A. Levine. To amnio or not to amnio: That is the decision for Bayes. *Chance*, 20(3):26–32, 2007. pages 9

[57] D. Faraggi and R. Simon. A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis. *Stat Med*, 15:2203–2213, 1996. pages 29

[58] J. J. Faraway. The cost of data analysis. *J Comp Graph Stat*, 1:213–229, 1992. pages 132, 153, 156

[59] V. Fedorov, F. Mannino, and R. Zhang. Consequences of dichotomization. *Pharm Stat*, 8:50–61, 2009. pages 8, 29

[60] S. E. Fienberg. *The Analysis of Cross-Classified Categorical Data*. Springer, New York, second edition, 2007. pages 287

[61] 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. pages 154

[62] J. H. Friedman. A variable span smoother. Technical Report 5, Laboratory for Computational Statistics, Department of Statistics, Stanford University, 1984. pages 116

[63] M. H. Gail and R. M. Pfeiffer. On criteria for evaluating models of absolute risk. *Biostatistics*, 6(2):227–239, 2005. pages 9

[64] J. C. Gardiner, Z. Luo, and L. A. Roman. Fixed effects, random effects and GEE: What are the differences? *Stat Med*, 28:221–239, 2009. pages 184

[65] A. Giannoni, R. Baruah, T. Leong, M. B. Rehman, L. E. Pastormerlo, F. E. Harrell, A. J. Coats, and D. P. Francis. Do optimal prognostic thresholds in continuous physiological variables really exist? Analysis of origin of apparent thresholds, with systematic review for peak oxygen consumption, ejection fraction and BNP. *PLoS ONE*, 9(1), 2014. pages 29, 33

[66] J. H. Giudice, J. R. Fieberg, and M. S. Lenarz. Spending degrees of freedom in a poor economy: A case study of building a sightability model for moose in northeastern minnesota. *J Wildlife Manage*, 2011. pages 87

[67] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *J Am Stat Assoc*, 102:359–378, 2007. pages 9

[68] H. Goldstein. Restricted unbiased iterative generalized least-squares estimation. *Biometrika*, 76(3):622–623, 1989. pages 180, 185

[69] U. S. Govindarajulu, D. Spiegelman, S. W. Thurston, B. Ganguli, and E. A. Eisen. Comparing smoothing techniques in Cox models for exposure-response relationships. *Stat Med*, 26:3735–3752, 2007. pages 44

[70] P. M. Grambsch and P. C. O'Brien. The effects of transformations and preliminary tests for non-linearity in regression. *Stat Med*, 10:697–709, 1991. pages 54, 93

[71] R. J. Gray. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *J Am Stat Assoc*, 87:942–951, 1992. pages 63, 103

[72] R. J. Gray. Spline-based tests in survival analysis. *Biometrics*, 50:640–652, 1994. pages 63

[73] M. J. Greenacre. Correspondence analysis of multivariate categorical data by weighted least-squares. *Biometrika*, 75:457–467, 1988. pages 110

[74] S. Greenland. When should epidemiologic regressions use random coefficients? *Biometrics*, 56:915–921, 2000. pages 93, 126

[75] J. Guo, G. James, E. Levina, G. Michailidis, and J. Zhu. Principal component analysis with sparse fused loadings. *J Comp Graph Stat*, 19(4):930–946, 2011. pages 109

[76] D. Hand and M. Crowder. *Practical Longitudinal Data Analysis*. Chapman & Hall, London, 1996. pages 175

[77] O. Harel and X. Zhou. Multiple imputation: Review of theory, implementation and software. *Stat Med*, 26:3057–3077, 2007. pages 66, 72, 77, 81

[78] F. E. Harrell. The LOGIST Procedure. In *SUGI Supplemental Library Users Guide*, pages 269–293. SAS Institute, Inc., Cary, NC, Version 5 edition, 1986. pages 98

