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

Questions/discussions: Slack channels `rms` and `bios330` in vbiostatcourse.slack.com
To be replaced by datamethods.org
Archive discussions: Google group `regmod`
General questions: stats.stackexchange.com, tag `regression-strategies`
Course notes: fharrell.com/doc/rms.pdf
Supplemental material: fharrell.com/doc/bbr.pdf:
Biostatistics for Biomedical Research

Blog: fharrell.com   Twitter: `@f2harrell` `#StatThink`

Sponsored by ASA Biometrics Section
Joint Statistical Meetings   Vancouver BC   2018-07-29

Contents

in the right margin indicates a hyperlink to a YouTube video related to the subject.

in the right margin is a hyperlink to an audio file elaborating on the notes. Red letters and numbers in the right margin are cues referred to within the audio recordings.

Rotated boxed blue text in the right margin at the start of a section represents the mnemonic key for linking to discussions about that section in vbiostatcourse.slack.com channel #rms. Anyone starting a new discussion about a topic related to the section should include the mnemonic somewhere in the posting, and the posting should be marked to slack as threaded. The mnemonic in the right margin is also a hyperlink to a search in the rms channel for messages containing the mnemonic. When you click on it the relevant messages will appear in the search results on the right side of the slack browser window. howto

Members of the slack group can also create submnemonics for subsections or other narrower-scope parts of the notes. When creating keys "on the fly," use names of the form chapterkey-sectionkey-yourkey where sectionkey is defined in the notes. That way a search on chapterkey-sectionkey will also bring up notes related to yourkey.

Several longer and more discussed subsections in the text have already been given short keys in these notes. For example, restricted cubic splines has the key genreg-rcs.

blog in the right margin is a link to a blog entry that further discusses the topic.

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 [140].

Chapter 1

Introduction

1.1 Hypothesis Testing, Estimation, and Prediction

intro-hep

A

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 B

- 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

intro-ex
C

- 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
- Checking that a previously developed summary index (e.g., BMI) adequately summarizes its component variables
- Developing new summary indexes by how variables predict an outcome
- 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

intro-classify

Blog: Classification vs. Prediction

D

- Many analysts desire to develop “classifiers” instead of predictions
- Outside of, for example, visual or sound pattern recognition, classification represents a premature decision
- See this blog for details
- 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

- Classification and its associated classification accuracy measure—the proportion classified "correctly"—are very sensitive to the relative frequencies of the outcome variable. If a classifier is applied to another dataset with a different outcome prevalence, the classifier may no longer apply.

- 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 [53]

- 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 [153, 20, 56, 16, 50, 60].

[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 $\mathrm{Prob}(Y = 1|X = x)$; we don't care about $\mathrm{Prob}(Y = 1|X > c)$, and an individual having $X = x$ would be quite puzzled if she were given $\mathrm{Prob}(X > c|\text{future unknown Y})$ 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.

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

In summary: E

- Classification is a forced choice — a decision.
- 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

F

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: G

- Are you going faster than 70?

An answer from a better dichotomizer: H

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

Better: I

- 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:

J

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

1.4 Planning for Modeling

intro-plan

- Chance that predictive model will be used [126] K
- 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: L

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? M

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

Iezzoni [82] lists these dimensions to capture, for patient outcome studies: N

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: O

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

intro-choice P

- 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

intro-uncertainty

- 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 [165]: "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 [102] for an approach involving adding noise to Y to improve variable selection

Chapter 2

General Aspects of Fitting Regression Models

Regression modeling meets many analytic needs:

genreg-intro

- Prediction, capitalizing on efficient estimation methods such as maximum likelihood and the predominant additivity in a variety of problems
  - E.g.: effects of age, smoking, and air quality add to predict lung capacity
  - When effects are predominantly additive, or when there aren't too many interactions and one knows the likely interacting variables in advance, regression can beat machine learning techniques that assume interaction effects are likely to be as strong as main effects
- Separate effects of variables (especially exposure and treatment)

- Hypothesis testing
- Deep understanding of uncertainties associated with all model components
  - Simplest example: confidence interval for the slope of a predictor
  - Confidence intervals for predicted values; simultaneous confidence intervals for a series of predicted values
    - E.g.: confidence band for Y over a series of values of X

Alternative: Stratification

- Cross-classify subjects on the basis of the Xs, estimate a property of Y for each stratum
- Only handles a small number of Xs
- Does not handle continuous X

Alternative: Single Trees (recursive partitioning/CART)

- Interpretable because they are over-simplified and usually wrong
- Cannot separate effects

- Finds spurious interactions
- Require huge sample size
- Do not handle continuous X effectively; results in very heterogeneous nodes because of incomplete conditioning
- Tree structure is unstable so insights are fragile

Alternative: Machine Learning

- E.g. random forests, bagging, boosting, support vector machines, neural networks, deep learning
- Allows for high-order interactions and does not require prespecification of interaction terms
- Almost automatic; can save analyst time and do the analysis in one step (long computing time)
- Uninterpretable black box
- Effects of individual predictors are not separable
- Interaction effects (e.g., differential treatment effect = precision medicine = personalized medicine) not available
- Because of not using prior information about dominance of additivity, can require 200 events per candidate predictor

when Y is binary [122]

- Logistic regression may require 20 events per candidate predictor
- Can create a demand for "big data" where additive statistical models can work on moderate-size data
- See this article in *Harvard Business Review* for more about regression vs. complex methods

2.1 Notation for Multivariable Regression Models

genreg-notation

A

- 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 B
$C(Y|X)$: 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

genreg-modform

General regression model

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

General linear regression model

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

Examples C

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

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

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

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

2.3 Interpreting Model Parameters

genreg-interp

Suppose that X_j is linear and doesn't interact with other X's[a]. E

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

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

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

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

where

$$X_1 = 1 \text{ if } T = K, \; 0 \text{ otherwise}$$

[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).

$$X_2 = 1 \text{ if } T = L, \; 0 \text{ otherwise}$$
$$X_3 = 1 \text{ if } T = M, \; 0 \text{ otherwise.}$$

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

G

genreg-interp-interact

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

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

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

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

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

2.3.3

Example: Inference for a Simple Model

Postulate 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 H

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

Interpretations of parameters: I

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

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

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

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

Hypotheses that are usually inappropriate: J

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 K

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

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

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

L

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.

2.3.4 Review of Composite (Chunk) Tests

genreg-chunk

- In the model

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

M

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)

2.4 Relaxing Linearity Assumption for Continuous Predictors

genreg-nonlinear

2.4.1 Avoiding Categorization

genreg-cat

Natura non facit saltus
(Nature does not make jumps)

Gottfried Wilhelm Leibniz

- Relationships seldom linear except when predicting one variable from itself measured earlier N
- Categorizing continuous predictors into intervals is a disaster; see references
  [128, 1, 78, 92, 3, 8, 51, 125, 145, 22]
  [105, 131, 4, 80, 108, 155, 53, 58, 32, 12] and Biostatistics for Biomedical Research, Chapter 18.
- Some problems caused by this approach:
  1. Estimated values have reduced precision, and associated tests have reduced power
  2. Categorization assumes relationship between predictor and response is flat within intervals; far less reasonable than a linearity assumption in most cases
  3. To make a continuous predictor be more accurately modeled when categorization is used, multiple intervals are required

4. Because of sample size limitations in the very low and very high range of the variable, the outer intervals (e.g., outer quintiles) will be wide, resulting in significant heterogeneity of subjects within those intervals, and residual confounding
5. Categorization assumes that there is a discontinuity in response as interval boundaries are crossed. Other than the effect of time (e.g., an instant stock price drop after bad news), there are very few examples in which such discontinuities have been shown to exist.
6. Categorization only seems to yield interpretable estimates. E.g. odds ratio for stroke for persons with a systolic blood pressure > 160 mmHg compared to persons with a blood pressure ≤ 160 mmHg $\rightarrow$ interpretation of OR depends on distribution of blood pressures in the sample (the proportion of subjects > 170, > 180, etc.). If blood pressure is modeled as a continuous variable (e.g., using a regression spline, quadratic, or linear effect) one can estimate the ratio of odds for *exact* settings of the predictor, e.g., the odds ratio for 200 mmHg compared to 120 mmHg.
7. Categorization does not condition on full information. When, for example, the risk of stroke is being assessed for a new subject with a known blood pressure (say 162 mmHg), the subject does not report to her physician "my blood pressure exceeds 160" but rather reports 162 mmHg. The risk for this subject will be much lower than that of a subject with a blood pressure of 200 mmHg.

8. If cutpoints are determined in a way that is not blinded to the response variable, calculation of P-values and confidence intervals requires special simulation techniques; ordinary inferential methods are completely invalid. E.g.: cutpoints chosen by trial and error utilizing Y, even informally $\rightarrow$ P-values too small and CLs not accurate[c].
9. Categorization not blinded to Y $\rightarrow$ biased effect estimates [3, 131]
10. "Optimal" cutpoints do not replicate over studies. Hollander *et al.* [80] 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 recent update of the American Society of Clinical Oncology." Giannoni *et al.* [58] 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

[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 [80].

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.

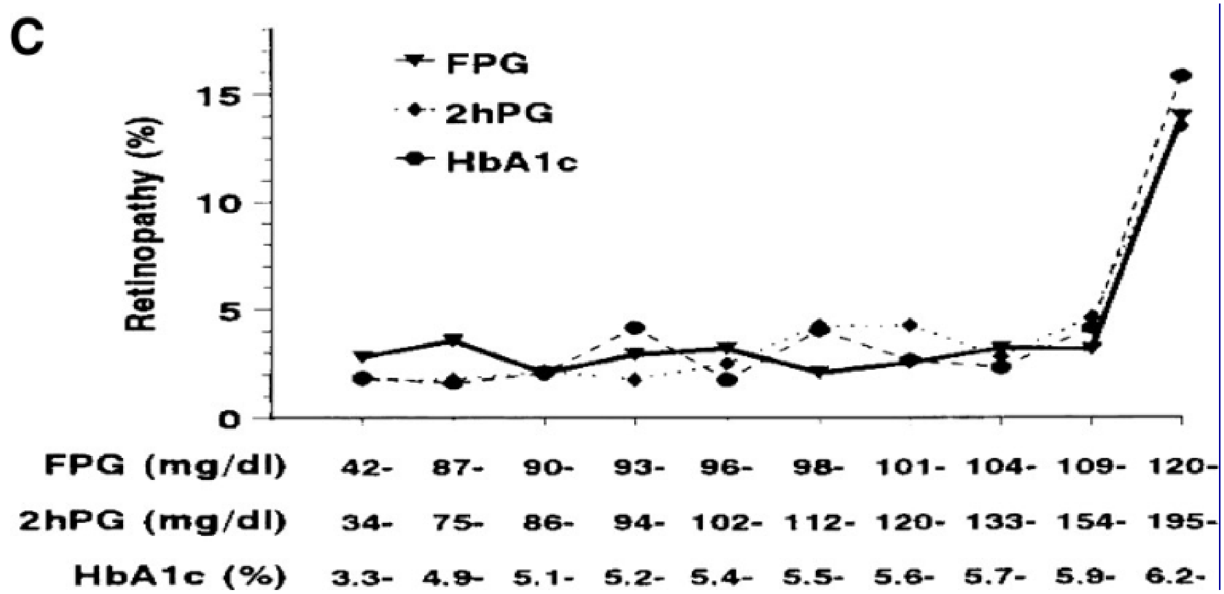
12. Cutpoints are arbitrary and manipulatable; cutpoints can be found that can result in both positive and negative associations [155].
13. If a confounder is adjusted for by categorization, there will be residual confounding that can be explained away by inclusion of the continuous form of the predictor in the model in addition to the categories.

- To summarize: The use of a (single) cutpoint c makes many assumptions, including: o
  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 (must run in `RStudio`)

```
require(Hmisc)
getRs('catgNoise.r')
```

Example[d] of misleading results from creating intervals (here, deciles) of a continuous predictor. Final interval is extremely heterogeneous and is greatly influenced by very large glycohemoglobin values, creating the false impression of an inflection point at 5.9.



See this for excellent graphical examples of the harm of categorizing predictors, especially when using quantile groups.

2.4.2

Simple Nonlinear Terms

genreg-poly

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

P

- H_0 : model is linear in X_1 vs. H_a : model is quadratic in $X_1 \equiv H_0 : \beta_2 = 0$.

[d] From NHANES III; Diabetes Care 32:1327-34; 2009 adapted from Diabetes Care 20:1183-1197; 1997.

- 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.

2.4.3

Splines for Estimating Shape of Regression Function and Determining Predictor Transformations

genreg-spline

Q

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

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

R

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

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

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

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

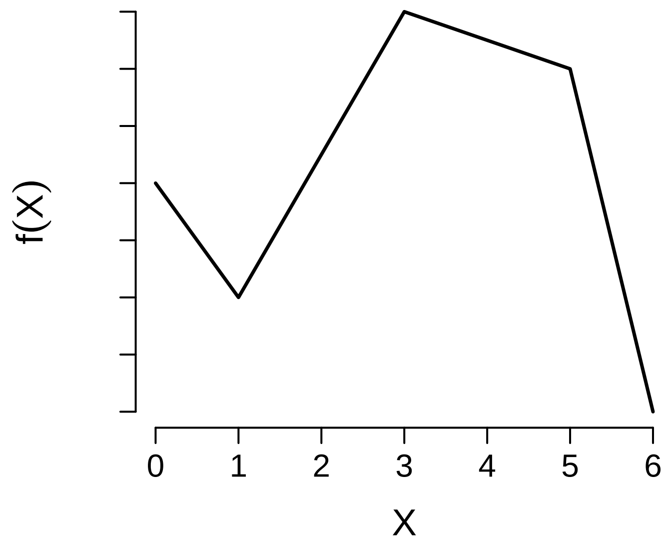


Figure 2.1: A linear spline function with knots at $a = 1, b = 3, c = 5$.

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

2.4.4

Cubic Spline Functions

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

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

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

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

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

2.4.5 Restricted Cubic Splines

genreg-rcs

Stone and Koo [144]: cubic splines poorly behaved in tails. Constrain function to be linear in tails.
$k+3 \rightarrow k-1$ parameters [43].

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 [43]

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

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)
```

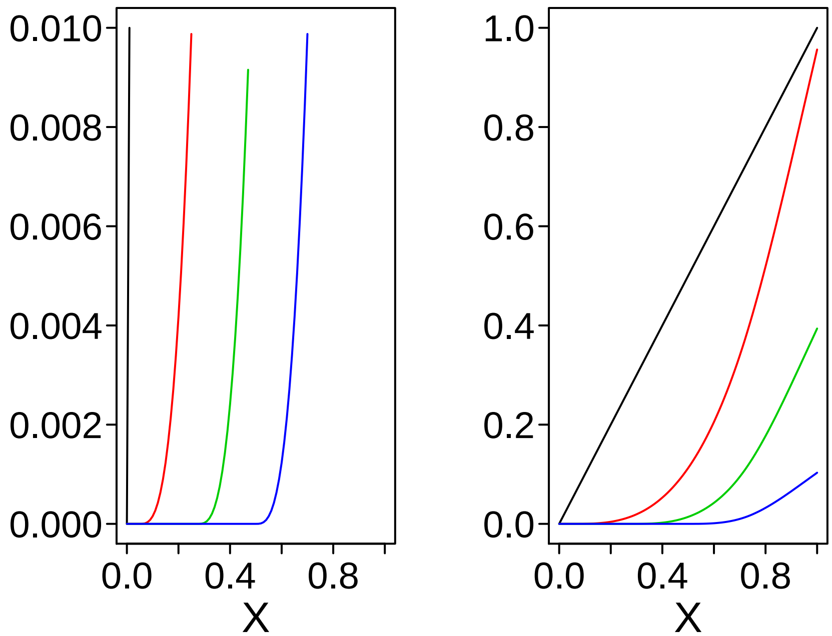


Figure 2.2: Restricted cubic spline component variables for $k = 5$ and knots at $X = .05, .275, .5, .725$, and $.95$. Nonlinear basis functions are scaled by τ. The left panel is a y–magnification of the right panel. Fitted functions such as those in Figure 2.3 will be linear combinations of these basis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

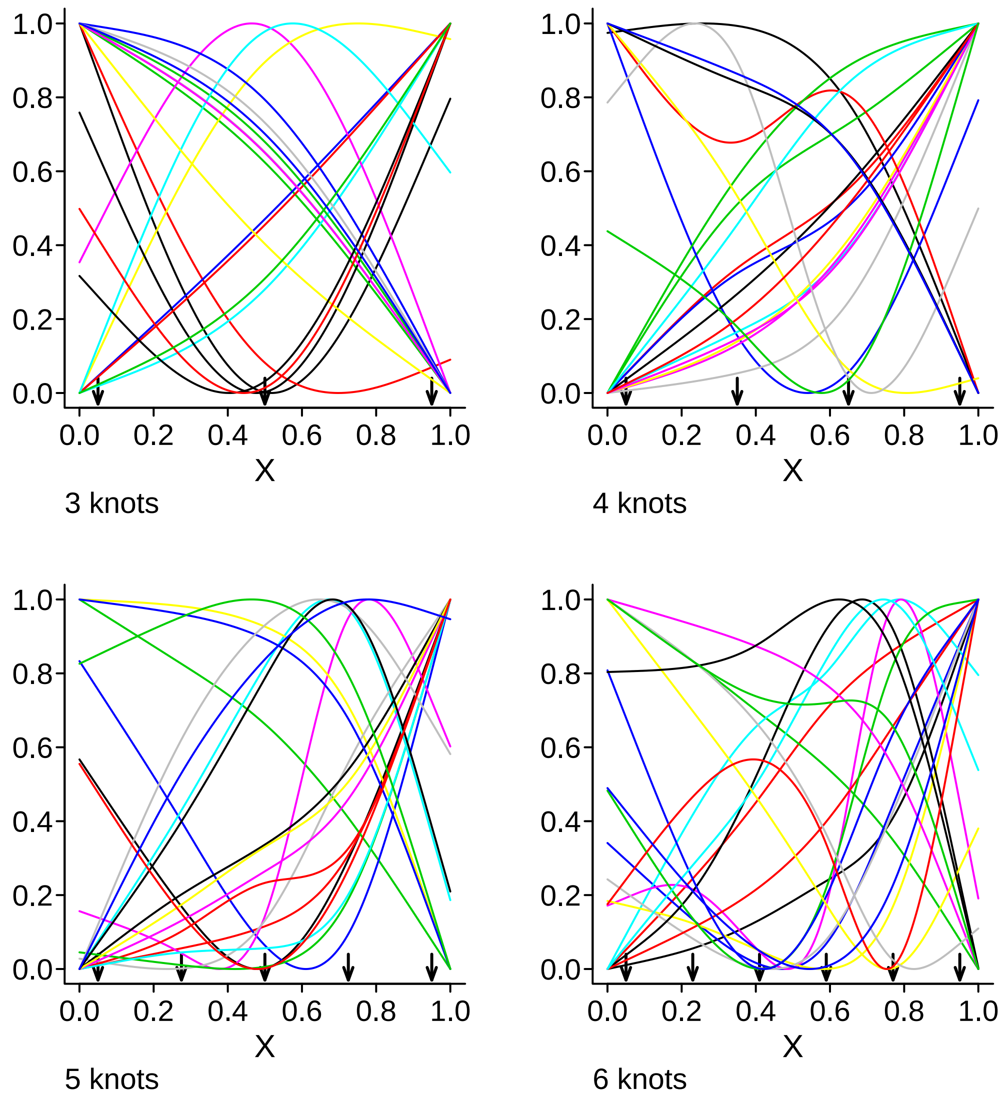


Figure 2.3: Some typical restricted cubic spline functions for $k = 3, 4, 5, 6$. The y–axis is $X\beta$. Arrows indicate knots. These curves were derived by randomly choosing values of β subject to standard deviations of fitted functions being normalized.

Interactive demonstration of linear and cubic spline fitting, plus ordinary 4^{th} order polynomial. This can be run with `RStudio` or in an ordinary R session.

```
require(Hmisc)
getRs('demoSpline.r')                  # if using RStudio
getRs('demoSpline.r', put='source')    # if not
```

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

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

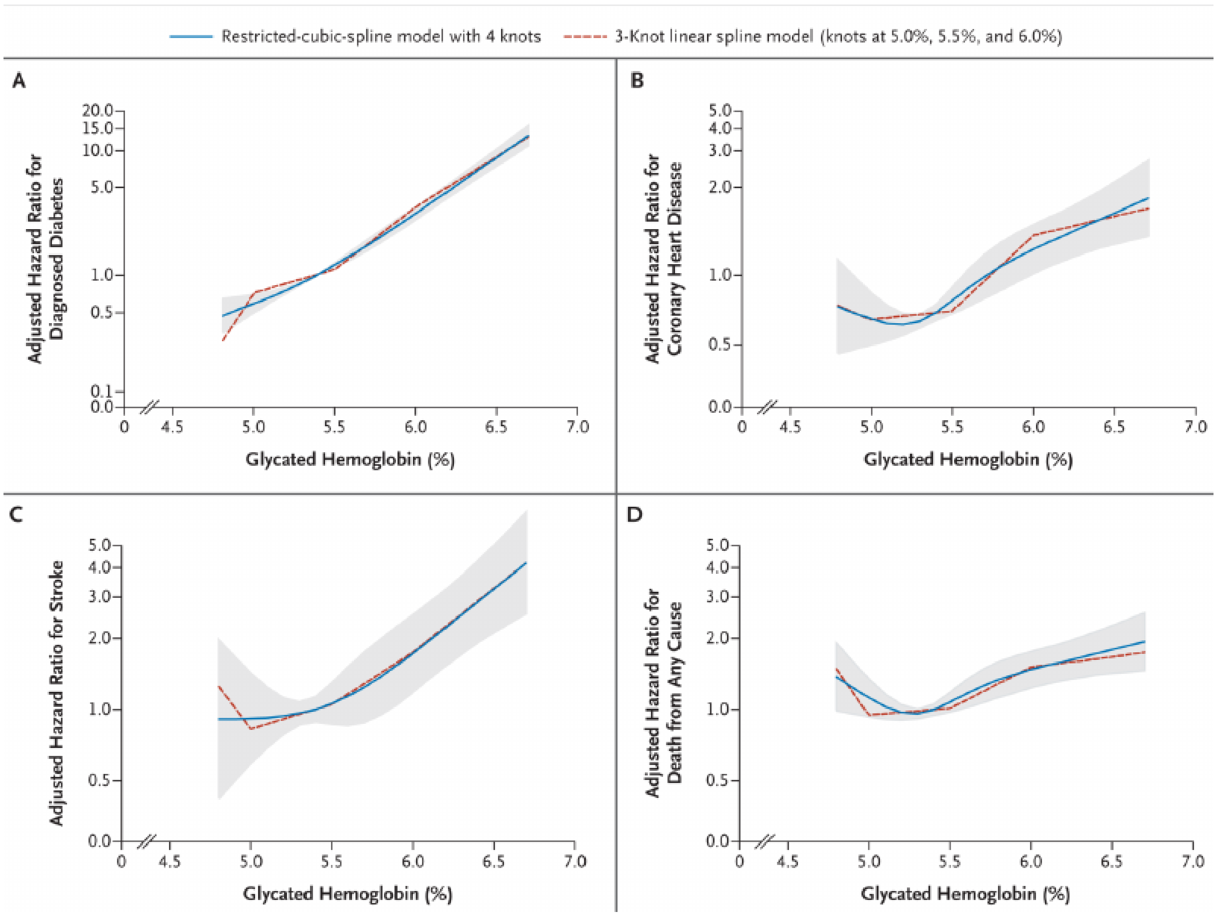
by dividing $\beta_2, \ldots, \beta_{k-1}$ by τ and computing

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

A test of linearity in X can be obtained by testing

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

Example: [132]



2.4.6 Choosing Number and Position of Knots

W

- Knots are specified in advance in regression splines
- Locations not important in most situations [143, 46]
- 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 [144].

Choice of k: x

- Flexibility of fit vs. n and variance
- Usually $k = 3, 4, 5$. Often $k = 4$
- Large n (e.g. $n \geq 100$) – $k = 5$
- Small n (< 30, say) – $k = 3$
- Can use Akaike's information criterion (AIC) [5, 149] to choose k
- This chooses k to maximize model likelihood ratio $\chi^2 - 2k$.

See [62] for a comparison of restricted cubic splines, fractional polynomials, and penalized splines.

2.4.7

Nonparametric Regression

genreg-nonpar

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

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

- Another alternative: smoothing splines[f]

- Smoothers are very useful for estimating trends in residual plots

2.4.8

Advantages of Regression Splines over Other Methods

Regression splines have several advantages [71]: B

- 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

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

burdensome computations.

2.5 Recursive Partitioning: Tree-Based Models

genreg-rpart

Breiman, Friedman, Olshen, and Stone [19]: CART (Classification and Regression Trees) — essentially model-free

Method: C

- 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: D

- 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 [33]; missing value imputation)

See [7].

2.5.1 New Directions in Predictive Modeling

genreg-newdir

The approaches recommended in this course are

E

- 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. F

- lasso (shrinkage using L1 norm favoring zero regression coefficients) [147, 142]
- elastic net (combination of L1 and L2 norms that handles the $p > n$ case better than the lasso) [169]
- adaptive lasso [167, 156]
- more flexible lasso to differentially penalize for variable selection and for regression coefficient estimation [124]
- 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 [157]
- sparse-group lasso using L1 and L2 norms to achieve spareness on groups and within groups of variables [135]
- adaptive group lasso (Wang & Leng)
- Breiman's nonnegative garrote [164]
- "preconditioning", i.e., model simplification after developing a "black box" predictive model [113, 112]
- sparse principal components analysis to achieve parsimony in data reduction [161, 168, 97, 96]
- bagging, boosting, and random forests [77]

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. G

When data reduction is not required, generalized additive models [76, 162] should also be considered.

2.5.2

Choosing Between Machine Learning and Statistical Modeling

- Statistical models allow for complexity (nonlinearity, interaction)
- Easy to allow every predictor to have nonlinear effect
- Easy to handle unlimited numbers of candidate predictors if assume additivity (e.g., using ridge regression, lasso, elastic net)
- Interactions should be pre-specified
- Machine learning is gaining attention but is oversold in some settings blog
- Researchers are under the mistaken impression that machine learning can be used on small samples blog

Considerations in Choosing One Approach over Another

A statistical model may be the better choice if

- Uncertainty is inherent and the signal:noise ratio is not large—

even with identical twins, one twin may get colon cancer and the other not; model tendencies instead of doing classification

- One could never have perfect training data, e.g., cannot repeatedly test one subject and have outcomes assessed without error
- One wants to isolate effects of a small number of variables
- Uncertainty in an overall prediction or the effect of a predictor is sought
- Additivity is the dominant way that predictors affect the outcome, or interactions are relatively small in number and can be pre-specified
- The sample size isn't huge
- One wants to isolate (with a predominantly additive effect) the effects of "special" variables such as treatment or a risk factor

Machine learning may be the better choice if

- The signal:noise ratio is large and the outcome being predicted doesn't have a strong component of randomness; e.g., in visual pattern recognition an object must be an "E" or not an "E"

- The learning algorithm can be trained on an unlimited number of exact replications (e.g., 1000 repetitions of each letter in the alphabet or of a certain word to be translated to German)

- Overall prediction is the goal, without being able to succinctly describe the impact of any one variable (e.g., treatment)

- One is not very interested in estimating uncertainty in forecasts or in effects of select predictors

- Non-additivity is expected to be strong and can't be isolated to a few pre-specified variables (e.g., in visual pattern recognition the letter "L" must have both a dominating vertical component and a dominating horizontal component)

- The sample size is huge [122]

- One does not need to isolate the effect of a special variable such as treatment

2.6 Multiple Degree of Freedom Tests of Association

genreg-multidf

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

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

In the 5-knot restricted cubic spline model

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

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

- 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 [63] elegantly described the hazards of pretesting J

- 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

2.7 Assessment of Model Fit

2.7.1 Regression Assumptions

genreg-gof

The general linear regression model is K

$$C(Y|X) = X\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k.$$

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.

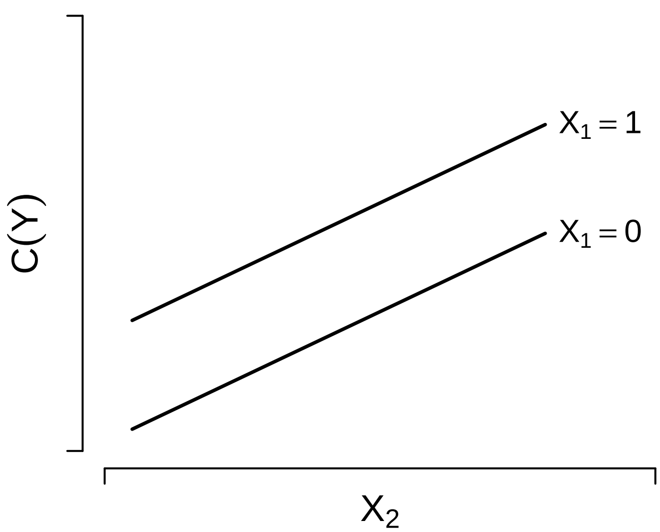


Figure 2.4: Regression assumptions for one binary and one continuous predictor

Methods for checking fit: L

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 M
   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 N
   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. O

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. P

- 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

Q

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

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

2.7.2

Modeling and Testing Complex Interactions

genreg-interact

X_1 binary or linear, X_2 continuous:

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

R

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

- 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 $X_2 g(X_1)$:

T

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

U

- 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: V

- Cover $X_1 \times X_2$ plane with grid and fit patch-wise cubic polynomial in two variables

- Restrict to be of form $aX_1 + bX_2 + cX_1X_2$ in corners

- Uses all $(k-1)^2$ cross-products of restricted cubic spline terms

- See Gray [64, 65, Section 3.2] for penalized splines allowing control of effective degrees of freedom. See Berhane *et al.* [13] for a good discussion of tensor splines.

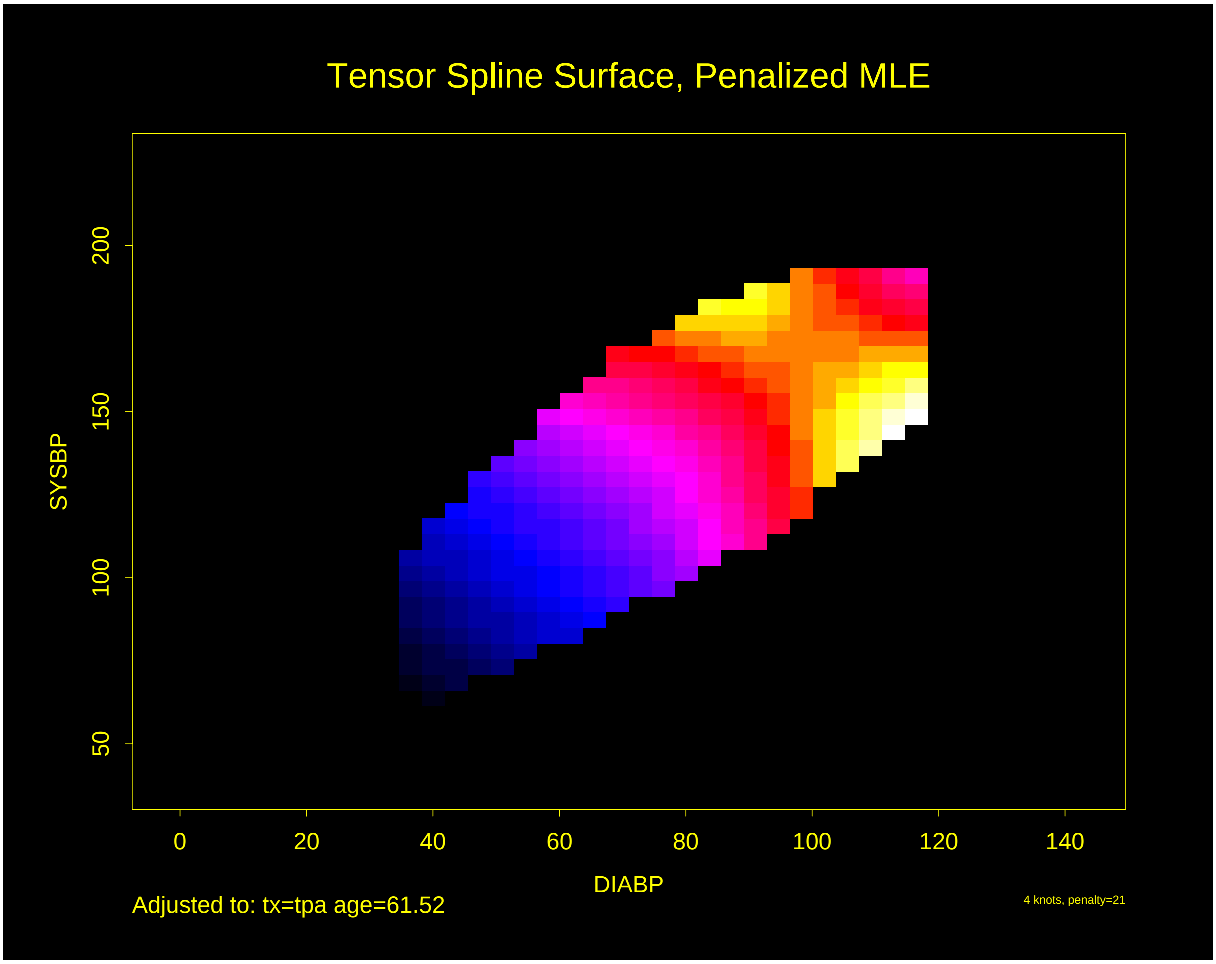


Figure 2.5: Logistic regression estimate of probability of a hemorrhagic stroke for patients in the GUSTO-I trial given t-PA, using a tensor spline of two restricted cubic splines and penalization (shrinkage). Dark (cold color) regions are low risk, and bright (hot) regions are higher risk.

Figure 2.5 is particularly interesting because the literature had suggested (based on approximately 24 strokes) that pulse pressure was the main cause of hemorrhagic stroke whereas this flexible modeling approach (based on approximately 230 strokes) suggests that mean arterial blood pressure (roughly a 45° line) is what is most important over a broad range of blood pressures. At the far right one can see that pulse pressure (axis perpendicular to 45° line) may have an impact although a non-monotonic one.

Other issues: W

- Y non-censored (especially continuous) $\rightarrow$ multi-dimensional scatterplot smoother [24]
- 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: X

- 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

- Measurement $\times$ amount of deterioration of the measurement

The section between the two horizontal blue lines was inserted after the audio narration was recorded.

The last example is worth expanding as an example in model formulation. Consider the following study.

- A sample of patients seen over several years have a blood sample taken at time of hospitalization
- Blood samples are frozen
- Long after the last patient was sampled, the blood samples are thawed all in the same week and a blood analysis is done
- It is known that the quality of the blood analysis deteriorates roughly logarithmically by the age of the sample; blood measurements made on old samples are assumed to be less predictive of outcome
- This is reflected in an interaction between a function of sample age and the blood measurement `B`[g]
- Patients were followed for an event, and the outcome variable of interest is the time from hospitalization to that event
- To not assume a perfect logarithmic relationship for sample

[g]For continuous Y one might need to model the residual variance of Y as increasing with sample age, in addition to modeling the mean function.

age on the effect of the blood measurement, a restricted cubic spline model with 3 default knots will be fitted for log sample age

- Sample age is assumed to not modify the effects of non-blood predictors patient `age` and `sex`

- Model may be specified the following way using the R `rms` package to fit a Cox proportional hazards model

- Test for nonlinearity of `sampleAge` tests the adequacy of assuming a plain logarithmic trend in sample age

