BIOSTATISTICAL MODELING

Frank E Harrell Jr
Department of Biostatistics
Vanderbilt University School of Medicine
Nashville TN USA
f.harrell@vanderbilt.edu
http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/BioMod
June 1, 2004

Contents

Chapter 1

Introduction

H 1

1.1 Hypothesis Testing, Estimation, and Prediction

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

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

mates

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

Statistical estimation is usually model-based

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

1.2 Examples of Uses of Predictive Multivariable Modeling

- Financial performance, consumer purchasing, loan pay-back
- Ecology
- Product life
- Employment discrimination
- Medicine, epidemiology, health services research
- Probability of diagnosis, time course of a disease
- Comparing non-randomized treatments

- Getting the correct estimate of relative effects in randomized studies requires covariable adjustment if model is nonlinear
  - Crude odds ratios biased towards 1.0 if sample heterogeneous

- Estimating absolute treatment effect (e.g., risk difference)
  - Use e.g. difference in two predicted probabilities

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

1.3 Planning for Modeling

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

- Sites

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

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

1.4 Choice of the Model

- In biostatistics and epidemiology we usually choose model empirically
- Model must use data efficiently
- Should model overall structure (e.g., acute vs. chronic)
- Robust models are better
- Should have correct mathematical structure (e.g., constraints on probabilities)

1.5 Model uncertainty / Data-driven Model Specification

- Standard errors, C.L., P-values, R^2 wrong if computed as if the model pre-specified

- Stepwise variable selection is widely used and abused

- Bootstrap can be used to repeat all analysis steps to properly penalize variances, etc.

- Ye : "generalized degrees of freedom" (GDF) for any "data mining" or model selection procedure based on least squares
  - Example: 20 candidate predictors, $n = 22$, forward stepwise, best 5-variable model: GDF=14.1

  - Example: CART, 10 candidate predictors, $n = 100$, 19 nodes: GDF=76

Chapter 2

Simple Linear Regression

Rosner 11.1-6

2.1 Notation

- y : random variable representing response variable
- x : random variable representing independent variable (subject descriptor, predictor, covariable
  - conditioned upon
  - treating as constants, measured without error
- What does conditioning mean?
  - holding constant
  - subsetting on
  - slicing scatterplot vertically

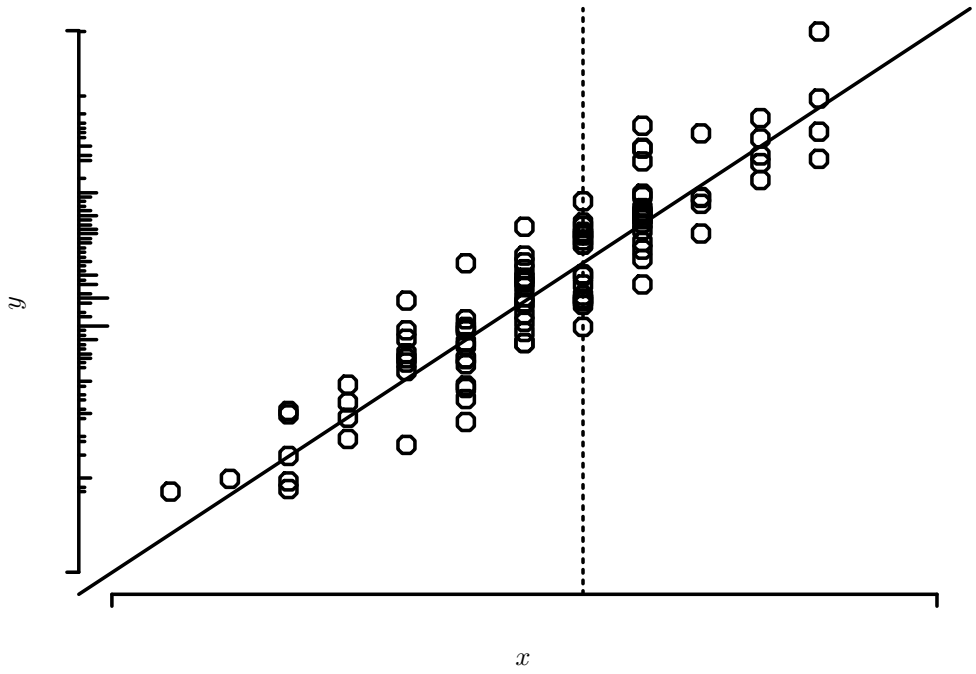

Figure 2.1: *Data from a sample of $n = 100$ points along with population linear regression line. The x variable is discrete. The conditional distribution of $y|x$ can be thought of as a vertical slice at x. The unconditional distribution of y is shown on the y-axis.*

- $E(y|x)$: population expected value or long-run average of y conditioned on the value of x
  Example: population average blood pressure for a 30-year old

- α : y-intercept

- β : slope of y on x ($\frac{\Delta y}{\Delta x}$)

Simple linear regression is used when

- Only two variables are of interest

- One variable is a response and one a predictor

- No adjustment is needed for confounding or other between-subject variation

- The investigator is interested in assessing the strength of the relationship between x and y in real data units, or in predicting y from x

- A linear relationship is assumed (why assume this? why not use nonparametric regression?)

- Not when one only needs to test for association (use Spearman's ρ rank correlation) or estimate a correlation index

2.2 Two Ways of Stating the Model

- $E(y|x) = \alpha + \beta x$

- $y = \alpha + \beta x + e$
  e is a random error (residual) representing variation between subjects in y

even if x is constant, e.g. variation in blood pressure for patients of the same age

2.3 Assumptions, If Inference Needed

- Conditional on x, y is normal with mean $\alpha+\beta x$ and constant variance σ^2, **or:**
- e is normal with mean 0 and constant variance σ^2
- $E(y|x) = E(\alpha+\beta x+e) = \alpha+\beta x+E(e)$,
  $E(e) = 0$.
- Observations are independent

2.4 How Can α and β be Estimated?

- Need a criterion for what are good estimates
- **One** criterion is to choose values of the two parameters that minimize the sum of squared errors in predicting individual subject responses
- Let a, b be guesses for α, β
- Sample of size n : $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$
- $SSE = \sum_{i=1}^{n}(y_i - a - bx_i)^2$
- Values that minimize SSE are *least squares estimates*

- These are obtained from

$$L_{xx} = \sum(x_i - \bar{x})^2 \quad L_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$$
$$\hat{\beta} = b = \frac{L_{xy}}{L_{xx}} \qquad \hat{\alpha} = a = \bar{y} - b\bar{x}$$

- Note: A term from L_{xy} will be positive when x and y are concordant in terms of both being above their means or both being below their means.

2.5 Inference about Parameters

- Residual: $d = y - \hat{y}$

- d large if line was not the proper fit to the data or if there is large variability across subjects for the same x

- Beware of that many authors combine both components when using the terms *goodness of fit* and *lack of fit*

- Might be better to think of lack of fit as being due to a structural defect in the model (e.g., nonlinearity)

- $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$
  $SSR = \sum(\hat{y}_i - \bar{y})^2$
  $SSE = \sum(y_i - \hat{y}_i)^2$
  $SST = SSR + SSE$
  $SSR = SST - SSE$

- SS increases in proportion to n

- Mean squares: normalized for for d.f.: $\frac{SS}{d.f.(SS)}$

- $MSR = SSR/p$, p = no. of parameters besides intercept (here, 1)
  $MSE = SSE/(n-p-1)$ (sample conditional variance of y)
  $MST = SST/(n-1)$ (sample unconditional variance of y)

- Brief review of ordinary ANOVA (analysis of variance):
  - Generalizes 2-sample t-test to > 2 groups
  - SSR is SS between treatment means
  - SSE is SS within treatments, summed over treatments

- ANOVA Table for Regression

Source	d.f.	SS	MS	F
Regression	p	SSR	$MSR = SSR/p$	MSR/MSE
Error	$n-p-1$	SSE	$MSE = SSE/(n-p-1)$	
Total	$n-1$	SST	$MST = SST/(n-1)$	

- Statistical evidence for large values of β can be summarized by $F = \frac{MSR}{MSE}$

- Has F distribution with p and $n-p-1$ d.f.

- Large values $\rightarrow |\beta|$ large

2.6 Estimating σ, S.E. of $\hat{\beta}$; t-test

- $s^2_{y\cdot x} = \hat{\sigma}^2 = MSE = \widehat{Var}(y|x) = \widehat{Var}(e)$

- $\widehat{se}(b) = s_{y\cdot x}/L_{xx}^{\frac{1}{2}}$

- $t = b/\widehat{se}(b), n-p-1$ d.f.

- $t^2 \equiv F$ when $p = 1$

- $t_{n-2} \equiv \sqrt{F_{1,n-2}}$

- t identical to 2-sample t-test (x has two values)

- If x takes on only the values 0 and 1, b equals $\bar{y}$ when $x = 1$ minus $\bar{y}$ when $x = 0$

2.7 Interval Estimation

Rosner 11.5

- 2-sided $1 - \alpha$ CI for β: $b \pm t_{n-2,1-\alpha/2}\widehat{se}(b)$

- CI for *predictions* depend on what you want to predict even though $\hat{y}$ estimates both y [a] and $E(y|x)$

- Notation for these two goals: $\hat{y}$ and $\hat{E}(y|x)$
  - Predicting y with $\hat{y}$:
    $\widehat{s.e.}(\hat{y}) = s_{y\cdot x}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{L_{xx}}}$
    Note: This s.e. $\rightarrow s_{y\cdot x}$ as $n \rightarrow \infty$.

  - Predicting $\hat{E}(y|x)$ with $\hat{y}$:
    $\widehat{s.e.}(\hat{E}(y|x)) = s_{y\cdot x}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{L_{xx}}}$ See footnote[b]
    Note: This s.e. shrinks to 0 as $n \rightarrow \infty$

- $1 - \alpha$ 2-sided CI for either one:
  $\hat{y} \pm t_{n-p-1,1-\alpha/2}\widehat{s.e.}$

[a] With a normal distribution, the least dangerous guess for an individual y is the estimated mean of y.

[b] n here is the grand total number of observations because we are borrowing information about neighboring x-points, i.e., using interpolation. If we didn't assume anything and just computed mean y at each separate x, the standard error would instead by estimated by $s_{y\cdot x}\sqrt{\frac{1}{m}}$, where m is the number of original observations with x exactly equal to the x for which we are obtaining the prediction. The latter s.e. is much larger than the one from the linear model.

- Wide CI (large $\widehat{s.e.}$) due to:
  - small n
  - large σ^2
  - being far from the data center ($\bar{x}$)
- Example usages:
  - Is a child of age x smaller than predicted for her age?
    Use $s.e.(\hat{y})$
  - What is the best estimate of the population mean blood pressure for patients on treatment A?
    Use $s.e.(\hat{E}(y|x))$
- Example pointwise 0.95 confidence bands:

x	1	3	5	6	7	9	11
y:	5	10	70	58	85	89	135

2.8 Assessing Goodness of Fit

Rosner 11.6

Assumptions:

1. Linearity
2. σ^2 is constant, independent of x
3. Observations (e's) are independent of each other
4. For proper statistical inference (CI, P-values), y (e) is normal conditional on x

Verifying some of the assumptions:

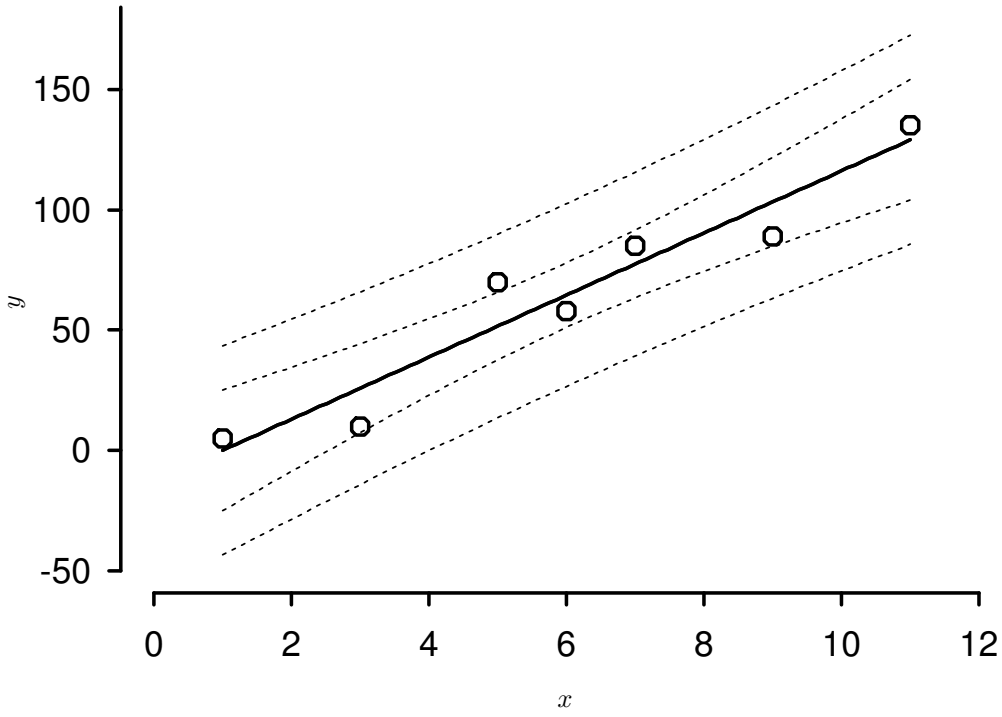

Figure 2.2: *Pointwise 0.95 confidence intervals for $\hat{y}$ (wider bands) and $\hat{E}(y|x)$ (narrower bands).*