[79] F. E. Harrell, K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati. Regression modeling strategies for improved prognostic prediction. *Stat Med*, 3:143–152, 1984. pages 100

[80] F. E. Harrell, K. L. Lee, and D. B. Mark. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*, 15:361–387, 1996. [1534 citations as of 1/1/11]. pages 87

[81] F. E. Harrell, K. L. Lee, D. B. Matchar, and T. A. Reichert. Regression models for prognostic prediction: Advantages, problems, and suggested solutions. *Ca Trt Rep*, 69:1071–1077, 1985. pages 100

[82] F. E. Harrell, K. L. Lee, and B. G. Pollock. Regression models in clinical studies: Determining relationships between predictors and response. *J Nat Cancer Inst*, 80:1198–1202, 1988. pages 47

[83] F. E. Harrell, P. A. Margolis, S. Gove, K. E. Mason, E. K. Mulholland, D. Lehmann, L. Muhe, S. Gatchalian, and H. F. Eichenwald. Development of a clinical prediction model for an ordinal outcome: The World Health Organization ARI Multicentre Study of clinical signs and etiologic agents of pneumonia, sepsis, and meningitis in young infants. *Stat Med*, 17:909–944, 1998. pages 103, 129

[84] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, New York, second edition, 2008. ISBN-10: 0387848576; ISBN-13: 978-0387848570. pages 52

[85] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC, Boca Raton, FL, 1990. ISBN 9780412343902. pages 53

[86] Y. He and A. M. Zaslavsky. Diagnosing imputation models by applying target analyses to posterior replicates of completed data. *Stat Med*, 31(1):1–18, 2012. pages 82

[87] S. G. Hilsenbeck and G. M. Clark. Practical p-value adjustment for optimally selected cutpoints. *Stat Med*, 15:103–112, 1996. pages 29

[88] W. Hoeffding. A non-parametric test of independence. *Ann Math Stat*, 19:546–557, 1948. pages 109

[89] N. Holländer, W. Sauerbrei, and M. Schumacher. Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint. *Stat Med*, 23:1701–1713, 2004. pages 29, 32

[90] N. J. Horton and K. P. Kleinman. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am Statistician*, 61(1):79–90, 2007. pages 80

[91] C. M. Hurvich and C. L. Tsai. The impact of model selection on inference in linear regression. *Am Statistician*, 44:214–217, 1990. pages 99

[92] 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. pages 15

[93] K. J. Janssen, A. R. Donders, F. E. Harrell, Y. Vergouwe, Q. Chen, D. E. Grobbee, and K. G. Moons. Missing covariate data in medical research: To impute is better than to ignore. *J Clin Epi*, 63:721–727, 2010. pages 85

[94] M. P. Jones. Indicator and stratification methods for missing explanatory variables in multiple linear regression. *J Am Stat Assoc*, 91:222–230, 1996. pages 69

[95] J. D. Kalbfleisch and R. L. Prentice. Marginal likelihood based on Cox's regression and life model. *Biometrika*, 60:267–278, 1973. pages 316

[96] J. Karvanen and F. E. Harrell. Visualizing covariates in proportional hazards model. *Stat Med*, 28:1957–1966, 2009. pages 135

[97] M. G. Kenward, I. R. White, and J. R. Carpener. Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? (letter to the editor). *Stat Med*, 29:1455–1456, 2010. pages 179

[98] W. A. Knaus, F. E. Harrell, J. Lynn, L. Goldman, R. S. Phillips, A. F. Connors, N. V. Dawson, W. J. Fulkerson, R. M. Califf, N. Desbiens, P. Layde, R. K. Oye, P. E. Bellamy, R. B. Hakim, and D. P. Wagner. The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Ann Int Med*, 122:191–203, 1995. pages 116, 332

[99] M. J. Knol, K. J. M. Janssen, R. T. Donders, A. C. G. Egberts, E. R. Heerding, D. E. Grobbee, K. G. M. Moons, and M. I. Geerlings. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *J Clin Epi*, 63:728–736, 2010. pages 69