```
f ← cph(Surv(etime, event) ~ rcs(log(sampleAge), 3) * rcs(B, 4) +
          rcs(age, 5) * sex, data=mydata)
```

The `B` $\times$ `sampleAge` interaction effects have 6 d.f. and tests whether the sample deterioration affects the effect of `B`. By not assuming that `B` has the same effect for old samples as for young samples, the investigator will be able to estimate the effect of `B` on outcome when the blood analysis is ideal by inserting `sampleAge` $= 1$ day when requesting predicted values as a function of `B`.

2.7.3 Fitting Ordinal Predictors

Y

- 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

2.7.4

Distributional Assumptions

Z

- 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 [36]):
  $C(Y|X) = C(Y = y|X) = d(y) + X\beta$
  $C =$ log hazard
- Check form of $d(y)$
- Show $d(y)$ does not interact with X

Chapter 3

Multivariable Modeling Strategies

A

- "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 B
- Decide where to spend them
- Spend them
- General references: [111, 141, 74, 59]

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

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

3.1 Prespecification of Predictor Complexity Without Later Simplification

strategy-complexity

D

- 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

Motivating examples:

```
# Overfitting a flat relationship
require(rms)
```

```
set.seed(1)
x ← runif(1000)
y ← runif(1000, -0.5, 0.5)
dd ← datadist(x, y); options(datadist='dd')
par(mfrow=c(2,2), mar=c(2, 2, 3, 0.5))
pp ← function(actual) {
  yhat  ← predict(f, data.frame(x=xs))
```

```
  yreal ← actual(xs)
  plot(0, 0, xlim=c(0,1),
       ylim=range(c(quantile(y, c(0.1, 0.9)), yhat,
                    yreal)),
       type='n', axes=FALSE)
  axis(1, labels=FALSE); axis(2, labels=FALSE)
  lines(xs, yreal)
  lines(xs, yhat, col='blue')
}
f ← ols(y ~ rcs(x, 5))
xs ← seq(0, 1, length=150)
pp(function(x) 0*x)
title('Mild Error:\nOverfitting a Flat Relationship',
      cex=0.5)
y ← x + runif(1000, -0.5, 0.5)
f ← ols(y ~ rcs(x, 5))
pp(function(x) x)
title('Mild Error:\nOverfitting a Linear Relationship',
      cex=0.5)
y ← x^4 + runif(1000, -1, 1)
f ← ols(y ~ x)
pp(function(x) x^4)
title('Serious Error:\nUnderfitting a Steep Relationship',
      cex=0.5)
y ← - (x - 0.5) ^ 2 + runif(1000, -0.2, 0.2)
f ← ols(y ~ x)
pp(function(x) - (x - 0.5) ^ 2)
title('Tragic Error:\nMonotonic Fit to\nNon-Monotonic Relationship',
      cex=0.5)
```

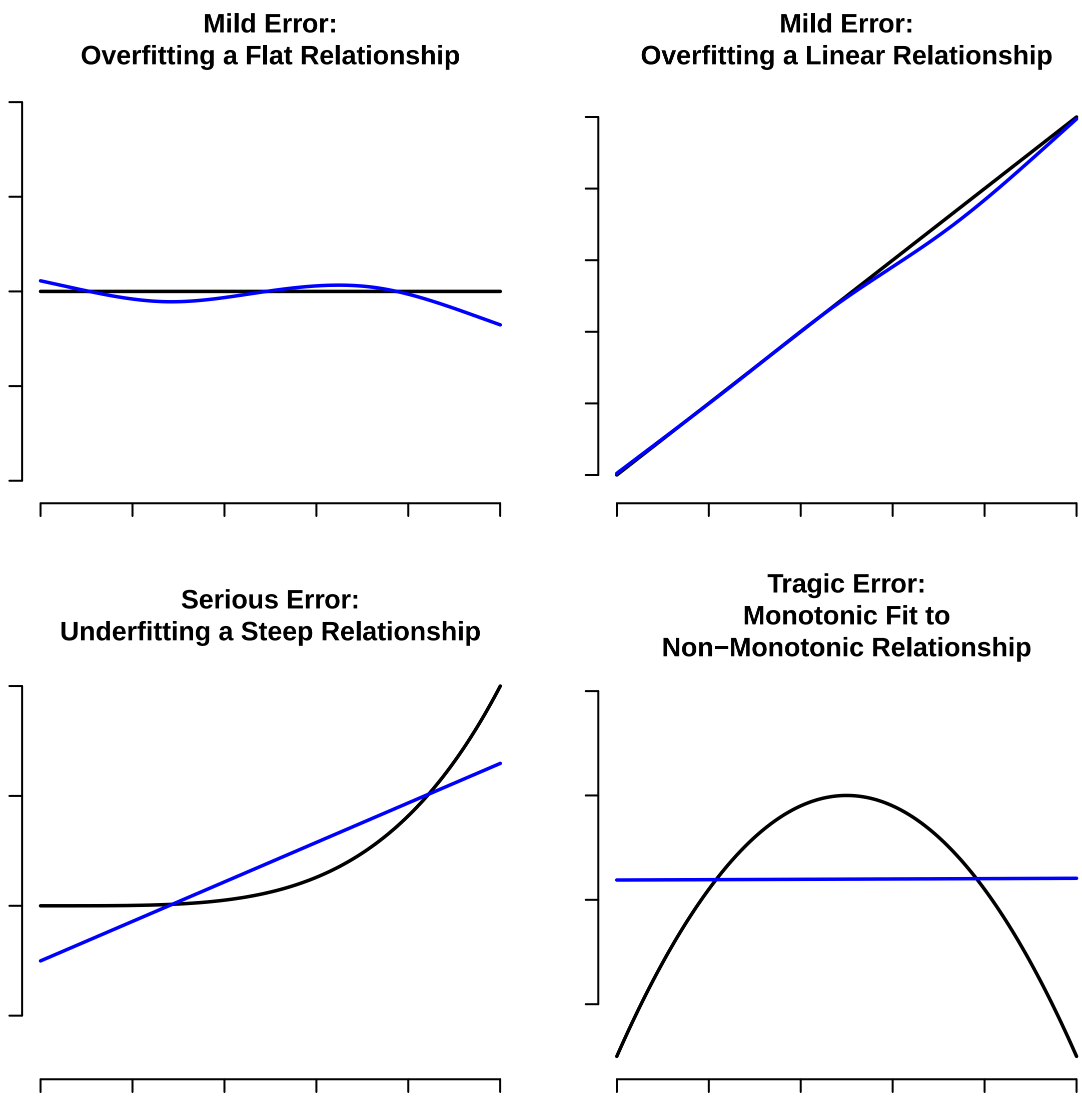


Table 3.1: Examples of Reducing the Number of Parameters

Categorical predictor with k levels	Collapse less frequent categories into "other"
Continuous predictor represented as k-knot r.c. spline	Reduce k to a number as low as 3, or 0 (linear)

E

3.1.1 Learning From a Saturated Model

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

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

- Let all categorical predictors retain their original categories except for pooling of very low prevalence categories (e.g., ones containing < 6 observations).

- Fit this general main effects model.

- Compute the partial χ^2 statistic for testing the association of each predictor with the response, adjusted for all other predictors. In the case of ordinary regression convert partial F statistics to χ^2 statistics or partial R^2 values.

- Make corrections for chance associations to "level the playing field" for predictors having greatly varying d.f., e.g., subtract the d.f. from the partial χ^2 (the expected value of χ^2_p is p under H_0).

- Make certain that tests of nonlinearity are not revealed as this would bias the analyst.

- Sort the partial association statistics in descending order.

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

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

3.1.2

Using Marginal Generalized Rank Correlations

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

- 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

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

ρ^2 is a fair procedure because H

- 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 simpler to more complex models but not the other way round

3.2 Checking Assumptions of Multiple Predictors Simultaneously

strategy-simult

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

3.3 Variable Selection

strategy-selection

J

- Series of potential predictors with no prior knowledge
- $\uparrow$ exploration $\rightarrow \uparrow$ shrinkage (overfitting)
- Summary of problem: $E(\hat{\beta}|\hat{\beta}$ "significant" $) \neq \beta$ [26]
- Biased R^2, $\hat{\beta}$, standard errors, P-values too small
- F and χ^2 statistics do not have the claimed distribution[b] [63]
- Will result in residual confounding if use variable selection to find confounders [67]
- Derksen and Keselman [42] 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: K
  1. "The degree of correlation between the predictor variables affected the frequency with which authentic predictor variables found their way into the final model.
  2. The number of candidate predictor variables affected the number of noise variables that gained entry to the model.

[b] Lockhart *et al.* [101] 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.

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. L

```
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  0.80  0.74 0.00 0.00
x2  0.89  1.00 -0.01 0.00  0.71  0.83 0.00 0.00
x3 -0.01 -0.01  1.00 0.56 -0.01 -0.01 0.74 0.73
x4  0.00  0.00  0.56 1.00  0.00  0.00 0.88 0.86
x5  0.80  0.71 -0.01 0.00  1.00  0.59 0.00 0.00
x6  0.74  0.83 -0.01 0.00  0.59  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. M

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

```
        n     vratio        gdf apparentdf  rmse.full  rmse.step
    20.00       0.70       8.09       3.65       1.62       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   rmse.step
     40.00        0.87         7.34        3.06        1.21        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   rmse.step
    150.00        0.97         9.08        3.81        0.59        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   rmse.step
    300.00        0.98         9.26        4.21        0.43        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   rmse.step
   2000.00        1.00         6.30        4.58        0.17        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 [70]:

N

- Forward selection, backward elimination
- Stopping rule: "residual χ^2" with d.f. $=$ no. candidates remaining at current step
- Test for significance or use Akaike's information criterion (AIC [5]), here $\chi^2 - 2 \times d.f.$
- Better to use subject matter knowledge!
- No currently available stopping rule was developed for stepwise, only for comparing a limited number of pre-specified models [18, Section 1.3]
- Roecker [127] studied forward selection (FS), all possible subsets selection (APS), full fits o
- APS more likely to select smaller, less accurate models than FS
- Neither as accurate as full model fit unless $> \frac{1}{2}$ candidate variables redundant or unnecessary
- Step-down is usually better than forward [103] and can be used efficiently with maximum likelihood estimation [93]
- Fruitless to try different stepwise methods to look for agreement [160]

- Bootstrap can help decide between full and reduced model P

- Full model fits gives meaningful confidence intervals with standard formulas, C.I. after stepwise does not [2, 81, 18]

- Data reduction (grouping variables) can help

- Using the bootstrap to select important variables for inclusion in the final model [130] is problematic [6]

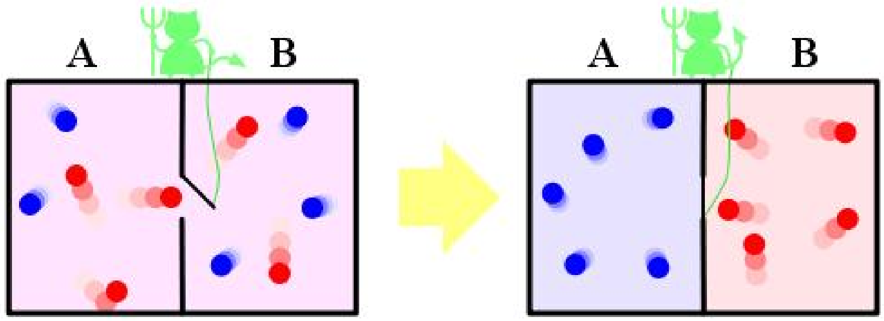
- It is not logical that a population regression coefficient would be exactly zero just because its estimate was "insignificant"

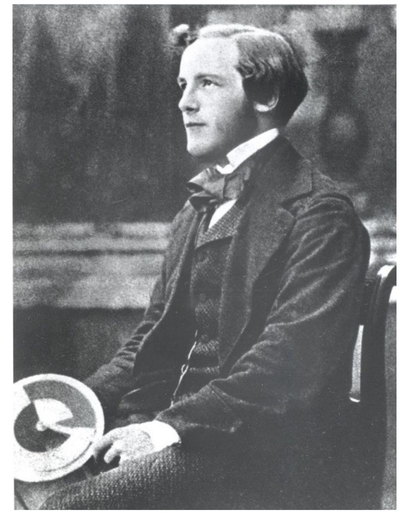
3.3.1 Maxwell's Demon as an Analogy to Variable Selection

Some of the information in the data is spent on variable selection instead of using all information for estimation.

Model *specification* is preferred to model *selection*.

Information content of the data usually insufficient for reliable variable selection.





James Clerk Maxwell

Maxwell imagines one container divided into two parts, A and B. Both parts are filled with the same gas at equal temperatures and placed next to each other. Observing the molecules on both sides, an imaginary demon guards a trapdoor between the two parts. When a faster-than-average molecule from A flies towards the trapdoor, the demon opens it, and the molecule will fly from A to B. Likewise, when a slower-than-average molecule from B flies towards the trapdoor, the demon will let it pass from B to A. The average speed of the molecules in B will have increased while in A they will have slowed down on average. Since average molecular speed corresponds to temperature, the temperature decreases in A and increases in B, contrary to the second law of thermodynamics.

Szilárd pointed out that a real-life Maxwell's demon would need to have some means of measuring molecular speed, and that the act of acquiring information would require an expenditure of energy. Since the demon and the gas are interacting, we must consider the total entropy of the gas and the demon combined. The expenditure of energy by the demon will cause an increase in the entropy of the demon, which will be larger than the lowering of the entropy of the gas.

Source: `commons.wikimedia.org/wiki/File:YoungJamesClerkMaxwell.jpg`

`en.wikipedia.org/wiki/Maxwell's_demon`

Peter Ellis' blog article contains excellent examples of issues discussed here but applied to time series modeling.

3.4 Overfitting and Limits on Number of Predictors

strategy-limits

Q

- 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 $< \frac{m}{15}$ [72, 73, 137, 115, 114, 154, 122]
- $p =$ 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
- Ignores sample size needed just to estimate the intercept or, in semiparametric models, the underlying distribution function[c]

[c] The sample size needed for these is model-dependent

- To just estimate σ in a linear model with a multiplicative margin of error of 1.2 with 0.95 confidence requires $n = 70$

R

Table 3.2: Limiting Sample Sizes for Various Response Variables

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

[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 [159, Eq. 10, 15]. Here n_1 and n_2 are the marginal frequencies of the two response levels [114].

[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) [159, Eq, 3]. If all cell sizes are equal, the relative efficiency of having k response categories compared to a continuous response is $1 - \frac{1}{k^2}$ [159, Eq. 14], e.g., a 5-level response is almost as efficient as a continuous one if proportional odds holds across category cutoffs.

[c] This is approximate, as the effective sample size may sometimes be boosted somewhat by censored observations, especially for non-proportional hazards methods such as Wilcoxon-type tests [11]. S

- 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 [146]

3.5 Shrinkage

strategy-shrinkage

T

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

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

- Prevent shrinkage by using pre-shrinkage

U

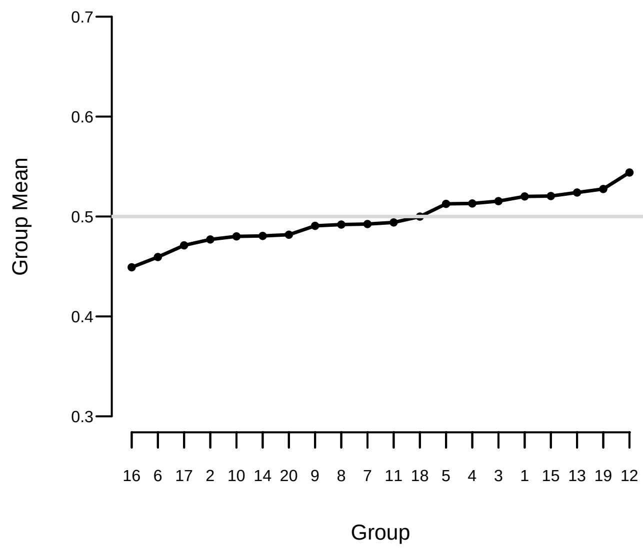


Figure 3.1: Sorted means from 20 samples of size 50 from a uniform $[0,1]$ distribution. The reference line at 0.5 depicts the true population value of all of the means.

- Spiegelhalter [139]: 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 [35]

- Ridge regression [94, 149]

- Penalized MLE [152, 64, 75]

- Heuristic shrinkage parameter of van Houwelingen and le Cessie [149, Eq. 77] v

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

- OLS[d]: $\hat{\gamma} = \frac{n-p-1}{n-1} R^2_{\text{adj}} / R^2$
  $R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$
- p close to no. candidate variables
- Copas [35, Eq. 8.5] adds 2 to numerator

[d]An excellent discussion about such indexes may be found here.

3.6 Collinearity

strategy-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 [27, Chap. 9]
- 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 [109, 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 [40, 158]
- See Belsley [10, pp. 28-30] for problems with VIF
- Easy approach: SAS `VARCLUS` procedure [129], R `varclus` function, other clustering techniques: group highly correlated variables Y
- Can score each group (e.g., first principal component, PC_1 [39]); summary scores not collinear

3.7

Data Reduction

strategy-reduction

Z

- 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.
- Data reduction is **completely masked to** Y, which is precisely why it does not distort estimates, standard errors, P-values, or confidence limits
- Data reduction = *unsupervised learning*
- Example: dataset with 40 events and 60 candidate predictors

- Use variable clustering to group variables by correlation structure
- Use clinical knowledge to refine the clusters
- Keep age and severity of disease as separate predictors because of their strength
- For others create clusters: socioeconomic, risk factors/history, and physiologic function
- Summarize each cluster with its first principal component PC_1, i.e., the linear combination of characteristics that maximizes variance of the score across subjects subject to an overall constraint on the coefficients
- Fit outcome model with 5 predictors

3.7.1

Redundancy Analysis

A

- 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 function `redun` in `Hmisc` package
- Related to *principal variables* [106] but faster

3.7.2

Variable Clustering

B

- 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 [129, 39] (oblique-rotation PC analysis)

→ separate variables so that first PC is representative of group

- Can also do hierarchical cluster analysis on similarity matrix based on squared Spearman or Pearson correlations, or more generally, Hoeffding's D [79].

- See [68] for a method related to variable clustering and sparse principal components.

- [28] implement many more variable clustering methods

Example: Figure 8.6

3.7.3 Transformation and Scaling Variables Without Using Y

strategy-transcan

C

- Reduce p by estimating transformations using associations with other predictors

- Purely categorical predictors – correspondence analysis [95, 38, 29, 66, 107]

- Mixture of qualitative and continuous variables: qualitative principal components

- Maximum total variance (MTV) of Young, Takane, de Leeuw [166, 107]

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 [91, 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` [91]

1. Allows flexible transformations including monotonic splines
2. Does not allow restricted cubic splines, so may be unstable unless monotonicity assumed

3. Allows simultaneous imputation but often yields wild estimates

3.7.4 Simultaneous Transformation and Imputation

R `transcan` Function for Data Reduction & Imputation D

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

1. Continuous $\rightarrow$ back-solve with linear interpolation
2. Categorical $\rightarrow$ classification tree (most freq. cat.) or match to category whose canonical score is closest to one predicted

- Multiple imputation — bootstrap or approx. Bayesian boot.
  1. Sample residuals multiple times (default $M = 5$)
  2. Are on "optimally" transformed scale
  3. Back-transform
  4. `fit.mult.impute` works with `aregImpute` and `transcan` output to easily get imputation-corrected variances and avg. $\hat{\beta}$

- Option to insert constants as imputed values (ignored during transformation estimation); helpful when a lab value may be missing because the patient returned to normal

- Imputations and transformed values may be easily obtained for new data

- An R function `Function` will create a series of R 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 3.2 and 3.3

```
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 3.2
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
R^2 achieved in predicting each variable:

    heart.rate blood.pressure
         0.259          0.259

Adjusted R^2:

    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)
```

```
plot(heart.rate, blood.pressure)    # Figure 3.3
plot(t[,'heart.rate'], t[,'blood.pressure'],
     xlab='Transformed hr', ylab='Transformed bp')
```

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

E

1. Uses nonparametric "super smoother" [55]
2. Allows monotonicity constraints, categorical vars.

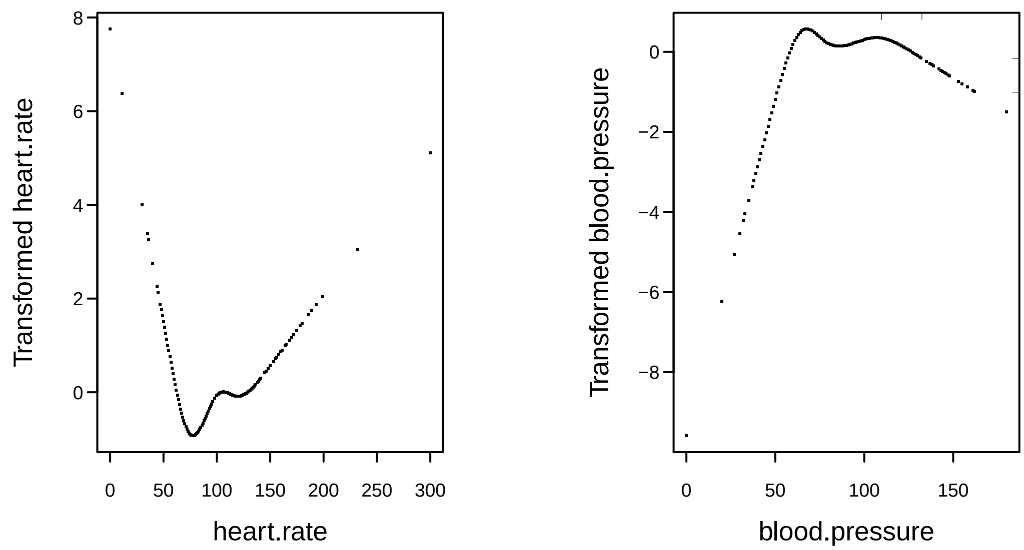


Figure 3.2: Transformations fitted using `transcan`. Tick marks indicate the two imputed values for blood pressure.

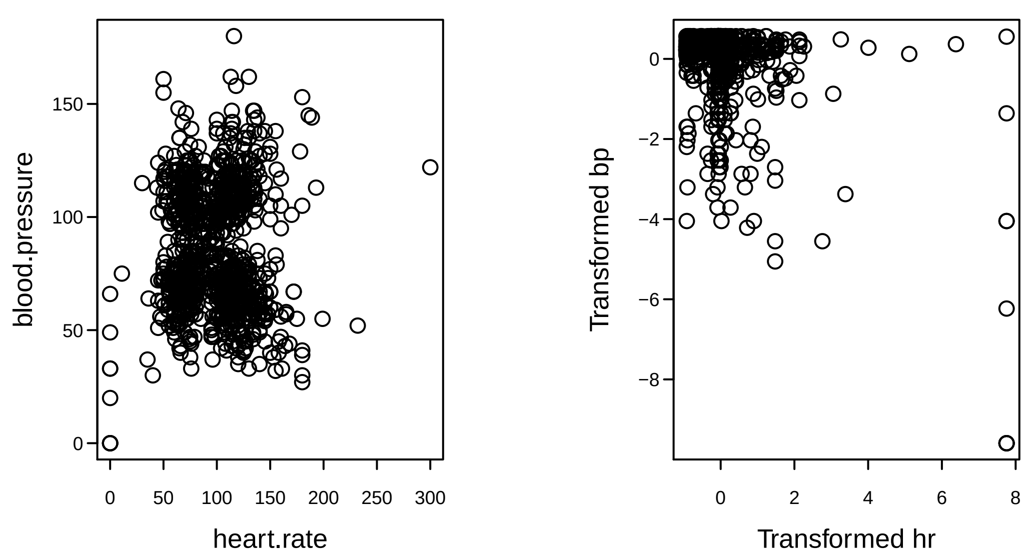


Figure 3.3: The lower left plot contains raw data (Spearman $\rho = -0.02$); the lower right is a scatterplot of the corresponding transformed values ($\rho = -0.13$). Data courtesy of the SUPPORT study [86].

3. Does not handle missing data

F

- 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

3.7.5

Simple Scoring of Variable Clusters

strategy-scoring

- Try to score groups of transformed variables with PC_1

- Reduces d.f. by pre-transforming var. and by combining multiple var. G

- 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: H

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

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

3.7.6 Simplifying Cluster Scores

3.7.7 How Much Data Reduction Is Necessary?

strategy-howmuch

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:

J

- 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

3.8 Other Approaches to Predictive Modeling

3.9 Overly Influential Observations

strategy-influence

K

- 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: L

- Too few observations for complexity of model (see Sections 3.7, 3.3)
- Data transcription or entry errors
- Extreme values of a predictor
  1. Sometimes subject so atypical should remove from dataset
  2. Sometimes truncate measurements where data density ends
  3. Example: $n = 4000$, 2000 deaths, white blood count range 500-100,000, .05,.95 quantiles=2755, 26700

4. Linear spline function fit
5. Sensitive to WBC> 60000 $(n = 16)$
6. Predictions stable if truncate WBC to 40000 ($n = 46$ above 40000)

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

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

Statistical Measures: M

- 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 [9]

- $h_{ii} > 2(p + 1)/n$ may signal a high leverage point [9]

- DFBETAS: change in $\hat{\beta}$ upon deletion of each obs, scaled by s.e.

- DFFIT: change in $X\hat{\beta}$ upon deletion of each obs

- DFFITS: DFFIT standardized by s.e. of $\hat{\beta}$

- Some classify obs as overly influential when $|\text{DFFITS}| > 2\sqrt{(p+1)/(n-p-1)}$ [9]

- Others examine entire distribution for "outliers"
- No substitute for careful examination of data [25, 138]
- Maximum likelihood estimation requires 1-step approximations

3.10 Comparing Two Models

strategy-compare

N

- Level playing field (independent datasets, same no. candidate d.f., careful bootstrapping)
- Criteria:
  1. calibration
  2. discrimination
  3. face validity
  4. measurement errors in required predictors
  5. use of continuous predictors (which are usually better defined than categorical ones)
  6. omission of "insignificant" variables that nonetheless make sense as risk factors
  7. simplicity (though this is less important with the availability of computers)
  8. lack of fit for specific types of subjects
- Goal is to rank-order: ignore calibration O
- Otherwise, dismiss a model having poor calibration
- Good calibration $\rightarrow$ compare discrimination (e.g., R^2 [110], 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 P
- Rank measures may not give enough credit to extreme predictions → model χ^2, R^2, examine extremes of distribution of $\hat{Y}$
- Examine differences in predicted values from the two models
- See [116, 119, 118, 117] for discussions and examples of low power for testing differences in ROC areas, and for other approaches.

3.11 Improving the Practice of Multivariable Prediction

strategy-improve

Q

See also Section 4.6.

Greenland [67] 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)”

Greenland’s example of inadequate adjustment for confounders as a result of using a bad modeling strategy: R

- Case-control study of diet, food constituents, breast cancer
- 140 cases, 222 controls
- 35 food constituent intakes and 5 confounders
- Food intakes are correlated
- Traditional stepwise analysis not adjusting simultaneously for all foods consumed $\rightarrow$ 11 foods had $P < 0.05$
- Full model with all 35 foods competing $\rightarrow$ 2 had $P < 0.05$
- Rigorous simultaneous analysis (hierarchical random slopes model) penalizing estimates for the number of associations examined $\rightarrow$ no foods associated with breast cancer

S

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

T

Preferred Strategy in a Nutshell

- Decide how many d.f. can be spent

- Decide where to spend them

- Spend them

- Don't reconsider, especially if inference needed

3.12 Summary: Possible Modeling Strategies

3.12.1 Developing Predictive Models

strategy-summary

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 [37]. For certain models it is effective to multiply impute Y.
4. Characterize and impute missing X. In most cases use multiple imputation based on X and Y
5. For each predictor specify complexity or degree of nonlinearity that should be allowed (more for important predictors or for large n) (Section 3.1)
6. Do data reduction if needed (pre-transformations, combinations), or use penalized estimation [75]
7. Use the entire sample in model development
8. Can do highly structured testing to simplify "initial" model
   (a) Test entire group of predictors with a single P-value

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

9. Make tests of linearity of effects in the model only to demonstrate to others that such effects are often statistically significant. Don't remove individual insignificant effects from the model.
10. Check additivity assumptions by testing pre-specified interaction terms. Use a global test and either keep all or delete all interactions.
11. Check to see if there are overly-influential observations.
12. Check distributional assumptions and choose a different model if needed.
13. Do limited backwards step-down variable selection if parsimony is more important that accuracy [139]. 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 [52] to penalize for the randomness inherent in the multiple steps.
20. Develop simplifications to the final model as needed.

3.12.2

Developing Models for Effect Estimation

V

1. Less need for parsimony; even less need to remove insignificant variables from model (otherwise CLs too narrow)
2. Careful consideration of interactions; inclusion forces estimates to be conditional and raises variances
3. If variable of interest is mostly the one that is missing, multiple imputation less valuable
4. Complexity of main variable specified by prior beliefs, compromise between variance and bias
5. Don't penalize terms for variable of interest
6. Model validation less necessary

3.12.3

Developing Models for Hypothesis Testing

W

1. Virtually same as previous strategy
2. Interactions require tests of effect by varying values of another variable, or "main effect + interaction" joint tests (e.g., is treatment effective for either sex, allowing effects to be different)
3. Validation may help quantify overadjustment

Chapter 4

Describing, Resampling, Validating, and Simplifying the Model

4.1

Describing the Fitted Model

4.1.1

Interpreting Effects

val-describe

A

- 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 inter-

acting factors

- Subtract estimates, anti-log, e.g., to get inter-quartile-range odds or hazards ratios. Base C.L. on s.e. of difference. See Figure 9.10. B

- Partial effect plot: Plot effect of each predictor on $X\beta$ or some transformation. See Figure 9.8. See also [84].

- Nomogram. See Figure 9.12.

- Use regression tree to approximate the full model

4.1.2

Indexes of Model Performance

Error Measures

C

- 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 D

- 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 E

- calibration–in–the–large: average $\hat{Y}$ vs. average Y
- high-resolution calibration curve (calibration–in–the–small). See Figure 7.7.
- calibration slope and intercept
- maximum absolute calibration error

- mean absolute calibration error
- 0.9 quantile of calibration error

See Van Calster *et al.* [148] for a nice discussion of different levels of calibration stringency and their relationship to likelihood of errors in decision making.

F

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
  $g_r = \exp(g)$
- For models in which $\hat{Y}$ can be turned into a probability

estimate (e.g., logistic regression):
$g_p =$ Gini's mean difference of $\hat{P}$

- These g–indexes represent e.g. "typical" odds ratios, "typical" risk differences

- Can define partial g

4.2 The Bootstrap

val-boot H

- If know population model, use simulation or analytic derivations to study behavior of statistical estimator
- Suppose Y has a cumulative dist. fctn. $F(y) = \text{Prob}\{Y \leq y\}$
- We have sample of size n from $F(y)$, $Y_1, Y_2, \ldots, Y_n$
- Steps:
  1. Repeatedly simulate sample of size n from F
  2. Compute statistic of interest
  3. Study behavior over B repetitions
- Example: 1000 samples, 1000 sample medians, compute their sample variance I
- 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)
```

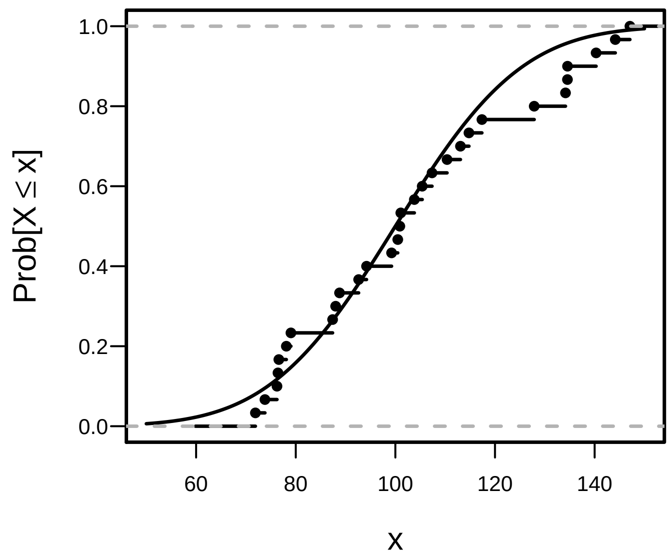


Figure 4.1: Empirical and population cumulative distribution function

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

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

- 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). M

```
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="")
```

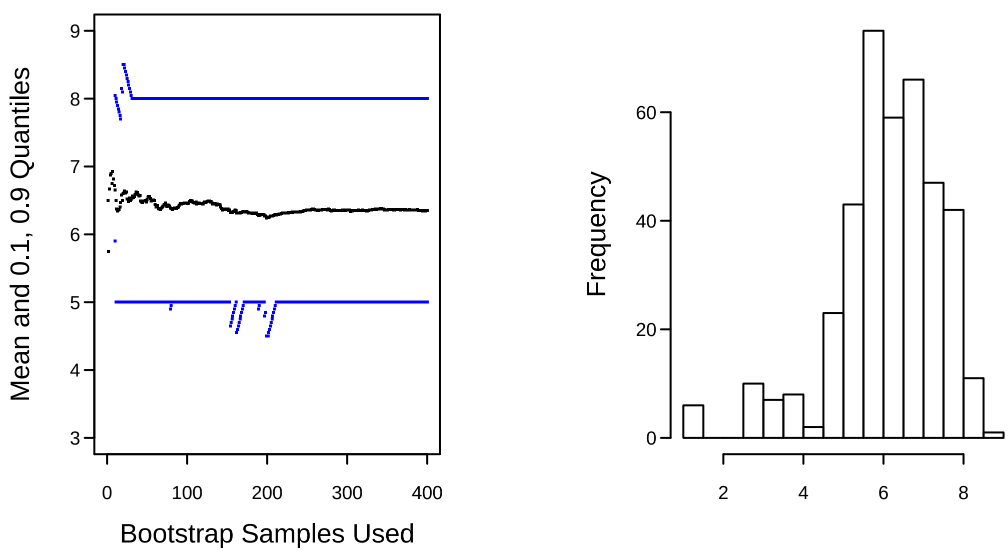


Figure 4.2: Estimating properties of sample median using the bootstrap

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

N

- 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 [15])

4.3

Model Validation

4.3.1

Introduction

val-how

O

- 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 P
- Predictions on another dataset will be inaccurate
- Need unbiased assessment of predictive accuracy

Working definition of external validation: Validation of a prediction tool on a sample that was not available at publi-

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

cation 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. Q

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- R
   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

Some points to consider: S

- Unless both sample sizes are huge, external validation can be low precision
- External validation can be costly and slow and may result in disappointment that would have been revealed earlier with rigorous internal validation
- External validation is sometimes *gamed*; researchers disappointed in the validation sometimes ask for a "do over"; resampling validation is harder to game as long as all analytical steps using Y are repeated each time.
- Instead of external validation to determine model applicability at a different time or place, and being disappointed if T the model does not work in that setting, consider building a unified model containing time and place as predictors
- When the model was fully pre-specified, external validation tests *the model*

- But when the model was fitted using machine learning, feature screening, variable selection, or model selection, the model developed using training data is usually only an example of a model, and the test sample validation could be called an *example validation*

- When resampling is used to repeat *all* modeling steps for each resample, rigorous internal validation tests the *process* used to develop the model and happens to also provide a high-precision estimate of the likely future performance of the "final" model developed using that process, properly penalizing for model uncertainty.