- In a scattergram the spread of y about the fitted line should be constant as x increases, and y vs. x should appear linear

- Easier to see this with a plot of $\hat{d} = y - \hat{y}$ vs. $\hat{y}$

- In this plot there are no systematic patterns (no trend in central tendency, no change in spread of points with x)

- Trend in central tendency indicates failure of linearity

- `qqnorm` plot of d

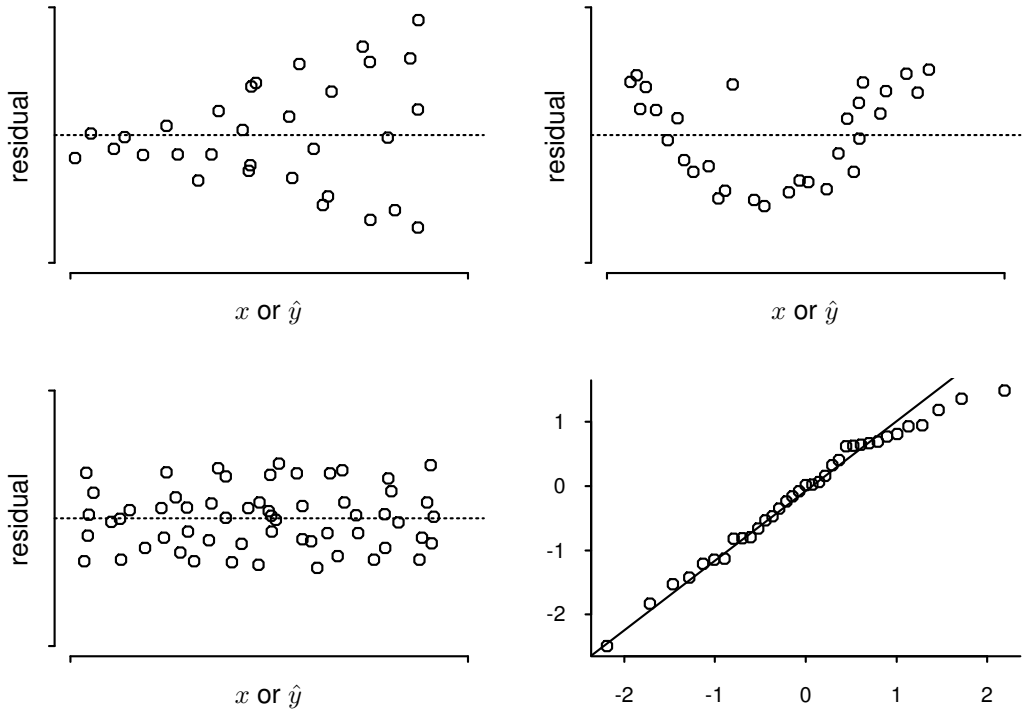

Figure 2.3: *Using residuals to check some of the assumptions of the simple linear regression model. Top left panel depicts non-constant σ^2, which might call for transforming y. Top right panel shows constant variance but the presence of a systemic trend which indicates failure of the linearity assumption. Bottom left panel shows the ideal situation of white noise (no trend, constant variance). Bottom right panel shows a $q - q$ plot that demonstrates approximate normality of residuals, for a sample of size $n = 35$. Horizontal reference lines are at zero, which is by definition the mean of all residuals.*

Chapter 3

Multiple Linear Regression

Rosner 11.9

3.1 The Model and How Parameters are Estimated

- p independent variables $x_1, x_2, \ldots, x_p$
- Examples: multiple risk factors, treatment plus patient descriptors when adjusting for non-randomized treatment selection in an observational study
- Each variable has its own effect (slope) representing *partial effects*: effect of increasing a variable by one unit, holding all others constant
- Initially assume that the different variables act in an additive fashion
- Assume the variables act linearly against y
- Model: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e$
- Or: $E(y|x) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

- For two x-variables: $y = \alpha + \beta_1 x_1 + \beta_2 x_2$

- Estimated equation: $\hat{y} = a + b_1 x_1 + b_2 x_2$

- Least squares criterion for fitting the model (estimating the parameters): $SSE = \sum_{i=1}^{n} [y - (a + b_1 x_1 + b_2 x_2)]^2$

- Solve for a, b_1, b_2 to minimize SSE

- When $p > 1$, least squares estimates require complex formulas; still all of the coefficient estimates are weighted combinations of the y's, $\sum w_i y_i$[a].

3.2 Interpretation of Parameters

- Regression coefficients are (b) are commonly called *partial regression coefficients*: effects of each variable holding all other variables in the model constant

- Examples of partial effects:
  - model containing x_1=age (years) and x_2=sex (0=male 1=female) Coefficient of age (β_1) is the change in the mean of y for males when age increases by 1 year. It is also the change in y per unit increase in age for females. β_2 is the female minus male mean difference in y for two subjects of the same age.
  - $E(y|x_1, x_2) = \alpha + \beta_1 x_1$ for males, $\alpha + \beta_1 x_1 + \beta_2 = (\alpha + \beta_2) + \beta_1 x_1$ for females [the sex effect is a shift effect or change in y-intercept]
  - model with age and systolic blood pressure measured when the study begins

[a] When $p = 1$, the w_i for estimating β are $\frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$

Coefficient of blood pressure is the change in mean y when blood pressure increases by 1mmHg for subjects of the same age

- What is meant by changing a variable?
  - We usually really mean a comparison of two subjects with different blood pressures
  - Or we can envision what would be the expected response had *this* subject's blood pressure been 1mmHg greater at the outset[b]
  - We are not speaking of longitudinal changes in a single person's blood pressure
  - We can use subtraction to get the adjusted (partial) effect of a variable, e.g., $E(y|x_1, x_2) - \beta_2 x_2 = \alpha + \beta_1 x_1$
- Example: $\hat{y} = 37 + .01 \times$ weight $+ 0.5 \times$ cigarettes smoked per day
  - .01 is the estimate of average increase y across subjects when weight is increased by 1lb. if cigarette smoking is unchanged
  - 0.5 is the estimate of the average increase in y across subjects per additional cigarette smoked per day if weight does not change
  - 37 is the estimated mean of y for a subject of zero weight who does not smoke
- Comparing regression coefficients:
  - Can't compare directly because of different units of measurement. Coefficients in units of $\frac{y}{x}$.

[b]This setup is the basis for randomized controlled trials. Drug effects may be estimated with between-patient group differences under a statistical model.

– Standardizing by standard deviations: not recommended. Standard deviations are not magic summaries of scale and they give the wrong answer when an x is categorical (e.g., sex).

3.3 Hypthesis Testing

Rosner 11.9.2

3.3.1 Testing Total Association (Global Null Hypotheses)

- ANOVA table is same as before for general p
- $F_{p,n-p-1}$ tests $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$
- This is a test of *total association*, i.e., a test of whether *any* of the predictors is associated with y
- To assess total association we accumulate partial effects of all variables in the model *even though* we are testing if *any* of the partial effects is nonzero
- H_a : at least one of the β's is non-zero. **Note**: This does not mean that all of the x variables are associated with y.
- Weight and smoking example: H_0 tests the null hypothesis that neither weight nor smoking is associated with y. H_a is that at least one of the two variables is associated with y. The other may or may not have a non-zero β.
- Test of total association does not test whether cigarette smoking is related to y holding weight constant.
- SSR can be called the model SS

3.3.2 Testing Partial Effects

- $H_0 : \beta_1 = 0$ is a test for the effect of x_1 on y holding x_2 and any other x's constant

- Note that β_2 is *not* part of the null or alternative hypothesis; we assume that we have adjusted for *whatever* effect x_2 has, *if any*

- One way to test β_1 is to use a t-test: $t_{n-p-1} = \frac{b_1}{\widehat{s.e.}(b_1)}$

- In multiple regression it is difficult to compute standard errors so we use a computer

- These standard errors, like the one-variable case, decrease when
  - $n \uparrow$
  - variance of the variable being tested $\uparrow$
  - σ^2 (residual y-variance) $\downarrow$

- Another way to get partial tests: the F test
  - Gives identical 2-tailed P-value to t test when one x being tested $t^2 \equiv$ partial F
  - Allows testing for > 1 variable
  - Example: is either systolic or diastolic blood pressure (or both) associated with the time until a stroke, holding weight constant

- To get a partial F define partial SS

- Partial SS is the change in SS when the variables **being tested** are dropped from the model and the model is re-fitted

- A general principle in regression models: a set of variables can be tested for their combined partial effects by removing that set of variables from the model and measuring the harm ($\uparrow SSE$) done to the model

- Let $full$ refer to computed values from the full model including all variables; $reduced$ denotes a reduced model containing only the adjustment variables and not the variables being tested

- Dropping variables $\uparrow SSE, \downarrow SSR$ unless the dropped variables had exactly zero slope estimates in the full model (which never happens)

- $SSE_{reduced} - SSE_{full} = SSR_{full} - SSR_{reduced}$
  Numberator of F test can use either SSE or SSR

- Form of partial F-test: change in SS when dropping the variables of interest divided by change in d.f., then divided by MSE;
  MSE is chosen as that which best estimates σ^2, namely the MSE from the full model

- Full model has p slopes; suppose we want to test q of the slopes

$$\begin{aligned} F_{q,n-p-1} &= \frac{(SSE_{reduced} - SSE_{full})/q}{MSE} \\ &= \frac{(SSR_{full} - SSR_{reduced})/q}{MSE} \end{aligned}$$

3.4 Assessing Goodness of Fit

Rosner 11.9.3

Assumptions:

- Linearity of each predictor against y holding others constant
- σ^2 is constant, independent of x
- Observations (e's) are independent of each other
- For proper statistical inference (CI, P-values), y (e) is normal conditional on x
- x's act additively; effect of x_j does not depend on the other x's (**But** note that the x's may be correlated with each other without affecting what we are doing.)

Verifying some of the assumptions:

1. When $p = 2$, x_1 is continuous, and x_2 is binary, the pattern of y vs. x_1, with points identified by x_2, is two straight, parallel lines. β_2 is the slope of y vs. x_2 holding x_1 constant, which is just the difference in means for $x_2 = 1$ vs. $x_2 = 0$ as $\Delta x_2 = 1$ in this simple case.
2. In a residual plot ($d = y - \hat{y}$ vs. $\hat{y}$) there are no systematic patterns (no trend in central tendency, no change in spread of points with $\hat{y}$). The same is true if one plots d vs. any of the x's (these are more stringent assessments). If x_2 is binary box plots of d stratified by x_2 are effective.
3. Partial residual plots reveal the partial (adjusted) relationship between a chosen x_j and y, controlling for all other $x_i, i \neq j$, without assuming linearity for x_j. In these plots, the following quantities appear on the axes:

   y **axis:** residuals from predicting y from all predictors except x_j

   x **axis:** residuals from predicting x_j from all predictors except x_j (y is ignored)

   Partial residual plots ask how does what we can't predict about y without knowing x_j depend on what we can't predict about x_j from the other x's.

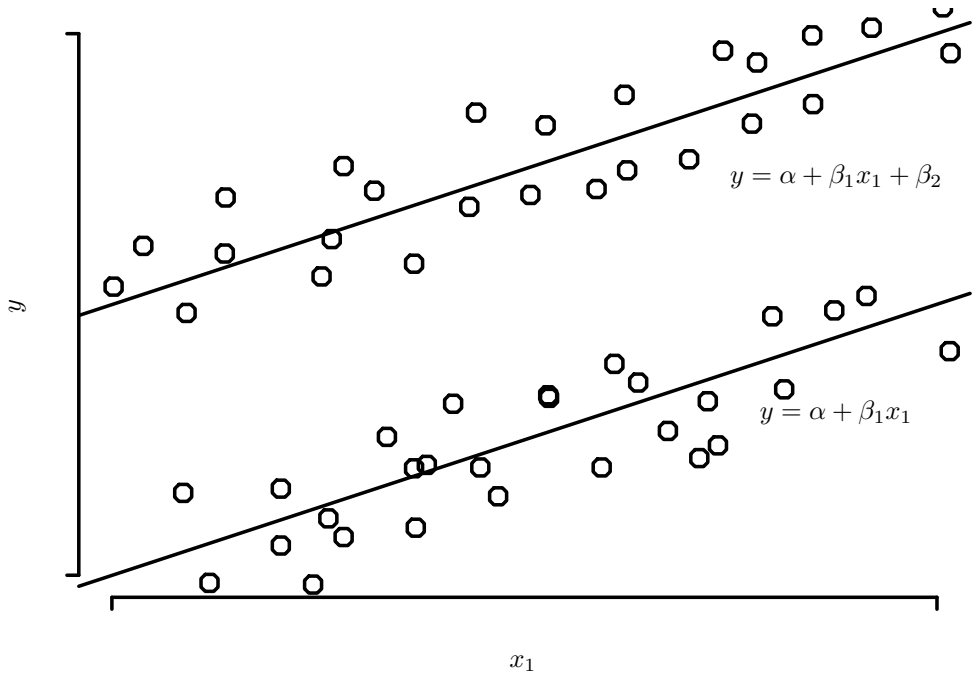

Figure 3.1: *Data satisfying all the assumptions of simple multiple linear regression in two predictors. Note equal spread of points around the population regression lines for the $x_2 = 1$ and $x_2 = 0$ groups (upper and lower lines, respectively) and the equal spread across x_1. The $x_2 = 1$ group has a new intercept, $\alpha + \beta_2$, as the x_2 effect is β_2.*

3.5 What are Degrees of Freedom

For a model : the total number of parameters not counting intercept(s)

For a hypothesis test : the number of parameters that are hypothesized to equal specified constants. The constants specified are usually zeros (for *null* hypotheses) but this is not always the case. Some tests involve combinations of multiple parameters but test this combination against a single constant; the d.f. in this case is still one. Example: $H_0 : \beta_3 = \beta_4$ is the same as $H_0 : \beta_3 - \beta_4 = 0$ and is a 1 d.f. test because it tests one parameter ($\beta_3 - \beta_4$) against a constant (0).