[100] R. Koenker. *Quantile Regression*. Cambridge University Press, New York, 2005. ISBN-10: 0-521-60827-9; ISBN-13: 978-0-521-60827-5. pages 300

[101] R. Koenker. *quantreg: Quantile Regression*, 2009. R package version 4.38. pages 301

[102] R. Koenker and G. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978. pages 300

[103] C. Kooperberg, C. J. Stone, and Y. K. Truong. Hazard regression. *J Am Stat Assoc*, 90:78–94, 1995. pages 351

[104] W. F. Kuhfeld. The PRINQUAL procedure. In *SAS/STAT 9.2 User's Guide*. SAS Publishing, Cary, NC, second edition, 2009. pages 110, 111

[105] J. M. Landwehr, D. Pregibon, and A. C. Shoemaker. Graphical methods for assessing logistic regression models (with discussion). *J Am Stat Assoc*, 79:61–83, 1984. pages 280

[106] B. Lausen and M. Schumacher. Evaluating the effect of optimized cutoff values in the assessment of prognostic factors. *Comp Stat Data Analysis*, 21(3):307–326, 1996. pages 29

[107] J. F. Lawless and K. Singhal. Efficient screening of nonnormal regression models. *Biometrics*, 34:318–327, 1978. pages 99

[108] S. le Cessie and J. C. van Houwelingen. Ridge estimators in logistic regression. *Appl Stat*, 41:191–201, 1992. pages 103

[109] A. Leclerc, D. Luce, F. Lert, J. F. Chastang, and P. Logeay. Correspondence analysis and logistic modelling: Complementary use in the analysis of a health survey among nurses. *Stat Med*, 7:983–995, 1988. pages 110

[110] S. Lee, J. Z. Huang, and J. Hu. Sparse logistic principal components analysis for binary data. *Ann Appl Stat*, 4(3):1579–1601, 2010. pages 52

[111] C. Leng and H. Wang. On general adaptive sparse principal component analysis. *J Comp Graph Stat*, 18(1):201–215, 2009. pages 52

[112] C. Li and B. E. Shepherd. A new residual for ordinal outcomes. *Biometrika*, 99(2):473–480, 2012. pages 281

[113] K.-Y. Liang and S. L. Zeger. Longitudinal data analysis of continuous and discrete responses for pre-post designs. *Sankhyā*, 62:134–148, 2000. pages 179

[114] J. K. Lindsey. *Models for Repeated Measurements*. Clarendon Press, 1997. pages 175

[115] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, New York, second edition, 2002. pages 68, 74, 82

[116] G. F. Liu, K. Lu, R. Mogg, M. Mallick, and D. V. Mehrotra. Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? *Stat Med*, 28:2509–2530, 2009. pages 179

[117] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. Technical report, arXiv, 2013. pages 93

[118] X. Luo, L. A. Stfanski, and D. D. Boos. Tuning variable selection procedures by adding noise. *Technometrics*, 48:165–175, 2006. pages 18

[119] N. Mantel. Why stepdown procedures in variable selection. *Technometrics*, 12:621–625, 1970. pages 99

[120] S. E. Maxwell and H. D. Delaney. Bivariate median splits and spurious statistical significance. *Psych Bull*, 113:181–190, 1993. pages 29

[121] G. P. McCabe. Principal variables. *Technometrics*, 26:137–144, 1984. pages 108

[122] G. Michailidis and J. de Leeuw. The Gifi system of descriptive multivariate analysis. *Statistical Sci*, 13:307–336, 1998. pages 110

[123] K. G. M. Moons, R. A. R. T. Donders, T. Stijnen, and F. E. Harrell. Using the outcome for imputation of missing predictor values was preferred. *J Clin Epi*, 59:1092–1101, 2006. pages 78

[124] B. K. Moser and L. P. Coombs. Odds ratios for a continuous outcome variable without dichotomizing. *Stat Med*, 23:1843–1860, 2004. pages 29