- Resampling also reveals the volatility of the model selection process

→ See BBR 10.11

Collins *et al.* [31] estimate that a typical sample size needed for externally validating a time-to-event model is 200 events.

4.3.2 Which Quantities Should Be Used in Validation?

U

- 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 v
- Two main types of quantities to validate
  1. Calibration or reliability: ability to make unbiased estimates of response ($\hat{Y}$ vs. Y)
  2. Discrimination: ability to separate responses
     OLS: R^2; g–index; binary logistic model: ROC area, equivalent to rank correlation between predicted probability of event and 0/1 event
- Unbiased validation nearly always necessary, to detect overfitting

4.3.3

Data-Splitting

W

- 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 X
- 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: Y
  1. Costly in $\downarrow n$ [127, 18]
  2. Requires *decision* to split at beginning of analysis
  3. Requires larger sample held out than cross-validation

4. Results vary if split again
5. Does not validate the final model (from recombined data)
6. Not helpful in getting CL corrected for var. selection

4.3.4 Improvements on Data-Splitting: Resampling

Z

- No sacrifice in sample size
- Work when modeling process automated
- Bootstrap excellent for studying arbitrariness of variable selection [130]. See P. ??.
- Cross-validation solves many problems of data splitting [149, 134, 163, 47]
- Example of $\times$-validation: A
  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 B
  1. Randomly permute Y
  2. Optimism = performance of fitted model compared to what expect by chance

4.3.5

Validation Using the Bootstrap

C

- 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 $\overline{optimism}$
  - $R^2_{overfitting\ corrected} = R^2_{app} - \overline{optimism}$
- Example: See P. **??**

- Is estimating unconditional (not conditional on X) distribution of R^2, etc. [52, p. 217] E

- 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 [47], [48, p. 253], [49]

Bootstrap useful for assessing calibration in addition to discrimination: F

- 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 [139]

- Use Efron's method to estimate optimism in $(0, 1)$, estimate (γ_0, γ_1) by subtracting optimism from $(0, 1)$

- See also Copas [34] and van Houwelingen and le Cessie [149,

p. 1318]

See [54] for warnings about the bootstrap, and [47] for variations on the bootstrap to reduce bias.

Use bootstrap to choose between full and reduced models: G

- Bootstrap estimate of accuracy for full model
- Repeat, using chosen stopping rule for each re-sample
- Full fit usually outperforms reduced model [139]
- 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 H

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 [149] for many interesting ideas
- Faraway [52] shows how bootstrap is used to penalize for choosing transformations for Y, outlier and influence checking, variable selection, etc. simultaneously
- Brownstone [21, 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 [14] and Chatfield [26] for more interesting examples of problems resulting from data-driven analyses.

4.4 Bootstrapping Ranks of Predictors

val-ranks

- 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)
```

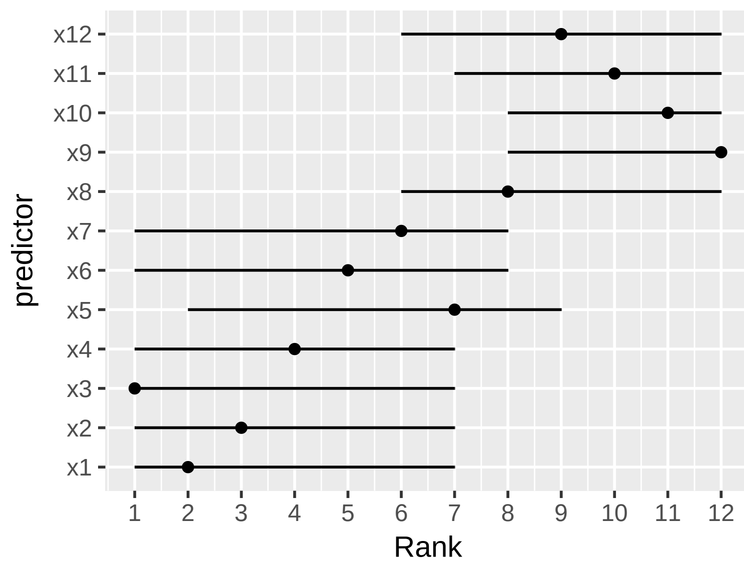


Figure 4.3: Bootstrap percentile 0.95 confidence limits for ranks of predictors in an OLS model. Ranking is on the basis of partial χ^2 minus d.f. Point estimates are original ranks

4.5 Simplifying the Final Model by Approximating It

val-approx

4.5.1 Difficulties Using Full Models

K

- 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.

4.5.2 Approximating the Full Model

L

- 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 M
  1. Uses unbiased estimate of σ from full fit
  2. Stopping rule less arbitrary
  3. Inheritance of shrinkage
- If estimates from full model are $\hat{\beta}$ and approx. model is based on a subset T of predictors X, coef. of approx. model are $W\hat{\beta}$, where
  $W = (T'T)^{-1}T'X$
- Variance matrix of reduced coef.: WVW'

4.6 How Do We Break Bad Habits?

val-habits N

- 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 5

R Software

rrms

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 $\rightarrow$ automatic hypothesis tests, predictions for new data
- Superior graphics
- Classes and generic functions

5.1 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.
```

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

5.2 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

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`)

5.3 The `rms` Package

- `datadist` function to compute predictor distribution summaries

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

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

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

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

Table 5.1: `rms` Fitting Functions

Function	Purpose	Related 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)

Below notice that there are three graphic models implemented

Table 5.2: 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

for depicting the effects of predictors in the fitted model: lattice graphics, a ggplot method using the ggplot2 package (which has an option to convert the result to plotly), and a direct plotly method. plotly is used to create somewhat interactive graphics with drill-down capability, and the rms package takes advantage of this capability. plotly graphics are best used with RStudio Rmarkdown html output.

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`	
`plotp.Predict`	Plot the result of `Predict` using `plotly`	
`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`
`survplotp`	Plot survival curves with `plotly` features	`survplot`
`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`

Global options `prType` and `grType` control printed and some graphical output, respectively as shown in example code below. The default is plain output and static graphics. If using `plotly` interactive graphics through `ggplot` or `plotp` or with `anova` or `summary` functions it is best to do so with RStudio `html` output or `html` notebooks. If using `html` output you must be producing an `html` document or notebook. When setting `grType` to use LaTeX or `html` it is highly recommended that you use the `knitr` package.

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
options(prType='latex')           # print, summary, anova LaTeX output
                                  # others: 'html', 'plain'
options(grType='plotly')          # plotly graphics for ggplot, anova, summary
                                  # default is 'base' for static graphics
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))
fit    # outputs plain, LaTeX, or html markup
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)                        # plain, LaTex, or html
anova(fit, treat, cholesterol)    # Test these 2 by themselves
```

```
plot(anova(fit))                    # Summarize anova graphically
summary(fit)                        # Estimate effects using default ranges
                                    # prints plain, LaTeX, or html
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), plotp(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())
                                    # can also use plotp() for plotly
# 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))
```

5.4 Other Functions

- `supsmu`: Friedman's "super smoother"
- `lowess`: Cleveland's scatterplot smoother
- `glm`: generalized linear models (see `Glm`)
- `gam`: Generalized additive models
- `rpart`: Like original CART with surrogate splits for missings, censored data extension (Atkinson & Therneau)
- `validate.rpart`: in `rms`; validates recursive partitioning with respect to certain accuracy indexes
- `loess`: multi-dimensional scatterplot smoother

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

Chapter 6

Modeling Longitudinal Responses using Generalized Least Squares

6.1

Notation

A

- 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 $g_1(t), g_2(t), \ldots,$

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

6.2.1 Common Basis Functions

B

- 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)

6.2.2 Model for Mean Profile

C

- 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

6.2.3 Model Specification for Treatment Comparisons

D

- 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 [133] 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. E
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 [98] and Liu *et al.* [100] are problematic[a].

[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 [133]. See [85] for a rebuke of [100].

6.3 Modeling Within-Subject Dependence

F

- Random effects and mixed effects models have become very popular
- Disadvantages:
  - Induced correlation structure for Y may be unrealistic
  - Numerically demanding
  - 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 [123, 61] 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 $t_{i1}, \ldots, t_{in_i}$ G
- 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: H
  - 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 different 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					$\times$	
Harder 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] Unlike 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.* [57] compared several longitudinal data models, especially with regard to assumptions and how regression coefficients are estimated. Peters *et al.* [120] have an empirical study confirming that the "use all available data" approach of likelihood–based longitudinal models makes imputation of follow-up measurements unnecessary.

6.4 Parameter Estimation Procedure

J

- 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 [44, Section 5.3], [121, Chapter 5], [61] (esp. to get more unbiased estimate of the covariance matrix)
- When imbalances are not severe, OLS fitted ignoring subject identifiers may be efficient K
  - 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 L
  - a subject contributing two distant observations receives $\frac{1}{5}$ the weight of a subject having 10 tightly-spaced observations

6.5 Common Correlation Structures

M

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

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 : $h = (1 - s/\rho)[s < \rho]$ — `corLin`

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

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

Linear exponent AR(1) : $h = \rho^{d_{min} + \delta \frac{s - d_{min}}{d_{max} - d_{min}}}$, 1 if $t_1 = t_2$ [136]

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

6.6

Checking Model Fit

O

- 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

6.7 R Software

P

- Nonlinear mixed effects model package of Pinheiro & Bates
- 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, ggplot, plotp` : 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

6.8 Case Study

Consider the dataset in Table 6.9 of Davis [41, 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.

Q

- 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

6.8.1 Graphical Exploration of Data

```
require(rms)
```

```
options(prType='latex')    # for model print, summary, anova
getHdata(cdystonia)
attach(cdystonia)

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

```
# Tabulate patterns 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    0 2 4 8 12
               1              1             3              1             1
0 2 4 8 12 16     0 2 4 8 16    0 2 8 12 16    0 4 8 12 16       0 4 8 16
              94              1             2              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. 6.1
```

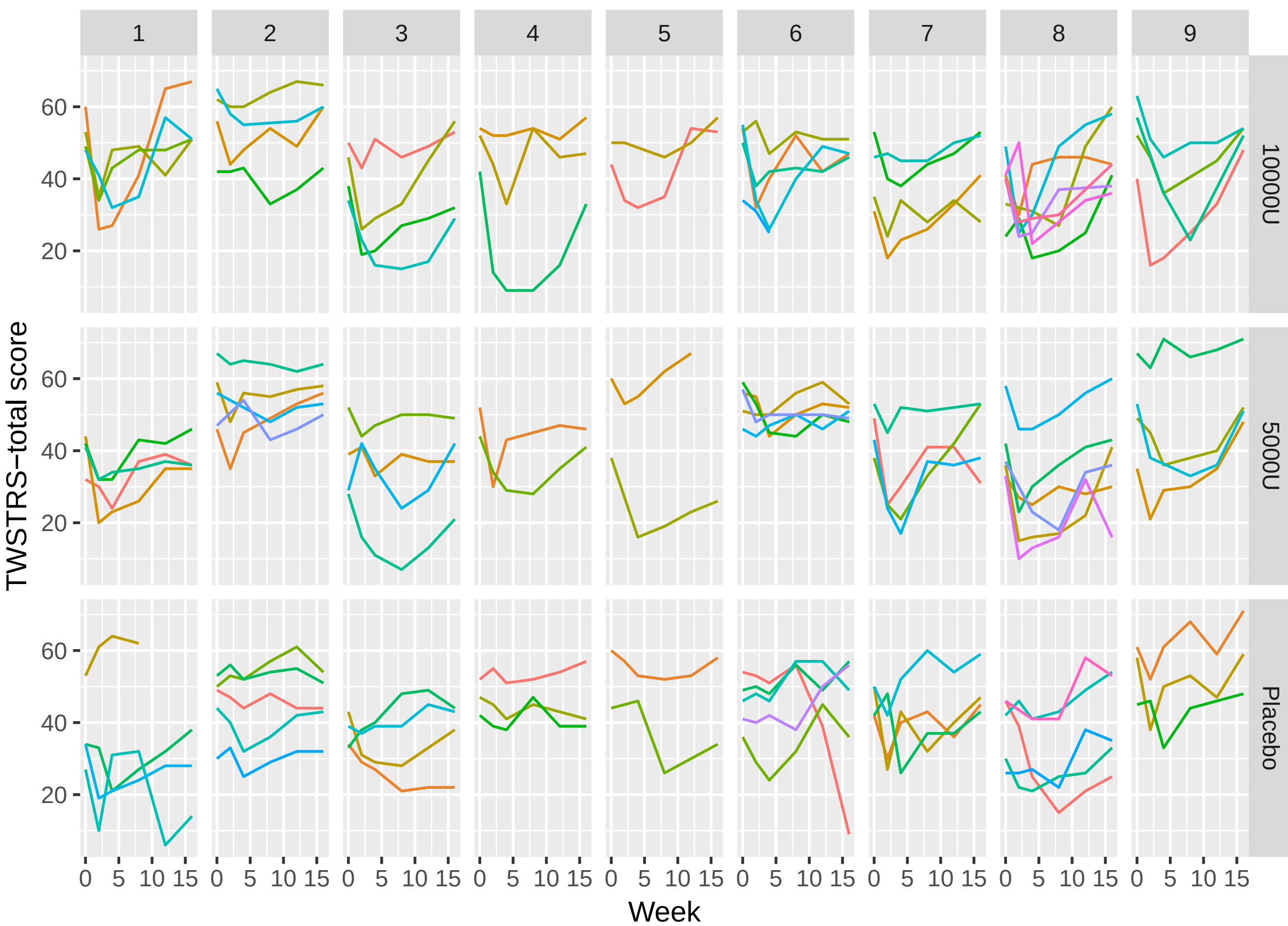


Figure 6.1: Time profiles for individual subjects, stratified by study site and dose

```
# Show quartiles
require(data.table)
```

```
cdystonia ← data.table(cdystonia)
cdys ← cdystonia[, j=as.list(quantile(twstrs, (1 : 3)/4)),
                  by = list(treat, week)]
cdys ← upData(cdys, rename=c('25%'='Q1', '50%'='Q2', '75%'='Q3'), print=FALSE)
ggplot(cdys, aes(x=week, y=Q2)) + xl + yl + ylim(0, 70) +
```

```
  geom_line() + facet_wrap(~ treat, nrow=2) +
  geom_ribbon(aes(ymin=Q1, ymax=Q3), alpha=0.2)   # Fig. 6.2
```

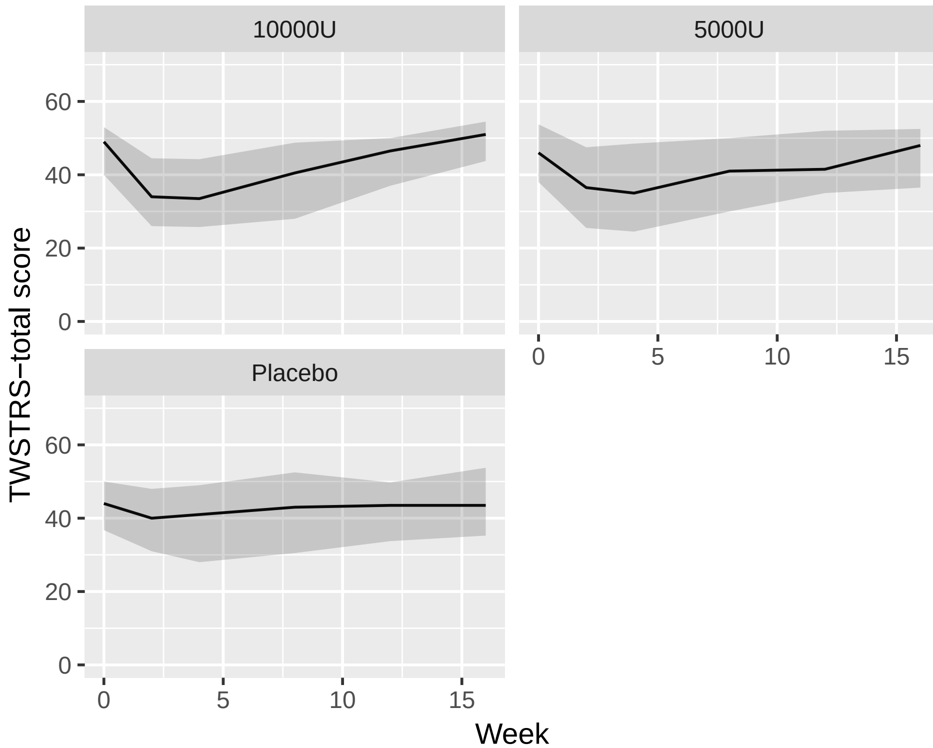


Figure 6.2: Quartiles of TWSTRS stratified by dose

```
# Show means with bootstrap nonparametric CLs
cdys ← cdystonia[, j=as.list(smean.cl.boot(twstrs)),
                    by = list(treat, week)]
ggplot(cdys, aes(x=week, y=Mean)) + xl + yl + ylim(0, 70) +
  geom_line() + facet_wrap(~ treat, nrow=2) +
  geom_ribbon(aes(x=week, ymin=Lower, ymax=Upper), alpha=0.2)   # Fig. 6.3
```

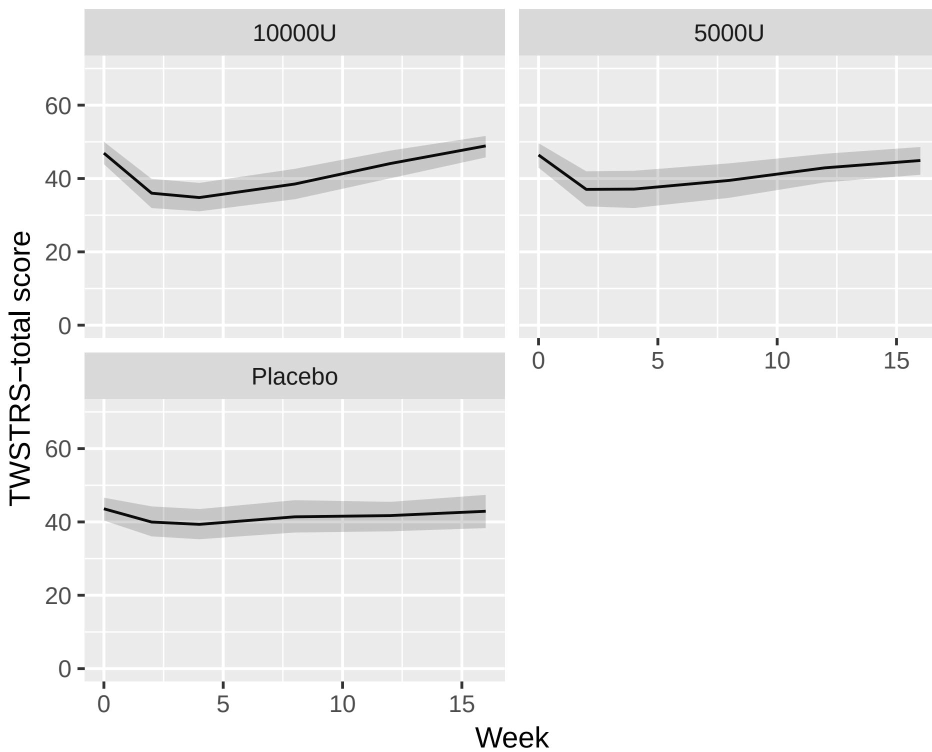


Figure 6.3: Mean responses and nonparametric bootstrap 0.95 confidence limits for population means, stratified by dose

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')
```

6.8.2 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`. R

```
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`.

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

```
a
```

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

	$\hat{\beta}$	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 $0.8672^{10} = 0.24$. S

```
v ← Variogram(a, form=∼ week | uid)
plot(v)   # Figure 6.4
```

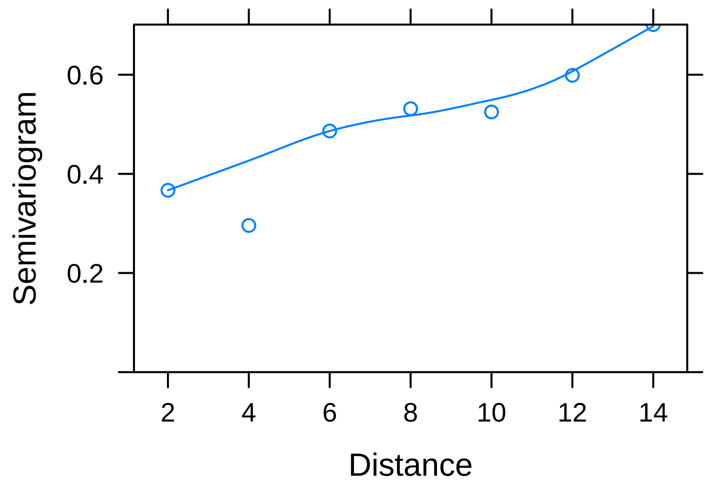


Figure 6.4: Variogram, with assumed correlation pattern superimposed

Check constant variance and normality assumptions: T

```
both$resid ← r ← resid(a); both$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 + ylim(-20,20) +
      stat_summary(fun.data="mean_sdl", geom='smooth')
p4 ← ggplot(both, aes(sample=resid)) + stat_qq() +
      geom_abline(intercept=mean(r), slope=sd(r)) + yl
gridExtra::grid.arrange(p1, p2, p3, p4, ncol=2)   # Figure 6.5
```

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

```
anova(a)
```

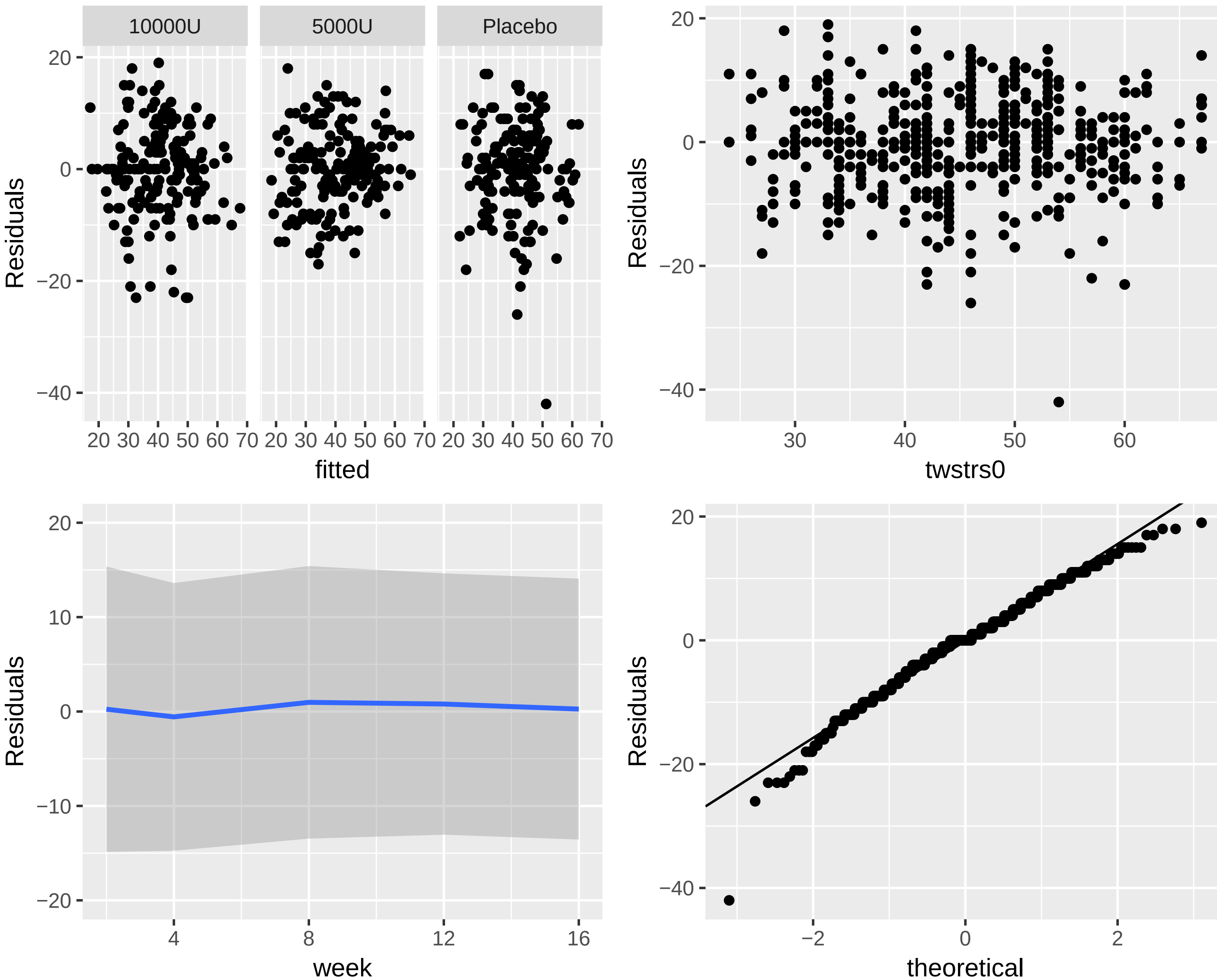


Figure 6.5: Three residual plots to check for absence of trends in central tendency and in variability. Upper right panel shows the baseline score on the x-axis. Bottom left panel shows the mean $\pm 2\times$SD. Bottom right panel is the QQ plot for checking normality of residuals from the GLS fit.

	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

```
plot(anova(a))      # Figure 6.6
```

```
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
gridExtra::grid.arrange(p1, p2, p3, ncol=2)    # Figure 6.7
```

```
summary(a)   # 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

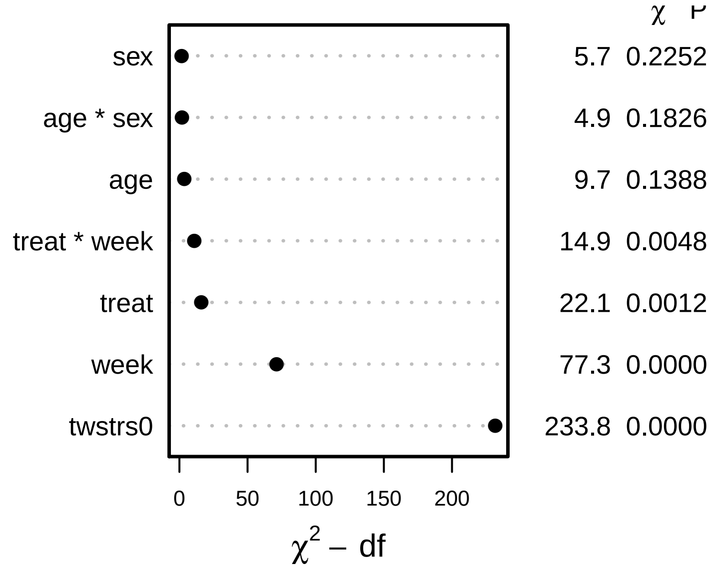


Figure 6.6: Results of `anova.rms` from generalized least squares fit with continuous time AR1 correlation structure

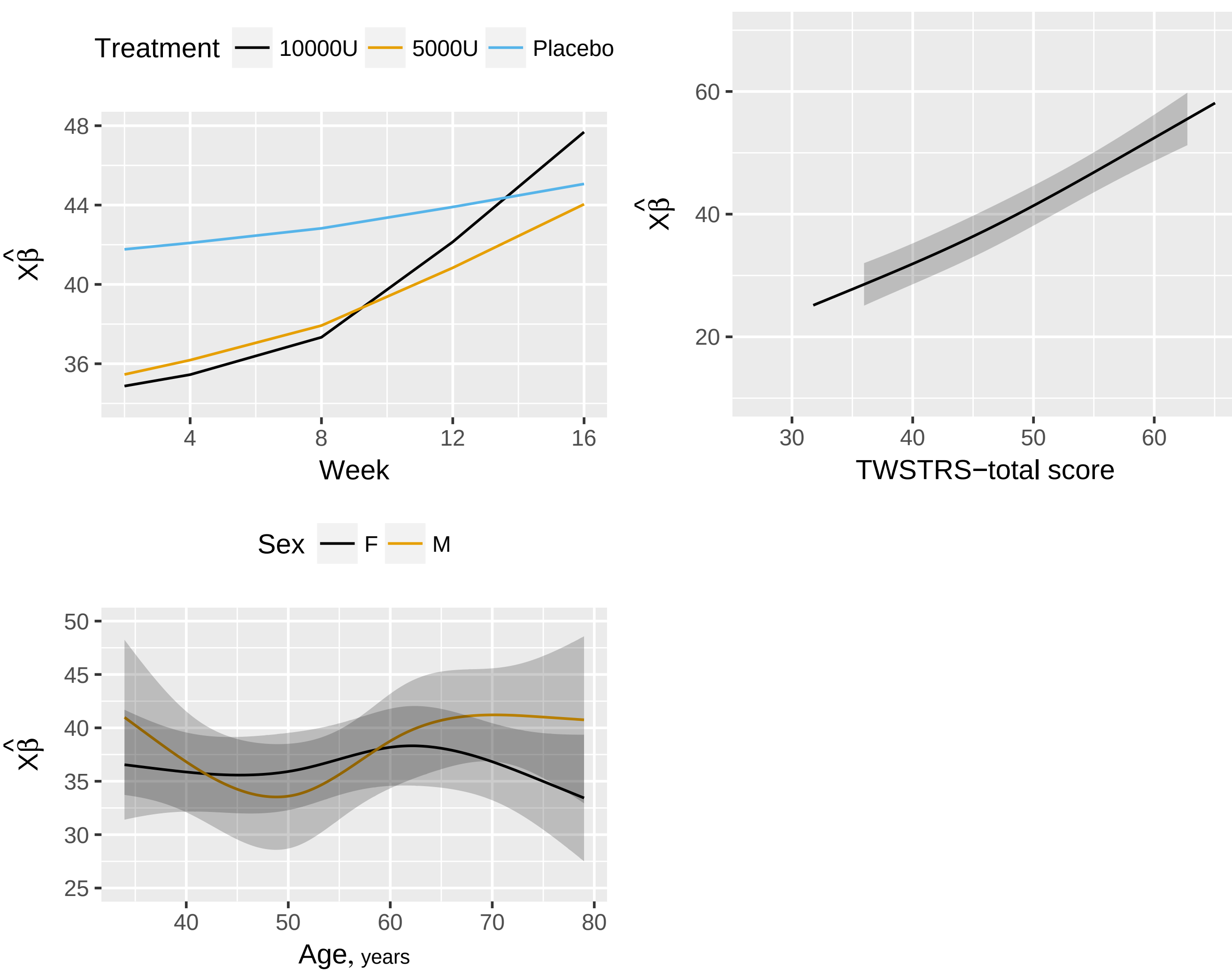


Figure 6.7: Estimated effects of time, baseline TWSTRS, age, and sex

```
# 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=3)
```

```
    week twstrs0 age sex Contrast S.E.  Lower  Upper     Z Pr(>|z|)
1      2      46  56   F    -6.31 2.10 -10.43 -2.186 -3.00   0.0027
2      4      46  56   F    -5.91 1.82  -9.47 -2.349 -3.25   0.0011
3      8      46  56   F    -4.90 2.01  -8.85 -0.953 -2.43   0.0150
4*    12      46  56   F    -3.07 1.75  -6.49  0.361 -1.75   0.0795
5*    16      46  56   F    -1.02 2.10  -5.14  3.092 -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=3)
```

```
    week twstrs0 age sex Contrast S.E.  Lower Upper     Z Pr(>|z|)
1      2      46  56   F    -6.89 2.07 -10.96 -2.83 -3.32   0.0009
2      4      46  56   F    -6.64 1.79 -10.15 -3.13 -3.70   0.0002
3      8      46  56   F    -5.49 2.00  -9.42 -1.56 -2.74   0.0061
4*    12      46  56   F    -1.76 1.74  -5.17  1.65 -1.01   0.3109
5*    16      46  56   F     2.62 2.09  -1.47  6.71  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() + ylab('Low Dose - Placebo') +
     geom_errorbar(aes(ymin=Lower, ymax=Upper), width=0)
k2 ← as.data.frame(k2[c('week', 'Contrast', 'Lower', 'Upper')])
p2 ← ggplot(k2, aes(x=week, y=Contrast)) + geom_point() +
     geom_line() + ylab('High Dose - Placebo') +
     geom_errorbar(aes(ymin=Lower, ymax=Upper), width=0)
gridExtra::grid.arrange(p1, p2, ncol=2)    # Figure 6.8
```

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 U

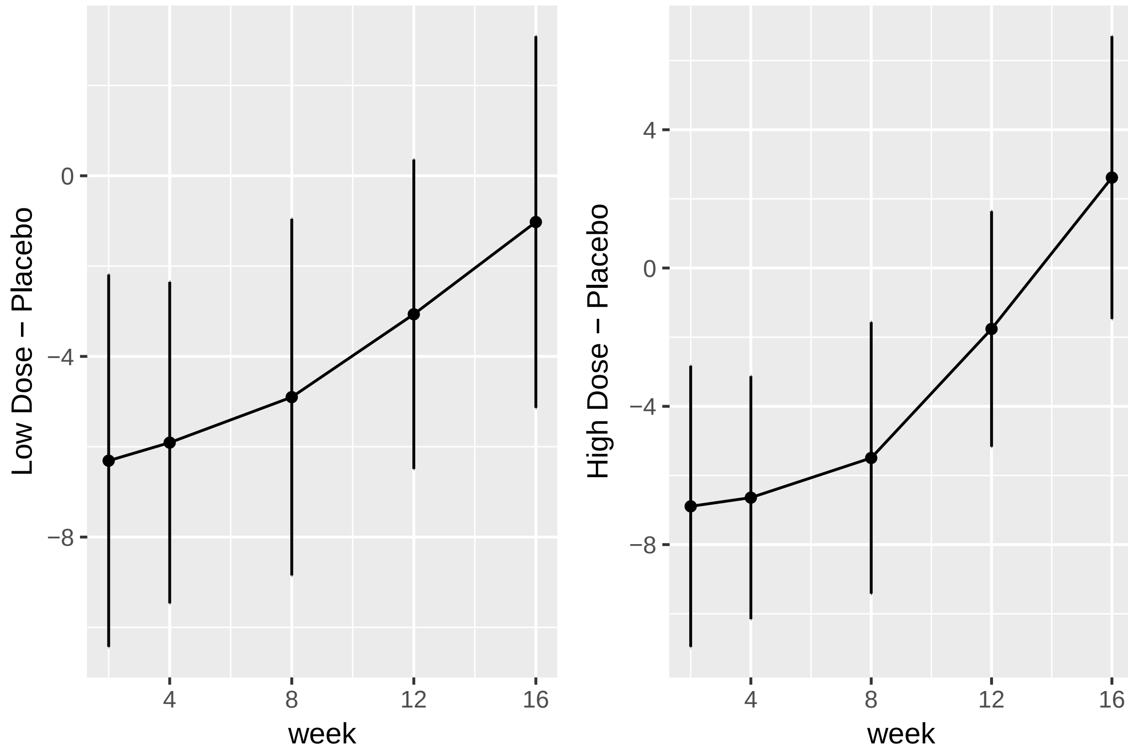


Figure 6.8: Contrasts and 0.95 confidence limits from GLS fit

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 6.9
```

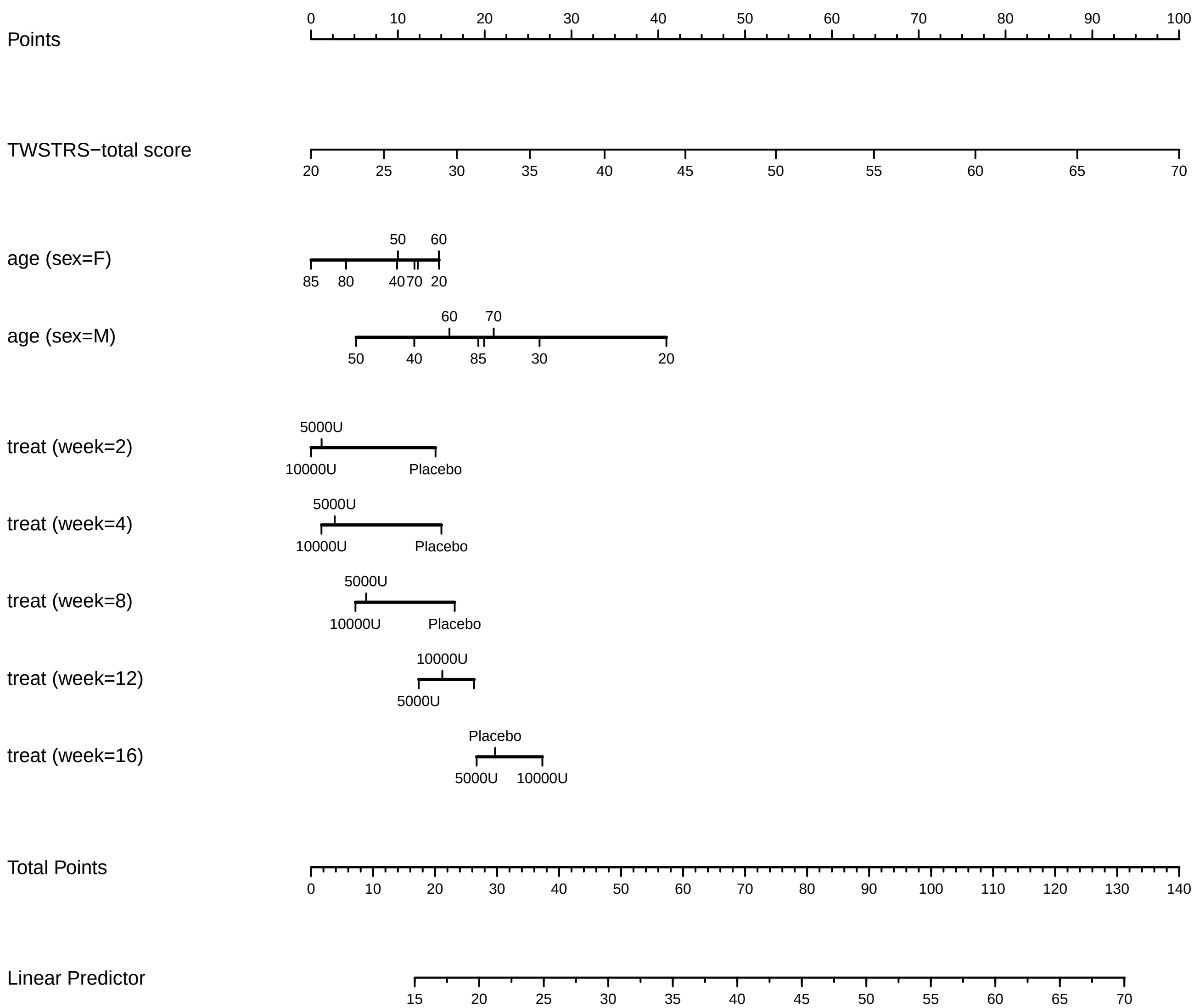


Figure 6.9: Nomogram from GLS fit. Second axis is the baseline score.