These are **numerator d.f.** in the sense of the F-test in multiple linear regression. The F-test also entails a second kind of d.f., the **denominator** or **error** d.f., $n - p - 1$, where p is the number of parameters aside from the intercept. The error d.f. is the denominator of the estimator for σ^2 that is used to unbias the estimator, penalizing for having estimated $p + 1$ parameters by minimizing the sum of squared errors used to estimate σ^2 itself. You can think of the error d.f. as the sample size penalized for the number of parameters estimated, or as a measure of the information base used to fit the model.

Chapter 4

S `Design` Library for Multiple Linear Regression

H 6

4.1 Formula Language and Fitting Function

- Statistical formula in S:

  `y ~ x1 + x2 + x3`

  `y` *is modeled as* $\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

- Formula is the first argument to a *fitting* function (just as it is the first argument to a `trellis` graphics function)

- `Design` library makes many aspects of regression modeling and graphical display of model results easier to do

- `Design` gets its name from the bookkeeping it does to remember details about the *design matrix* for the model and to use these details in making automatic hypothesis tests, estimates, and plots. The design matrix is the matrix of independent variables after coding them numerically and adding nonlinear and product terms if needed.

- To use `Design` you need to first get access to the `Hmisc` library. It is imperative that access to the libraries is does as below, with the *order indicated*. You can attach libraries with `File ... Load Libraries` but it is laborious and too easy to forget to check the box `Attach at Top of Search List`. So define a `.First` function once and for all in your project area:

```
.First ← function(...) {
  library(Hmisc,T)
  library(Design,T)
  invisible()
  }
```

  `.First` will be executed the next time you start S from that project directory. If you don't want to exit a current S session to make this happen, just type `.First()`.

- `Design` library fitting function for ordinary least squares regression: `ols`

- Example:

```
f ← ols(y ∼ age + sys.bp)
```

  `f` is an S list object, containing coefficients, variances, and many other quantities. `f` is called the *fit object*. Below the fit object will be `f` throughout. In practice, use any legal S name, e.g. `fit.full.model`.

4.2 Operating on the Fit Object

- Regression coefficient estimates may be obtained by any of the methods listed below

```
f$coefficients
f$coef            # abbreviation
coef(f)           # use the coef extractor function
coef(f)[1]        # get intercept
f$coef[2]         # get 2nd coefficient (1st slope)
f$coef['age']     # get coefficient of age
coef(f)['age']    # ditto
```

- But often we use *methods* which do something more interesting with the model fit.

  `print(f)` : print coefficients, standard errors, t-test, other statistics; can also just type `f` to print

  `fitted(f)` : compute $\hat{y}$

  `predict(f, newdata)` : get predicted values, for subjects described in data frame `newdata`[a]

  `r ←resid(f)` : compute the vector of n residuals (here, store it in `r`)

  `formula(f)` : print the regression formula fitted

  `anova(f)` : print ANOVA table for all total and partial effects

  `summary(f)` : print estimates partial effects using meaningful changes in predictors

  `plot(f)` : plot partial effects, with predictor ranging over the x-axis

  `g ←Function(f)` : create an S function that evaluates the analytic form of the fitted function

  `nomogram(f)` : draw a nomogram of the model

4.3 The `Design datadist` Function

To use `plot, summary`, or `nomogram` in the `Design` library, you need to let `Design` first compute summaries of the distributional characteristics of the predictors:

```
dd ← datadist(x1,x2,x3,...)    # generic form
dd ← datadist(age, sys.bp, sex)
options(datadist='dd')
```

Note that the name `dd` can be any name you choose as long as you use the same name in quotes to `options` that you specify (unquoted) to the left of `←datadist(...)`. It is best to invoke `datadist` early in your program before fitting any models.

[a] You can get confidence limits for predicted means or predicted individual responses using the `conf.int` and `conf.type` arguments to `predict`. `predict(f)` without the `newdata` argument yields the same result as `fitted(f)`.

That way the datadist information is stored in the fit object so the model is self-contained. That allows you to make plots in later sessions without worrying about datadist.
datadist must be re-run if you add a new predictor or recode an old one. You can update it using for example

```
dd ← datadist(dd, cholesterol, height)
# Adds or replaces cholesterol, height summary stats in dd
```

With the usual S setup objects such as dd are stored in _Data and will be automatically available in future S sessions. You will just need to re-issue options(datadist='dd') in future sessions.

4.4 Operating on Residuals

Residuals may be summarized and plotted just like any raw data variable.

- To see detailed residual plots use the plot.lm function builtin to S:

  ```
  f ← ols(y ~ age + sys.bp)
  plot.lm(f)
  ```

  If there are NAs in the data, the last graph plot.lm tries to produce will bomb. The others should be OK. plot.lm will not plot against each predictor separately[b].

- To take control of the plots and to plot residuals vs. each predictor, use these examples:

  ```
  r ← resid(f)
  plot(fitted(f), r); abline(h=0)   # yhat vs. r
  plot(age,    r);    abline(h=0)
  plot(sys.bp, r);    abline(h=0)
  bwplot(sex ~ r)                   # box plot stratified by sex
  ```

[b] Add the argument smooths=T to plot.lm to get trend lines for the residuals. These lines should be flat if model assumptions hold.

```
qqnorm(r); qqline(r)                # linearity indicates normality
```

4.5 Plotting Partial Effects

- `plot(f)` makes one plot for each predictor
- Predictor is on x-axis, $\hat{y}$ on the y-axis
- Predictors not shown in plot are set to constants
  - median for continuous predictors
  - mode for categorical ones
- For categorical predictor, estimates are shown only at data values
- 0.95 pointwise confidence limits for $\hat{E}(y|x)$ are shown (add `conf.int=F` to suppress CLs)
- Example:

```
par(mfrow=c(2,2))
plot(f)
```

  Makes 3 plots on one page if 3 predictors are in the model.

- To take control of which predictors are plotted, or to specify customized options:

```
plot(f, age=NA)     # plot age effect, using default range,
                    # 10th smallest to 10th largest age
plot(f, age=20:70)# plot age=20,21,...,70
plot(f, age=seq(10,80,length=150))     # plot age=10-80, 150 points
```

- To get confidence limits for $\hat{y}$:

```
plot(f, age=NA, conf.type='individual')
```

- To show both types of 0.99 confidence limits on one plot:

```
# Draw the wider CLs first so all will fit on the plot
plot(f, age=NA, conf.int=0.99, conf.type='individual')
plot(f, age=NA, conf.int=0.99, conf.type='mean', add=T)
# add=T means to add to an existing plot
```

- Non-plotted variables are set to reference values (median and mode by default)

- To control the settings of non-plotted values use e.g.

```
plot(f, age=NA, sex='female')
```

- To make separate lines for the two sexes:

```
plot(f, age=NA, sex=NA)  # add ,conf.int=F to suppress conf. bands
```

- To plot a 3-d surface for two continuous predictors against $\hat{y}$:

```
plot(f, age=NA, cholesterol=NA)
```

4.6 Getting Predicted Values

- Using `predict`

```
predict(f, data.frame(age=30, sex='male'))
# assumes that age and sex are the only variables in the model
```

```
predict(f, data.frame(age=c(30,50), sex=c('female','male')))
# predictions for 30 y.o. female and 50 y.o. male

newdat ← expand.grid(age=c(30,50), sex=levels(sex))
predict(f, newdat)      # 4 predictions

predict(f, newdat, conf.int=0.95)   # also get CLs for mean
predict(f, newdat, conf.int=0.95, conf.type='individual')   # CLs for indiv.
```

See also `gendata` and `Dialog`.

- The brute-force way

```
f ← ols(later.sys.bp ~ age + sex)
# Model is a + b1*age + b2*(sex=='female') if
# levels(sex) = c('male','female') in that order
b ← coef(f)
b[1] + b[2]*30  # prediction for 30 y.o. male, assuming
                # the reference category is sex='male'

b[1] + b[2]*30 + b[3]   # for 30 y.o. female

sexes ← c('female','male')
b[1] + b[2]*30 + b[3]*(sexes=='female')   # for both sexes, 30 y.o.

ages ← 10:20
b[1] + b[2]*ages        # for 10-20 y.o. males
```

- Using `Function` function

```
g ← Function(f)
g(age=10:20, sex='female')                  # 21 predictions
g(age=17)               # for 17 year old of most prevalent sex
```

4.7 ANOVA

- Use `anova(fitobject)` to get all total effects and individual partial effects

- Use `anova(f,age,sex)` to get combined partial effects of `age` and `sex`, for example

- Store result of `anova` in an object in you want to print it various ways, or to plot it:

```
an ← anova(f)
print(an, 'names')      # print names of variables being tested
print(an, 'subscripts')# print subscripts in coef(f) (ignoring
                        # the intercept) being tested
print(an, 'dots')       # a dot in each position being tested
```

- Example:

```
f ← ols(y ∼ x1 + x2 + x3)
an ← anova(f)
print(an, 'subscripts')

               Analysis of Variance          Response: y

Factor     d.f. Partial SS        MS     F      P Tested
x1          1    0.008772   0.008772  0.01 0.9198 1
x2          1    0.017749   0.017749  0.02 0.8861 2
x3          1   23.002598  23.002598 26.76 <.0001 3
REGRESSION  3   23.361519   7.787173  9.06 <.0001 1-3
ERROR      95   81.668570   0.859669

Subscripts correspond to:
[1] x1 x2 x3
print(an, 'dots')

               Analysis of Variance          Response: y

Factor     d.f. Partial SS        MS     F      P Tested
x1          1    0.008772   0.008772  0.01 0.9198 .
x2          1    0.017749   0.017749  0.02 0.8861  .
x3          1   23.002598  23.002598 26.76 <.0001   .
REGRESSION  3   23.361519   7.787173  9.06 <.0001 ...
ERROR      95   81.668570   0.859669
```

```
print(an, 'names')

               Analysis of Variance          Response: y

Factor     d.f. Partial SS        MS     F      P   Tested
x1          1    0.008772   0.008772  0.01 0.9198 x1
x2          1    0.017749   0.017749  0.02 0.8861 x2
x3          1   23.002598  23.002598 26.76 <.0001 x3
REGRESSION  3   23.361519   7.787173  9.06 <.0001 x1,x2,x3
ERROR      95   81.668570   0.859669

an ← anova(f, x2, x3)    # get combined x2,x3 effects
print(an, 'names')

               Analysis of Variance          Response: y

Factor     d.f. Partial SS       MS     F      P Tested
x2          1    0.01775    0.01775  0.02 0.8861 x2
x3          1   23.00260   23.00260 26.76 <.0001 x3
REGRESSION  2   23.03582   11.51791 13.40 <.0001 x2,x3
ERROR      95   81.66857    0.85967
```

Chapter 5

Case Study: Lead Exposure and Neuro-Psychological Function

Rosner 11.10

5.1 Dummy Variable for Two-Level Categorical Predictors

- Categories of predictor: A, B (for example)
- First category = reference cell, gets a zero
- Second category gets a 1.0
- Formal definition of dummy variable: $x = I[category = B]$
  $I[w] = 1$ if w is true, 0 otherwise
- $\alpha + \beta x = \alpha + \beta I[category = B]$ =
  α for category A subjects
  $\alpha + \beta$ for category B subjects
  β = mean difference ($B - A$)

5.2 Two-Sample t-test vs. Simple Linear Regression

- They are equivalent in every sense:
  - P-value
  - Estimates and C.L.s after rephrasing the model
  - Assumptions (equal variance assumption of two groups in t-test is the same as constant variance of $y|x$ for every x)

- $a = \bar{Y}_A$
  $b = \bar{Y}_B - \bar{Y}_A$

- $\widehat{s.e.}(b) = \widehat{s.e.}(\bar{Y}_B - \bar{Y}_A)$

5.3 Analysis of Covariance

- Multiple regression can extend the t-test
  - More than 2 groups (multiple dummy variables can do multiple-group ANOVA)
  - Allow for categorical or continuous adjustment variables (covariates, co-variables)

- Model: $MAXFWT = \alpha + \beta_1 age + \beta_2 sex + e$

- Rosner coded $sex = 1, 2$ for male, female
  Does not affect interpretation of β_2 but makes interpretation of α more tricky (mean $MAXFWT$ when $age = 0$ and $sex = 0$ which is impossible by this coding.