[125] R. H. Myers. *Classical and Modern Regression with Applications*. PWS-Kent, Boston, 1990. pages 105

[126] N. J. D. Nagelkerke. A note on a general definition of the coefficient of determination. *Biometrika*, 78:691–692, 1991. pages 125

[127] T. G. Nick and J. M. Hardin. Regression modeling strategies: An illustrative case study from medical rehabilitation outcomes research. *Am J Occ Ther*, 53:459–470, 1999. pages 87

[128] D. Paul, E. Bair, T. Hastie, and R. Tibshirani. "Preconditioning" for feature selection and regression in high-dimensional problems. *Ann Stat*, 36(4):1595–1619, 2008. pages 52

[129] P. Peduzzi, J. Concato, A. R. Feinstein, and T. R. Holford. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. *J Clin Epi*, 48:1503–1510, 1995. pages 100

[130] P. Peduzzi, J. Concato, E. Kemper, T. R. Holford, and A. R. Feinstein. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epi*, 49:1373–1379, 1996. pages 100, 101

[131] N. Peek, D. G. T. Arts, R. J. Bosman, P. H. J. van der Voort, and N. F. de Keizer. External validation of prognostic models for critically ill patients required substantial sample sizes. *J Clin Epi*, 60:491–501, 2007. pages 126

[132] M. J. Pencina, R. B. D'Agostino, and O. V. Demler. Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models. *Stat Med*, 31(2):101–113, 2012. pages 126, 238

[133] M. J. Pencina, R. B. D'Agostino, and E. W. Steyerberg. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. *Stat Med*, 30:11–21, 2011. pages 126

[134] M. J. Pencina, R. B. D'Agostino Sr, R. B. D'Agostino Jr, and R. S. Vasan. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat Med*, 27:157–172, 2008. pages 126, 238

[135] S. A. Peters, M. L. Bots, H. M. den Ruijter, M. K. Palmer, D. E. Grobbee, J. R. Crouse, D. H. O'Leary, G. W. Evans, J. S. Raichlen, K. G. Moons, H. Koffijberg, and METEOR study group. Multiple imputation of missing repeated outcome measurements did not add to linear mixed-effects models. *J Clin Epi*, 65(6):686–695, 2012. pages 184

[136] J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, 2000. pages 175, 185, 186

[137] R. F. Potthoff and S. N. Roy. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51:313–326, 1964. pages 180

[138] D. B. Pryor, F. E. Harrell, K. L. Lee, R. M. Califf, and R. A. Rosati. Estimating the likelihood of significant coronary artery disease. *Am J Med*, 75:771–780, 1983. pages 248

[139] P. Radchenko and G. M. James. Variable inclusion and shrinkage algorithms. *J Am Stat Assoc*, 103(483):1304–1315, 2008. pages 51

[140] D. R. Ragland. Dichotomizing continuous outcome variables: Dependence of the magnitude of association and statistical power on the cutpoint. *Epi*, 3:434–440, 1992. See letters to editor May 1993 P. 274-, Vol 4 No. 3. pages 29

[141] B. M. Reilly and A. T. Evans. Translating clinical research into clinical practice: Impact of using prediction rules to make decisions. *Ann Int Med*, 144:201–209, 2006. pages 12

[142] E. B. Roecker. Prediction error and its estimation for subset-selected models. *Technometrics*, 33:459–468, 1991. pages 98, 150

[143] P. Royston, D. G. Altman, and W. Sauerbrei. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med*, 25:127–141, 2006. pages 29

[144] D. Rubin and N. Schenker. Multiple imputation in health-care data bases: An overview and some applications. *Stat Med*, 10:585–598, 1991. pages 77

[145] W. 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. pages 106, 109

[146] W. Sauerbrei and M. Schumacher. A bootstrap resampling procedure for model building: Application to the Cox regression model. *Stat Med*, 11:2093–2109, 1992. pages 99, 150