Chapter 7

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 Rand spreadsheet formats from `biostat.mc.vanderbilt.edu/DataSets` under the name `titanic3`.

7.1 Descriptive Statistics

```
require(rms)
```

```
options(prType='latex')    # for print, summary, anova
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	distinct
1309	0	3

Value	1st	2nd	3rd
Frequency	323	277	709
Proportion	0.247	0.212	0.542

---

survived : Survived

n	missing	distinct	Info	Sum	Mean	Gmd
1309	0	2	0.708	500	0.382	0.4725

---

age : Age [years]

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1046	263	98	0.999	29.88	16.06	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	distinct
1309	0	2

Value	female	male
Frequency	466	843
Proportion	0.356	0.644

---

sibsp : Number of Siblings/Spouses Aboard

n	missing	distinct	Info	Mean	Gmd
1309	0	7	0.67	0.4989	0.777

Value	0	1	2	3	4	5	8
Frequency	891	319	42	20	22	6	9
Proportion	0.681	0.244	0.032	0.015	0.017	0.005	0.007

---

parch : Number of Parents/Children Aboard

n	missing	distinct	Info	Mean	Gmd
1309	0	8	0.549	0.385	0.6375

Value	0	1	2	3	4	5	6	9
Frequency	1002	170	113	8	6	6	2	2

```
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 7.1
```

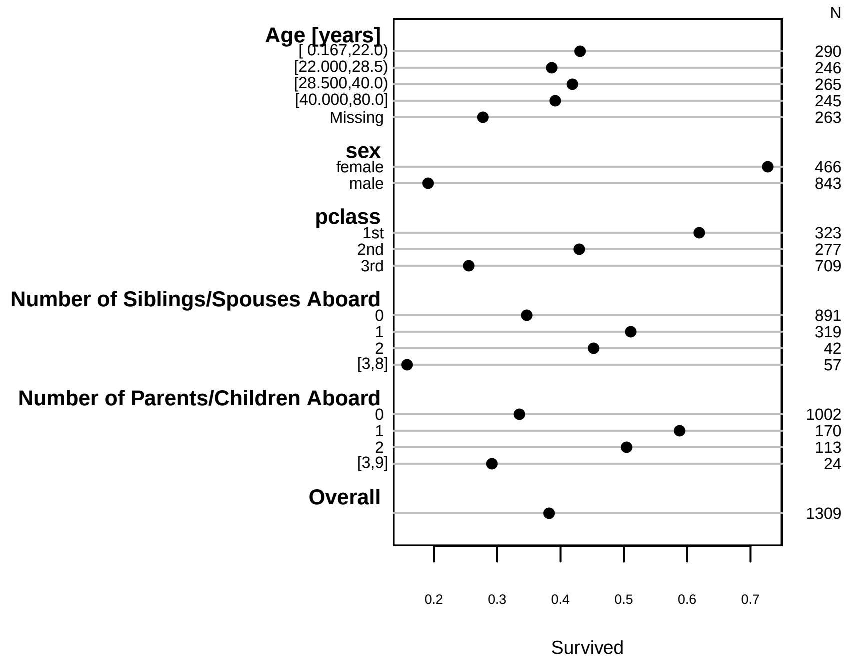


Figure 7.1: Univariable summaries of Titanic survival

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

A

```
tn ← transform(t3,
  agec = ifelse(age < 21, 'child', 'adult'),
  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 7.2:
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))
```

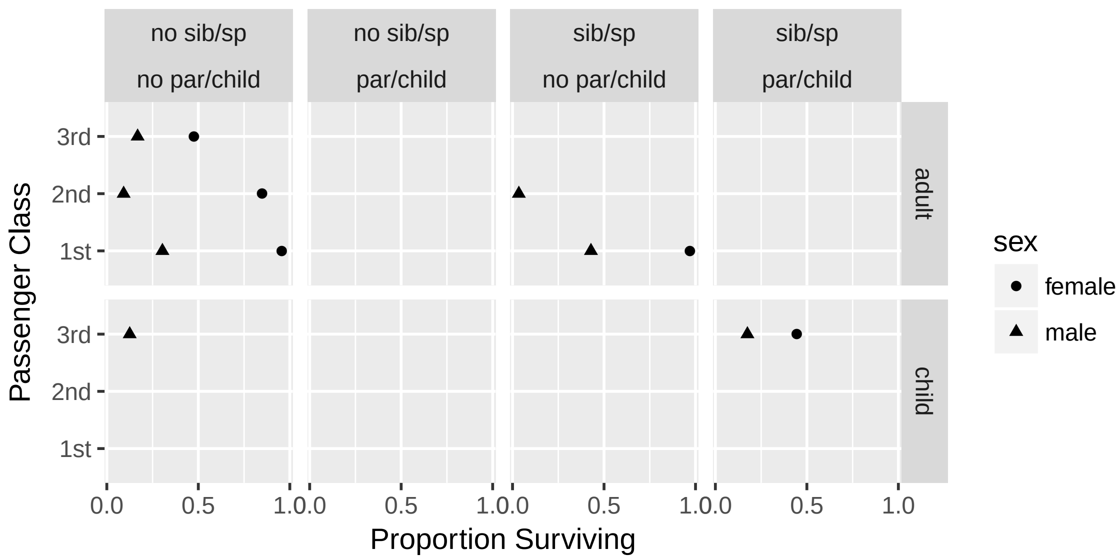


Figure 7.2: Multi-way summary of Titanic survival

7.2 Exploring Trends with Nonparametric Regression

```
# Figure 7.3
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,
                 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 7.4
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)
```

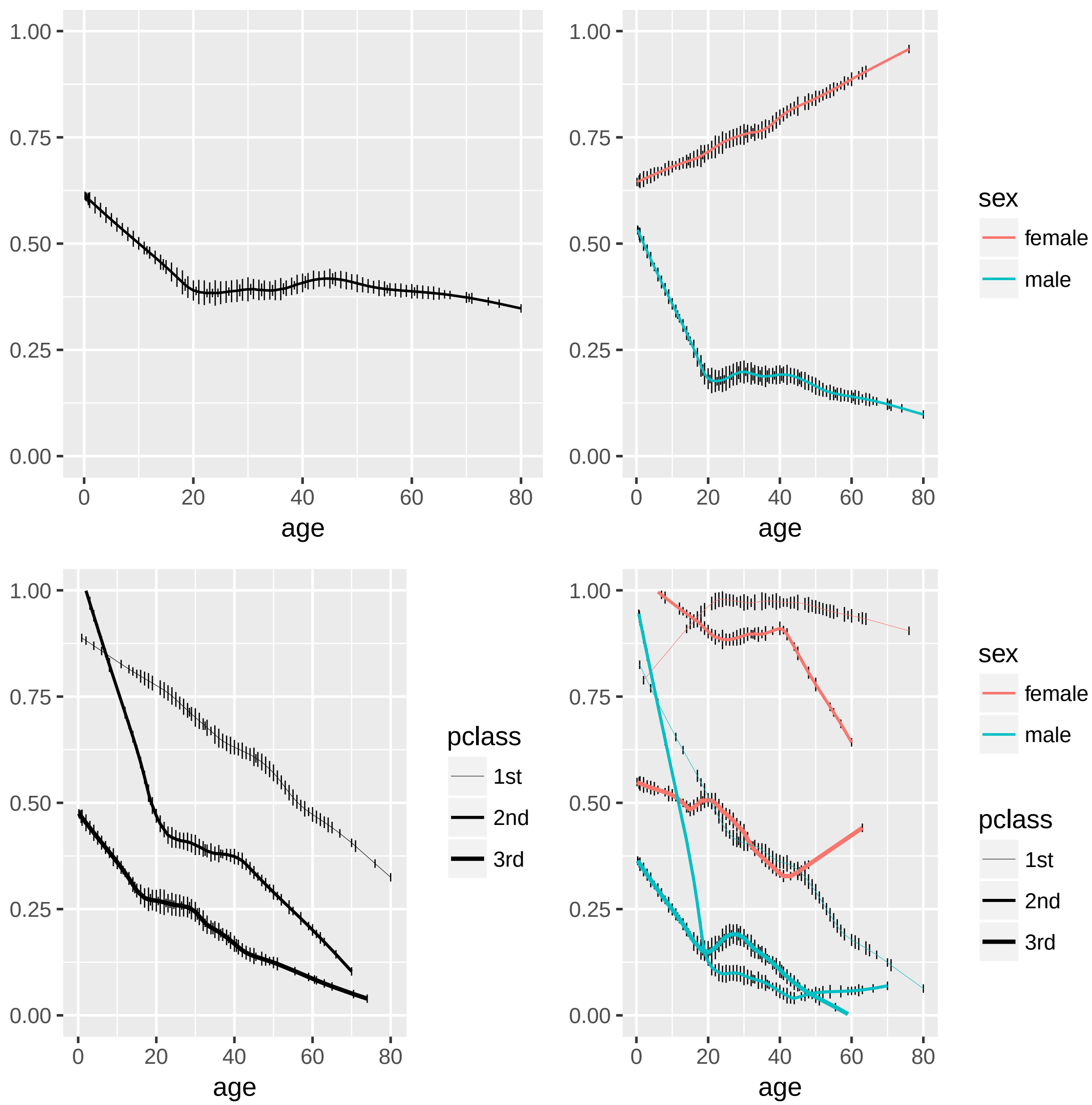


Figure 7.3: Nonparametric regression (`loess`) estimates of the relationship between age and the probability of surviving the Titanic, with tick marks depicting the age distribution. The top left panel shows unstratified estimates of the probability of survival. Other panels show nonparametric estimates by various stratifications.

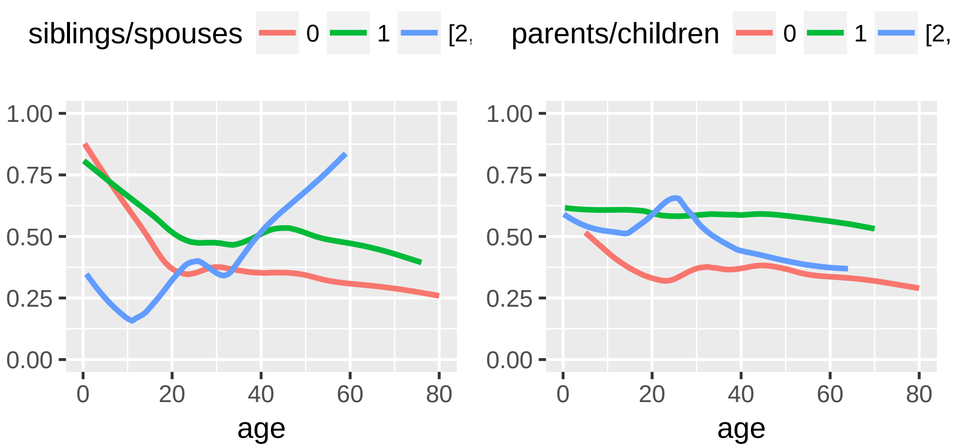


Figure 7.4: Relationship between age and survival stratified by the number of siblings or spouses on board (left panel) or by the number of parents or children of the passenger on board (right panel).

Table 7.1: 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

7.3 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 7.1
print(anova(f1), table.env=TRUE, 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)
```

Logistic Regression Model

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

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		

	$\hat{\beta}$	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

Table 7.2: 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

	$\hat{\beta}$	S.E.	Wald Z	Pr$(>\|Z\|)$
age'' × sibsp	0.5574	1.6680	0.33	0.7382
age''' × sibsp	-1.1937	2.5711	-0.46	0.6425

```
print(anova(f), table.env=TRUE, label='titanic-anova2',size='small') #7.2
```

Show the many effects of predictors. B

```
p ← Predict(f, age, sex, pclass, sibsp=0, fun=plogis)
ggplot(p)         # Fig. 7.5
```

```
ggplot(Predict(f, sibsp, age=c(10,15,20,50), conf.int=FALSE))
## Figure 7.6
```

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. C

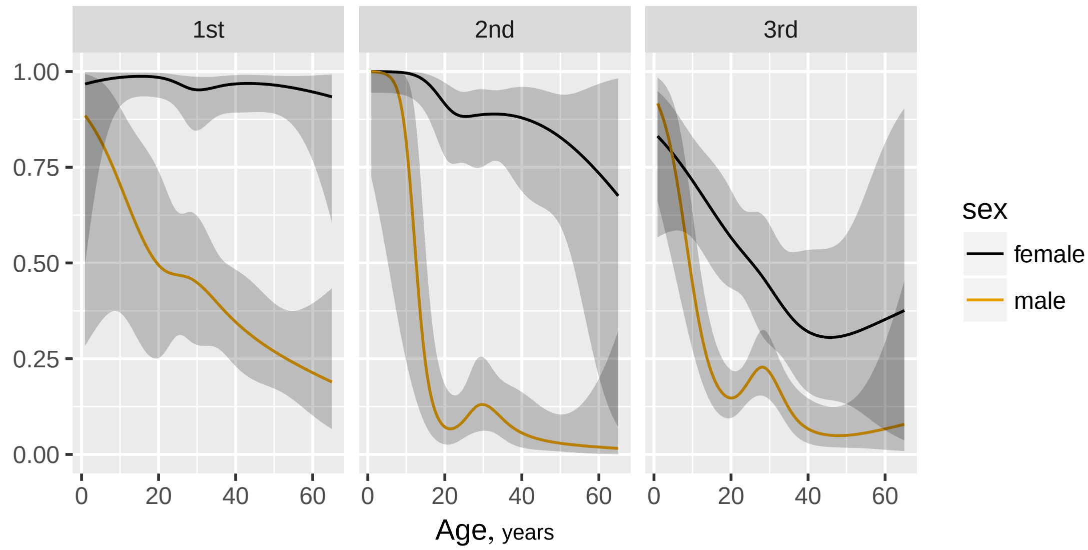


Figure 7.5: Effects of predictors on probability of survival of Titanic passengers, estimated for zero siblings or spouses

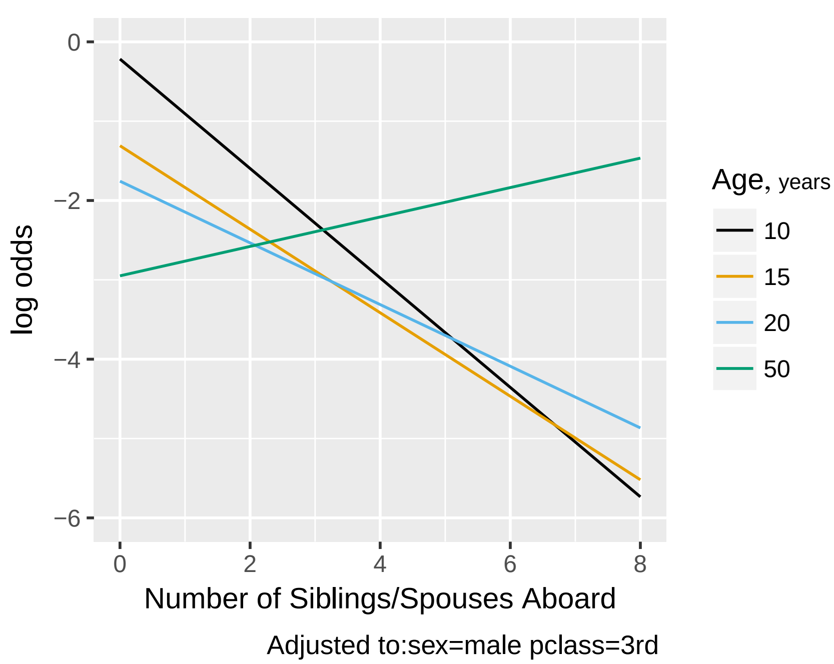


Figure 7.6: Effect of number of siblings and spouses on the log odds of surviving, for third class males

```
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 7.7
plot(cal, subtitles=FALSE)
```

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

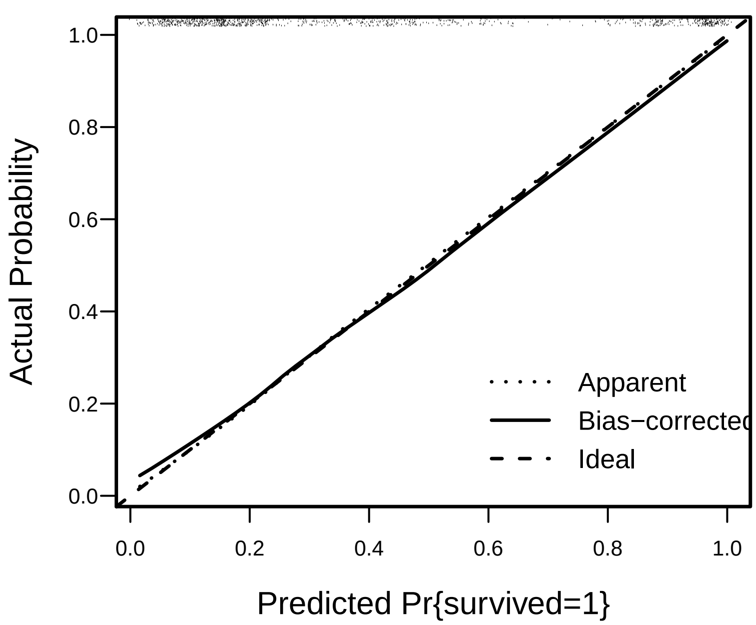


Figure 7.7: Bootstrap overfitting-corrected loess nonparametric calibration curve for casewise deletion model

But moderate problem with missing data

7.4 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 7.8
plot(na.patterns)
```

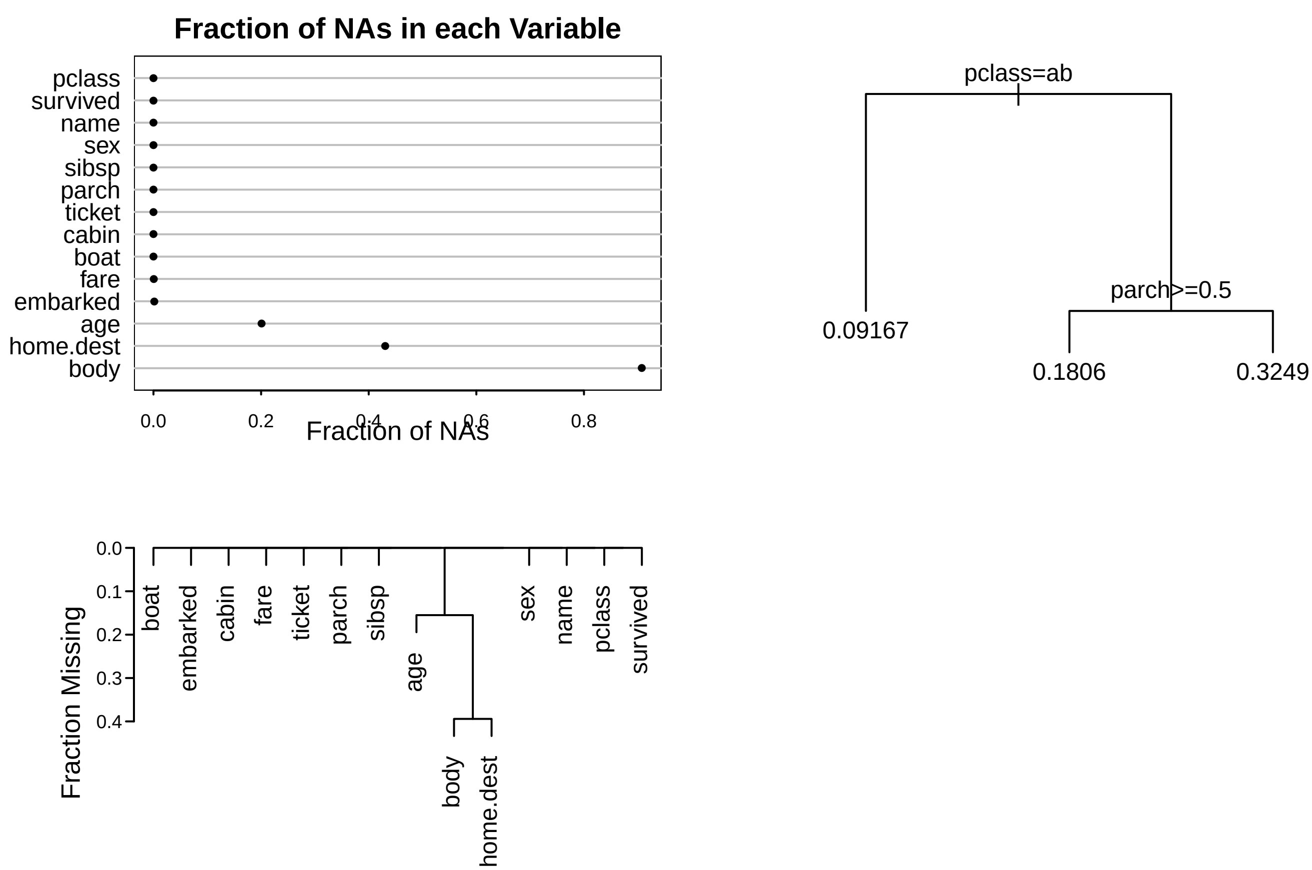


Figure 7.8: Patterns of missing data. Upper left panel shows the fraction of observations missing on each predictor. Lower panel depicts a hierarchical cluster analysis of missingness combinations. The similarity measure shown on the Y-axis is the fraction of observations for which both variables are 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 7.9
```

```
m ← lrm(is.na(age) ~ sex * pclass + survived + sibsp + parch,
        data=t3)
print(m, needspace='3.5in')
```

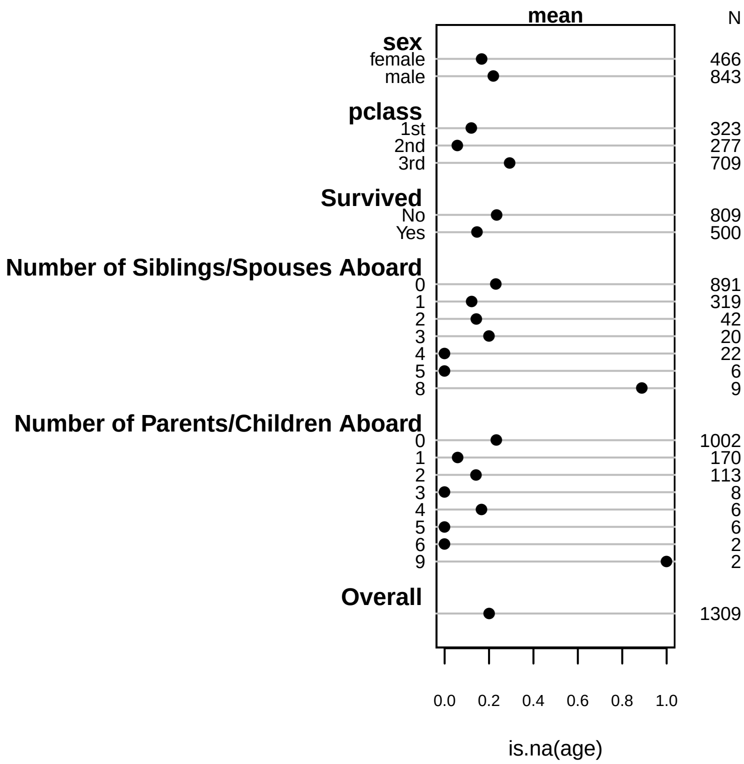


Figure 7.9: Univariable descriptions of proportion of passengers with missing age

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 7.3: Wald Statistics for `is.na(age)`

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

	$\hat{\beta}$	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

```
print(anova(m), table.env=TRUE, label='titanic-anova.na') # Table 7.3
```

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

7.5 Single Conditional Mean Imputation

D

First try: conditional mean imputation
Default spline transformation for age caused distribution of imputed values to be much different from non-imputed ones; constrain to linear

```
xtrans ← transcan(∼ I(age) + sex + pclass + sibsp + parch,
                  imputed=TRUE, pl=FALSE, pr=FALSE, data=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 distinct     Info     Mean      Gmd      .05      .10
     263        0       24     0.91    28.53    6.925    17.34    21.77
     .25      .50      .75      .90      .95
   26.17    28.10    28.10    42.77    42.77

lowest :  9.82894 11.75710 13.22440 15.15250 17.28300
highest: 33.24650 34.73840 38.63790 40.83950 42.76770

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)
```

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
$\max\lvert\frac{\partial \log L}{\partial \beta}\rvert$ 0.0004		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 7.10
```

```
print(anova(f.si), table.env=TRUE, label='titanic-anova.si')   # Table 7.4
```

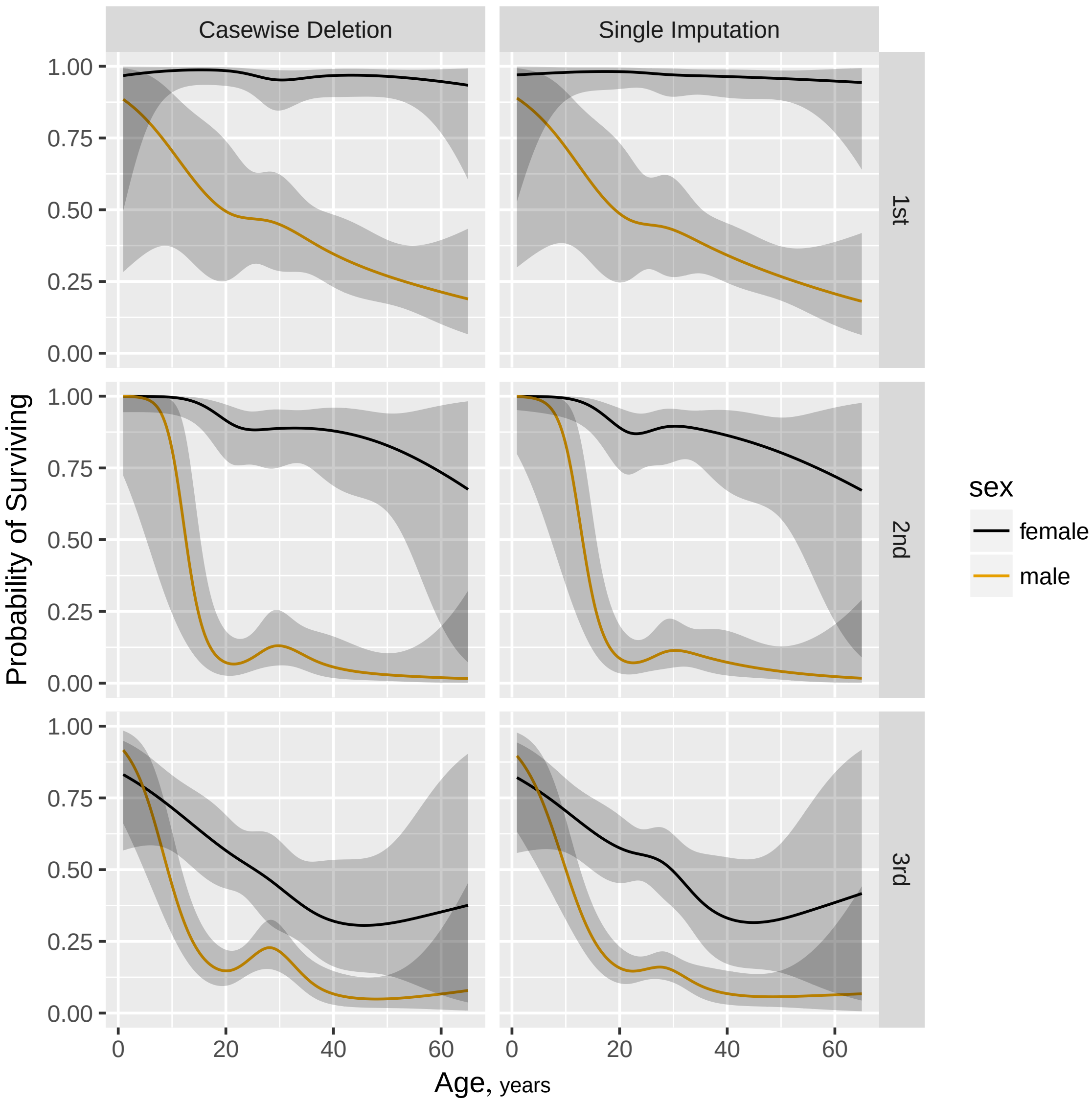


Figure 7.10: Predicted probability of survival for males from fit using casewise deletion (bottom) and single conditional mean imputation (top). `sibsp` is set to zero for these predicted values.

Table 7.4: 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

7.6 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).

E

```
plot(mi)
Ecdf(t3$age, add=TRUE, col='gray', lwd=2,
     subtitles=FALSE) #Fig. 7.11
```

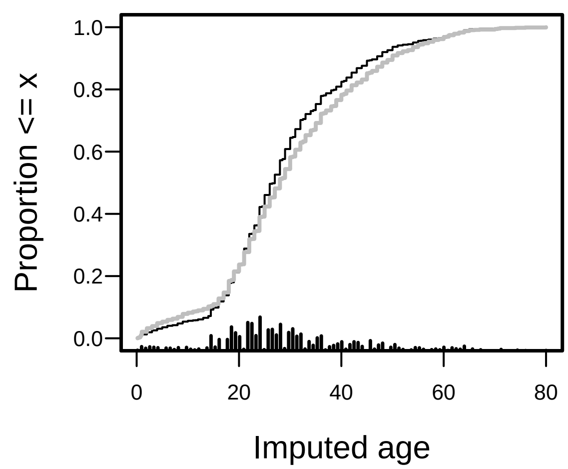


Figure 7.11: Distributions of imputed and actual ages for the Titanic dataset. Imputed values are in black and actual ages in gray.

Fit logistic models for 20 completed datasets and print the ratio of imputation-corrected variances to average ordinary variances

F

```
f.mi ← fit.mult.impute(
  survived ∼ (sex + pclass + rcs(age,5))^2 +
  rcs(age,5)*sibsp,
  lrm, mi, data=t3, pr=FALSE)
print(anova(f.mi), table.env=TRUE, label='titanic-anova.mi',
      size='small')    # Table 7.5
```

The Wald χ^2 for age is reduced by accounting for imputation

Table 7.5: 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

but is increased by using patterns of association with survival status to impute missing age. G

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 7.12
```

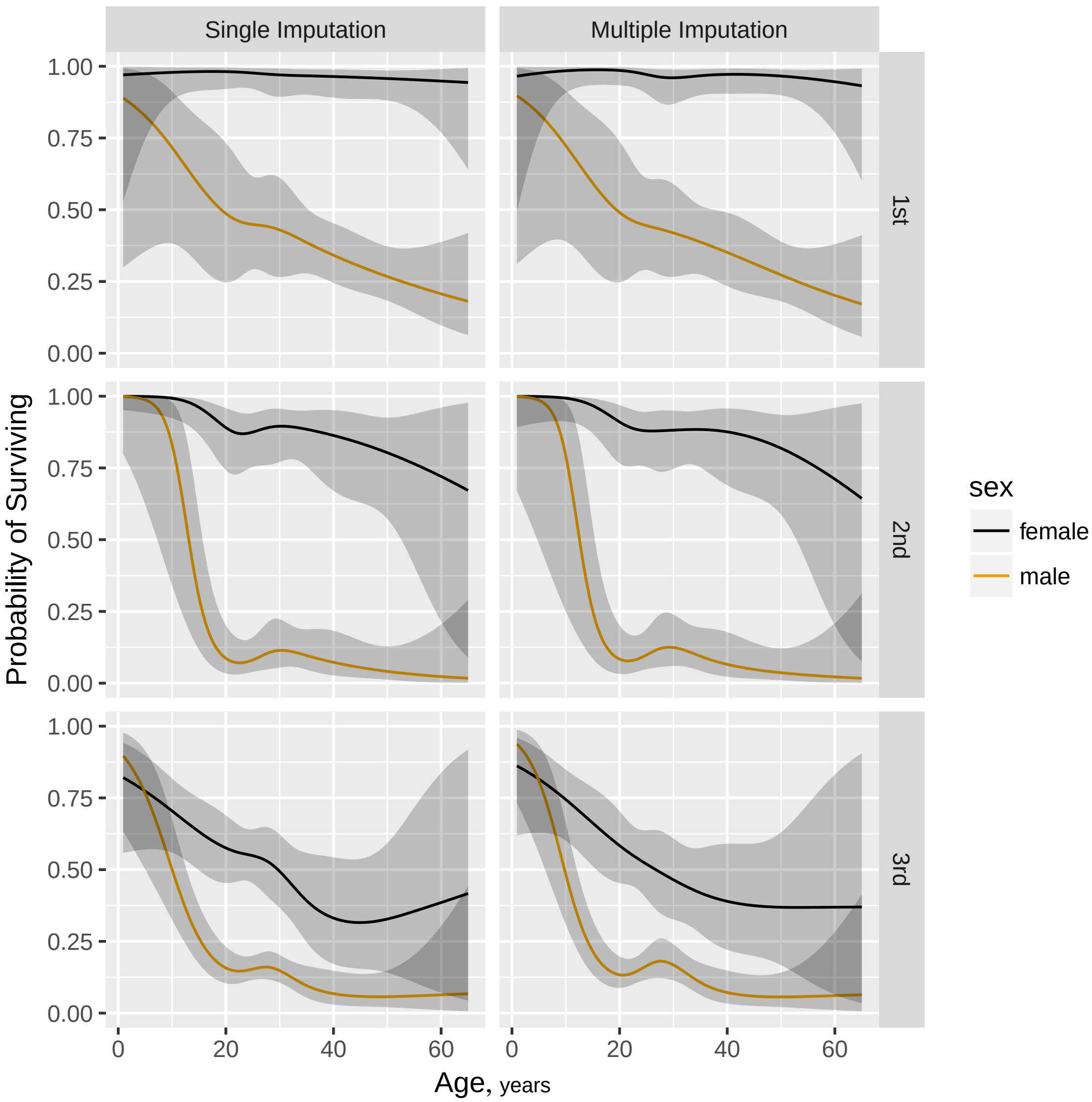


Figure 7.12: Predicted probability of survival for males from fit using single conditional mean imputation again (top) and multiple random draw imputation (bottom). Both sets of predictions are for `sibsp`=0.