- Better coding would have been $sex = 0, 1$ for male, female
  - α = mean $MAXFWT$ for a zero year-old male
  - β_1 = increase in mean $MAXFWT$ per 1-year increase in age
  - $\beta_2 =$ mean $MAXFWT$ for females minus mean $MAXFWT$ for males, holding age constant
- Model: $MAXFWT = \alpha + \beta_1 CSCN2 + \beta_2 age + \beta_3 sex + e$
  $CSCN2$ = 1 for exposed, 0 for unexposed
- β_1 = mean $MAXFWT$ for exposed minus mean for unexposed, holding age and sex constant
- Pay attention to Rosner's
  - t and F statistics and what they test
  - Figure 11.28 for checking for trend and equal variability of residuals (don't worry about standardizing residuals)

Chapter 6

The Correlation Coefficient

Rosner 11.7

Pearson product-moment linear correlation coefficient:

$$\begin{aligned} r &= \frac{Lxy}{\sqrt{L_{xx}L_{yy}}} \\ &= \frac{s_{xy}}{s_x s_y} \\ &= b\sqrt{\frac{L_{xx}}{L_{yy}}} \end{aligned}$$

- r is unitless
- r estimates the population correlation coefficient ρ (not to be confused with Spearman ρ rank correlation coefficient)
- $-1 \leq r \leq 1$
- $r = -1$: perfect negative correlation
- $r = 1$: perfect positive correlation

- $r = 0$: no correlation (no association)

- $t - test$ for r is identical to t-test for b

- r^2 is the proportion of variation in y explained by conditioning on x

- $(n-2)\frac{r^2}{1-r^2} = F_{1,n-2} = \frac{MSR}{MSE}$

- For multiple regression in general we use R^2 to denote the fraction of variation in y explained jointly by all the x's (variation in y explained by the whole model)

- $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1$ minus fraction of unexplained variation

- R^2 is called the *coefficient of determination*

- R^2 is between 0 and 1
  - 0 when $\hat{y}_i = \bar{y}$ for all i; $SSE = SST$
  - 1 when $\hat{y}_i = y_i$ for all i; SSE=0

- $R^2 \equiv r^2$ in the one-predictor case

6.1 Using r to Compute Sample Size

- Without knowledge of population variances, etc., r can be useful for planning studies

- Choose n so that margin for error (half-width of C.L.) for r is acceptable

- Precision of r in estimating ρ is generally worst when $\rho = 0$
- This margin for error is shown in the figure below

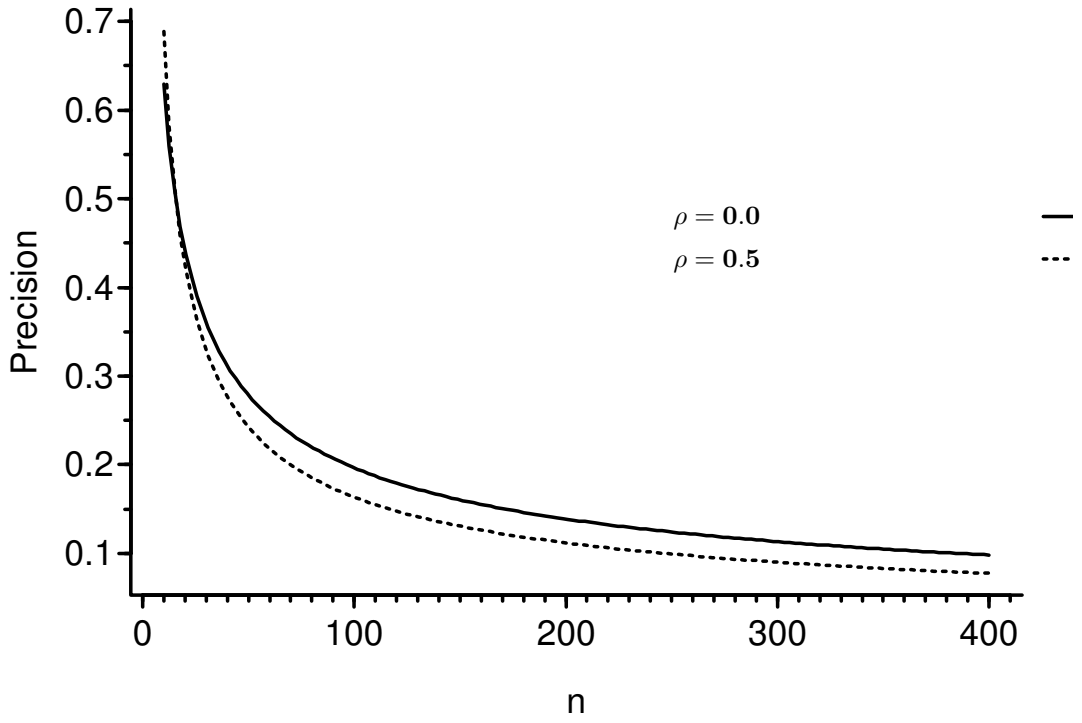

Figure 6.1: *Margin for error (length of longer side of asymmetric 0.95 confidence interval) for r in estimating ρ, when $\rho = 0$ (solid line) and $\rho = 0.5$ (dotted line). Calculations are based on Fisher's z transformation of r.*

6.2 Comparing Two r's

- Rarely appropriate
- Two r's can be the same even though b's may differ
- Usually better to compare effects on a real scale (b)

Chapter 7

Using Regression for ANOVA

Rosner 12.5.2

7.1 Dummy Variables

`Lead Exposure Group`:

control : normal in both 1972 and 1973

currently exposed : elevated serum lead level in 1973, normal in 1972

previously exposed : elevated lead in 1972, normal in 1973

- Requires two dummy variables (and 2 d.f.) to perfectly describe 3 categories
- $x_1 = I[\text{currently exposed}]$
- $x_2 = I[\text{previously exposed}]$
- Reference cell is control

- Model:

$$\begin{aligned} E(y|exposure) &= \alpha + \beta_1 x_1 + \beta_2 x_2 \\ &= \alpha, \text{controls} \\ &= \alpha + \beta_1, \text{currently exposed} \\ &= \alpha + \beta_2, \text{previously exposed} \end{aligned}$$

  α : mean `maxfwt` for controls

  β_1 : mean `maxfwt` for currently exposed minus mean for controls

  β_2 : mean `maxfwt` for previously exposed minus mean for controls

  $\beta_2 - \beta_1$: mean for previously exposed minus mean for currently exposed

- In general requires $k-1$ dummies to describe k categories

- For testing or prediction, choice of reference cell is irrelevant

- Does matter for interpreting individual coefficients

- Modern statistical programs automatically generate dummy variables from categorical or character predictors[a]

- In S never generate dummy variables yourself; just tell the functions you are using the name of the categorical predictor

7.2 Obtaining ANOVA with Multiple Regression

- Estimate α, β_j using standard least squares

- F-test for overall regression is exactly F for ANOVA

[a] In S dummies are generated automatically any time a `factor` or `category` variable is in the model. For SAS you must list such variables in a `CLASS` statement.

- In ANOVA, SSR is call sum of squares between treatments
- SSE is called sum of squares within treatments
- Don't need to learn formulas specifically for ANOVA

7.3 One-Way Analysis of Covariance

Rosner 12.5.3

- Just add other variables (covariates) to the model
- Example: predictors age and treatment
  age is the covariate (adjustment variable)
- Global F test tests the global null hypothesis that neither age nor treatment is associated with response
- To test the adjusted treatment effect, use the partial F test for treatment based on the partial SS for treatment adjusted for age
- If treatment has only two categories, the partial t-test is an easier way to get the age-adjusted treatment test
- In S you can use

```
full ← ols(y ∼ age + treat)
anova(full)      # actually gives you everything needed
reduced ← ols(y ∼ age)
anova(reduced)
# Subtract SSR or SSE from these two models to get treat effect
```

7.4 Two-Way ANOVA

Rosner 12.6

- Two categorical variables as predictors
- Each variable is expanded into dummy variables
- One of the predictor variables may not be time or episode within subject; two-way ANOVA is often misused for analyzing repeated measurements within subject
- Example: 3 diet groups (NOR, SV, LV) and 2 sex groups
- $E(y|diet, sex) = \alpha + \beta_1 I[SV] + \beta_2 I[LV] + \beta_3 I[male]$
- Assumes effects of diet and sex are additive (separable) and not synergistic
- $\beta_1 = SV - NOR$ mean difference holding sex constant
  $\beta_3 = male - female$ effect holding diet constant
- Test of diet effect controlling for sex effect:
  $H_0 : \beta_1 = \beta_2 = 0$
  $H_a : \beta_1 \neq 0$ or $\beta_2 \neq 0$
- This is a 2 d.f. partial F-test, best obtained by taking difference in SS between this full model and a model that excludes all diet terms.
- Test for significant difference in mean y for males vs. females, controlling for diet:
  $H_0 : \beta_3 = 0$
- For a model that has m categorical predictors (only), none of which inter-

act, with numbers of categories given by $k_1, k_2, \ldots, k_m$, the total numerator regression d.f. is $\sum_{i=1}^{m}(k_i - 1)$

7.5 Two-way ANOVA and Interaction

Example: sex (F,M) and treatment (A,B)
Reference cells: F, A Model:

$$\begin{aligned} E(y|sex, treatment) &= \alpha + \beta_1 I[sex = M] \\ &+ \beta_2 I[treatment = B] + \beta_3 I[sex = M \cap treatment = B] \end{aligned}$$

Note that $I[sex = M \cap treatment = B] = I[sex = M] \times I[treatment = B]$.

α : mean y for female on treatment A (all variables at reference values)

β_1 : mean y for males minus mean for females, both on treatment A = sex effect holding treatment constant at A

β_2 : mean for female subjects on treatment B minus mean for females on treatment A = treatment effect holding sex constant at $female$

β_3 : $B - A$ treatment difference for males minus $B - A$ treatment difference for females
Same as $M - F$ difference for treatment B minus $M - F$ difference for treatment A

In this setting think of interaction as a "double difference". To understand the parameters:

Group	$E(y\|Group)$
F A	α
M A	$\alpha + \beta_1$
F B	$\alpha + \beta_2$
M B	$\alpha + \beta_1 + \beta_2 + \beta_3$

Thus $MB - MA - [FB - FA] = \beta_2 + \beta_3 - \beta_2 = \beta_3$.

7.6 Interaction Between Categorical and Continuous Variables

This is how one allows the slope of a predictor to vary by categories of another variable. Example: separate slope for males and females:

$$\begin{aligned} E(y|x) &= \alpha + \beta_1 age + \beta_2 I[sex = m] \\ &+ \beta_3 age \times I[sex = m] \\ E(y|age, sex = f) &= \alpha + \beta_1 age \\ E(y|age, sex = m) &= \alpha + \beta_1 age + \beta_2 + \beta_3 age \\ &= (\alpha + \beta_2) + (\beta_1 + \beta_3) age \end{aligned}$$

α : mean y for zero year-old female

β_1 : slope of age for females

β_2 : mean y for males minus mean y for females, for zero year-olds

β_3 : increment in slope in going from females to males

7.7 Specifying Interactions in S

Asterisk in formula means "include all main effects and interactions involving these variables."

```
y ~ race + age*treatment
```

If race has levels B, W, O in that order and treatment has levels A, B in that order, this specifies the model

$$\begin{aligned} Y &= \alpha + \beta_1 I[race = W] + \beta_2 I[race = O] \\ &+ \beta_3 age \\ &+ \beta_4 I[treatment = B] \\ &+ \beta_5 age \times I[treatment = B] \end{aligned}$$

The last term equals $\beta_5 age$ if $treatment = B$, zero if $treatment = A$.

If you run the `Design` library `anova(fit object)` command you will notice that meaningless "main effects" are not tested by default. In the above model the tests that are provided are

1. `race` main effect (2 d.f.)
2. combined `age` and `age`×`treatment` effect (2 d.f.); tests whether `age` is associated with Y for either treatment ($H_0 : \beta_3 = \beta_5 = 0$)
3. combined `treatment` and `treatment`×`age` effect (2 d.f.); tests whether `treatment` is associated with Y for any `age` ($H_0 : \beta_4 = \beta_5 = 0$)
4. `age`×`treatment` interaction (1 d.f., $H_0 : \beta_5 = 0$)
5. global test (5 d.f.)

Chapter and section numbers from this point on are numbered according to REGRESSION MODELING STRATEGIES.