[147] G. Schulgen, B. Lausen, J. Olsen, and M. Schumacher. Outcome-oriented cutpoints in quantitative exposure. *Am J Epi*, 120:172–184, 1994. pages 29, 32

[148] S. Senn. Change from baseline and analysis of covariance revisited. *Stat Med*, 25:4334–4344, 2006. pages 178, 179

[149] J. Shao. Linear model selection by cross-validation. *J Am Stat Assoc*, 88:486–494, 1993. pages 151

[150] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *J Comp Graph Stat*, 22(2):231–245, 2013. pages 52

[151] S. L. Simpson, L. J. Edwards, K. E. Muller, P. K. Sen, and M. A. Styner. A linear exponent AR(1) family of correlation structures. *Stat Med*, 29:1825–1838, 2010. pages 187

[152] 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. pages 100

[153] A. Spanos, F. E. Harrell, and D. T. Durack. Differential diagnosis of acute meningitis: An analysis of the predictive value of initial observations. *JAMA*, 262:2700–2707, 1989. pages 246, 249

[154] I. Spence and R. F. Garrison. A remarkable scatterplot. *Am Statistician*, 47:12–19, 1993. pages 124

[155] D. J. Spiegelhalter. Probabilistic prediction in patient management and clinical trials. *Stat Med*, 5:421–433, 1986. pages 103, 131, 154, 155

[156] E. W. Steyerberg. *Clinical Prediction Models*. Springer, New York, 2009. pages 2, 383

[157] E. W. Steyerberg, M. J. C. Eijkemans, F. E. Harrell, and Habbema. Prognostic modelling with logistic regression analysis: A comparison of selection and estimation methods in small data sets. *Stat Med*, 19:1059–1079, 2000. pages 51

[158] E. W. Steyerberg, M. J. C. Eijkemans, F. E. Harrell, and Habbema. Prognostic modeling with logistic regression analysis: In search of a sensible strategy in small data sets. *Med Decis Mak*, 21:45–56, 2001. pages 87

[159] C. J. Stone. Comment: Generalized additive models. *Statistical Sci*, 1:312–314, 1986. pages 43

[160] 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. pages 39, 43

[161] S. Suissa and L. Blais. Binary regression with continuous outcomes. *Stat Med*, 14:247–255, 1995. pages 29

[162] G. Sun, T. L. Shook, and G. L. Kay. Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis. *J Clin Epi*, 49:907–916, 1996. pages 101

[163] R. Tibshirani. Regression shrinkage and selection via the lasso. *J Roy Stat Soc B*, 58:267–288, 1996. pages 51

[164] T. Tjur. Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *Am Statistician*, 63(4):366–372, 2009. pages 238

[165] J. Twisk, M. de Boer, W. de Vente, and M. Heymans. Multiple imputation of missing values was not necessary before performing a longitudinal mixed-model analysis. *J Clin Epi*, 66(9):1022–1028, 2013. pages 67

[166] W. Vach and M. Blettner. Missing Data in Epidemiologic Studies. In *Ency of Biostatistics*, pages 2641–2654. Wiley, New York, 1998. pages 69

[167] S. van Buuren, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin. Fully conditional specification in multivariate imputation. *J Stat Computation Sim*, 76(12):1049–1064, 2006. pages 81, 82

[168] G. J. M. G. van der Heijden, Donders, T. Stijnen, and K. G. M. Moons. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. *J Clin Epi*, 59:1102–1109, 2006. pages 69

[169] T. van der Ploeg, P. C. Austin, and E. W. Steyerberg. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, 14(1):137+, Dec. 2014. pages 100

[170] J. C. van Houwelingen and S. le Cessie. Predictive value of statistical models. *Stat Med*, 9:1303–1325, 1990. pages 44, 103, 104, 151, 154, 156

[171] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, New York, fourth edition, 2003. pages 175, 291

[172] G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, New York, 2000. pages 175