7.7 Summarizing the Fitted Model

Show odds ratios for changes in predictor values

H

```
# 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 7.13
```

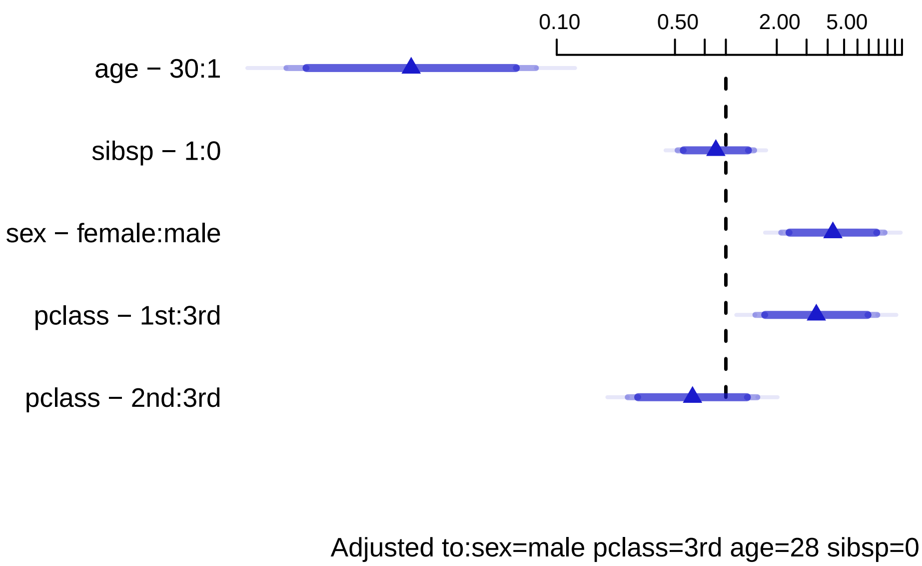


Figure 7.13: Odds ratios for some predictor settings

```
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)
```

```
   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. J

R Software Used

Package	Purpose	Functions
Hmisc	Miscellaneous functions	summary,plsmo,naclus,llist,latex
		summarize,Dotplot,describe
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 8

Regression Models for Continuous Y and Case Study in Ordinal Regression

A

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 8.3)
- other robust regression models that, like quantile regres-

sion, use an objective different from minimizing the sum of squared errors [150]

- 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. B

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.

8.1 Dataset and Descriptive Statistics

C

- 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 D
  - 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[23] E
- age ≥ 80 coded as 80 by CDC
- Subset with age ≥ 21, neither diagnosed nor treated for DM

```
require(rms)
```

```
options(prType='latex')   # for print, summary, anova
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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4629	0	4629	1	56902	3501	52136	52633	54284	56930	59495	61079	61641

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

sex

n	missing	distinct
4629	0	2

Value	male	female
Frequency	2259	2370
Proportion	0.488	0.512

age : Age [years]

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4629	0	703	1	48.57	19.85	23.33	26.08	33.92	46.83	61.83	74.83	80.00

lowest : 21.00000 21.08333 21.16667 21.25000 21.33333, highest: 79.66667 79.75000 79.83333 79.91667 80.00000

re : Race/Ethnicity

n	missing	distinct
4629	0	5

```
Mexican American (832, 0.180), Other Hispanic (474, 0.102), Non-Hispanic White (2318, 0.501),
Non-Hispanic Black (756, 0.163), Other Race Including Multi-Racial (249, 0.054)
```

---

income : Family Income

n	missing	distinct
4389	240	14

```
[0,5000) (162, 0.037), [5000,10000) (216, 0.049), [10000,15000) (371, 0.085), [15000,20000)
(300, 0.068), [20000,25000) (374, 0.085), [25000,35000) (535, 0.122), [35000,45000) (421,
0.096), [45000,55000) (346, 0.079), [55000,65000) (257, 0.059), [65000,75000) (188, 0.043), >
20000 (149, 0.034), < 20000 (52, 0.012), [75000,100000) (399, 0.091), >= 100000 (619, 0.141)
```

---

wt : Weight [kg]

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4629	0	890	1	80.49	22.34	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4629	0	512	1	167.5	11.71	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4629	0	1994	1	28.59	6.965	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4474	155	216	1	38.39	4.301	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4502	127	156	1	37.01	3.116	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4499	130	290	1	32.87	5.475	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4465	164	716	1	97.62	17.18	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4295	334	342	1	18.94	9.463	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
3974	655	329	1	20.8	9.124	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4629	0	63	0.994	5.533	0.5411	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
```

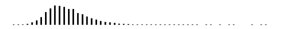
---

albumin : Albumin [g/dL]

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4576	53	26	0.99	4.261	0.3528	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4576	53	50	0.995	13.03	5.309	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
4576	53	167	1	0.8887	0.2697	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')
```

8.2 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 $\Sigma(Y_i - X\hat{\beta})^2$. F

- 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]

8.2.1 Checking Assumptions of OLS and Other Models

G

- First see if `gh` would make a Gaussian residuals model fit
- 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

[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.

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

```
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
```

H

- Upper right curves are not linear, implying that a normal conditional distribution cannot work for gh[b]

[b] They are not parallel either.

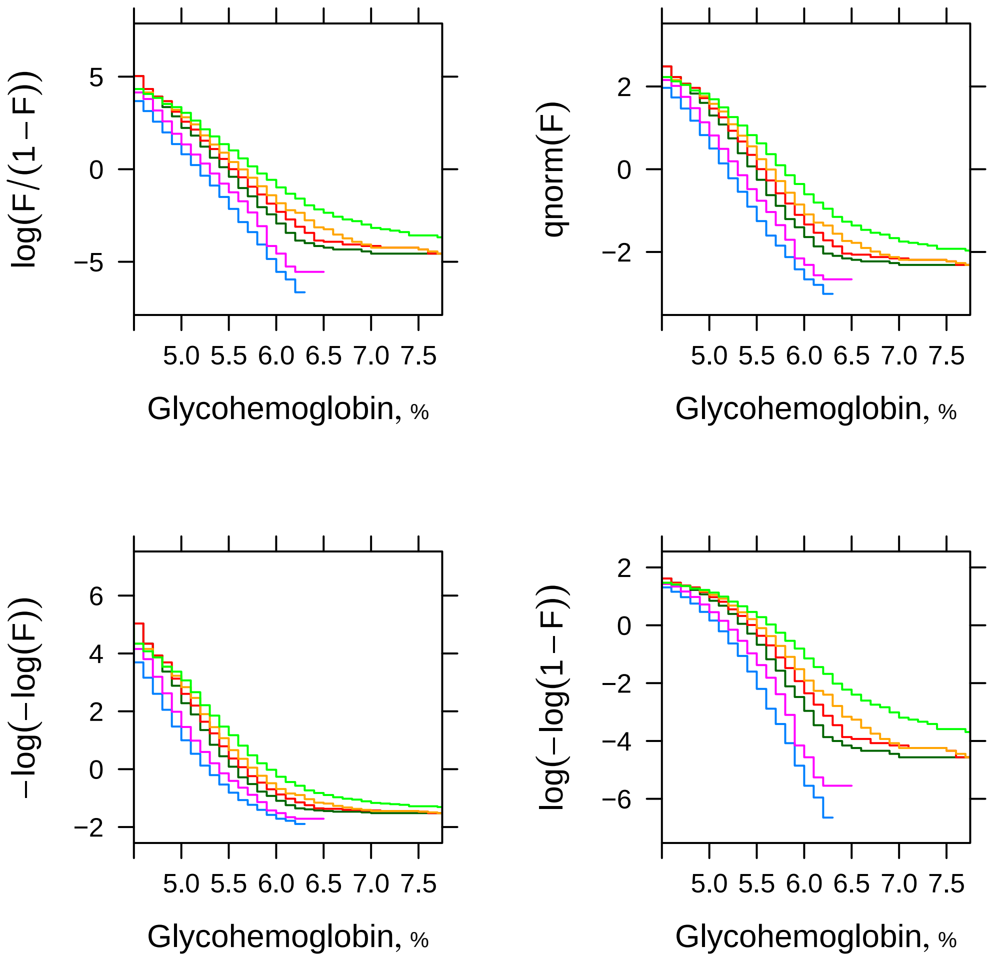


Figure 8.1: Examination of normality and constant variance assumption, and assumptions for various ordinal models

- There is non-parallelism for the logit model
- Other graphs will be used to guide selection of an ordinal model below

8.3

Quantile Regression

- Ruled out OLS and semiparametric proportional odds model
- Quantile regression [88, 87] 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 [89] 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.

8.4 Ordinal Regression Models for Continuous Y

K

- 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
- Manuguerra [104] devloped an ordinal model for continuous Y which they incorrectly labeled semi-parametric and is actually a lower-dimensional flexible parametric model that instead of having intercepts has a spline function of 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.

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 $\alpha_y = \alpha_i, i : y_i = y$. 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,

$$\begin{aligned} F^{-1}(\text{Prob}[Y \geq y | X_2]) - F^{-1}(\text{Prob}[Y \geq y | X_1]) \\ = (X_2 - X_1)\beta \end{aligned}$$

independent of the αs (parallelism assumption). If $F = [1 + \exp(-y)]^{-1}$, this is the proportional odds assumption.

[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.

[f] $\hat{\alpha_y}$ are unchanged if a constant is added to all y.

Table 8.1: 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	

Common choices of F, implemented in the `rms` `orm` function, are shown in Table 8.1. M

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 N

$$\sum_{i=1}^{k} y_i \hat{\text{Prob}}[Y = y_i|X]$$

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

[g] The intercepts have to be shifted to the left one position in solving this equation because the quantile is such that $\text{Prob}[Y \leq y] = q$ whereas the model is stated in terms of $\text{Prob}[Y \geq y]$.

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 O

- estimation of quantiles as efficiently as quantile regression if the parallel slopes assumptions hold
- efficient estimation of mean 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 semiparametric models, consider the ordinary linear model or its special case the equal variance two-sample P

[h] But it is not sensible to estimate quantiles of Y when there are heavy ties in Y in the area containing the quantile.

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 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 the following:

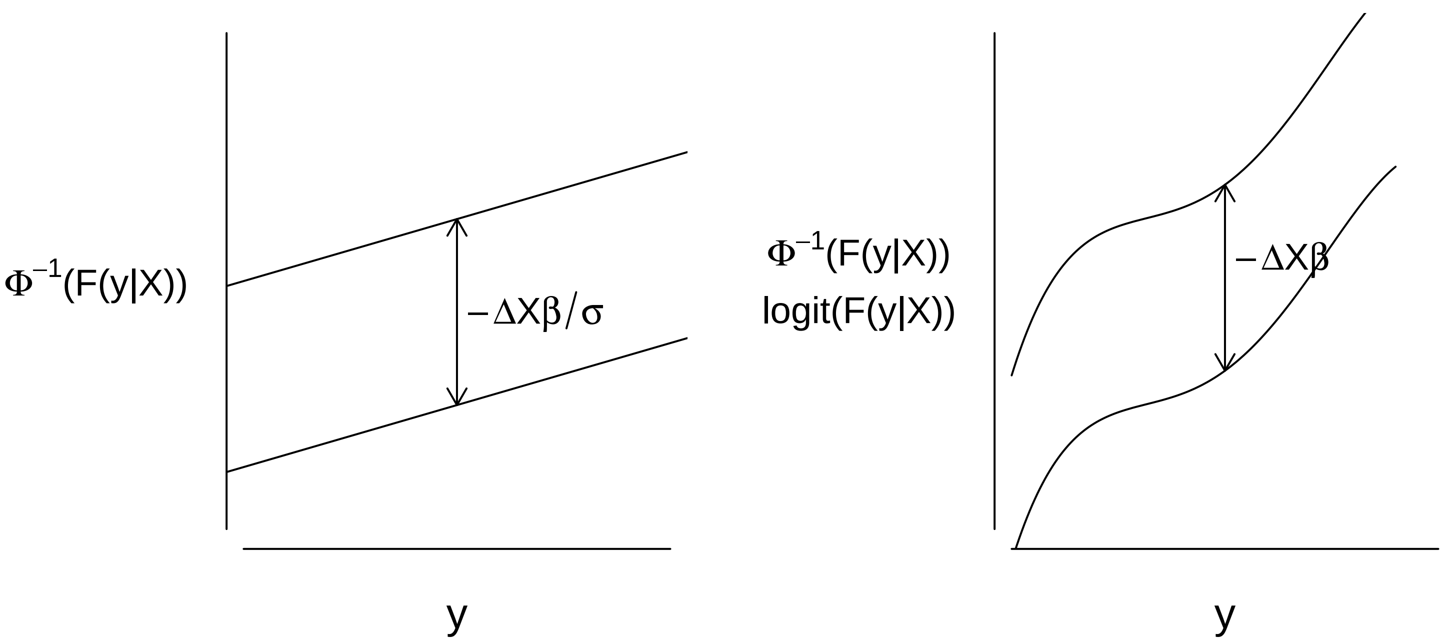


Figure 8.2: Assumptions of the linear model (left panel) and semiparametric ordinal probit or logit (proportional odds) models (right panel). Ordinal models do not assume any shape for the distribution of Y for a given X; they only assume parallelism.

$$\mathrm{Prob}[Y \leq y|X] = F(g(y) - X\beta),$$

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.

Global Modeling Implications

Q

- Ordinal regression invariant to choice of transformation of Y
- Y needs to be ordinal
- Difference in two ordinal variables is not necessarily ordinal
- $\rightarrow$ Never analyze differences in regression
- Y=final value, adjust for baseline values as covariates

8.5

Ordinal Regression Applied to HbA$_{1c}$

R

- In Figure 8.1, 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)

8.5.1

Checking Fit for Various Models Using Age

S

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 8.3
       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 8.3 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 T

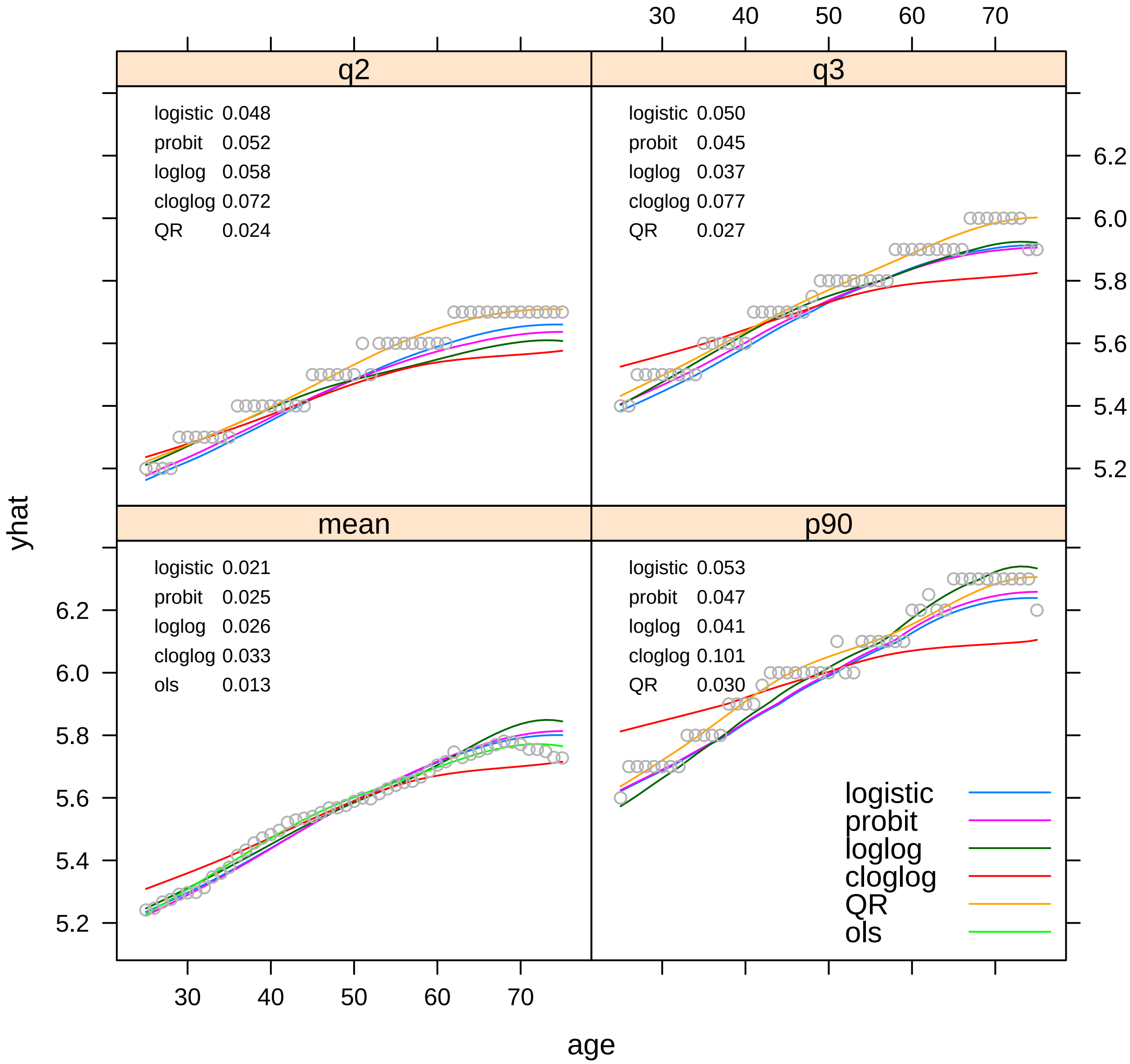


Figure 8.3: Three estimated quantiles and estimated mean using 6 methods, compared against caliper-matched sample quantiles/means (circles). Numbers are mean absolute differences between predicted and sample quantities using overlapping intervals of age and caliper matching. QR:quantile regression.

probability model did not estimate the median as accurately as some other methods, 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 8.1, let's look at quantile groups of predicted HbA$_{1c}$ by OLS and plot predicted distributions of actual HbA$_{1c}$ against empirical distributions. U

```
w$pghg ← cut2(pgh, g=6)
f   ← orm(gh ~ pghg, family=loglog, data=w)
lp  ← predict(f, newdata=data.frame(pghg=levels(w$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. 8.4
```

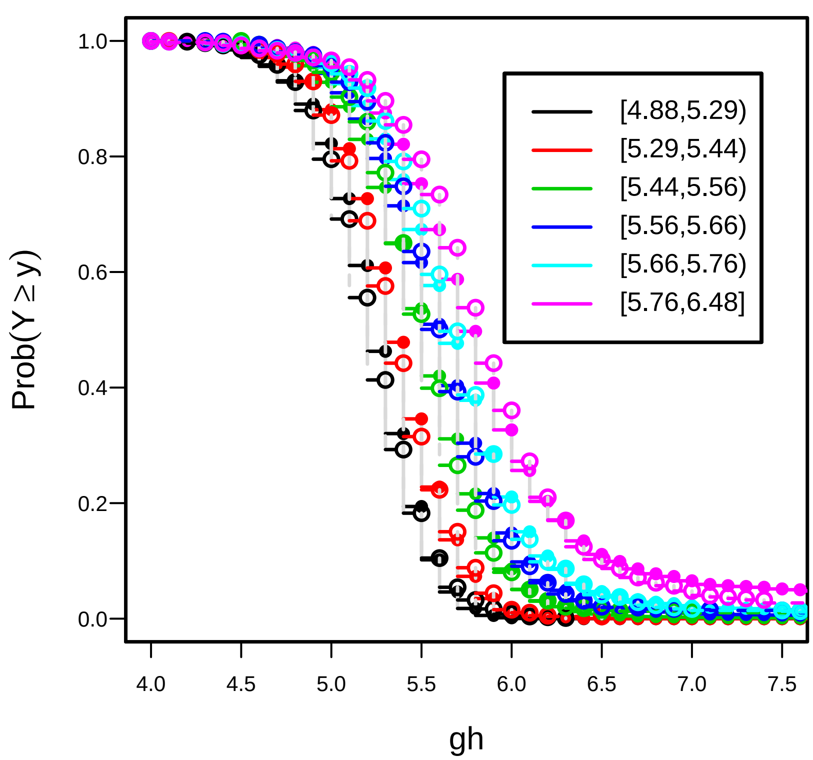


Figure 8.4: Observed (dashed lines, open circles) and predicted (solid lines, closed circles) exceedance probability distributions from a model using 6-tiles of OLS-predicted HbA$_{1c}$. Key shows quantile group intervals of predicted mean HbA$_{1c}$.

Agreement between predicted and observed exceedance probability distributions is excellent in Figure 8.4.

To return to the initial look at a linear model with assumed Gaussian residuals, fit a probit ordinal model and compare the estimated intercepts to the linear relationship with `gh` that is assumed by the normal distribution. v

```
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. 8.5
     xlab=expression(-y/hat(sigma)), ylab=expression(alpha[y]))
```

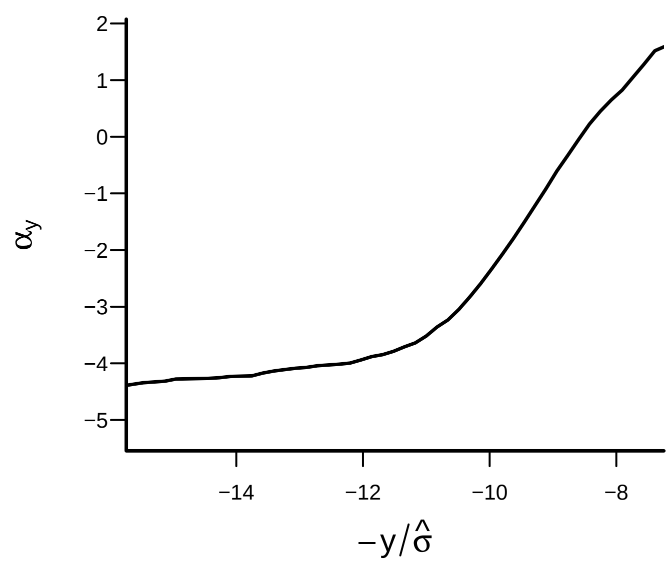


Figure 8.5: Estimated intercepts from probit model

Figure 8.5 depicts a significant departure from that implied by Gaussian residuals.

8.5.2

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`. w

- 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)
f
```

-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		
$\max\lvert\frac{\partial \log L}{\partial \beta}\rvert$	1×10^{-6}	Score χ^2	1262.81	$\lvert\Pr(Y \geq Y_{0.5}) - \frac{1}{2}\rvert$	0.153		
		Pr($>\chi^2$)	<0.0001				

	$\hat{\beta}$	S.E.	Wald Z	Pr($>\lvert Z\rvert$)
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. x

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 [83].

```
print(cph(Surv(gh) ~ rcs(age,5) + log(ht) + log(wt), data=w))
```

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		

	$\hat{\beta}$	S.E.	Wald Z	Pr$(>\|Z\|)$
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

	$\hat{\beta}$	S.E.	Wald Z	Pr($>|Z|$)
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. Y

```
v ← varclus(~ wt + ht + bmi + leg + arml + armc + waist +
            tri + sub + age + sex + re, data=w)
plot(v)   # Figure 8.6
# 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.

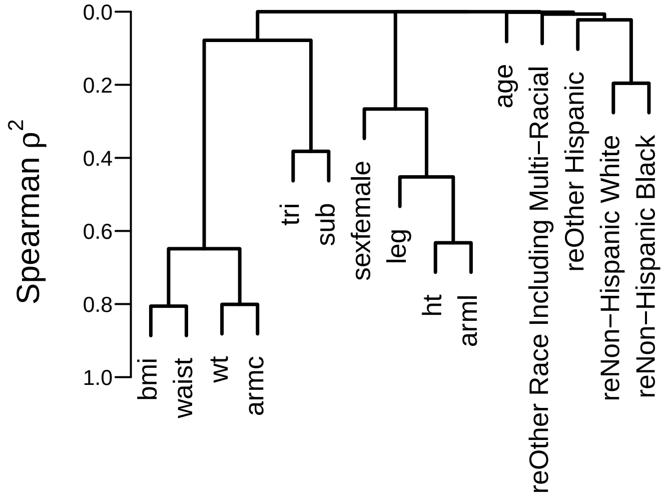


Figure 8.6: Variable clustering for all potential predictors

An advantage of removing height is that it is age-dependent in the elderly: z

```
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')
```

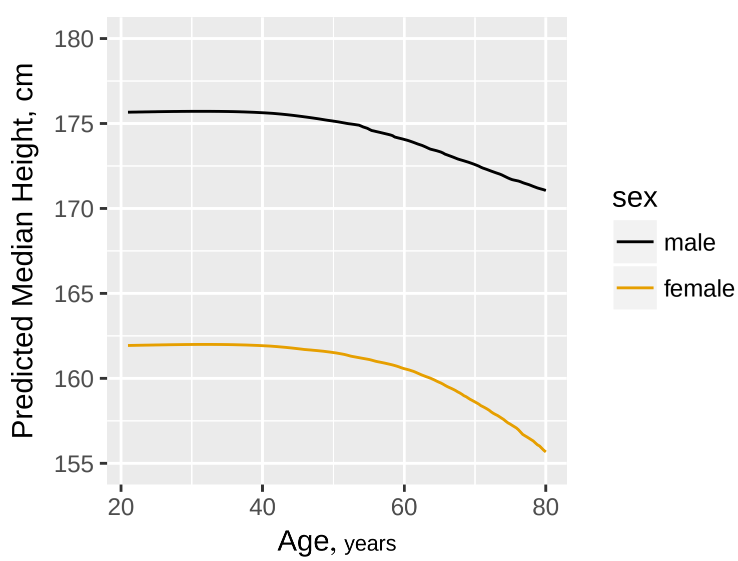


Figure 8.7: Estimated median height as a smooth function of age, allowing age to interact with sex, from a proportional odds model

But also see a change in leg length:

```
f ← orm(leg ~ rcs(age,4)*sex, data=w)
qu ← Quantile(f); med ← function(x) qu(.5, x)
ggplot(Predict(f, age, sex, fun=med, conf.int=FALSE),
```

```
       ylab='Predicted Median Upper Leg Length, cm')
```

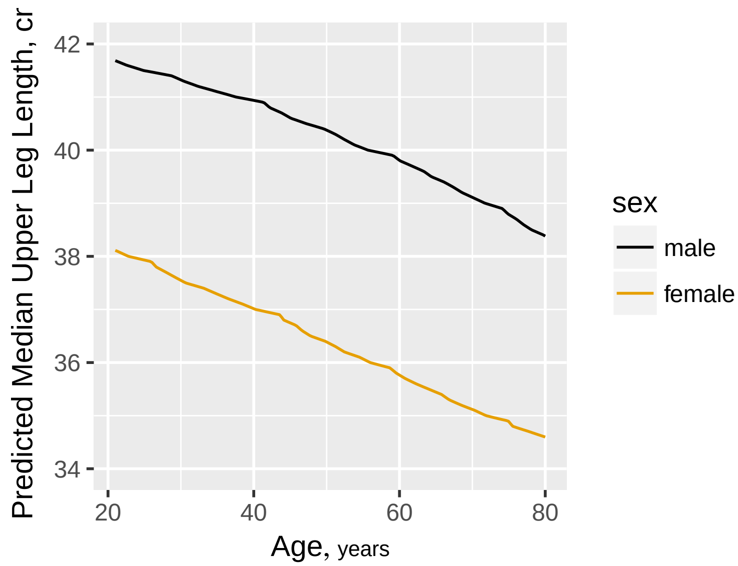


Figure 8.8: Estimated median upper leg length as a smooth function of age, allowing age to interact with sex, from a proportional odds model

Next allocate d.f. according to generalized Spearman ρ^2[i].

A

```
s ← spearman2(gh ~ age + sex + re + wt + leg + arml + armc +
                 waist + tri + sub, data=w, p=2)
plot(s)
```

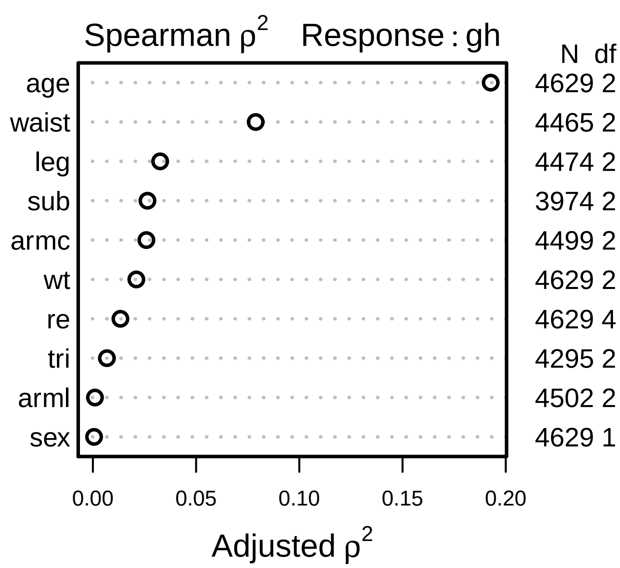


Figure 8.9: Generalized squared rank correlations

Parameters will be allocated in descending order of ρ^2. But note that subscapular skinfold has a large number of `NA`s and

[i] Competition between collinear size measures hurts interpretation of partial tests of association in a saturated additive model.

other predictors also have `NA`s. Suboptimal casewise deletion will be used until the final model is fitted.