Chapter 2

General Aspects of Fitting Regression Models

2.1 Notation for Multivariable Regression Models

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

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

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

2.2 Model Formulations

General regression model

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

General linear regression model

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

Examples

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

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

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

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

2.3 Interpreting Model Parameters

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

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

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

2.3.1 Nominal Predictors

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

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

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

where

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

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

2.3.2 Interactions

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

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

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

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

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

2.3.3 Example: Inference for a Simple Model

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

Model assumes

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

Interpretations of parameters:

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

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

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

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

Hypotheses that are usually inappropriate:

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

More useful hypotheses follow. For any hypothesis need to

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

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

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

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

2.4 Review of Composite (Chunk) Tests

- In the model

  `y` $\sim$ `age + sex + weight + waist + tricep`

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

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

2.5 Relaxing Linearity Assumption for Continuous Predictors

2.5.1 Simple Nonlinear Terms

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

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

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

Draftman's *spline* : flexible strip of metal or rubber used to trace curves.

Spline Function : piecewise polynomial

Linear Spline Function : piecewise linear function

- Bilinear regression: model is $\beta_0 + \beta_1 X$ if $X \leq a$, $\beta_2 + \beta_3 X$ if $X > a$.
- Problem with this notation: two lines not constrained to join
- To force simple continuity: $\beta_0 + \beta_1 X + \beta_2 (X - a) \times I[X > a] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, where $X_2 = (X_1 - a) \times I[X_1 > a]$.
- Slope is $\beta_1, X \leq a$, $\beta_1 + \beta_2, X > a$.
- β_2 is the slope increment as you pass a

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

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

where

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

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

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

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

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

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

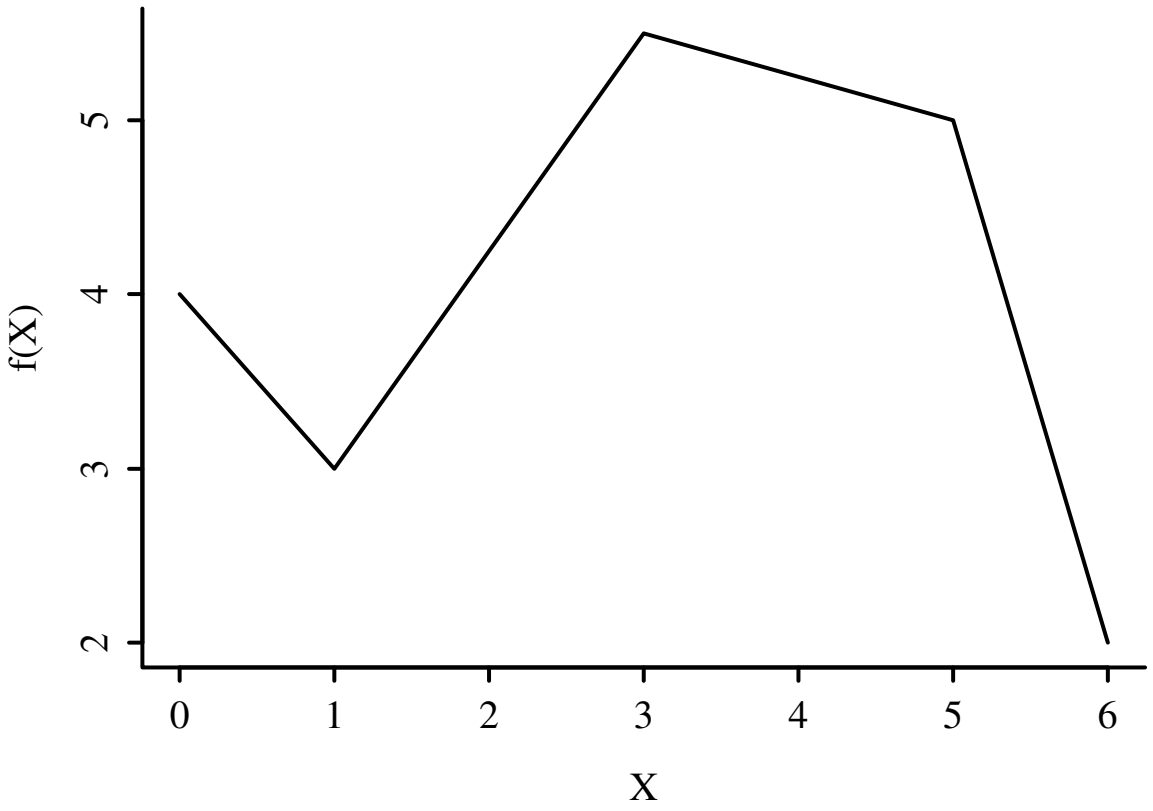

Figure 2.1: *A linear spline function with knots at a=1, b=3, c=5*

2.5.3 Cubic Spline Functions

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

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

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

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

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

2.5.4 Restricted Cubic Splines

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

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

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

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

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

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

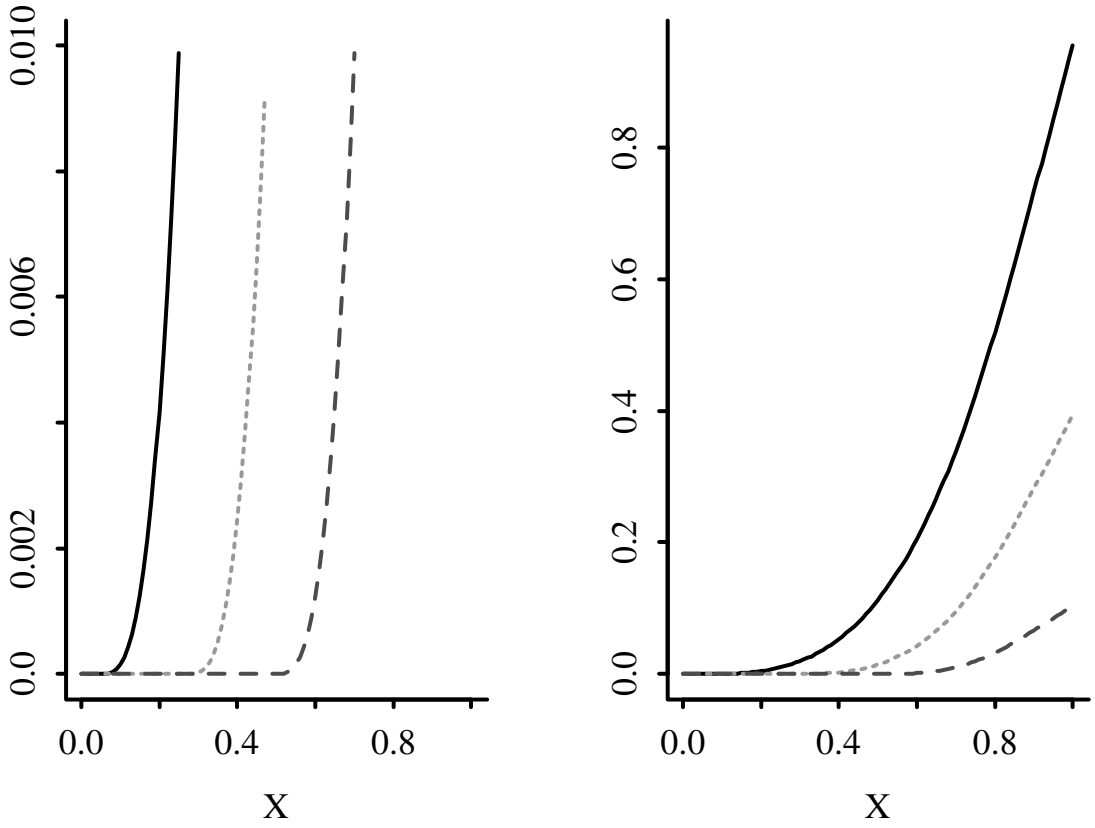

Figure 2.2: *Restricted cubic spline component variables for k=5 and knots at* $X = .05, .275, .5, .725$*, and* $.95$*. Left panel is a magnification of the right. Fitted functions such as those in Figure 2.3 will be linear combinations of these basis functions as long as knots are at the same locations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

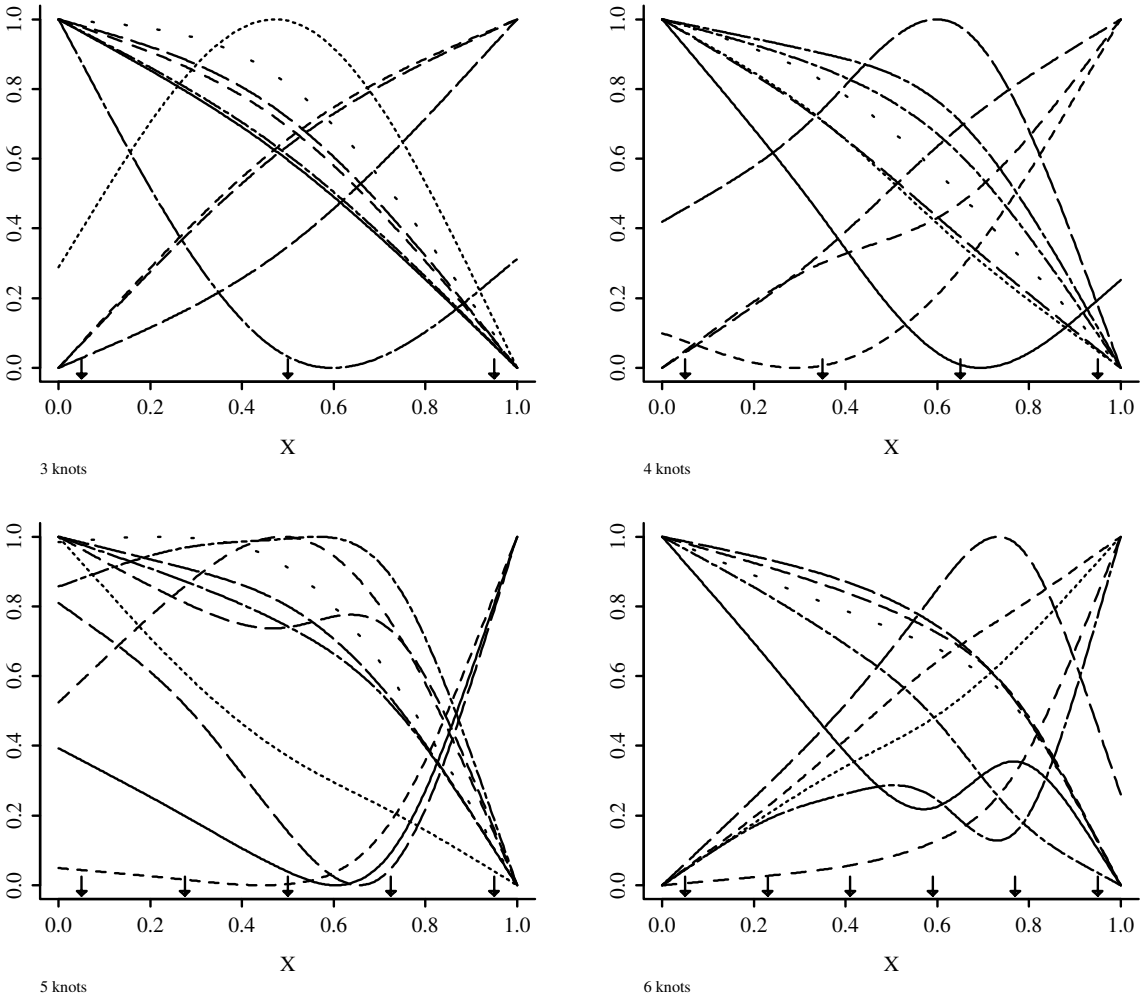

Figure 2.3: *Some typical restricted cubic spline functions for $k = 3, 4, 5, 6$. The y-axis is $X\beta$. Arrows indicate knots. These curves were derived by randomly choosing values of β subject to standard deviations of fitted functions being normalized. See the Web site for a script to create more random spline functions, for $k = 3, \ldots, 7$.*

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

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

by computing

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

A test of linearity in X can be obtained by testing

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

2.5.5 Choosing Number and Position of Knots

- Knots are specified in advance in regression splines
- Locations not important in most situations
- 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 .

Choice of k:

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

2.5.6 Nonparametric Regression

- Estimate tendency (mean or median) of Y as a function of X
- Few assumptions
- Especially handy when there is a single X

- Plotted trend line may be the final result of the analysis

- Simplest smoother: moving average

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

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

  - overlap OK

  - problem in estimating $E(Y)$ at outer X-values

  - estimates very sensitive to bin width

- Moving linear regression far superior to moving avg. (moving flat line)

- Cleveland's moving linear regression smoother *loess* (locally weighted least squares) is the most popular smoother. To estimate central tendency of Y at $X = x$:
  - take all the data having X values within a suitable interval about x (default is $\frac{2}{3}$ of the data)

  - fit weighted least squares linear regression within this neighborhood

  - points near x given the most weight[b]

  - points near extremes of interval receive almost no weight

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

  - loess works much better at extremes of X than moving avg.
  - provides an estimate at each observed X; other estimates obtained by linear interpolation
  - outlier rejection algorithm built-in

- loess works great for binary Y — just turn off outlier detection
- Other popular smoother: Friedman's "super smoother"
- For loess or supsmu amount of smoothing can be controlled by analyst
- Another alternative: smoothing splines[c]
- Smoothers are very useful for estimating trends in residual plots

2.5.7 Advantages of Regression Splines over Other Methods

Regression splines have several advantages :

- 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 $C(Y|X)$.