[173] P. Verweij and H. C. van Houwelingen. Penalized likelihood in Cox regression. *Stat Med*, 13:2427–2436, 1994. pages 103

[174] A. J. Vickers. Decision analysis for the evaluation of diagnostic tests, prediction models, and molecular markers. *Am Statistician*, 62(4):314–320, 2008. pages 9

[175] E. Vittinghoff and C. E. McCulloch. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epi*, 165:710–718, 2006. pages 100

[176] P. T. von Hippel. Regression with missing ys: An improved strategy for analyzing multiple imputed data. *Soc Meth*, 37(1):83–117, 2007. pages 68

[177] H. Wainer. Finding what is not there through the unfortunate binning of results: The Mendel effect. *Chance*, 19(1):49–56, 2006. pages 29, 33

[178] S. H. Walker and D. B. Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54:167–178, 1967. pages 278

[179] H. Wang and C. Leng. Unified LASSO estimation by least squares approximation. *J Am Stat Assoc*, 102:1039–1048, 2007. pages 51

[180] S. Wang, B. Nan, N. Zhou, and J. Zhu. Hierarchically penalized Cox regression with grouped variables. *Biometrika*, 96(2):307–322, 2009. pages 52

[181] Y. Wax. Collinearity diagnosis for a relative risk regression analysis: An application to assessment of diet-cancer relationship in epidemiological studies. *Stat Med*, 11:1273–1287, 1992. pages 105

[182] 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. *Can J Phys Pharm*, 62:224–228, 1984. pages 247

[183] I. R. White and J. B. Carlin. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. *Stat Med*, 29:2920–2931, 2010. pages 81

[184] I. R. White and P. Royston. Imputing missing covariate values for the Cox model. *Stat Med*, 28:1982–1998, 2009. pages 67

[185] I. R. White, P. Royston, and A. M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med*, 30(4):377–399, 2011. pages 66, 75, 81, 84

[186] J. Whitehead. Sample size calculations for ordered categorical data. *Stat Med*, 12:2257–2271, 1993. See letter to editor SM 15:1065-6 for binary case;see errata in SM 13:871 1994;see kol95com, jul96sam. pages 101

[187] R. E. Wiegand. Performance of using multiple stepwise algorithms for variable selection. *Stat Med*, 29:1647–1659, 2010. pages 99

[188] D. M. Witten and R. Tibshirani. Testing significance of features by lassoed principal components. *Ann Appl Stat*, 2(3):986–1012, 2008. pages 52

[189] S. N. Wood. *Generalized Additive Models: An Introduction with R*. Chapman & Hall/CRC, Boca Raton, FL, 2006. ISBN 9781584884743. pages 53

[190] C. F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann Stat*, 14(4):1261–1350, 1986. pages 151

[191] S. Xiong. Some notes on the nonnegative garrote. *Technometrics*, 52(3):349–361, 2010. pages 52

[192] J. Ye. On measuring and correcting the effects of data mining and model selection. *J Am Stat Assoc*, 93:120–131, 1998. pages 17

[193] 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. pages 110

[194] R. M. Yucel and A. M. Zaslavsky. Using calibration to improve rounding in imputation. *Am Statistician*, 62(2):125–129, 2008. pages 82

[195] H. H. Zhang and W. Lu. Adaptive lasso for Cox's proportional hazards model. *Biometrika*, 94:691–703, 2007. pages 51

[196] H. Zhou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *J Comp Graph Stat*, 15:265–286, 2006. pages 52

[197] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J Roy Stat Soc B*, 67(2):301–320, 2005. pages 51

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

To obtain a 588-page book with detailed examples and case studies and notes on the theory and applications of survival analysis, logistic regression, and linear models, order REGRESSION MODELING STRATEGIES *with Applications to Linear Models, Logistic Regression, and Survival Analysis* by FE Harrell from Springer NY (2001). Steyerberg[156] and Dupont[51] are excellent texts for accompanying the book.

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