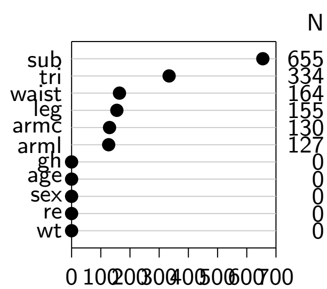
Because there are many competing body measures, we use backwards stepdown to arrive at a set of predictors. The bootstrap will be used to penalize predictive ability for variable selection. First the full model is fit using casewise deletion, then we do a composite test to assess whether any of the frequently–missing predictors is important. B

```
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, 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:
anova(f, leg, arml, armc, waist, tri, sub)
```

Wald Statistics for `gh`

	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). C

```
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)
```

Next do fast backward step-down in an attempt to get a model without so much competition amoung variables. The stepwise selection will be penalized for in the model validation. D

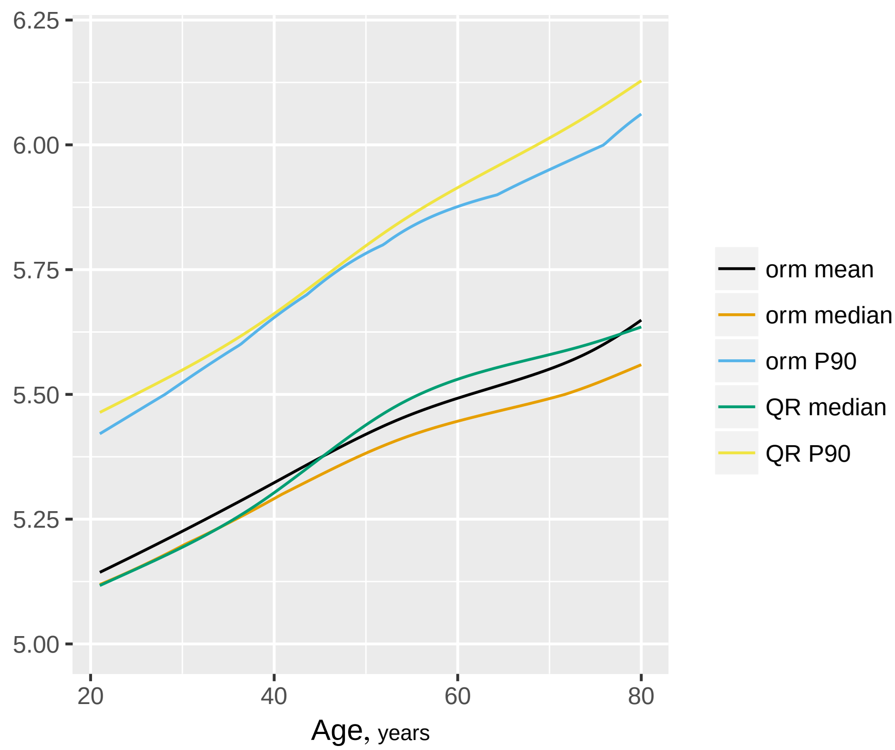


Figure 8.10: Estimated mean and 0.5 and 0.9 quantiles from the log-log ordinal model using casewise deletion, along with predictions of 0.5 and 0.9 quantiles from quantile regression (QR). Age is varied and other predictors are held constant to medians/modes.

```
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
```

Validate the model, properly penalizing for variable selection

E

```
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. F

Do an ANOVA for the reduced model, taking imputation into account. G

```
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, 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		

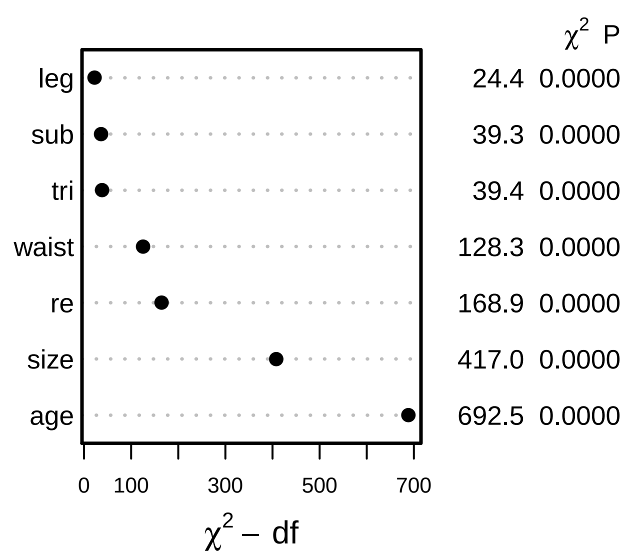
	$\hat{\beta}$	S.E.	Wald Z	Pr($>\lvert Z\rvert$)
age	0.0404	0.0055	7.29	<0.0001
age'	-0.0228	0.0279	-0.82	0.4137
age''	0.0126	0.0876	0.14	0.8857
age'''	0.0424	0.1148	0.37	0.7116
re=Other Hispanic	-0.0766	0.0597	-1.28	0.1992
re=Non-Hispanic White	-0.4121	0.0449	-9.17	<0.0001
re=Non-Hispanic Black	0.0645	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)
print(an, caption='ANOVA for reduced model after multiple imputation, with
      addition of a combined effect for four size variables')
```

ANOVA for reduced model after multiple imputation, with addition of a combined effect for four size variables

	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)
```




```
ggplot(Predict(g), abbrev=TRUE, ylab=NULL)   # Figure 8.11
```

```
M ← Mean(g)
ggplot(Predict(g, fun=M), abbrev=TRUE, ylab=NULL)   # Figure 8.12
```

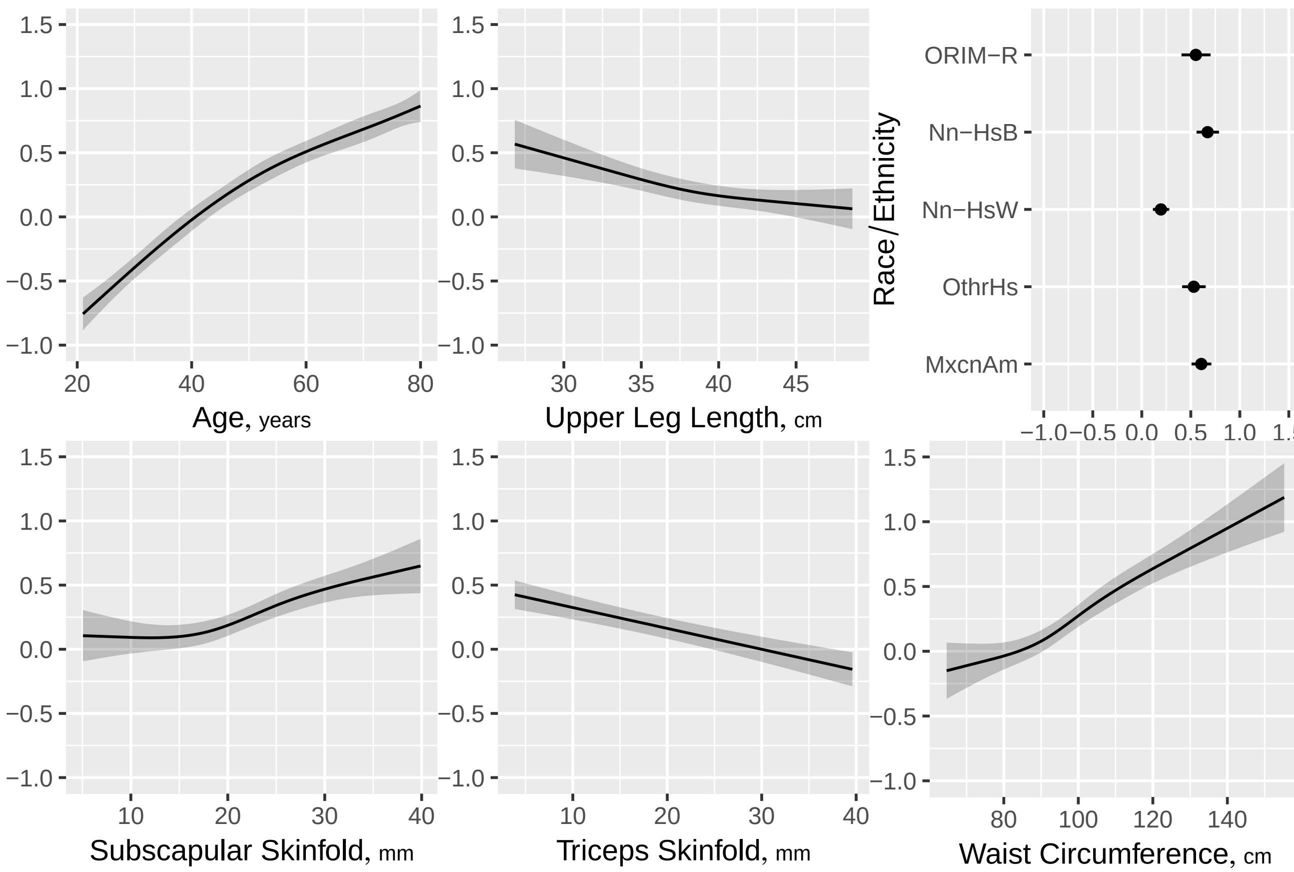


Figure 8.11: Partial effects (log hazard or log-log cumulative probability scale) of all predictors in reduced model, after multiple imputation

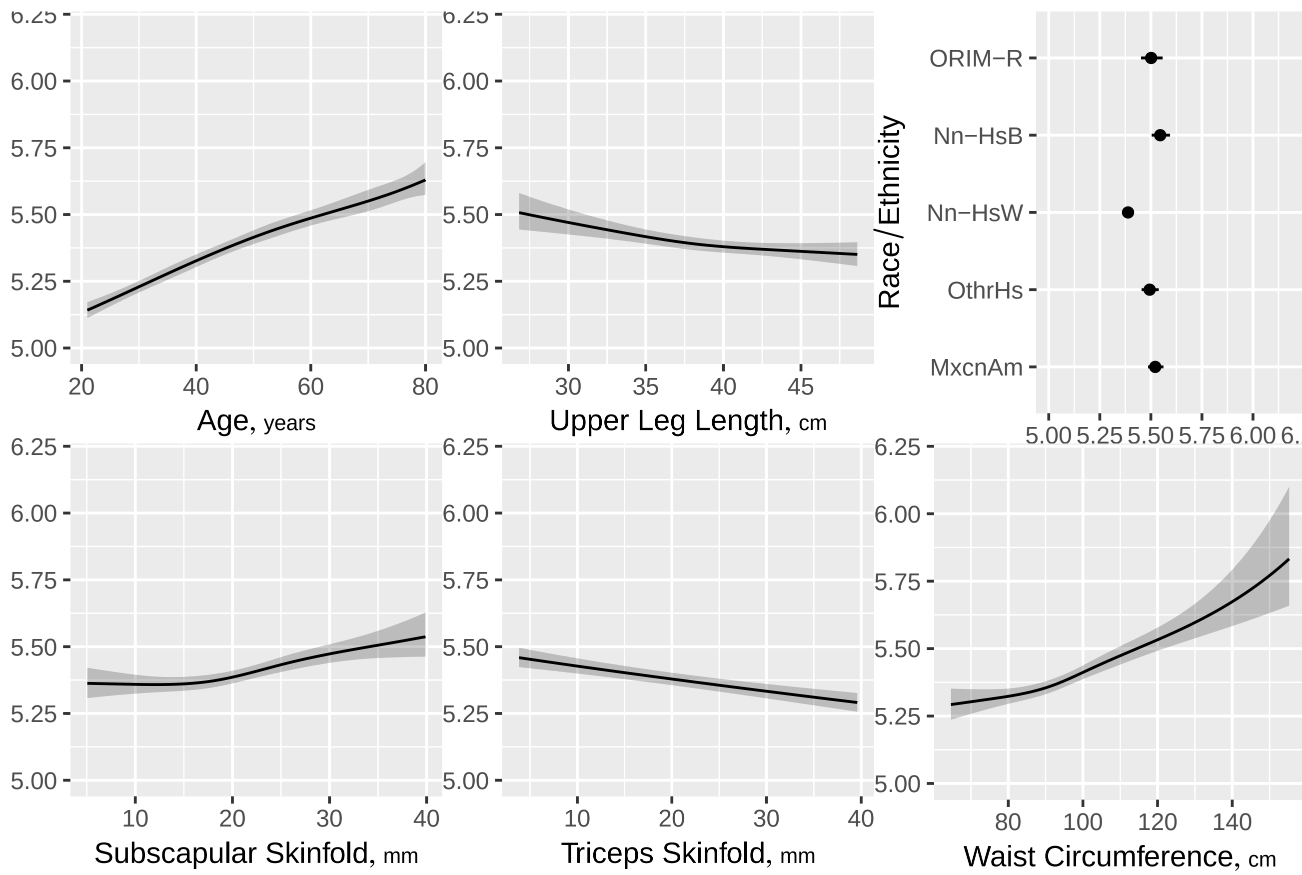


Figure 8.12: Partial effects (mean scale) of all predictors in reduced model, after multiple imputation

Compare the estimated age partial effects and confidence intervals with those from a model using casewise deletion, and with bootstrap nonparametric confidence intervals (also with casewise deletion). J

```
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 8.13
```

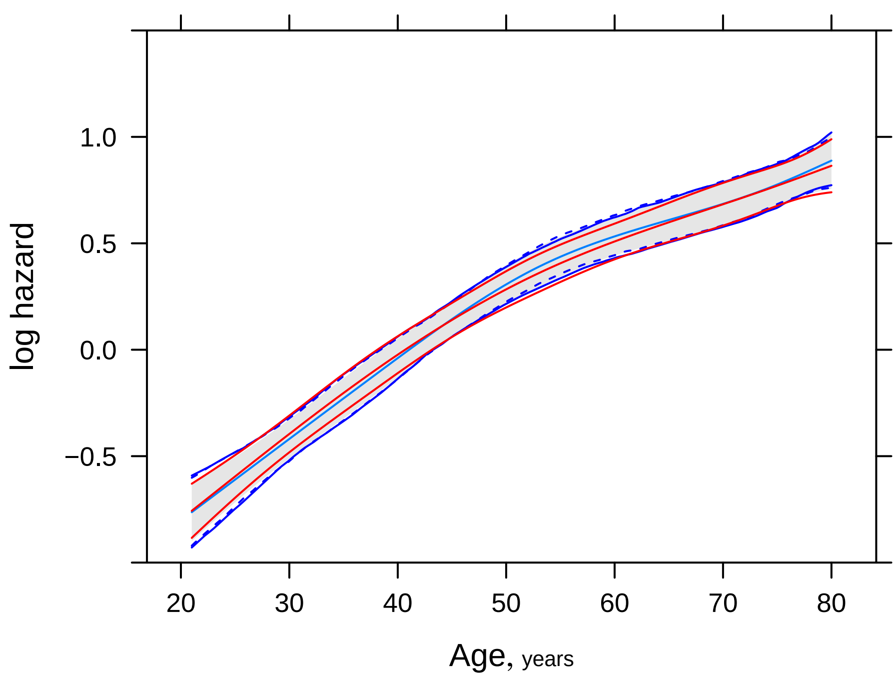


Figure 8.13: Partial effect for age from multiple imputation (center red line) and casewise deletion (center blue line) with symmetric Wald 0.95 confidence bands using casewise deletion (gray shaded area), basic bootstrap confidence bands using casewise deletion (blue lines), percentile bootstrap confidence bands using casewise deletion (dashed blue lines), and symmetric Wald confidence bands accounting for multiple imputation (red lines).

In OLS the mean equals the median and both are linearly related to any other quantiles. Semiparametric models are not this restrictive:

```
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 8.14
     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)
```

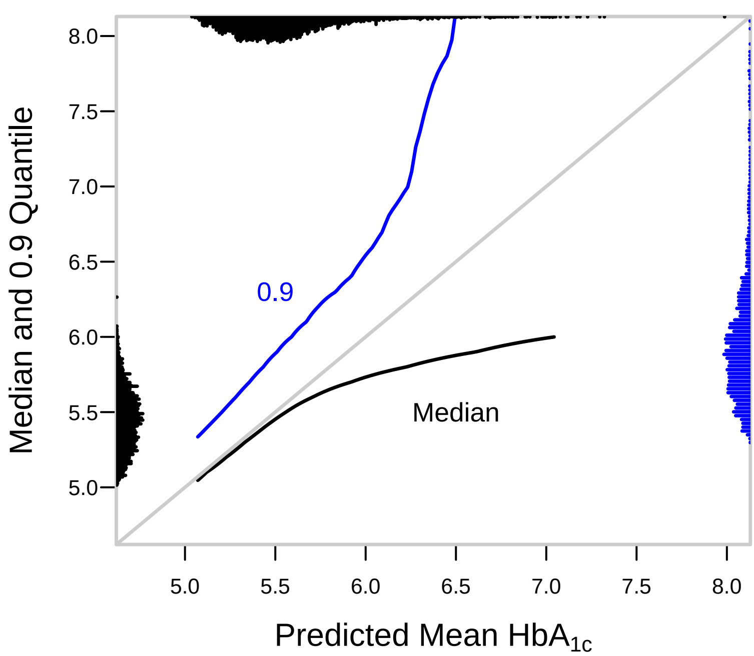


Figure 8.14: Predicted mean HbA_{1c}vs. predicted median and 0.9 quantile along with their marginal distributions

Draw a nomogram to compute 7 different predicted values for each subject. L

```
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 8.15
```

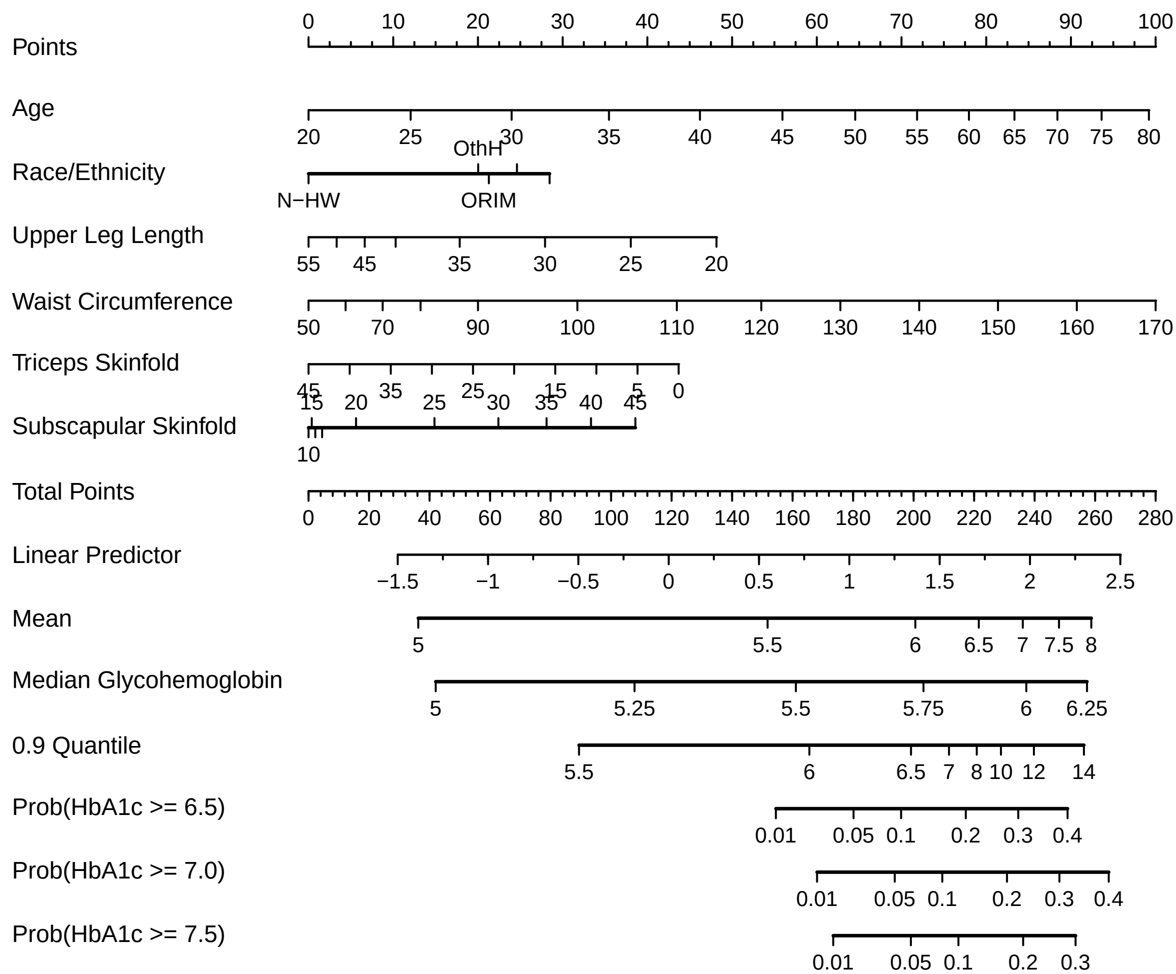


Figure 8.15: Nomogram for predicting median, mean, and 0.9 quantile of glycohemoglobin, along with the estimated probability that HbA$_{1c} \geq 6.5, 7$, or 7.5, all from the log-log ordinal model

Chapter 9

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

- 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

9.1

Descriptive Statistics

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

```
require(rms)
```

```
options(prType='latex')      # for print, summary, anova
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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
537	0	529	1	60.7	19.98	28.49	35.22	47.93	63.67	74.49	81.54	85.56

```
lowest : 18.04199 18.41499 19.76500 20.29599 20.31200, highest: 91.61896 91.81696 91.93396 92.73895 95.50995
```

death : Death at any time up to NDI date:31DEC94

n	missing	distinct	Info	Sum	Mean	Gmd
537	0	2	0.67	356	0.6629	0.4477

sex

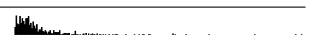
n	missing	distinct
537	0	2

Value	female	male
Frequency	251	286
Proportion	0.467	0.533

hospdead : Death in Hospital

n	missing	distinct	Info	Sum	Mean	Gmd
537	0	2	0.703	201	0.3743	0.4693

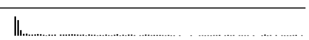
slos : Days from Study Entry to Discharge



n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
537	0	85	0.999	23.44	22.24	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
```

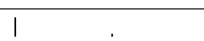
d.time : Days of Follow-Up



n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
537	0	340	1	446.1	566.1	4	6	16	182	724	1421	1742

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

dzgroup



n	missing	distinct
537	0	3

Value	ARF/MOSF w/Sepsis	Coma	MOSF w/Malig
Frequency	391	60	86
Proportion	0.728	0.112	0.160

dzclass

n	missing	distinct
537	0	2

Value	ARF/MOSF	Coma
Frequency	477	60
Proportion	0.888	0.112

num.co : number of comorbidities

n	missing	distinct	Info	Mean	Gmd
537	0	7	0.926	1.525	1.346

```
Value          0     1     2     3     4     5     6
Frequency    111   196   133    51    31    10     5
Proportion 0.207 0.365 0.248 0.095 0.058 0.019 0.009
```

---

edu : Years of Education

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
411	126	22	0.957	12.03	3.581	7	8	10	12	14	16	17

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

---

income

n	missing	distinct
335	202	4

```
Value      under $11k  $11-$25k  $25-$50k     >$50k
Frequency         158        79        63        35
Proportion      0.472     0.236     0.188     0.104
```

---

scoma : SUPPORT Coma Score based on Glasgow D3

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
537	0	11	0.822	19.24	27.87	0	0	0	0	37	55	100

```
Value          0     9    26    37    41    44    55    61    89    94   100
Frequency    301    50    44    19    17    43    11     6     8     6    32
Proportion 0.561 0.093 0.082 0.035 0.032 0.080 0.020 0.011 0.015 0.011 0.060
```

---

charges : Hospital Charges

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
517	20	516	1	86652	90079	11075	15180	27389	51079	100904	205562	283411

```
lowest :   3448.0   4432.0   4574.0   5555.0   5849.0, highest: 504659.5 538323.0 543761.0 706577.0 740010.0
```

---

totcst : Total RCC cost

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
471	66	471	1	46360	46195	6359	8449	15412	29308	57028	108927	141569

```
lowest :      0.000   2071.109   2522.451   3190.625   3325.350
highest: 269057.000 269131.250 338955.000 357918.750 390460.500
```

---

totmcst : Total micro-cost

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
331	206	328	1	39022	36200	6131	8283	14415	26323	54102	87495	111920

```
lowest :      0.000   1561.619   2477.510   2626.270   3421.068
highest: 144234.000 154709.000 198047.000 234875.875 271467.250
```

---

avtisst : Average TISS, Days 3-25

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
536	1	205	1	29.83	14.19	12.46	14.50	19.62	28.00	39.00	47.17	50.37

```
lowest :  4.000000  5.666664  8.000000  9.000000  9.500000
highest: 58.500000 59.000000 60.000000 61.000000 64.000000
```

---

race

n	missing	distinct
535	2	5

```
Value         white    black    asian    other hispanic
Frequency       417       84        4        8       22
Proportion    0.779    0.157    0.007    0.015    0.041
```

---

meanbp : Mean Arterial Blood Pressure Day 3

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
537	0	109	1	83.28	35	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
532	5	241	1	14.1	9.984	0.8999	4.5000	7.9749	12.3984	18.1992	25.1891	30.1873

```
lowest :   0.04999542   0.06999207   0.09999084   0.14999390   0.19998169
highest:  51.39843750  58.19531250  61.19531250  79.39062500 100.00000000
```

---

hrt : Heart Rate Day 3

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
537	0	111	0.999	105	38.59	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
537	0	45	0.997	23.72	12.65	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
537	0	61	0.999	37.52	1.505	35.50	35.80	36.40	37.80	38.50	39.09	39.50

```
lowest : 32.50000 34.00000 34.09375 34.89844 35.00000, highest: 40.19531 40.59375 40.89844 41.00000 41.19531
```

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

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
500	37	357	1	227.2	125	86.99	105.08	137.88	202.56	290.00	390.49	433.31

```
lowest :  45.00000  48.00000  53.32812  54.00000  55.00000
highest: 574.00000 595.12500 640.00000 680.00000 869.37500
```

alb : Serum Albumin Day 3

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
346	191	34	0.997	2.668	0.7219	1.700	1.900	2.225	2.600	3.100	3.400	3.800

```
lowest : 1.099854 1.199951 1.299805 1.399902 1.500000, highest: 4.099609 4.199219 4.500000 4.699219 4.799805
```

bili : Bilirubin Day 3

```
       n  missing distinct     Info     Mean      Gmd      .05      .10      .25      .50
     386      151       88    0.997    2.678    3.507   0.3000   0.4000   0.6000   0.8999
     .75      .90      .95
  2.0000   6.5996  13.1743

lowest :  0.09999084  0.19998169  0.29998779  0.39996338  0.50000000
highest: 22.59765620 30.00000000 31.50000000 35.00000000 39.29687500
```

crea : Serum creatinine Day 3

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
537	0	84	0.998	2.232	1.997	0.6000	0.7000	0.8999	1.3999	2.5996	5.2395	7.3197

```
lowest :  0.2999878  0.3999634  0.5000000  0.5999756  0.6999512
highest: 10.3984375 10.5996094 11.1992188 11.5996094 11.7988281
```

sod : Serum sodium Day 3

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
537	0	38	0.997	138.1	7.471	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
500	37	49	0.998	7.416	0.08775	7.270	7.319	7.380	7.420	7.470	7.510	7.529

```
lowest : 6.959961 6.989258 7.069336 7.119141 7.129883, highest: 7.559570 7.569336 7.589844 7.599609 7.659180
```

glucose : Glucose Day 3

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
297	240	179	1	167.7	92.13	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
304	233	100	1	38.91	31.12	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	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
303	234	262	1	2095	1579	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	distinct	Info	Mean	Gmd
104	433	8	0.875	1.577	2.152

```
Value          0     1     2     3     4     5     6     7
Frequency     51    19     7     6     4     7     8     2
Proportion 0.490 0.183 0.067 0.058 0.038 0.067 0.077 0.019
```

adls : ADL Surrogate Day 3

```
   n  missing  distinct     Info     Mean      Gmd
 392      145         8    0.888     1.86    2.466


Value          0     1     2     3     4     5     6     7
Frequency    185    68    22    18    17    20    39    23
Proportion 0.472 0.173 0.056 0.046 0.043 0.051 0.099 0.059
```

sfdm2

```
   n  missing  distinct
 468       69         5


Value      no(M2 and SIP pres) adl>=4 (>=5 if sur)              SIP>=30        Coma or Intub
Frequency                  134                  78                   30                    5
Proportion               0.286               0.167                0.064                0.011

Value         <2 mo. follow-up
Frequency                  221
Proportion               0.472
```

adlsc : Imputed ADL Calibrated to Surrogate

```
   n  missing  distinct     Info     Mean      Gmd      .05      .10      .25      .50      .75      .90      .95
 537        0       144    0.956    2.119    2.386    0.000    0.000    0.000    1.839    3.375    6.000    6.000


lowest : 0.0000000 0.4947510 0.4947999 1.0000000 1.1667481
highest: 5.7832031 6.0000000 6.3398438 6.4658203 7.0000000
```

B

```
# Show patterns of missing data
plot(naclus(support[acute,]))      # Figure 9.1
```

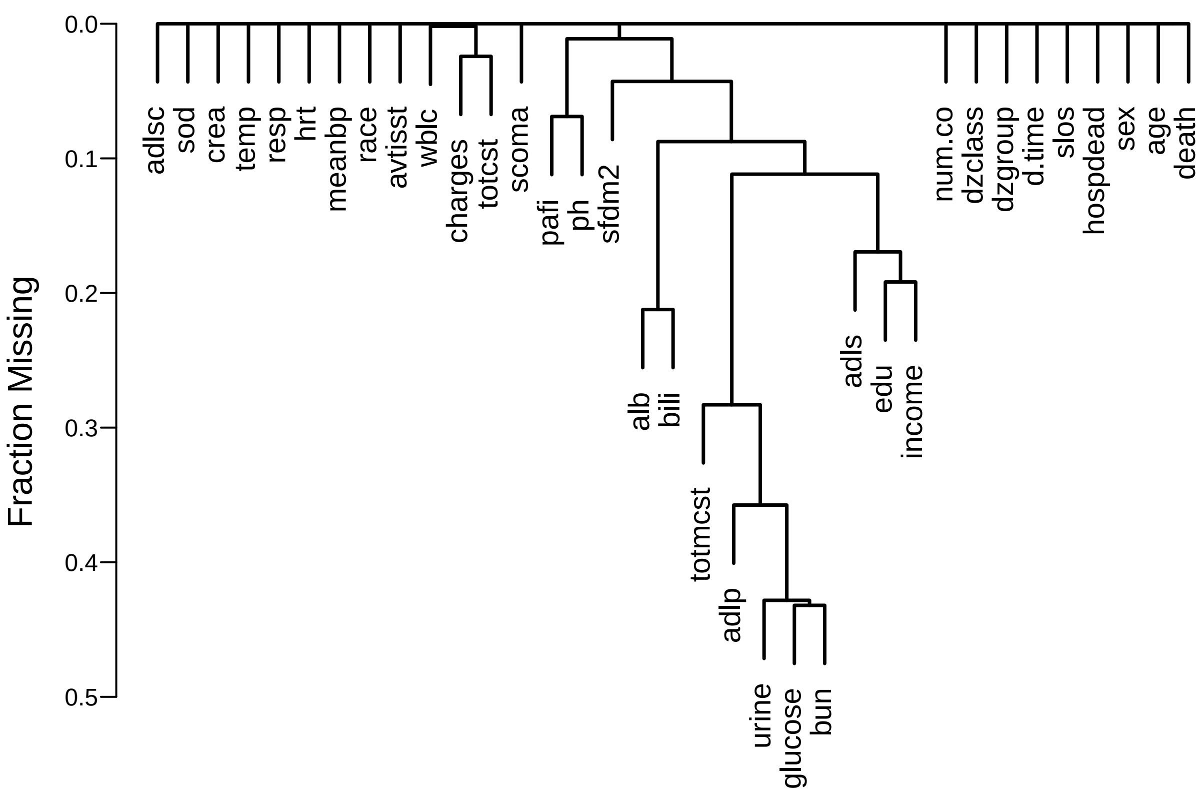


Figure 9.1: Cluster analysis showing which predictors tend to be missing on the same patients

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

C

```
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 9.2
```

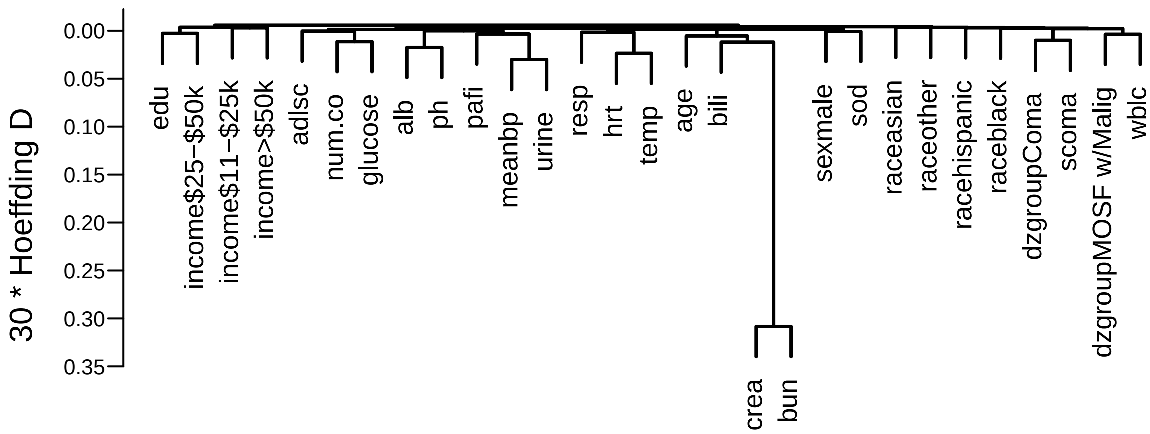


Figure 9.2: Hierarchical clustering of potential predictors using Hoeffding D as a similarity measure. Categorical predictors are automatically expanded into dummy variables.

9.2 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 9.3
```

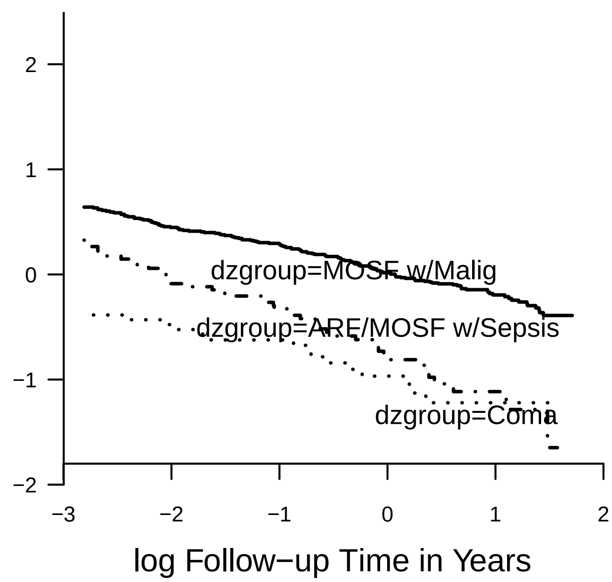


Figure 9.3: $\Phi^{-1}(S_{KM}(t))$ stratified by dzgroup. Linearity and semi-parallelism indicate a reasonable fit to the log-normal accelerated failure time model with respect to one predictor.

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

D

```
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))
```

```r
survplot(r, random.number, label.curve=FALSE)    # Figure 9.4
```

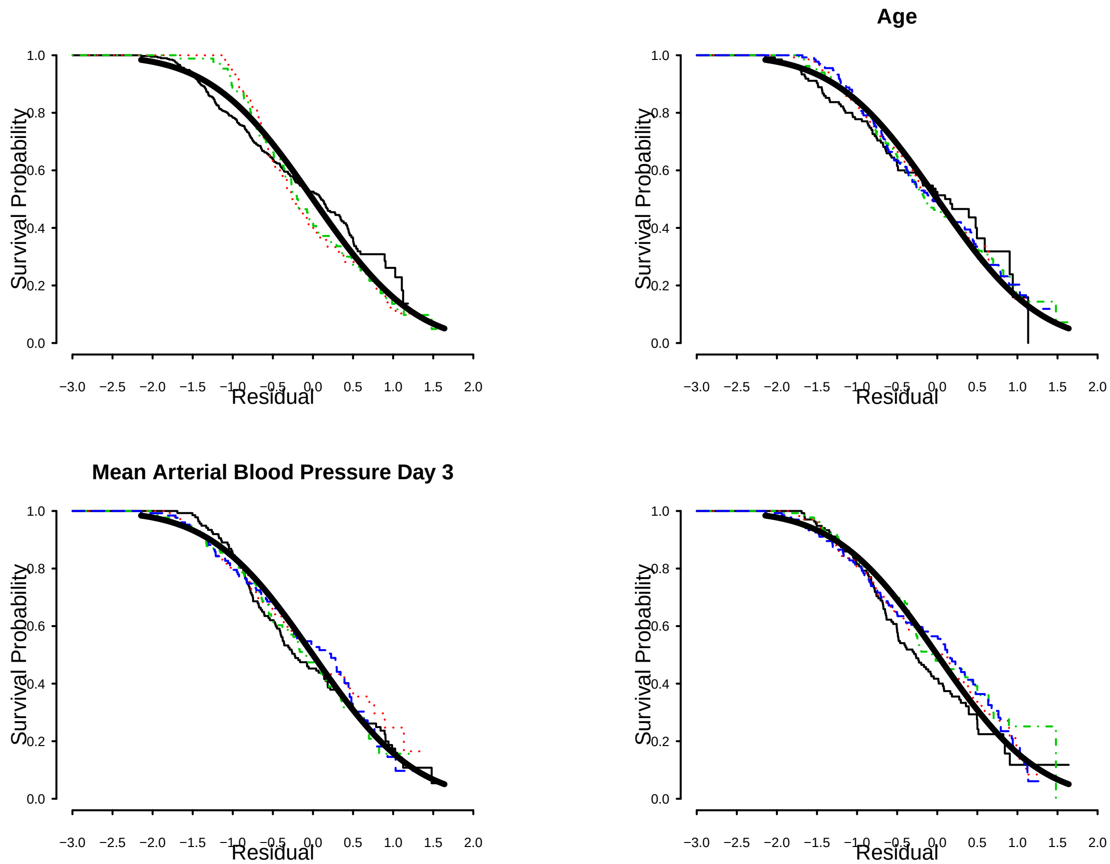


Figure 9.4: Kaplan-Meier estimates of distributions of normalized, right-censored residuals from the fitted log-normal survival model. Residuals are stratified by important variables in the model (by quartiles of continuous variables), plus a random variable to depict the natural variability (in the 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.

The fit for `dzgroup` is not great but overall fit is good.

Remove from consideration predictors that are missing in > 0.2 of the patients. Many of these were only collected for the second phase of SUPPORT.

Of those variables to be included in the model, find which ones have enough potential predictive power to justify allowing for nonlinear relationships or multiple categories, which spend more d.f. For each variable compute Spearman ρ^2 based on multiple linear regression of rank(x), rank(x)^2 and the survival time, 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 9.5
```

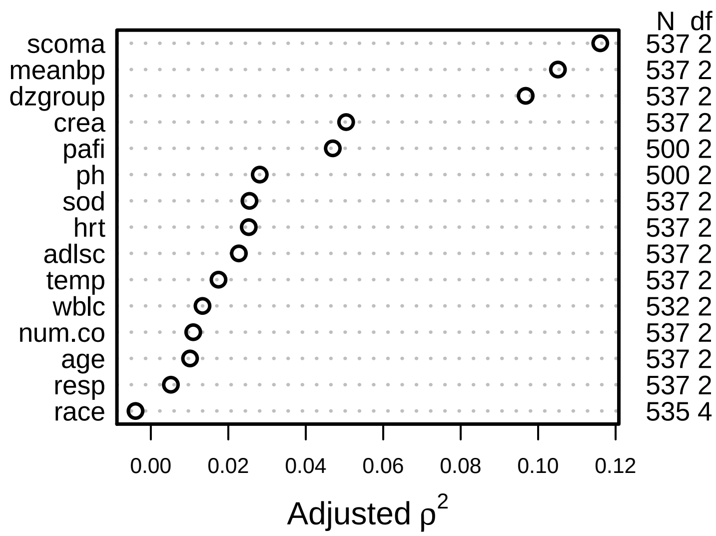


Figure 9.5: Generalized Spearman ρ^2 rank correlation between predictors and truncated survival time

A better approach is to use the complete information in the failure and censoring times by computing Somers' D_{xy} rank correlation allowing for censoring. E

```
w ← rcorrcens(S ~ age + num.co + scoma + meanbp + hrt + resp +
              temp + crea + sod + adlsc + wblc + pafi + ph +
              dzgroup + race)
plot(w, main='')          # Figure 9.6
```

```
# 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    sod  adlsc
     0      0      0      0      0      0      0      0      0      0
  wblc   pafi     ph
     5     37     37
```

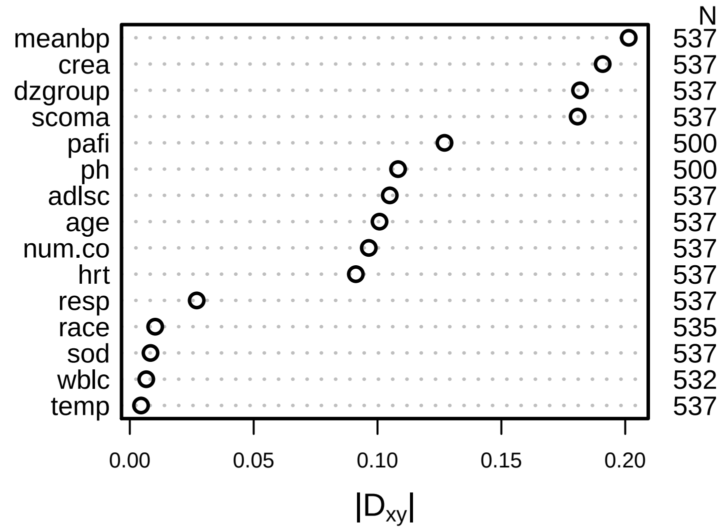


Figure 9.6: Somers' D_{xy} rank correlation between predictors and original survival time. For `dzgroup` or `race`, the correlation coefficient is the maximum correlation from using a dummy variable to represent the most frequent or one to represent the second most frequent category.',scap='Somers' D_{xy} rank correlation between predictors and original survival time

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

Do a formal redundancy analysis using more than pairwise associations, and allow for non-monotonic transformations in predicting each predictor from all other predictors. This analysis requires missing values to be imputed so as to not greatly reduce the sample size. F

```
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

R2 with which each variable can be predicted from all other variables:

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

No redundant variables
```

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

- 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 9.7
```

H

- Figure 9.7 properly blinds the analyst to the form of effects

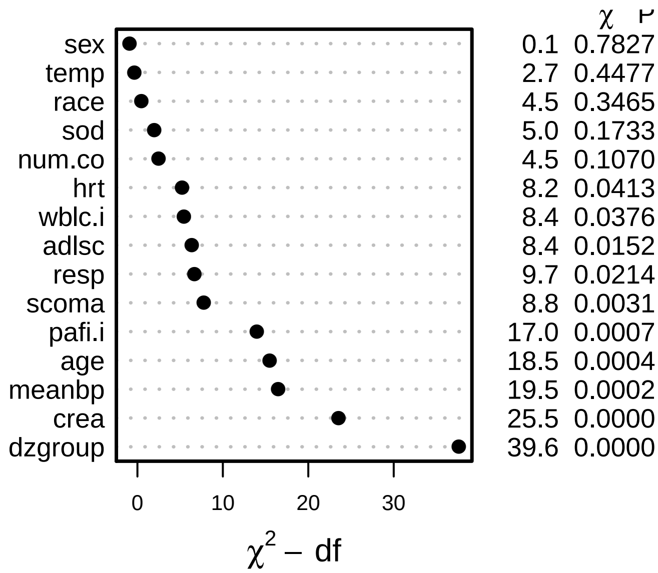


Figure 9.7: Partial χ^2 statistics for association of each predictor with response from saturated main effects model, penalized for d.f.

(tests of linearity).

- Fit a log-normal survival model with number of parameters corresponding to nonlinear effects determined from Figure 9.7. For the most promising predictors, five knots can be allocated, as there are fewer singularity problems once less promising predictors are simplified.

Note: Since the audio was recorded, a bug in `psm` was fixed on 2017-03-12. Discrimination indexes shown in the table below are correct but the audio is incorrect for g and g_r.