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

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

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

2.6 Recursive Partitioning: Tree-Based Models

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

Method:

- Find predictor so that best possible binary split has maximum value of some statistic for comparing 2 groups

- Within previously formed subsets, find best predictor and split maximizing criterion in the subset

- Proceed in like fashion until $< k$ obs. remain to split

- Summarize Y for the terminal node (e.g., mean, modal category)

- Prune tree backward until it cross-validates as well as its "apparent" accuracy, or use shrinkage

Advantages/disadvantages of recursive partitioning:

- Does not require functional form for predictors
- Does not assume additivity — can identify complex interactions
- Can deal with missing data flexibly
- Interactions detected are frequently spurious
- Does not use continuous predictors effectively
- Penalty for overfitting in 3 directions
- Often tree doesn't cross-validate optimally unless pruned back very conservatively
- Very useful in messy situations or those in which overfitting is not as problematic (confounder adjustment using propensity scores ; missing value imputation)

2.7 Multiple Degree of Freedom Tests of Association

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

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

In the 5-knot restricted cubic spline model

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

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

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

Grambsch and O'Brien elegantly described the hazards of pretesting

- Studied quadratic regression
- Showed 2 d.f. test of association is nearly optimal even when regression is linear if nonlinearity **entertained**
- Considered ordinary regression model $E(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2$
- Two ways to test association between X and Y
- Fit quadratic model and test for linearity ($H_0 : \beta_2 = 0$)
- F-test for linearity significant at $\alpha = 0.05$ level $\rightarrow$ report as the final test of association the 2 d.f. F test of $H_0 : \beta_1 = \beta_2 = 0$
- If the test of linearity insignificant, refit without the quadratic term and final test of association is 1 d.f. test, $H_0 : \beta_1 = 0 | \beta_2 = 0$
- Showed that type I error $> \alpha$
- Fairly accurate P-value obtained by instead testing against F with 2 d.f. even at second stage

- Cause: are retaining the most significant part of F

- **BUT** if test against 2 d.f. can only lose power when compared with original F for testing both βs

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

2.8 Assessment of Model Fit

2.8.1 Regression Assumptions

The general linear regression model is

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

Verify linearity and additivity. Special case:

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

where X_1 is binary and X_2 is continuous. Methods for checking fit:

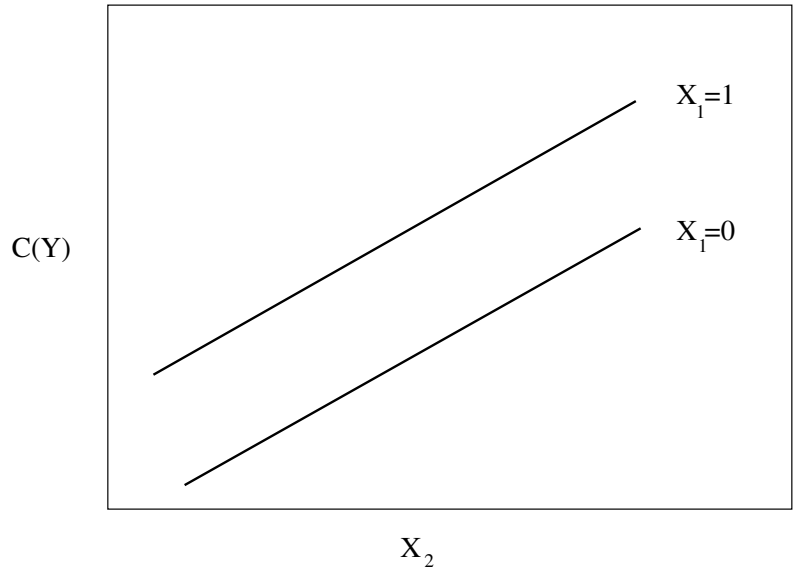

Figure 2.4: *Regression assumptions for one binary and one continuous predictor*

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

   Advantage: Simplicity

   Disadvantages:

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

2. Scatterplot of Y vs. X_2 using different symbols according to values of X_1
   Advantages: Simplicity, can see interaction
   Disadvantages:
   - Scatterplots cannot be drawn for binary, categorical, or censored Y
   - Patterns difficult to see if relationships are weak or n large
3. Stratify the sample by X_1 and quantile groups (e.g. deciles) of X_2; estimate $C(Y|X_1, X_2)$ for each stratum
   Advantages: Simplicity, can see interactions, handles censored Y (if you are careful)
   Disadvantages:
   - Requires large n
   - Does not use continuous var. effectively (no interpolation)
   - Subgroup estimates have low precision
   - Dependent on binning method
4. Separately for levels of X_1 fit a nonparametric smoother relating X_2 to Y
   Advantages: All regression aspects of the model can be summarized efficiently with minimal assumptions
   Disadvantages:

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

5. Fit flexible nonlinear parametric model

   Advantages:

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

   Disadvantages:

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

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

Restricted cubic spline works well for method 5.

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

where

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

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

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

- Tentative transformation $g(X_2) \rightarrow$ check adequacy by expanding $g(X_2)$ in spline function and testing linearity

- Can find transformations by plotting $g(X_2)$ vs. $\hat{f}(X_2)$ for variety of g

- Multiple continuous predictors $\rightarrow$ expand each using spline

- Example: assess linearity of X_2, X_3

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

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

2.8.2 Modeling and Testing Complex Interactions

X_1 binary or linear, X_2 continuous:

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

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

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

- Transformations depend on whether interaction terms adjusted for

- Fit interactions of the form $X_1 f(X_2)$ and $X_2 g(X_1)$:

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

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

General spline surface:

- Cover $X_1 \times X_2$ plane with grid and fit patch-wise cubic polynomial in two variables
- Restrict to be of form $aX_1 + bX_2 + cX_1X_2$ in corners
- Uses all $(k-1)^2$ cross-products of restricted cubic spline terms
- See Gray for penalized splines allowing control of effective degrees of freedom

Other issues:

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

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

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

Chapter 3

Missing Data

3.1 Types of Missing Data

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

3.2 Prelude to Modeling

- Quantify extent of missing data
- Characterize types of subjects with missing data
- Find sets of variables missing on same subjects

3.3 Missing Values for Different Types of Response Variables

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

3.4 Problems With Simple Alternatives to Imputation

Deletion of records—

- Badly biases parameter estimates when missingness is related to Y in a way that is unexplained by non-missing Xs
- Deletion because of a subset of X being missing always results in inefficient estimates
- Deletion of records with missing Y may result in serious biases
- Only discard obs. when
  - Rarely missing predictor of overriding importance that can't be imputed from other data
  - Fraction of obs. with missings small and n is large
- No advantage of deletion except savings of analyst time

- Making up missing data better than throwing away real data

Adding extra categories of categorical predictors—

- Including missing data but adding a category 'missing' causes serious biases
- Problem acute when values missing because subject too sick
- Difficult to interpret

3.5 Strategies for Developing Imputation Algorithms

Exactly how are missing values estimated?

- Could ignore all other information — random or grand mean fill-in
- Can use external info not used in response model (e.g., zip code for income)
- Need to utilize reason for non-response if possible
- Use statistical model with sometimes-missing X as response variable
- Ignoring imputation results in biased $\hat{V}(\hat{\beta})$
- `transcan` function in Hmisc library: "optimal" transformations of all variables to make residuals more stable and to allow non-monotonic transformations
- `aregImpute` function in Hmisc: good approximation to full Bayesian multiple imputation procedure using the bootstrap

- `aregImpute` and `transcan` work with `fit.mult.impute` to make final analysis of response variable relatively easy

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

- Recursive partitioning with surrogate splits — handles case where a predictor of a variable needing imputation is missing itself

3.6 Single Conditional Mean Imputation

- Can fill-in using unconditional mean or median if number of missings low and X is unrelated to other Xs

- Otherwise, first approximation to good imputation uses other Xs to predict a missing X

- This is a single "best guess" conditional mean

- $\hat{X}_j = Z\hat{\theta}, Z = X_{\bar{j}}$
  Cannot include Y in Z without adding random errors to imputed values (would steal info from Y)

- Recursive partitioning is very helpful for nonparametrically estimating conditional means

3.7 Multiple Imputation

- Single imputation using a random draw from the conditional distribution for an individual
  $\hat{X}_j = Z\hat{\theta} + \hat{\epsilon}, Z = [X\bar{j}, Y]$
  $\hat{\epsilon} = n(0, \hat{\sigma})$ or a random draw from the calculated residuals
  - bootstrap
  - approximate Bayesian bootstrap : sample with replacement from sample with replacement of residuals

- Multiple imputations (M) with random draws
  - Draw sample of M residuals for each missing value to be imputed
  - Average M $\hat{\beta}$
  - In general can provide least biased estimates of β
  - Simple formula for imputation-corrected var($\hat{\beta}$)
    Function of average "apparent" variances and between-imputation variances of $\hat{\beta}$
  - **BUT** full multiple imputation needs to account for uncertainty in the imputation models by refitting these models for each of the M draws

  - `transcan` does not do that; `aregImpute` does

- `aregImpute` algorithm
  - Takes all aspects of uncertainty into account using the bootstrap
  - Different bootstrap resamples used for each imputation by fitting a flexible additive model on a sample with replacement from the original data
  - This model is used to predict all of the original missing and non-missing values for the target variable for the current imputation
  - Uses `ace` or `avas` semiparametric regression models to impute
  - For continuous variables, monotonic transformations of the target variable are assumed when `avas` used
  - For `ace`, the default allows nonmonotonic transformations of target variables
  - Uses predictive mean matching for imputation; no residuals required
  - By default uses weighted PMM; option for just using closest match
  - When a predictor of the target variable is missing, it is first imputed from its last imputation when it was a target variable
  - First 3 iterations of process are ignored (“burn-in”)
  - Compares favorably to S `MICE` approach
  - Example:

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

3.8 Summary and Rough Guidelines

Table 3.1: *Summary of Methods for Dealing with Missing Values*

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

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

Proportion of missings ≤ 0.05 : Method of imputing and computing variances doesn't matter much

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

Proportion of missings > 0.15 : Multiple imputation, adjust variances for imputation

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

Reason for missings more important than number of missing values.

Chapter 4

Multivariable Modeling Strategies

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

4.1 Prespecification of Predictor Complexity Without Later Simplification

- Rarely expect linearity
- Can't always use graphs or other devices to choose transformation
- If select from among many transformations, results biased
- Need to allow flexible nonlinearity to potentially strong predictors not *known* to predict linearly
- Once decide a predictor is "in" can choose no. of parameters to devote to it using a general association index with Y
- Need a measure of "potential predictive punch" (ignoring collinearity and interaction for now)
- Measure needs to mask analyst to true form of regression to preserve statistical properties
- 2 d.f. generalization of Spearman ρ—R^2 based on $rank(X)$ and $rank(X)^2$ vs. $rank(Y)$
- ρ^2 can detect U-shaped relationships
- For categorical X, ρ^2 is R^2 from dummy variables regressed against $rank(Y)$; this is tightly related to the Wilcoxon–Mann–Whitney–Kruskal–Wallis rank test for group differences[a]
- Sort variables by descending order of ρ^2

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

- Specify number of knots for continuous X, combine infrequent categories of categorical X based on ρ^2

- Allocating d.f. based on sorting ρ^2 fair procedure because
  - already decided to keep variable in model no matter what ρ^2
  - ρ^2 does not reveal degree of nonlinearity; high value may be due solely to strong linear effect
  - low ρ^2 for a categorical variable might lead to collapsing the most disparate categories

- Initial simulations show the procedure to be conservative

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

4.2 Checking Assumptions of Multiple Predictors Simultaneously

- Sometimes failure to adjust for other variables gives wrong transformation of an X, or wrong significance of interactions

- Sometimes unwieldy to deal simultaneously with all predictors at each stage $\rightarrow$ assess regression assumptions separately for each predictor

4.3 Variable Selection

- Series of potential predictors with no prior knowledge

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

- Summary of problem: $E(\hat{\beta}|\hat{\beta}$ "significant" $) \neq \beta$

- F and χ^2 statistics do not have the claimed distribution

- Derksen and Keselman found that in stepwise analyses the final model represented noise 0.20-0.74 of time, final model usually contained $< \frac{1}{2}$ actual number of authentic predictors. Also:

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

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

Variable selection methods :

- 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), here $\chi^2 - 2 \times d.f.$

- Better to use subject matter knowledge!