```
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)
```

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	1.959
				g_r	7.095

	$\hat{\beta}$	S.E.	Wald Z	Pr($>\|Z\|$)
(Intercept)	-5.6883	3.7851	-1.50	0.1329
age	-0.0148	0.0309	-0.48	0.6322
age'	-0.0412	0.1078	-0.38	0.7024
age''	0.1670	0.5594	0.30	0.7653
age'''	-0.2099	1.3707	-0.15	0.8783
sex=male	-0.0737	0.2181	-0.34	0.7354
dzgroup=Coma	-2.0676	0.4062	-5.09	<0.0001
dzgroup=MOSF w/Malig	-1.4664	0.3112	-4.71	<0.0001
num.co	-0.1917	0.0858	-2.23	0.0255
scoma	-0.0142	0.0044	-3.25	0.0011
adlsc	-0.3735	0.1520	-2.46	0.0140
adlsc^2	0.0442	0.0243	1.82	0.0691
race2=other	0.2979	0.2658	1.12	0.2624
meanbp	0.0702	0.0210	3.34	0.0008
meanbp'	-0.3080	0.2261	-1.36	0.1732
meanbp''	0.8438	0.8556	0.99	0.3241
meanbp'''	-0.5715	0.7707	-0.74	0.4584
hrt	-0.0171	0.0069	-2.46	0.0140
hrt'	0.0064	0.0063	1.02	0.3090
resp	0.0454	0.0230	1.97	0.0483
resp'	-0.0851	0.0291	-2.93	0.0034
temp	0.0523	0.0834	0.63	0.5308
crea	-0.4585	0.6727	-0.68	0.4955
crea'	-11.5176	19.0027	-0.61	0.5444
crea''	21.9840	31.0113	0.71	0.4784
sod	0.0044	0.0157	0.28	0.7792
wblc.i	0.0746	0.0331	2.25	0.0242
wblc.i'	-0.0880	0.0377	-2.34	0.0195
pafi.i	0.0169	0.0055	3.07	0.0021
pafi.i'	-0.0569	0.0239	-2.38	0.0173
pafi.i''	0.1088	0.0482	2.26	0.0239
Log(scale)	0.8024	0.0401	19.99	<0.0001

```
a ← anova(f)
```

9.3 Summarizing the Fitted Model

J

- 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', anova=a)  # Figure 9.8
```

K

```
print(a, size='tsz')
```

Wald Statistics for `S`

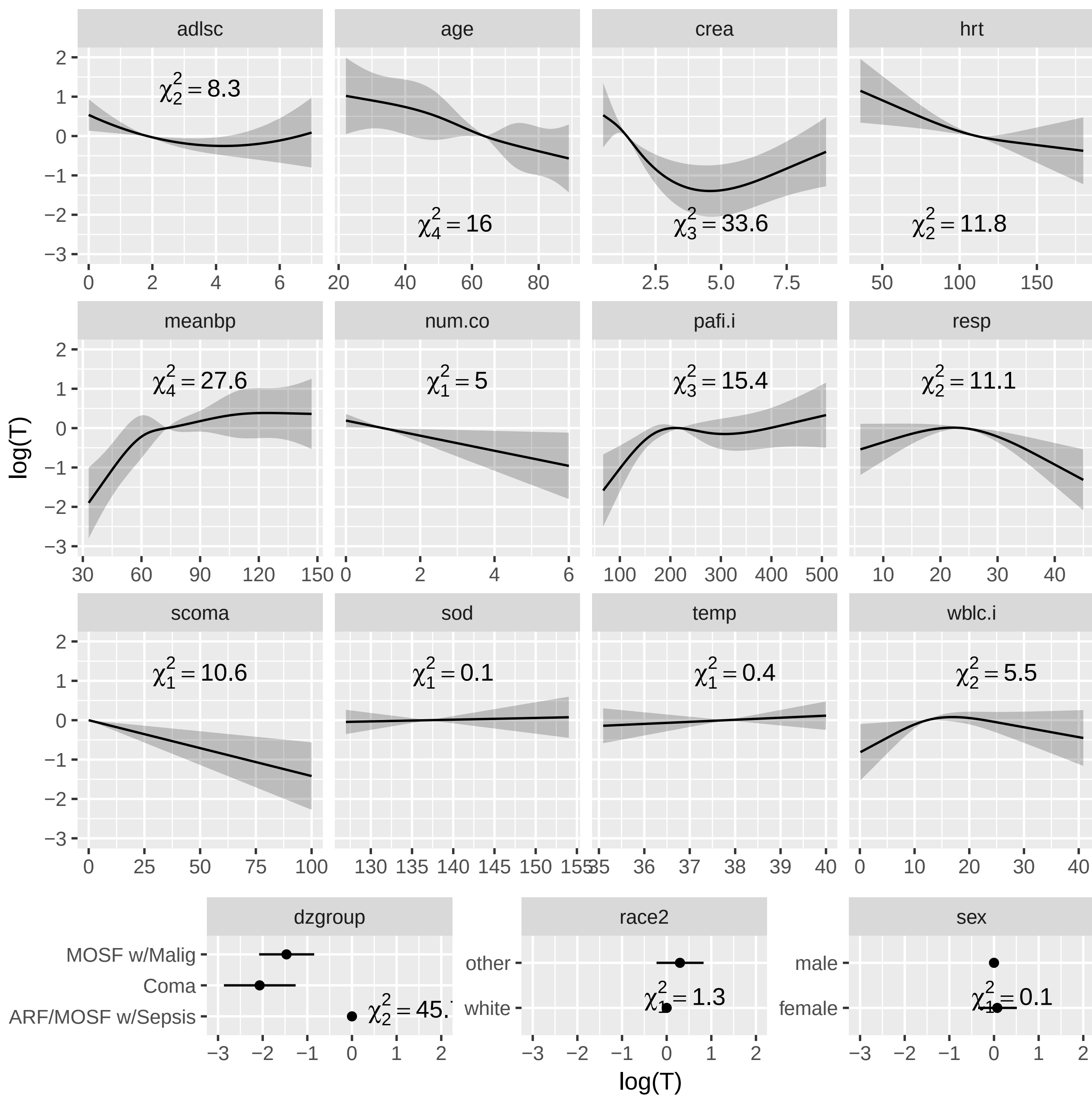


Figure 9.8: Effect of each predictor on log survival time. Predicted values have been centered so that predictions at predictor reference values are zero. Pointwise 0.95 confidence bands are also shown. As all Y-axes have the same scale, it is easy to see which predictors are strongest.

	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

```
plot(a)              # Figure 9.9
```

L

```
options(digits=3)
plot(summary(f), log=TRUE, main='')    # Figure 9.10
```

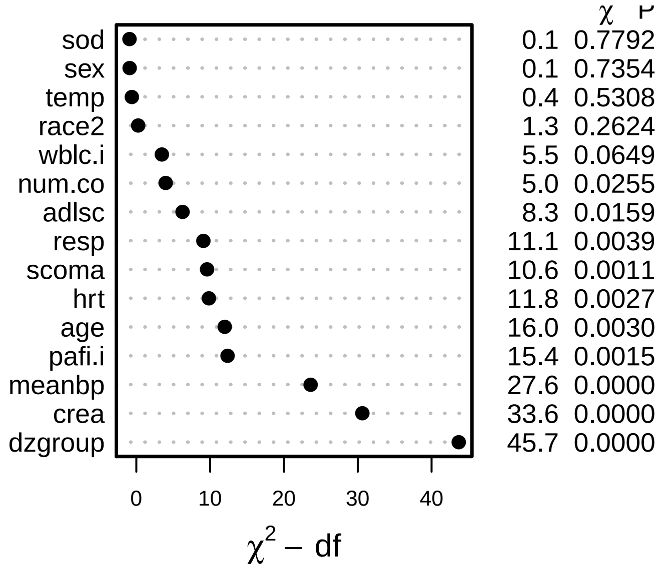


Figure 9.9: Contribution of variables in predicting survival time in log-normal model

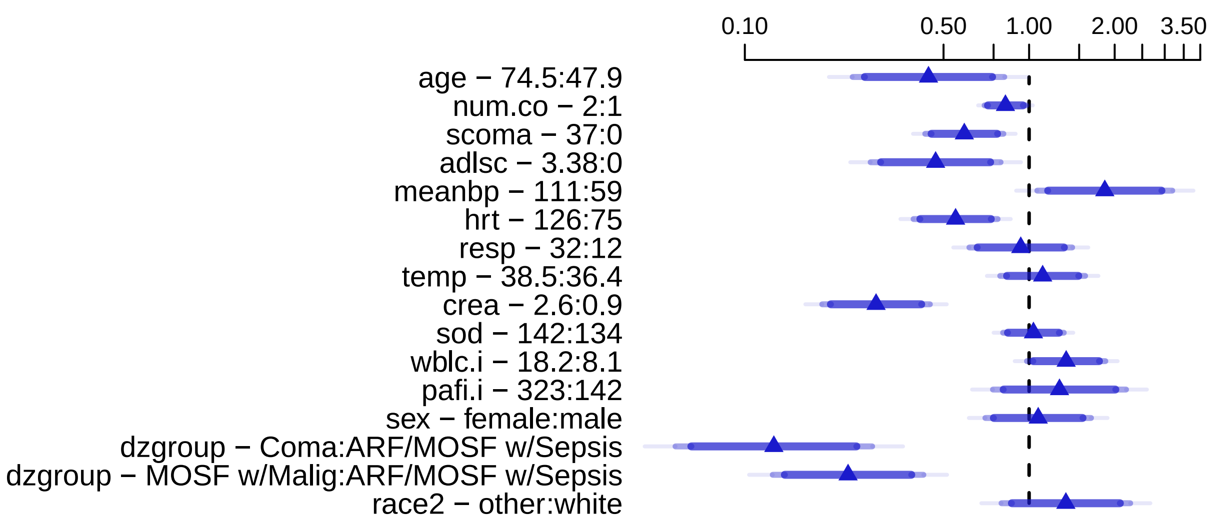


Figure 9.10: Estimated survival time ratios for default settings of predictors. For example, when age changes from its lower quartile to the upper quartile (47.9y to 74.5y), median survival time decreases by more than half. Different shaded areas of bars indicate different confidence levels (0.9, 0.95, 0.99).

9.4 Internal Validation of the Fitted Model Using the Bootstrap

Validate indexes describing the fitted model.

```
# 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

M

- From D_{xy} and R^2 there is a moderate amount of overfitting.
- Slope shrinkage factor (0.90) is not troublesome
- Almost unbiased estimate of future predictive discrimination on similar patients is the corrected D_{xy} of 0.43.

Validate predicted 1-year survival probabilities. Use a smooth approach that does not require binning [90] and use less precise Kaplan-Meier estimates obtained by stratifying patients by the predicted probability, with at least 60 patients per group. N

```
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 9.11
```

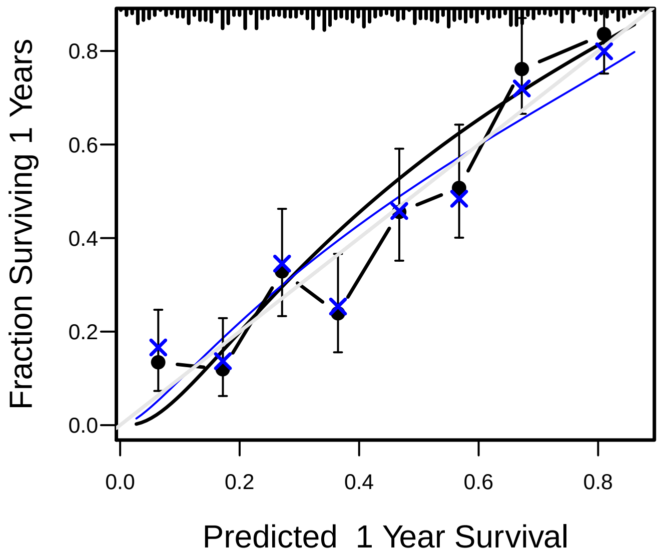


Figure 9.11: Bootstrap validation of calibration curve. Dots represent apparent calibration accuracy; × are bootstrap estimates corrected for overfitting, based on binning predicted survival probabilities and and computing Kaplan-Meier estimates. Black curve is the estimated observed relationship using `hare` and the blue curve is the overfitting-corrected `hare` estimate. The gray-scale line depicts the ideal relationship.

9.5 Approximating the Full Model

The fitted log-normal model is perhaps too complex for routine use and for routine data collection. Let us develop a simplified model that can predict the predicted values of the full model with high accuracy ($R^2 = 0.96$). The simplification is done using a fast backward stepdown against the full model predicted values.

```
Z ← predict(f)    # X*beta hat
a ← ols(Z ∼ rcs(age,5)+sex+dzgroup+num.co+
             scoma+pol(adlsc,2)+race2+
             rcs(meanbp,5)+rcs(hrt,3)+rcs(resp,3)+
             temp+rcs(crea,4)+sod+rcs(wblc.i,3)+
             rcs(pafi.i,4), sigma=1)
# sigma=1 is used to prevent sigma hat from being zero when
# R2=1.0 since we start out by approximating Z with all
#   component variables
fastbw(a, aics=10000)    # fast backward stepdown
```

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

Approximate Estimates after Deleting Factors

        Coef    S.E. Wald Z P
[1,] -0.5928 0.04315 -13.74 0

Factors in Final Model

None
```

```
f.approx ← ols(Z ~ dzgroup + rcs(meanbp,5) + rcs(crea,4) + rcs(age,5) +
                rcs(hrt,3) + scoma + rcs(pafi.i,4) + pol(adlsc,2)+
                rcs(resp,3), x=TRUE)
f.approx$stats
```

```
        n  Model L.R.       d.f.         R2          g
  537.000    1688.225     23.000      0.957      1.915
    Sigma
    0.370
```

O

- 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)
```

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
```

P

```
f.approx$var ← v
print(anova(f.approx, test='Chisq', ss=FALSE), size='tsz')
```

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

Equation for simplified model:

```
# Typeset mathematical form of approximate model
latex(f.approx)
```

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

$$
\begin{aligned}
X\hat{\beta} = \\
& -2.51 \\
& -1.94[\text{Coma}] - 1.75[\text{MOSF w/Malig}] \\
& +0.068\text{meanbp} - 3.08\times10^{-5}(\text{meanbp} - 41.8)_+^3 + 7.9\times10^{-5}(\text{meanbp} - 61)_+^3 \\
& -4.91\times10^{-5}(\text{meanbp} - 73)_+^3 + 2.61\times10^{-6}(\text{meanbp} - 109)_+^3 - 1.7\times10^{-6}(\text{meanbp} - 135)_+^3 \\
& -0.553\text{crea} - 0.229(\text{crea} - 0.6)_+^3 + 0.45(\text{crea} - 1.1)_+^3 - 0.233(\text{crea} - 1.94)_+^3 \\
& +0.0131(\text{crea} - 7.32)_+^3 \\
& -0.0165\text{age} - 1.13\times10^{-5}(\text{age} - 28.5)_+^3 + 4.05\times10^{-5}(\text{age} - 49.5)_+^3 \\
& -2.15\times10^{-5}(\text{age} - 63.7)_+^3 - 2.68\times10^{-5}(\text{age} - 72.7)_+^3 + 1.9\times10^{-5}(\text{age} - 85.6)_+^3 \\
& -0.0136\text{hrt} + 6.09\times10^{-7}(\text{hrt} - 60)_+^3 - 1.68\times10^{-6}(\text{hrt} - 111)_+^3 + 1.07\times10^{-6}(\text{hrt} - 140)_+^3 \\
& -0.0135\,\text{scoma}
\end{aligned}
$$

$$+0.0161\text{pafi.i} - 4.77\times10^{-7}(\text{pafi.i} - 88)_+^3 + 9.11\times10^{-7}(\text{pafi.i} - 167)_+^3$$
$$-5.02\times10^{-7}(\text{pafi.i} - 276)_+^3 + 6.76\times10^{-8}(\text{pafi.i} - 426)_+^3 - 0.369\,\text{adlsc} + 0.0409\,\text{adlsc}^2$$
$$+0.0394\text{resp} - 9.11\times10^{-5}(\text{resp} - 10)_+^3 + 0.000176(\text{resp} - 24)_+^3 - 8.5\times10^{-5}(\text{resp} - 39)_+^3$$