- No currently available stopping rule was developed for stepwise, only for comparing **2** pre-specified models

- Roecker studied forward selection (FS), all possible subsets selection (APS), full fits

- APS more likely to select smaller, less accurate models than FS

- Neither as accurate as full model fit unless $> \frac{1}{2}$ candidate variables redundant or unnecessary

- Step-down is usually better than forward and can be used efficiently with maximum likelihood estimation

- Bootstrap can help decide between full and reduced model

- Full model fits gives meaningful confidence intervals with standard formulas, C.I. after stepwise does not

- Data reduction (grouping variables) can help

- Using the bootstrap to select important variables for inclusion in the final model is problematic

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

4.4 Overfitting and Limits on Number of Predictors

- Concerned with avoiding overfitting

- p should be $< \frac{m}{15}$

- $p =$ number of parameters in full model or number of *candidate* parameters in a stepwise analysis

Table 4.1: *Limiting Sample Sizes for Various Response Variables*

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

- Narrowly distributed predictor $\rightarrow$ even higher n

- p includes *all* variables screened for association with response, including interactions

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

4.5 Shrinkage

- Slope of calibration plot; regression to the mean

- Statistical estimation procedure — "pre-shrunk" models

[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 . Here n_1 and n_2 are the marginal frequencies of the two response levels .

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

- 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. $X\hat{\beta}$

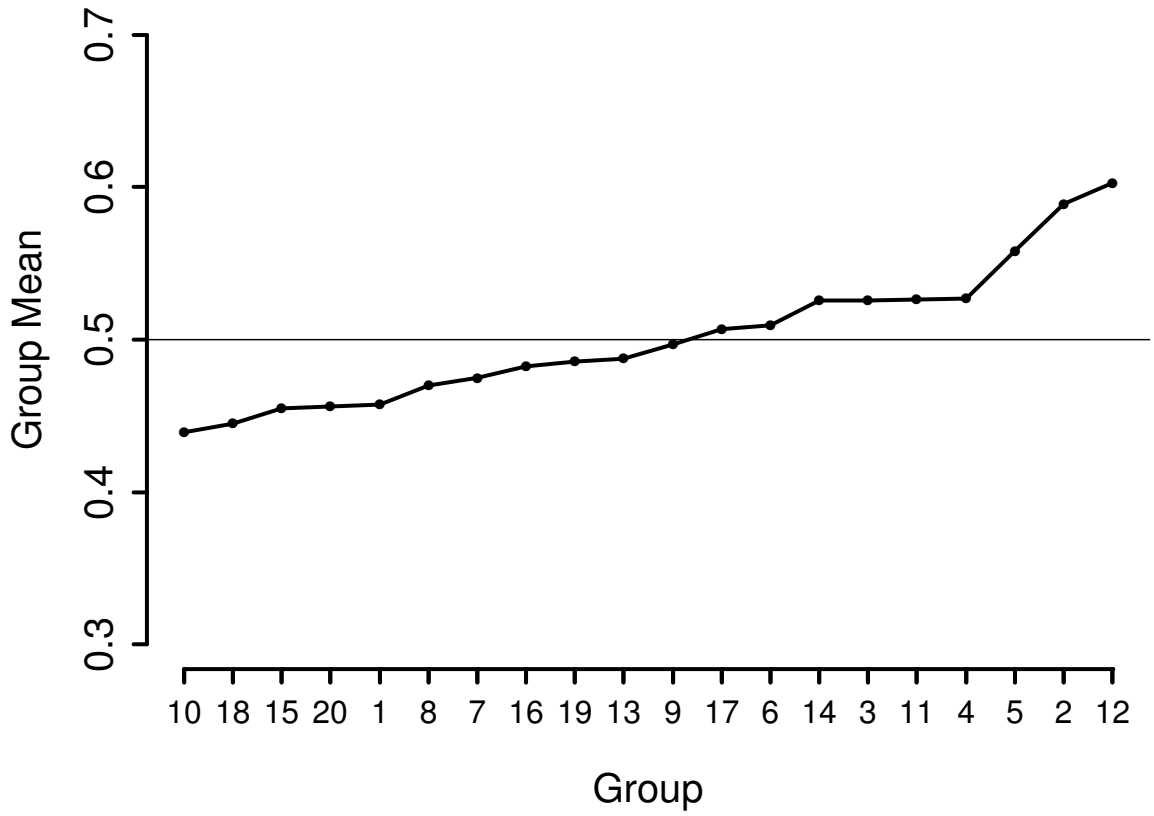

Figure 4.1: *Sorted means from 20 samples of size 50 from a uniform $[0, 1]$ distribution. The reference line at 0.5 depicts the true population value of all of the means.*

- Prevent shrinkage by using pre-shrinkage
- Spiegelhalter : 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
- Ridge regression
- Penalized MLE
- Heuristic shrinkage parameter of van Houwelingen and le Cessie

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

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

4.6 Collinearity

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

- This is appropriate $\rightarrow$ asking too much of the data
- 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 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
- See Belsley for problems with VIF
- Easy approach: SAS `VARCLUS` procedure , S-PLUS `varclus` function, other clustering techniques: group highly correlated variables
- Can score each group (e.g., first principal component, PC_1); summary scores not collinear

4.7 Data Reduction

- Unless $n >> p$, model unlikely to validate
- Data reduction: $\downarrow p$
- Use the literature to eliminate unimportant variables.
- Eliminate variables whose distributions are too narrow.
- Eliminate candidate predictors that are missing in a large number of subjects, especially if those same predictors are likely to be missing for future applications of the model.
- Use a statistical data reduction method such as incomplete principal components regression, nonlinear generalizations of principal components such as principal surfaces, sliced inverse regression, variable clustering, or ordinary cluster analysis on a measure of similarity between variables.

4.7.1 Variable Clustering

- Goal: Separate variables into groups
  - variables within group correlated with each other
  - variables not correlated with non-group members
- Score each dimension, stop trying to separate effects of factors measuring same phenomenon
- Variable clustering (oblique-rotation PC analysis) $\rightarrow$ separate variables so that first PC is representative of group

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

4.7.2 Transformation and Scaling Variables Without Using Y

- Reduce p by estimating transformations using associations with other predictors

- Purely categorical predictors – correspondence analysis

- Mixture of qualitative and continuous variables: qualitative principal components

- Maximum generalized variance (MGV) method of Sarle
  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

4.7.3 Simultaneous Transformation and Imputation

S-PLUS `transcan` Function for Data Reduction & Imputation

- Initialize missings to medians (or most frequent category)

- Initialize transformations to original variables

- Take each variable in turn as Y
- Exclude obs. missing on Y
- Expand Y (spline or dummy variables)
- Score (transform Y) using first canonical variate
- Missing $Y \rightarrow$ predict canonical variate from Xs
- The imputed values can optionally be shrunk to avoid overfitting for small n or large p
- Constrain imputed values to be in range of non-imputed ones
- Imputations on original scale
  1. Continuous $\rightarrow$ back-solve with linear interpolation
  2. Categorical $\rightarrow$ classification tree (most freq. cat.) or match to category whose canonical score is closest to one predicted
- Multiple imputation — bootstrap or approx. Bayesian boot.
  1. Sample residuals multiple times (default $M = 5$)
  2. Are on "optimally" transformed scale
  3. Back-transform
  4. `fit.mult.impute` works with `aregImpute` and `transcan` output to easily get imputation-corrected variances and avg. $\hat{\beta}$
- Example: $n = 415$ acutely ill patients
  1. Relate heart rate to mean arterial blood pressure
  2. Two blood pressures missing

3. Heart rate not monotonically related to blood pressure
4. See Figure 4.2

- These methods find *marginal* transformations

4.7.4 Simple Scoring of Variable Clusters

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

Series of dichotomous variables:

- Construct X_1 = 0-1 according to whether any variables positive
- Construct X_2 = number of positives
- Test whether original variables add to X_1 or X_2

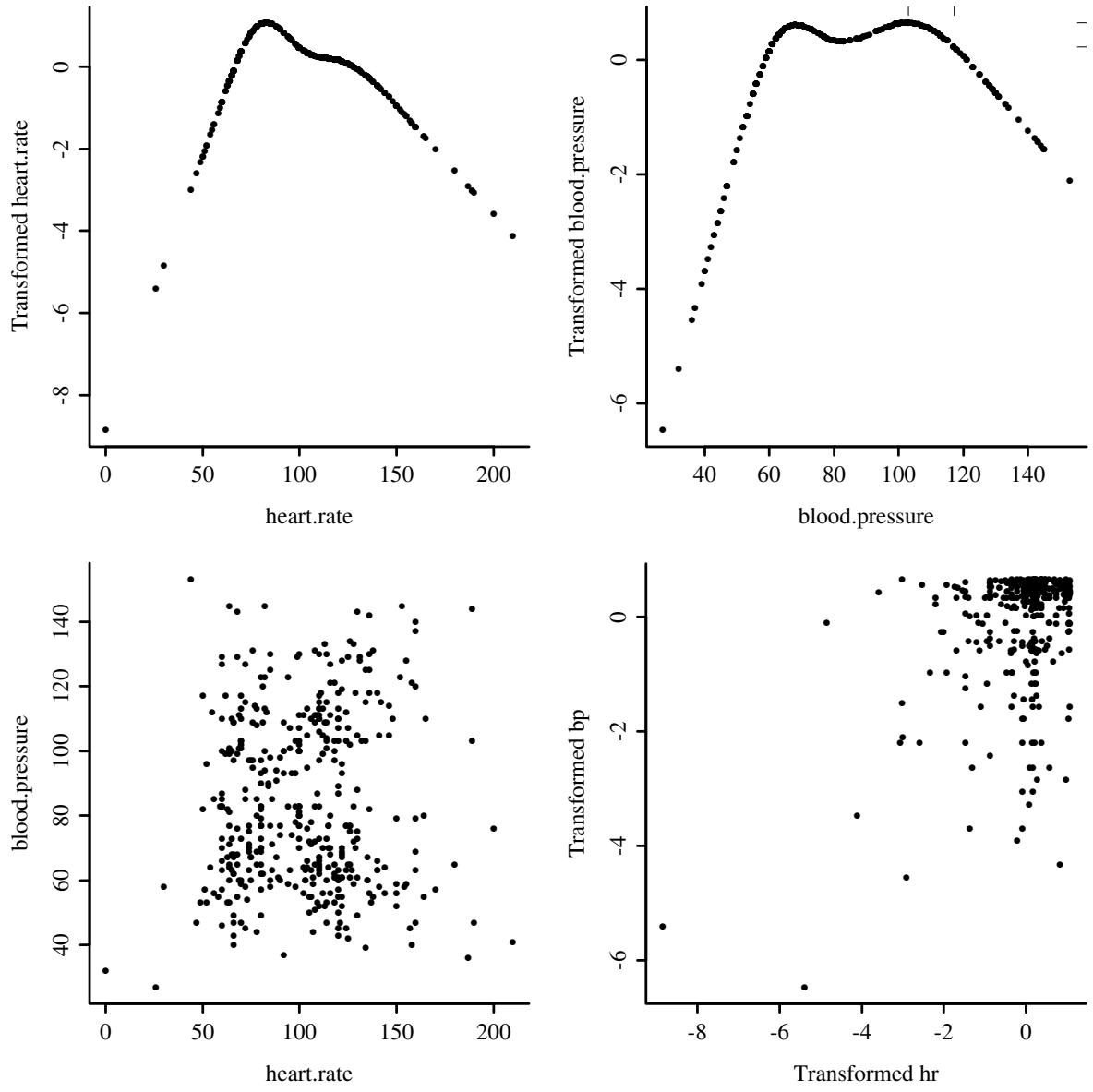

Figure 4.2: *Transformations fitted using* `transcan`*. Tick marks indicate the two imputed values for blood pressure. The lower left plot contains raw data ($D_{xy} = 0.02$); the lower right is a scatterplot of the corresponding transformed values ($D_{xy} = 0.14$). Data courtesy of the SUPPORT study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

4.7.5 Simplifying Cluster Scores

4.7.6 How Much Data Reduction Is Necessary?

Summary of Some Data Reduction Methods

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

4.8 Overly Influential Observations

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

Reasons for influence:

- Too few observations for complexity of model (see Sections 4.7, 4.3)
- Data transcription or entry errors
- Extreme values of a predictor
  1. Sometimes subject so atypical should remove from dataset
  2. Sometimes truncate measurements where data density ends
  3. Example: $n = 4000$, 2000 deaths, white blood count range 500-100,000, .05,.95 quantiles=2755, 26700
  4. Linear spline function fit
  5. Sensitive to WBC> 60000 $(n = 16)$
  6. Predictions stable if truncate WBC to 40000 ($n = 46$ above 40000)
- Disagreements between predictors and response. Ignore unless extreme values or another explanation
- Example: $n = 8000$, one extreme predictor value not on straight line relationship with other $(X, Y) \rightarrow \chi^2 = 36$ for H_0 : linearity

4.9 Comparing Two Models

4.10 Summary: Possible Modeling Strategies

Strategy in a nutshell:

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

4.10.1 Developing Predictive Models

1. Assemble accurate, pertinent data and lots of it, with wide distributions for X.
2. Formulate good hypotheses — specify relevant candidate predictors and possible interactions. Don't use Y to decide which X's to include.
3. Characterize subjects with missing Y. Delete such subjects in rare circumstances . For certain models it is effective to multiply impute Y.
4. Characterize and impute missing X. In most cases use multiple imputation based on X and Y
5. For each predictor specify complexity or degree of nonlinearity that should be allowed (more for important predictors or for large n) (Section 4.1)
6. Do data reduction if needed (pre-transformations, combinations), or use penalized estimation
7. Use the entire sample in model development
8. Can do highly structured testing to simplify "initial" model

   (a) Test entire group of predictors with a single P-value
   (b) Make each continuous predictor have same number of knots, and select the number that optimizes AIC
9. Check linearity assumptions and make transformations in Xs as needed but beware.
10. Check additivity assumptions by testing pre-specified interaction terms. Use a global test and either keep all or delete all interactions.
11. Check to see if there are overly-influential observations.
12. Check distributional assumptions and choose a different model if needed.
13. Do limited backwards step-down variable selection if parsimony is more important that accuracy . 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 to penalize for the randomness inherent in the multiple steps.
20. Develop simplifications to the final model as needed.