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 9.12
```

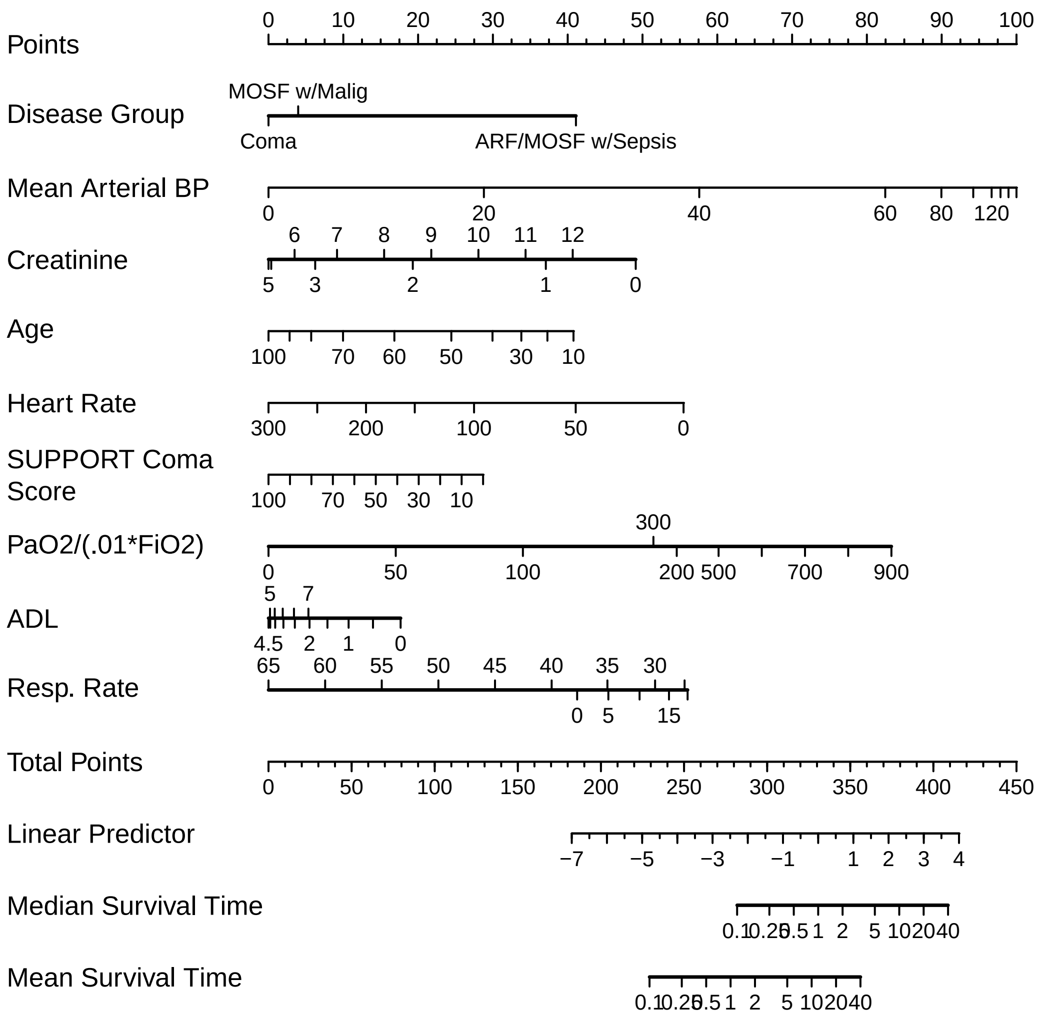


Figure 9.12: Nomogram for predicting median and mean survival time, based on approximation of full model

S Packages and Functions Used

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

Note: All packages are available from `CRAN`

Annotated Bibliography

[1] D. G. Altman. "Categorising continuous covariates (letter to the editor)". In: *Brit J Cancer* 64 (1991), p. 975 (cit. on p. 2-13).

[2] D. G. Altman and P. K. Andersen. "Bootstrap investigation of the stability of a Cox regression model". In: *Stat Med* 8 (1989), pp. 771–783 (cit. on p. 3-15).

[3] D. G. Altman et al. "Dangers of using 'optimal' cutpoints in the evaluation of prognostic factors". In: *J Nat Cancer Inst* 86 (1994), pp. 829–835 (cit. on pp. 2-13, 2-15).

[4] Douglas G. Altman. "Suboptimal analysis using 'optimal' cutpoints". In: *Brit J Cancer* 78 (1998), pp. 556–557 (cit. on p. 2-13).

[5] A. C. Atkinson. "A note on the generalized information criterion for choice of a model". In: *Biometrika* 67 (1980), pp. 413–418 (cit. on pp. 2-26, 3-14).

[6] Peter C. Austin. "Bootstrap model selection had similar performance for selecting authentic and noise variables compared to backward variable elimination: a simulation study". In: *J Clin Epi* 61 (2008), pp. 1009–1017 (cit. on p. 3-15).

"in general, a bootstrap model selection method had comparable performance to conventional backward variable elimination for identifying the true regression model. In most settings, both methods performed poorly at correctly identifying the correct regression model."

.

[7] Peter C. Austin, Jack V. Tu, and Douglas S. Lee. "Logistic regression had superior performance compared with regression trees for predicting in-hospital mortality in patients hospitalized with heart failure". In: *J Clin Epi* 63 (2010), pp. 1145–1155 (cit. on p. 2-32).

ROC areas for logistic models varied from 0.747 to 0.775 whereas they varied from 0.620-0.651 for recursive partitioning;repeated data simulation showed large variation in tree structure

.

[8] Heiko Belcher. "The concept of residual confounding in regression models and some applications". In: *Stat Med* 11 (1992), pp. 1747–1758 (cit. on p. 2-13).

[9] D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley, 1980 (cit. on p. 3-38).

[10] David A. Belsley. *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. New York: Wiley, 1991 (cit. on p. 3-23).

[11] Jacqueline K. Benedetti et al. "Effective sample size for tests of censored survival data". In: *Biometrika* 69 (1982), pp. 343–349 (cit. on p. 3-18).

[12] Caroline Bennette and Andrew Vickers. "Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents". In: *BMC Medical Research Methodology* 12.1 (Feb. 2012), pp. 21+. issn: 1471-2288. url: `http://dx.doi.org/10.1186/1471-2288-12-21` (cit. on p. 2-13).

terrific graphical examples; nice display of outcome heterogeneity within quantile groups of PSA

.

[13] Kiros Berhane, Michael Hauptmann, and Bryan Langholz. "Using tensor product splines in modeling exposure–time–response relationships: Application to the Colorado Plateau Uranium Miners cohort". In: *Stat Med* 27 (2008), pp. 5484–5496 (cit. on p. 2-46).

discusses taking product of all univariate spline basis functions

.

[14] Maria Blettner and Willi Sauerbrei. "Influence of model-building strategies on the results of a case-control study". In: *Stat Med* 12 (1993), pp. 1325–1338 (cit. on p. 4-22).

[15] James G. Booth and Somnath Sarkar. "Monte Carlo approximation of bootstrap variances". In: *Am Statistician* 52 (1998), pp. 354–357 (cit. on p. 4-10).

number of resamples required to estimate variances, quantiles; 800 resamples may be required to guarantee with 0.95 confidence that the relative error of a variance estimate is 0.1;Efron's original suggestions for as low as 25 resamples were based on comparing stability of bootstrap estimates to sampling error, but small relative effects can significantly change P-values;number of bootstrap resamples
.

[16] Robert Bordley. "Statistical decisionmaking without math". In: *Chance* 20.3 (2007), pp. 39–44 (cit. on p. 1-7).

[17] L. Breiman and J. H. Friedman. "Estimating optimal transformations for multiple regression and correlation (with discussion)". In: *J Am Stat Assoc* 80 (1985), pp. 580–619 (cit. on p. 3-31).

[18] Leo Breiman. "The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error". In: *J Am Stat Assoc* 87 (1992), pp. 738–754 (cit. on pp. 3-14, 3-15, 4-16).

[19] Leo Breiman et al. *Classification and Regression Trees*. Pacific Grove, CA: Wadsworth and Brooks/Cole, 1984 (cit. on p. 2-31).

[20] William M. Briggs and Russell Zaretzki. "The skill plot: A graphical technique for evaluating continuous diagnostic tests (with discussion)". In: *Biometrics* 64 (2008), pp. 250–261 (cit. on p. 1-7).

"statistics such as the AUC are not especially relevant to someone who must make a decision about a particular x_c. ... ROC curves lack or obscure several quantities that are necessary for evaluating the operational effectiveness of diagnostic tests. ... ROC curves were first used to check how radio *receivers* (like radar receivers) operated over a range of frequencies. ... This is not how most ROC curves are used now, particularly in medicine. The receiver of a diagnostic measurement ... wants to make a decision based on some x_c, and is not especially interested in how well he would have done had he used some different cutoff."; in the discussion David Hand states "when integrating to yield the overall AUC measure, it is necessary to decide what weight to give each value in the integration. The AUC implicitly does this using a weighting derived empirically from the data. This is nonsensical. The relative importance of misclassifying a case as a noncase, compared to the reverse, cannot come from the data itself. It must come externally, from considerations of the severity one attaches to the different kinds of misclassifications."; see Lin, Kvam, Lu Stat in Med 28:798-813;2009
.

[21] David Brownstone. "Regression strategies". In: *Proceedings of the 20th Symposium on the Interface between Computer Science and Statistics*. Washington, DC: American Statistical Association, 1988, pp. 74–79 (cit. on p. 4-22).

[22] Petra Buettner, Claus Garbe, and Irene Guggenmoos-Holzmann. "Problems in defining cutoff points of continuous prognostic factors: Example of tumor thickness in primary cutaneous melanoma". In: *J Clin Epi* 50 (1997), pp. 1201–1210 (cit. on p. 2-13).

choice of cut point depends on marginal distribution of predictor
.

[23] Centers for Disease Control and Prevention CDC. National Center for Health Statistics NCHS. *National Health and Nutrition Examination Survey*. Hyattsville, MD, 2010. url: `http://www.cdc.gov/nchs/nhanes/nhanes2009-2010/nhanes09%5C_10.htm` (cit. on p. 8-4).

[24] John M. Chambers and Trevor J. Hastie, eds. *Statistical Models in S*. Pacific Grove, CA: Wadsworth and Brooks/Cole, 1992 (cit. on p. 2-47).

[25] C. Chatfield. "Avoiding statistical pitfalls (with discussion)". In: *Statistical Sci* 6 (1991), pp. 240–268 (cit. on p. 3-39).

[26] C. Chatfield. "Model uncertainty, data mining and statistical inference (with discussion)". In: *J Roy Stat Soc A* 158 (1995), pp. 419–466 (cit. on pp. 3-10, 4-22).

bias by selecting model because it fits the data well; bias in standard errors;P. 420: ... need for a better balance in the literature and in statistical teaching between *techniques* and problem solving *strategies*. P. 421: It is 'well known' to be 'logically unsound and practically misleading' (Zhang, 1992) to make inferences as if a model is known to be true when it has, in fact, been selected from the *same* data to be used for estimation purposes. However, although statisticians may admit this privately (Breiman (1992) calls it a 'quiet scandal'), they (we) continue to ignore the difficulties because it is not clear what else could or should be done. P. 421: Estimation errors for regression coefficients are usually smaller than errors from failing to take into account model specification. P. 422: Statisticians must stop pretending that model uncertainty does not exist and begin to find ways of coping with it. P. 426: It is indeed strange that we often admit model uncertainty by searching for a best model but then ignore this uncertainty by making inferences and predictions as if certain that the best fitting model is actually true. P. 427: The analyst needs to assess the model selection *process* and not just the best fitting model. P. 432: The use of subset selection methods is well known to introduce alarming biases. P. 433: ... the AIC can be highly biased in data-driven model selection situations. P. 434: Prediction intervals will generally be too narrow. In the discussion, Jamal R. M. Ameen states that a model should be (a) satisfactory in performance relative to the stated objective, (b) logically sound, (c) representative, (d) questionable and subject to on-line interrogation, (e) able to accommodate external or expert information and (f) able to convey information.
.

[27] Samprit Chatterjee and Ali S. Hadi. *Regression Analysis by Example*. Fifth. New York: Wiley, 2012. isbn: 0470905840 (cit. on p. 3-22).

[28] Marie Chavent et al. "ClustOfVar: An R package for the clustering of variables". In: *J Stat Software* 50.13 (Sept. 2012), pp. 1–16 (cit. on p. 3-27).

[29] A. Ciampi et al. "Stratification by stepwise regression, correspondence analysis and recursive partition". In: *Comp Stat Data Analysis* 1986 (1986), pp. 185–204 (cit. on p. 3-27).

[30] W. S. Cleveland. "Robust locally weighted regression and smoothing scatterplots". In: *J Am Stat Assoc* 74 (1979), pp. 829–836 (cit. on p. 2-28).

[31] Gary S. Collins, Emmanuel O. Ogundimu, and Douglas G. Altman. "Sample size considerations for the external validation of a multivariable prognostic model: a resampling study". In: *Stat Med* 35.2 (Jan. 2016), pp. 214–226. issn: 02776715. url: http://dx.doi.org/10.1002/sim.6787 (cit. on p. 4-14).

[32] Gary S. Collins et al. "Quantifying the impact of different approaches for handling continuous predictors on the performance of a prognostic model". In: *Stat Med* 35.23 (Oct. 2016), pp. 4124–4135. issn: 02776715. url: http://dx.doi.org/10.1002/sim.6986 (cit. on p. 2-13).

used rms package hazard regression method (hare) for survival model calibration

.

[33] E. Francis Cook and Lee Goldman. "Asymmetric stratification: An outline for an efficient method for controlling confounding in cohort studies". In: *Am J Epi* 127 (1988), pp. 626–639 (cit. on p. 2-32).

[34] J. B. Copas. "Cross-validation shrinkage of regression predictors". In: *J Roy Stat Soc B* 49 (1987), pp. 175–183 (cit. on p. 4-20).

[35] J. B. Copas. "Regression, prediction and shrinkage (with discussion)". In: *J Roy Stat Soc B* 45 (1983), pp. 311–354 (cit. on pp. 3-20, 3-21).

[36] David R. Cox. "Regression models and life-tables (with discussion)". In: *J Roy Stat Soc B* 34 (1972), pp. 187–220 (cit. on p. 2-50).

[37] Sybil L. Crawford, Sharon L. Tennstedt, and John B. McKinlay. "A comparison of analytic methods for non-random missingness of outcome data". In: *J Clin Epi* 48 (1995), pp. 209–219 (cit. on p. 3-45).

[38] N. J. Crichton and J. P. Hinde. "Correspondence analysis as a screening method for indicants for clinical diagnosis". In: *Stat Med* 8 (1989), pp. 1351–1362 (cit. on p. 3-27).

[39] Ralph B. D'Agostino et al. "Development of health risk appraisal functions in the presence of multiple indicators: The Framingham Study nursing home institutionalization model". In: *Stat Med* 14 (1995), pp. 1757–1770 (cit. on pp. 3-23, 3-26).

[40] C. E. Davis et al. "An example of dependencies among variables in a conditional logistic regression". In: *Modern Statistical Methods in Chronic Disease Epi*. Ed. by S. H. Moolgavkar and R. L. Prentice. New York: Wiley, 1986, pp. 140–147 (cit. on p. 3-23).

[41] Charles S. Davis. *Statistical Methods for the Analysis of Repeated Measurements*. New York: Springer, 2002 (cit. on p. 6-16).

[42] S. Derksen and H. J. Keselman. "Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables". In: *British J Math Stat Psych* 45 (1992), pp. 265–282 (cit. on p. 3-10).

[43] T. F. Devlin and B. J. Weeks. "Spline functions for logistic regression modeling". In: *Proceedings of the Eleventh Annual SAS Users Group International Conference*. Cary, NC: SAS Institute, Inc., 1986, pp. 646–651 (cit. on p. 2-21).

[44] Peter J. Diggle et al. *Analysis of Longitudinal Data*. second. Oxford UK: Oxford University Press, 2002 (cit. on p. 6-10).

[45] William D. Dupont. *Statistical Modeling for Biomedical Researchers*. second. Cambridge, UK: Cambridge University Press, 2008 (cit. on p. 10-11).

[46] S. Durrleman and R. Simon. "Flexible regression models with cubic splines". In: *Stat Med* 8 (1989), pp. 551–561 (cit. on p. 2-25).

[47] B. Efron. "Estimating the error rate of a prediction rule: Improvement on cross-validation". In: *J Am Stat Assoc* 78 (1983), pp. 316–331 (cit. on pp. 4-17, 4-20, 4-21).

suggested need at least 200 models to get an average that is adequate, i.e., 20 repeats of 10-fold cv

.

[48] Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. New York: Chapman and Hall, 1993 (cit. on p. 4-20).

[49] Bradley Efron and Robert Tibshirani. "Improvements on cross-validation: The .632+ bootstrap method". In: *J Am Stat Assoc* 92 (1997), pp. 548–560 (cit. on p. 4-20).

[50] Juanjuan Fan and Richard A. Levine. "To amnio or not to amnio: That is the decision for Bayes". In: *Chance* 20.3 (2007), pp. 26–32 (cit. on p. 1-7).

[51] David Faraggi and Richard Simon. "A simulation study of cross-validation for selecting an optimal cutpoint in univariate survival analysis". In: *Stat Med* 15 (1996), pp. 2203–2213 (cit. on p. 2-13).

bias in point estimate of effect from selecting cutpoints based on P-value; loss of information from dichotomizing continuous predictors
.

[52] J. J. Faraway. "The cost of data analysis". In: *J Comp Graph Stat* 1 (1992), pp. 213–229 (cit. on pp. 3-47, 4-20, 4-22).

[53] Valerii Fedorov, Frank Mannino, and Rongmei Zhang. "Consequences of dichotomization". In: *Pharm Stat* 8 (2009), pp. 50–61. url: `http://dx.doi.org/10.1002/pst.331` (cit. on pp. 1-6, 2-13).

optimal cutpoint depends on unknown parameters;should only entertain dichotomization when "estimating a value of the cumulative distribution and when the assumed model is very different from the true model";nice graphics
.

[54] D. Freedman, W. Navidi, and S. Peters. "On the Impact of Variable Selection in Fitting Regression Equations". In: Lecture Notes in Economics and Mathematical Systems. New York: Springer-Verlag, 1988, pp. 1–16 (cit. on p. 4-21).

[55] J. H. Friedman. *A variable span smoother*. Technical Report 5. Laboratory for Computational Statistics, Department of Statistics, Stanford University, 1984 (cit. on p. 3-31).

[56] Mitchell H. Gail and Ruth M. Pfeiffer. "On criteria for evaluating models of absolute risk". In: *Biostatistics* 6.2 (2005), pp. 227–239 (cit. on p. 1-7).

[57] Joseph C. Gardiner, Zhehui Luo, and Lee A. Roman. "Fixed effects, random effects and GEE: What are the differences?" In: *Stat Med* 28 (2009), pp. 221–239 (cit. on p. 6-9).

nice comparison of models; econometrics; different use of the term "fixed effects model"
.

[58] A. Giannoni et al. "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". In: *PLoS ONE* 9.1 (2014). url: `http://dx.doi.org/10.1371/journal.pone.0081699` (cit. on pp. 2-13, 2-15).

[59] John H. Giudice, John R. Fieberg, and Mark S. Lenarz. "Spending degrees of freedom in a poor economy: A case study of building a sightability model for moose in northeastern minnesota". In: *J Wildlife Manage* (2011). url: `http://dx.doi.org/10.1002/jwmg.213` (cit. on p. 3-1).

[60] Tilmann Gneiting and Adrian E. Raftery. "Strictly proper scoring rules, prediction, and estimation". In: *J Am Stat Assoc* 102 (2007), pp. 359–378 (cit. on p. 1-7).

wonderful review article except missing references from Scandanavian and German medical decision making literature
.

[61] Harvey Goldstein. "Restricted unbiased iterative generalized least-squares estimation". In: *Biometrika* 76.3 (1989), pp. 622–623 (cit. on pp. 6-6, 6-10).

derivation of REML
.

[62] Usha S. Govindarajulu et al. "Comparing smoothing techniques in Cox models for exposure-response relationships". In: *Stat Med* 26 (2007), pp. 3735–3752 (cit. on p. 2-26).

authors wrote a SAS macro for restricted cubic splines even though such a macro as existed since 1984; would have gotten more useful results had simulation been used so would know the true regression shape;measure of agreement of two estimated curves by computing the area between them, standardized by average of areas under the two;penalized spline and rcs were closer to each other than to fractional polynomials
.

[63] P. M. Grambsch and P. C. O'Brien. "The effects of transformations and preliminary tests for non-linearity in regression". In: *Stat Med* 10 (1991), pp. 697–709 (cit. on pp. 2-38, 3-10).

[64] Robert J. Gray. "Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis". In: *J Am Stat Assoc* 87 (1992), pp. 942–951 (cit. on pp. 2-46, 3-20).

[65] Robert J. Gray. "Spline-based tests in survival analysis". In: *Biometrics* 50 (1994), pp. 640–652 (cit. on p. 2-46).

[66] Michael J. Greenacre. "Correspondence analysis of multivariate categorical data by weighted least-squares". In: *Biometrika* 75 (1988), pp. 457–467 (cit. on p. 3-27).

[67] Sander Greenland. "When should epidemiologic regressions use random coefficients?" In: *Biometrics* 56 (2000), pp. 915–921. url: http://dx.doi.org/10.1111/j.0006-341X.2000.00915.x (cit. on pp. 3-10, 3-42).

use of statistics in epidemiology is largely primitive;stepwise variable selection on confounders leaves important confounders uncontrolled;composition matrix;example with far too many significant predictors with many regression coefficients absurdly inflated when overfit;lack of evidence for dietary effects mediated through constituents;shrinkage instead of variable selection;larger effect on confidence interval width than on point estimates with variable selection;uncertainty about variance of random effects is just uncertainty about prior opinion;estimation of variance is pointless;instead the analysis should be repeated using different values;"if one feels compelled to estimate τ^2, I would recommend giving it a proper prior concentrated amount contextually reasonable values";claim about ordinary MLE being unbiased is misleading because it assumes the model is correct and is the only model entertained;shrinkage towards compositional model;"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)'". See also gus06per.
.

[68] Jian Guo et al. "Principal component analysis with sparse fused loadings". In: *J Comp Graph Stat* 19.4 (2011), pp. 930–946 (cit. on p. 3-27).

incorporates blocking structure in the variables;selects different variables for different components;encourages loadings of highly correlated variables to have same magnitude, which aids in interpretation
.

[69] D. Hand and M. Crowder. *Practical Longitudinal Data Analysis*. London: Chapman & Hall, 1996.

[70] F. E. Harrell. "The LOGIST Procedure". In: *SUGI Supplemental Library Users Guide*. Version 5. Cary, NC: SAS Institute, Inc., 1986, pp. 269–293 (cit. on p. 3-13).

[71] F. E. Harrell, K. L. Lee, and B. G. Pollock. "Regression models in clinical studies: Determining relationships between predictors and response". In: *J Nat Cancer Inst* 80 (1988), pp. 1198–1202 (cit. on p. 2-29).

[72] F. E. Harrell et al. "Regression modeling strategies for improved prognostic prediction". In: *Stat Med* 3 (1984), pp. 143–152 (cit. on p. 3-17).

[73] F. E. Harrell et al. "Regression models for prognostic prediction: Advantages, problems, and suggested solutions". In: *Ca Trt Rep* 69 (1985), pp. 1071–1077 (cit. on p. 3-17).

[74] Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. "Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors". In: *Stat Med* 15 (1996), pp. 361–387 (cit. on p. 3-1).

[75] Frank E. Harrell et al. "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". In: *Stat Med* 17 (1998), pp. 909–944. url: http://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0258(19980430)17:8%5C%3C909::AID-SIM753%5C%3E3.0.CO;2-O/abstract (cit. on pp. 3-20, 3-45).

[76] Trevor J. Hastie and Robert J. Tibshirani. *Generalized Additive Models*. ISBN 9780412343902. Boca Raton, FL: Chapman & Hall/CRC, 1990 (cit. on p. 2-35).

[77] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman. *The Elements of Statistical Learning*. second. ISBN-10: 0387848576; ISBN-13: 978-0387848570. New York: Springer, 2008 (cit. on p. 2-34).

[78] S. G. Hilsenbeck and G. M. Clark. "Practical p-value adjustment for optimally selected cutpoints". In: *Stat Med* 15 (1996), pp. 103–112 (cit. on p. 2-13).

[79] W. Hoeffding. "A non-parametric test of independence". In: *Ann Math Stat* 19 (1948), pp. 546–557 (cit. on p. 3-27).

[80] Norbert Holländer, Willi Sauerbrei, and Martin Schumacher. "Confidence intervals for the effect of a prognostic factor after selection of an 'optimal' cutpoint". In: *Stat Med* 23 (2004), pp. 1701–1713. url: http://dx.doi.org/10.1002/sim.1611 (cit. on pp. 2-13, 2-15).

true type I error can be much greater than nominal level;one example where nominal is 0.05 and true is 0.5;minimum P-value method;CART;recursive partitioning;bootstrap method for correcting confidence interval;based on heuristic shrinkage coefficient;"It should be noted, however, 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 recent update of the American Society of Clinical Oncology."; dichotomization; categorizing continuous variables; refs alt94dan, sch94out, alt98sub
.

[81] C. M. Hurvich and C. L. Tsai. "The impact of model selection on inference in linear regression". In: *Am Statistician* 44 (1990), pp. 214–217 (cit. on p. 3-15).

[82] Lisa I. Iezzoni. "Dimensions of Risk". In: *Risk Adjustment for Measuring Health Outcomes*. Ed. by Lisa I. Iezzoni. Ann Arbor, MI: Foundation of the American College of Healthcare Executives, 1994. Chap. 2, pp. 29–118 (cit. on p. 1-12).

dimensions of risk factors to include in models

.

[83] J. D. Kalbfleisch and R. L. Prentice. "Marginal likelihood based on Cox's regression and life model". In: *Biometrika* 60 (1973), pp. 267–278 (cit. on p. 8-25).

[84] Juha Karvanen and Frank E. Harrell. "Visualizing covariates in proportional hazards model". In: *Stat Med* 28 (2009), pp. 1957–1966 (cit. on p. 4-2).

[85] Michael G. Kenward, Ian R. White, and James R. Carpener. "Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials? (letter to the editor)". In: *Stat Med* 29 (2010), pp. 1455–1456 (cit. on p. 6-5).

sharp rebuke of liu09sho

.

[86] W. A. Knaus et al. "The SUPPORT prognostic model: Objective estimates of survival for seriously ill hospitalized adults". In: *Ann Int Med* 122 (1995), pp. 191–203. url: http://dx.doi.org/10.7326/0003-4819-122-3-199502010-00007 (cit. on pp. 3-32, 9-1).

[87] R. Koenker and G. Bassett. "Regression quantiles". In: *Econometrica* 46 (1978), pp. 33–50 (cit. on p. 8-11).

[88] Roger Koenker. *Quantile Regression*. ISBN-10: 0-521-60827-9; ISBN-13: 978-0-521-60827-5. New York: Cambridge University Press, 2005 (cit. on p. 8-11).

[89] Roger Koenker. *quantreg: Quantile Regression*. R package version 4.38. 2009. url: http://CRAN.R-project.org/package=quantreg (cit. on p. 8-11).

[90] Charles Kooperberg, Charles J. Stone, and Young K. Truong. "Hazard regression". In: *J Am Stat Assoc* 90 (1995), pp. 78–94 (cit. on p. 9-18).

[91] Warren F. Kuhfeld. "The PRINQUAL Procedure". In: *SAS/STAT 9.2 User's Guide*. Second. Cary, NC: SAS Publishing, 2009. url: http://support.sas.com/documentation/onlinedoc/stat (cit. on p. 3-28).

[92] B. Lausen and M. Schumacher. "Evaluating the effect of optimized cutoff values in the assessment of prognostic factors". In: *Comp Stat Data Analysis* 21.3 (1996), pp. 307–326. url: http://dx.doi.org/10.1016/0167-9473(95)00016-X (cit. on p. 2-13).

[93] J. F. Lawless and K. Singhal. "Efficient screening of nonnormal regression models". In: *Biometrics* 34 (1978), pp. 318–327 (cit. on p. 3-14).

[94] S. le Cessie and J. C. van Houwelingen. "Ridge estimators in logistic regression". In: *Appl Stat* 41 (1992), pp. 191–201 (cit. on p. 3-20).

[95] A. Leclerc et al. "Correspondence analysis and logistic modelling: Complementary use in the analysis of a health survey among nurses". In: *Stat Med* 7 (1988), pp. 983–995 (cit. on p. 3-27).

[96] Seokho Lee, Jianhua Z. Huang, and Jianhua Hu. "Sparse logistic principal components analysis for binary data". In: *Ann Appl Stat* 4.3 (2010), pp. 1579–1601 (cit. on p. 2-34).

[97] Chenlei Leng and Hansheng Wang. "On general adaptive sparse principal component analysis". In: *J Comp Graph Stat* 18.1 (2009), pp. 201–215 (cit. on p. 2-34).

[98] Kung-Yee Liang and Scott L. Zeger. "Longitudinal data analysis of continuous and discrete responses for pre-post designs". In: *Sankhyā* 62 (2000), pp. 134–148 (cit. on p. 6-5).

makes an error in assuming the baseline variable will have the same univariate distribution as the response except for a shift;baseline may have for example a truncated distribution based on a trial's inclusion criteria;if correlation between baseline and response is zero, ANCOVA will be twice as efficient as simple analysis of change scores;if correlation is one they may be equally efficient

.

[99] James K. Lindsey. *Models for Repeated Measurements*. Clarendon Press, 1997.

[100] Guanghan F. Liu et al. "Should baseline be a covariate or dependent variable in analyses of change from baseline in clinical trials?" In: *Stat Med* 28 (2009), pp. 2509–2530 (cit. on p. 6-5).

seems to miss several important points, such as the fact that the baseline variable is often part of the inclusion/exclusion criteria and so has a truncated distribution that is different from that of the follow-up measurements;sharp rebuke in ken10sho

.

[101] Richard Lockhart et al. *A significance test for the lasso*. Tech. rep. arXiv, 2013. arXiv: 1301.7161. url: http://arxiv.org/abs/1301.7161 (cit. on p. 3-10).

[102] Xiaohui Luo, Leonard A. Stfanski, and Dennis D. Boos. "Tuning variable selection procedures by adding noise". In: *Technometrics* 48 (2006), pp. 165–175 (cit. on p. 1-14).

adding a known amount of noise to the response and studying σ^2 to tune the stopping rule to avoid overfitting or underfitting;simulation setup
.

[103] Nathan Mantel. "Why stepdown procedures in variable selection". In: *Technometrics* 12 (1970), pp. 621–625 (cit. on p. 3-14).

[104] Maurizio Manuguerra and Gillian Z. Heller. "Ordinal Regression Models for Continuous Scales". In: *The International Journal of Biostatistics* 6.1 (Jan. 2010). issn: 1557-4679. url: http://dx.doi.org/10.2202/1557-4679.1230 (cit. on p. 8-13).

mislabeled a flexible parametric model as semi-parametric; does not cover semi-parametric approach with lots of intercepts
.

[105] S. E. Maxwell and H. D. Delaney. "Bivariate median splits and spurious statistical significance". In: *Psych Bull* 113 (1993), pp. 181–190. url: http://dx.doi.org/10.1037//0033-2909.113.1.181 (cit. on p. 2-13).

[106] George P. McCabe. "Principal variables". In: *Technometrics* 26 (1984), pp. 137–144 (cit. on p. 3-26).

[107] George Michailidis and Jan de Leeuw. "The Gifi system of descriptive multivariate analysis". In: *Statistical Sci* 13 (1998), pp. 307–336 (cit. on p. 3-27).

[108] Barry K. Moser and Laura P. Coombs. "Odds ratios for a continuous outcome variable without dichotomizing". In: *Stat Med* 23 (2004), pp. 1843–1860 (cit. on p. 2-13).

large loss of efficiency and power;embeds in a logistic distribution, similar to proportional odds model;categorization;dichotomization of a continuous response in order to obtain odds ratios often results in an inflation of the needed sample size by a factor greater than 1.5
.

[109] Raymond H. Myers. *Classical and Modern Regression with Applications*. Boston: PWS-Kent, 1990 (cit. on p. 3-22).

[110] N. J. D. Nagelkerke. "A note on a general definition of the coefficient of determination". In: *Biometrika* 78 (1991), pp. 691–692 (cit. on p. 3-40).

[111] Todd G. Nick and J. Michael Hardin. "Regression modeling strategies: An illustrative case study from medical rehabilitation outcomes research". In: *Am J Occ Ther* 53 (1999), pp. 459–470 (cit. on p. 3-1).

[112] David J. Nott and Chenlei Leng. "Bayesian projection approaches to variable selection in generalized linear models". In: *Computational Statistics & Data Analysis* 54.12 (Dec. 2010), pp. 3227–3241. issn: 01679473. url: http://dx.doi.org/10.1016/j.csda.2010.01.036 (cit. on p. 2-34).

[113] Debashis Paul et al. ""Preconditioning" for feature selection and regression in high-dimensional problems". In: *Ann Stat* 36.4 (2008), pp. 1595–1619. url: http://dx.doi.org/10.1214/009053607000000578 (cit. on p. 2-34).

develop consistent $\hat{Y}$ using a latent variable structure, using for example supervised principal components. Then run stepwise regression or lasso predicting $\hat{Y}$ (lasso worked better). Can run into problems when a predictor has importance in an adjusted sense but has no marginal correlation with Y;model approximation;model simplification
.

[114] Peter Peduzzi et al. "A simulation study of the number of events per variable in logistic regression analysis". In: *J Clin Epi* 49 (1996), pp. 1373–1379 (cit. on pp. 3-17, 3-18).

[115] Peter Peduzzi et al. "Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates". In: *J Clin Epi* 48 (1995), pp. 1503–1510 (cit. on p. 3-17).

[116] N. Peek et al. "External validation of prognostic models for critically ill patients required substantial sample sizes". In: *J Clin Epi* 60 (2007), pp. 491–501 (cit. on p. 3-41).

large sample sizes need to obtain reliable external validations;inadequate power of DeLong, DeLong, and Clarke-Pearson test for differences in correlated ROC areas (p. 498);problem with tests of calibration accuracy having too much power for large sample sizes
.

[117] Michael J. Pencina, Ralph B. D'Agostino, and Olga V. Demler. "Novel metrics for evaluating improvement in discrimination: net reclassification and integrated discrimination improvement for normal variables and nested models". In: *Stat Med* 31.2 (2012), pp. 101–113. url: http://dx.doi.org/10.1002/sim.4348 (cit. on p. 3-41).

[118] Michael J. Pencina, Ralph B. D'Agostino, and Ewout W. Steyerberg. "Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers". In: *Stat Med* 30 (2011), pp. 11–21. url: http://dx.doi.org/10.1002/sim.4085 (cit. on p. 3-41).

lack of need for NRI to be category-based;arbitrariness of categories;"category-less or continuous NRI is the most objective and versatile measure of improvement in risk prediction;authors misunderstood the inadequacy of three categories if categories are used;comparison of NRI to change in C index;example of continuous plot of risk for old model vs. risk for new model

.

[119] Michael J. Pencina et al. "Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond". In: *Stat Med* 27 (2008), pp. 157–172 (cit. on p. 3-41).

small differences in ROC area can still be very meaningful;example of insignificant test for difference in ROC areas with very significant results from new method;Yates' discrimination slope;reclassification table;limiting version of this based on whether and amount by which probabilities rise for events and lower for non-events when compare new model to old;comparing two models;see letter to the editor by Van Calster and Van Huffel, Stat in Med 29:318-319, 2010 and by Cook and Paynter, Stat in Med 31:93-97, 2012

.

[120] Sanne A. Peters et al. "Multiple imputation of missing repeated outcome measurements did not add to linear mixed-effects models." In: *J Clin Epi* 65.6 (2012), pp. 686–695. url: http://dx.doi.org/10.1016/j.jclinepi.2011.11.012 (cit. on p. 6-9).

[121] José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. New York: Springer, 2000 (cit. on pp. 6-10, 6-12).

[122] Tjeerd van der Ploeg, Peter C. Austin, and Ewout W. Steyerberg. "Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints." In: *BMC medical research methodology* 14.1 (Dec. 2014), pp. 137+. issn: 1471-2288. url: http://dx.doi.org/10.1186/1471-2288-14-137 (cit. on pp. 2-4, 2-37, 3-17).

Would be better to use proper accuracy scores in the assessment. Too much emphasis on optimism as opposed to final discrimination measure. But much good practical information. Recursive partitioning fared poorly.

.

[123] Richard F. Potthoff and S. N. Roy. "A generalized multivariate analysis of variance model useful especially for growth curve problems". In: *Biometrika* 51 (1964), pp. 313–326 (cit. on p. 6-6).

included an AR1 example

.

[124] Peter Radchenko and Gareth M. James. "Variable inclusion and shrinkage algorithms". In: *J Am Stat Assoc* 103.483 (2008), pp. 1304–1315 (cit. on p. 2-33).

solves problem caused by lasso using the same penalty parameter for variable selection and shrinkage which causes lasso to have to keep too many variables in the model to avoid overshrinking the remaining predictors;does not handle scaling issue well

.

[125] D. R. Ragland. "Dichotomizing continuous outcome variables: Dependence of the magnitude of association and statistical power on the cutpoint". In: *Epi* 3 (1992). See letters to editor May 1993 P. 274-, Vol 4 No. 3, pp. 434–440. url: http://dx.doi.org/10.1097/00001648-199209000-00009 (cit. on p. 2-13).

[126] Brendan M. Reilly and Arthur T. Evans. "Translating clinical research into clinical practice: Impact of using prediction rules to make decisions". In: *Ann Int Med* 144 (2006), pp. 201–209 (cit. on p. 1-10).

impact analysis;example of decision aid being ignored or overruled making MD decisions worse;assumed utilities are constant across subjects by concluding that directives have more impact than predictions;Goldman-Cook clinical prediction rule in AMI

.

[127] Ellen B. Roecker. "Prediction error and its estimation for subset-selected models". In: *Technometrics* 33 (1991), pp. 459–468 (cit. on pp. 3-14, 4-16).

[128] Patrick Royston, Douglas G. Altman, and Willi Sauerbrei. "Dichotomizing continuous predictors in multiple regression: a bad idea". In: *Stat Med* 25 (2006), pp. 127–141. url: http://dx.doi.org/10.1002/sim.2331 (cit. on p. 2-13).

destruction of statistical inference when cutpoints are chosen using the response variable; varying effect estimates when change cutpoints;difficult to interpret effects when dichotomize;nice plot showing effect of categorization; PBC data

.

[129] Warren Sarle. "The VARCLUS Procedure". In: *SAS/STAT User's Guide*. fourth. Vol. 2. Cary, NC: SAS Institute, Inc., 1990. Chap. 43, pp. 1641–1659. url: http://support.sas.com/documentation/onlinedoc/stat (cit. on pp. 3-23, 3-26).

[130] Willi Sauerbrei and Martin Schumacher. "A bootstrap resampling procedure for model building: Application to the Cox regression model". In: *Stat Med* 11 (1992), pp. 2093–2109 (cit. on pp. 3-15, 4-17).

[131] G. Schulgen et al. "Outcome-oriented cutpoints in quantitative exposure". In: *Am J Epi* 120 (1994), pp. 172–184 (cit. on pp. 2-13, 2-15).

[132] E. Selvin et al. "Glycated hemoglobin, diabetes, and cardiovascular risk in nondiabetic adults". In: *NEJM* 362.9 (Mar. 2010), pp. 800–811. url: http://dx.doi.org/10.1056/NEJMoa0908359 (cit. on p. 2-24).

[133] Stephen Senn. "Change from baseline and analysis of covariance revisited". In: *Stat Med* 25 (2006), pp. 4334–4344 (cit. on p. 6-5).

shows that claims that in a 2-arm study it is not true that ANCOVA requires the population means at baseline to be identical;refutes some claims of lia00lon;problems with counterfactuals;temporal additivity ("amounts to supposing that despite the fact that groups are difference at baseline they would show the same evolution over time");causal additivity;is difficult to design trials for which simple analysis of change scores is unbiased, ANCOVA is biased, and a causal interpretation can be given;temporally and logically, a "baseline cannot be a *response* to treatment", so baseline and response cannot be modeled in an integrated framework as Laird and Ware's model has been used;"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.";"You cannot establish necessary conditions for an estimator to be valid by nominating a model and seeing what the model implies unless the model is universally agreed to be impeccable. On the contrary it is appropriate to start with the estimator and see what assumptions are implied by valid conclusions.";this is in distinction to lia00lon
.

[134] Jun Shao. "Linear model selection by cross-validation". In: *J Am Stat Assoc* 88 (1993), pp. 486–494 (cit. on p. 4-17).

[135] Noah Simon et al. "A sparse-group lasso". In: *J Comp Graph Stat* 22.2 (2013), pp. 231–245. eprint: `http://www.tandfonline.com/doi/pdf/10.1080/10618600.2012.681250`. url: `http://www.tandfonline.com/doi/abs/10.1080/10618600.2012.681250` (cit. on p. 2-34).

sparse effects both on a group and within group levels;can also be considered special case of group lasso allowing overlap between groups
.

[136] Sean L. Simpson et al. "A linear exponent AR(1) family of correlation structures". In: *Stat Med* 29 (2010), pp. 1825–1838 (cit. on p. 6-13).

[137] L. R. Smith, F. E. Harrell, and L. H. Muhlbaier. "Problems and potentials in modeling survival". In: *Medical Effectiveness Research Data Methods (Summary Report), AHCPR Pub. No. 92-0056*. Ed. by Mary L. Grady and Harvey A. Schwartz. Rockville, MD: US Dept. of Health, Human Services, Agency for Health Care Policy, and Research, 1992, pp. 151–159. url: `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/FrankHarrell/smi92pro.pdf` (cit. on p. 3-17).

[138] Ian Spence and Robert F. Garrison. "A remarkable scatterplot". In: *Am Statistician* 47 (1993), pp. 12–19 (cit. on p. 3-39).

[139] D. J. Spiegelhalter. "Probabilistic prediction in patient management and clinical trials". In: *Stat Med* 5 (1986), pp. 421–433 (cit. on pp. 3-20, 3-46, 4-20, 4-21).

z-test for calibration inaccuracy (implemented in Stata, and R Hmisc package's `val.prob` function)
.

[140] Ewout W. Steyerberg. *Clinical Prediction Models*. New York: Springer, 2009 (cit. on pp. x, 10-11).

[141] Ewout W. Steyerberg et al. "Prognostic modeling with logistic regression analysis: In search of a sensible strategy in small data sets". In: *Med Decis Mak* 21 (2001), pp. 45–56 (cit. on p. 3-1).

[142] Ewout W. Steyerberg et al. "Prognostic modelling with logistic regression analysis: A comparison of selection and estimation methods in small data sets". In: *Stat Med* 19 (2000), pp. 1059–1079 (cit. on p. 2-33).

[143] C. J. Stone. "Comment: Generalized additive models". In: *Statistical Sci* 1 (1986), pp. 312–314 (cit. on p. 2-25).

[144] C. J. Stone and C. Y. Koo. "Additive splines in statistics". In: *Proceedings of the Statistical Computing Section ASA*. Washington, DC, 1985, pp. 45–48 (cit. on pp. 2-21, 2-26).

[145] Samy Suissa and Lucie Blais. "Binary regression with continuous outcomes". In: *Stat Med* 14 (1995), pp. 247–255. url: `http://dx.doi.org/10.1002/sim.4780140303` (cit. on p. 2-13).

[146] Guo-Wen Sun, Thomas L. Shook, and Gregory L. Kay. "Inappropriate use of bivariable analysis to screen risk factors for use in multivariable analysis". In: *J Clin Epi* 49 (1996), pp. 907–916 (cit. on p. 3-18).

[147] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *J Roy Stat Soc B* 58 (1996), pp. 267–288 (cit. on p. 2-33).

[148] Ben Van Calster et al. "A calibration hierarchy for risk models was defined: from utopia to empirical data". In: *Journal of Clinical Epi* 74 (June 2016), pp. 167–176. issn: 08954356. url: `http://dx.doi.org/10.1016/j.jclinepi.2015.12.005` (cit. on p. 4-4).

[149] J. C. van Houwelingen and S. le Cessie. "Predictive value of statistical models". In: *Stat Med* 9 (1990), pp. 1303–1325 (cit. on pp. 2-26, 3-20, 4-17, 4-20, 4-22).

[150] William N. Venables and Brian D. Ripley. *Modern Applied Statistics with S*. Fourth. New York: Springer-Verlag, 2003. isbn: 0387954570 (cit. on p. 8-2).

[151] Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data*. New York: Springer, 2000.

[152] Pierre J. Verweij and Hans C. van Houwelingen. "Penalized likelihood in Cox regression". In: *Stat Med* 13 (1994), pp. 2427–2436 (cit. on p. 3-20).

[153] Andrew J. Vickers. "Decision analysis for the evaluation of diagnostic tests, prediction models, and molecular markers". In: *Am Statistician* 62.4 (2008), pp. 314–320 (cit. on p. 1-7).

limitations of accuracy metrics;incorporating clinical consequences;nice example of calculation of expected outcome;drawbacks of conventional decision analysis, especially because of the difficulty of eliciting the expected harm of a missed diagnosis;use of a threshold on the probability of disease for taking some action;decision curve;has other good references to decision analysis
.

[154] Eric Vittinghoff and Charles E. McCulloch. "Relaxing the rule of ten events per variable in logistic and Cox regression". In: *Am J Epi* 165 (2006), pp. 710–718 (cit. on p. 3-17).

the authors may have not been quite stringent enough in their assessment of adequacy of predictions;letter to the editor submitted
.

[155] Howard Wainer. "Finding what is not there through the unfortunate binning of results: The Mendel effect". In: *Chance* 19.1 (2006), pp. 49–56 (cit. on pp. 2-13, 2-16).

can find bins that yield either positive or negative association;especially pertinent when effects are small;"With four parameters, I can fit an elephant; with five, I can make it wiggle its trunk." - John von Neumann
.

[156] Hansheng Wang and Chenlei Leng. "Unified LASSO estimation by least squares approximation". In: *J Am Stat Assoc* 102 (2007), pp. 1039–1048. url: `http://dx.doi.org/10.1198/016214507000000509` (cit. on p. 2-33).

[157] S. Wang et al. "Hierarchically penalized Cox regression with grouped variables". In: *Biometrika* 96.2 (2009), pp. 307–322 (cit. on p. 2-34).

[158] Yohanan Wax. "Collinearity diagnosis for a relative risk regression analysis: An application to assessment of diet-cancer relationship in epidemiological studies". In: *Stat Med* 11 (1992), pp. 1273–1287 (cit. on p. 3-23).

[159] John Whitehead. "Sample size calculations for ordered categorical data". In: *Stat Med* 12 (1993). See letter to editor SM 15:1065-6 for binary case;see errata in SM 13:871 1994;see kol95com, jul96sam, pp. 2257–2271 (cit. on p. 3-18).

[160] Ryan E. Wiegand. "Performance of using multiple stepwise algorithms for variable selection". In: *Stat Med* 29 (2010), pp. 1647–1659 (cit. on p. 3-14).

fruitless to try different stepwise methods and look for agreement;the methods will agree on the wrong model
.

[161] Daniela M. Witten and Robert Tibshirani. "Testing significance of features by lassoed principal components". In: *Ann Appl Stat* 2.3 (2008), pp. 986–1012 (cit. on p. 2-34).

reduction in false discovery rates over using a vector of t-statistics;borrowing strength across genes;"one would not expect a single gene to be associated with the outcome, since, in practice, many genes work together to effect a particular phenotype. LPC effectively down-weights individual genes that are associated with the outcome but that do not share an expression pattern with a larger group of genes, and instead favors large groups of genes that appear to be differentially-expressed.";regress principal components on outcome;sparse principal components
.

[162] S. N. Wood. *Generalized Additive Models: An Introduction with R*. ISBN 9781584884743. Boca Raton, FL: Chapman & Hall/CRC, 2006 (cit. on p. 2-35).

[163] C. F. J. Wu. "Jackknife, bootstrap and other resampling methods in regression analysis". In: *Ann Stat* 14.4 (1986), pp. 1261–1350 (cit. on p. 4-17).

[164] Shifeng Xiong. "Some notes on the nonnegative garrote". In: *Technometrics* 52.3 (2010), pp. 349–361 (cit. on p. 2-34).

"... to select tuning parameters, it may be unnecessary to optimize a model selectin criterion repeatedly";natural selection of penalty function
.

[165] Jianming Ye. "On measuring and correcting the effects of data mining and model selection". In: *J Am Stat Assoc* 93 (1998), pp. 120–131 (cit. on p. 1-14).

[166] 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". In: *Psychometrika* 43 (1978), pp. 279–281 (cit. on p. 3-27).

[167] Hao H. Zhang and Wenbin Lu. "Adaptive lasso for Cox's proportional hazards model". In: *Biometrika* 94 (2007), pp. 691–703 (cit. on p. 2-33).

penalty function has ratios against original MLE;scale-free lasso
.

[168] Hui Zhou, Trevor Hastie, and Robert Tibshirani. "Sparse principal component analysis". In: *J Comp Graph Stat* 15 (2006), pp. 265–286 (cit. on p. 2-34).

principal components analysis that shrinks some loadings to zero
.

[169] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net". In: *J Roy Stat Soc B* 67.2 (2005), pp. 301–320 (cit. on p. 2-33).

R packages written by FE Harrell are freely available from CRAN, and are managed at github.com/harrelfe.

To obtain a book with detailed examples and case studies and notes on the theory and applications of survival analysis, logistic regression, ordinal regression, linear models, and longitudinal models, order Regression Modeling Strategies *with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, 2nd Edition* by FE Harrell from Springer NY (2015). Steyerberg [140] and Dupont [45] 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`.