4.10.2 Developing Models for Effect Estimation

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

4.10.3 Developing Models for Hypothesis Testing

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

Chapter 5

Resampling, Validating, Describing, and Simplifying the Model

5.1 The Bootstrap

- If know population model, use simulation or analytic derivations to study behavior of statistical estimator

- Suppose Y has a cumulative dist. fctn. $F(y) = \text{Prob}\{Y \leq y\}$

- We have sample of size n from $F(y)$,
  $Y_1, Y_2, \ldots, Y_n$

- Steps:
  1. Repeatedly simulate sample of size n from F
  2. Compute statistic of interest
  3. Study behavior over B repetitions

- Example: 1000 samples, 1000 sample medians, compute their sample variance

- F unknown $\rightarrow$ estimate by empirical dist. fctn.

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

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

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

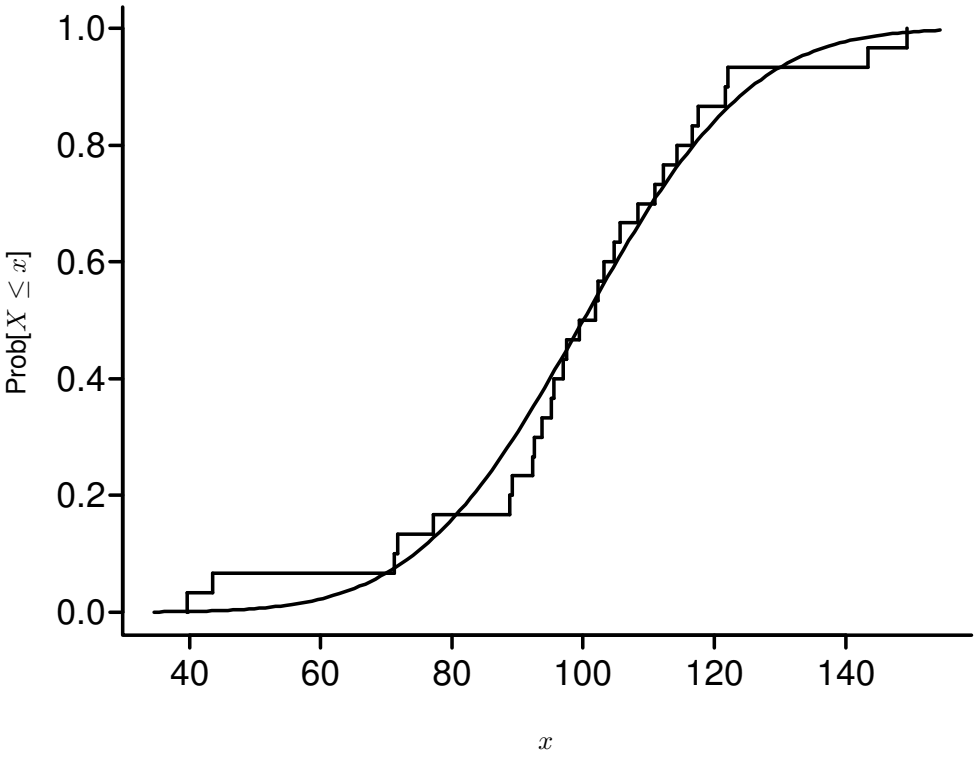

Figure 5.1: *Empirical and population cumulative distribution functions*

- 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

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

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

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

First 20 samples:

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

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

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

5.2 Model Validation

5.2.1 Introduction

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

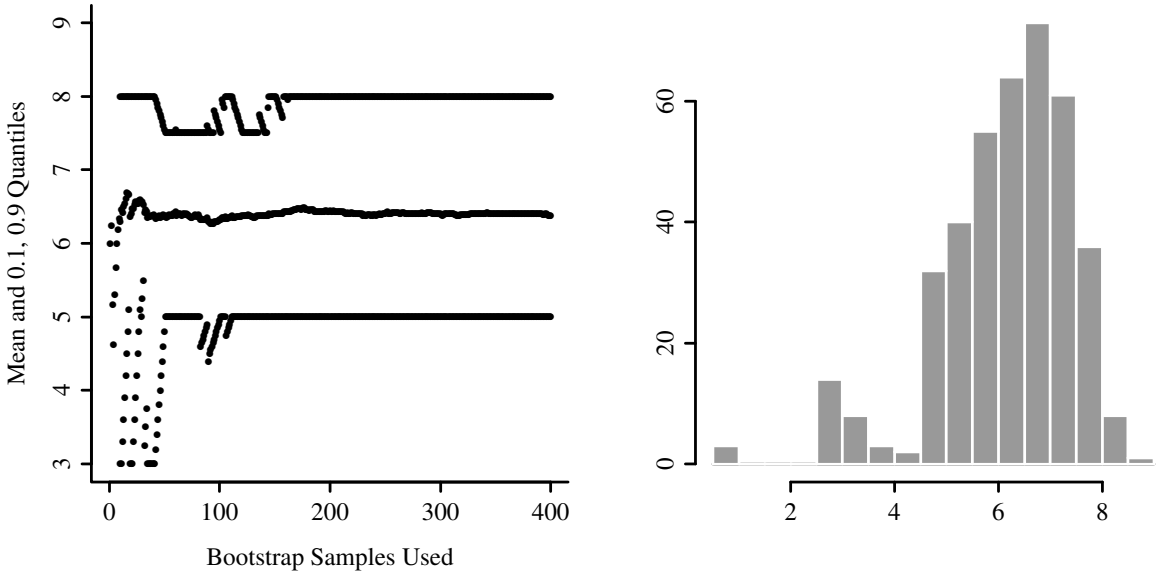

Figure 5.2: *Estimating properties of sample median using the bootstrap*

- Internal
  - apparent (evaluate fit on same data used to create fit)
  - data splitting
  - cross-validation
  - bootstrap: get overfitting-corrected accuracy index
- Best way to make model fit data well is to discard much of the data
- Predictions on another dataset will be inaccurate
- Need unbiased assessment of predictive accuracy

5.2.2 Which Quantities Should Be Used in Validation?

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

- Example: $n = 10, p = 9$, apparent $R^2 = 1$ but R^2 will be close to zero on new subjects

- Example: $n = 20, p = 10$, apparent $R^2 = .9$, R^2 on new data 0.7, $R^2_{adj} = 0.79$

- Adjusted R^2 solves much of the bias problem assuming p in its formula is the largest number of parameters ever examined against Y

- Few other adjusted indexes exist

- Also need to validate models with phantom d.f.

- Cross-validation or bootstrap can provide unbiased estimate of any index; bootstrap has higher precision

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

- Unbiased validation nearly always necessary, to detect overfitting

5.2.3 Data-Splitting

- Split data into *training* and *test* sets

- Interesting to compare index of accuracy in training and test

- Freeze parameters from training

- Make sure you allow $R^2 = 1 - SSE/SST$ for test sample to be < 0
- Don't compute ordinary R^2 on $X\hat{\beta}$ vs. Y; this allows for linear recalibration $aX\hat{\beta} + b$ vs. Y
- Test sample must be large enough to obtain very accurate assessment of accuracy
- Training sample is what's left
- Example: overall sample $n = 300$, training sample $n = 200$, develop model, freeze $\hat{\beta}$, predict on test sample ($n = 100$), $R^2 = 1 - \frac{\sum(Y_i - X_i\hat{\beta})^2}{\sum(Y_i - Y)^2}$.
- Disadvantages of data splitting:
  1. Costly in $\downarrow n$
  2. Requires *decision* to split at beginning of analysis
  3. Requires larger sample held out than cross-validation
  4. Results vary if split again
  5. Does not validate the final model (from recombined data)
  6. Not helpful in getting CL corrected for var. selection

5.2.4 Improvements on Data-Splitting: Resampling

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

- Example of $\times$-validation:
  1. Split data at random into 10 tenths
  2. Leave out $\frac{1}{10}$ of data at a time
  3. Develop model on $\frac{9}{10}$, including any variable selection, pre-testing, etc.
  4. Freeze coefficients, evaluate on $\frac{1}{10}$
  5. Average R^2 over 10 reps

- Drawbacks:
  1. Choice of number of groups and repetitions
  2. Doesn't show full variability of var. selection
  3. Does not validate full model
  4. Lower precision than bootstrap

5.2.5 Validation Using the Bootstrap

- Estimate optimism of *final whole sample fit* without holding out data

- From original X and Y select sample of size n with replacement

- Derive model from bootstrap sample

- Apply to original sample

- Simple bootstrap uses average of indexes computed on original sample

- Estimated optimism = difference in indexes

- Repeat about $B = 100$ times, get average expected optimism

- Subtract average optimism from apparent index in final model

- Example: $n = 1000$, have developed a final model that is hopefully ready to publish. Call estimates from this final model $\hat{\beta}$.
  - final model has apparent R^2 (R^2_{app}) =0.4
  - how inflated is R^2_{app}?
  - get resamples of size 1000 with replacement from original 1000
  - for each resample compute R^2_{boot} = apparent R^2 in bootstrap sample
  - freeze these coefficients (call them $\hat{\beta}_{boot}$), apply to original (whole) sample (X_{orig}, Y_{orig}) to get $R^2_{orig} = R^2(X_{orig}\hat{\beta}_{boot}, Y_{orig})$
  - optimism = $R^2_{boot} - R^2_{orig}$
  - average over $B = 100$ optimisms to get $\overline{optimism}$
  - $R^2_{overfitting\ corrected} = R^2_{app} - \overline{optimism}$

Use bootstrap to choose between full and reduced models:

- Bootstrap estimate of accuracy for full model
- Repeat, using chosen stopping rule for each re-sample
- Full fit usually outperforms reduced model
- Stepwise modeling often reduces optimism but this is not offset by loss of information from deleting marginal var.

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

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

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

5.3 Describing the Fitted Model

- Regression coefficients if 1 d.f. per factor, no interaction
- **Not** standardized regression coefficients
- Many programs print meaningless estimates such as effect of increasing age^2 by one unit, holding age constant
- Need to account for nonlinearity, interaction, and use meaningful ranges
- For monotonic relationships, estimate $X\hat{\beta}$ at quartiles of continuous variables, separately for various levels of interacting factors
- Subtract estimates, anti-log, e.g., to get inter-quartile-range odds or hazards ratios. Base C.L. on s.e. of difference.

- Plot effect of each predictor on $X\beta$ or some transformation of $X\beta$
- Nomogram
- Use regression tree to approximate the full model

5.4 Simplifying the Final Model by Approximating It

5.4.1 Difficulties Using Full Models

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

5.4.2 Approximating the Full Model

- Full model is gold standard
- Approximate it to any desired degree of accuracy
- If approx. with a tree, best c-v tree will have 1 obs./node
- Can use least squares to approx. model by predicting $\hat{Y} = X\hat{\beta}$
- When original model also fit using least squares, coef. of approx. model against $\hat{Y} \equiv$ coef. of subset of variables fitted against Y (as in stepwise)

- Model approximation still has some advantages
  1. Uses unbiased estimate of σ from full fit
  2. Stopping rule less arbitrary
  3. Inheritance of shrinkage

Chapter 6

S Software

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

S is comprehensive:

- Easy to write S functions for new models → wide variety of modern regression models implemented (trees, nonparametric, ACE, AVAS, survival models for multiple events)
- Designs can be generated for any model → all handle "class" var, interactions, nonlinear expansions
- Single S objects (e.g., fit object) can be self-documenting → automatic hypothesis tests, predictions for new data
- Superior graphics
- Classes and generic functions

6.1 The S Modeling Language

S statistical modeling language:

```
response ~ terms

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

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

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

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

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

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

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

6.2 User-Contributed Functions

- S is high-level object-oriented language.
- S-PLUS 3.4, 4.5, 2000, 6.0 (UNIX, Linux, Microsoft Windows)
- R (UNIX, Linux, Mac, Windows)
- Multitude of user-contributed functions on `StatLib`
- International community of users through `S-news`

Some S functions:

- See Venables and Ripley
- Hierarchical clustering: `hclust`
- Principal components: `princomp, prcomp`
- Canonical correlation: `cancor`
- ACE: `ace`
- `areg.boot` (Harrell)
- Rank correlation methods:
  `rcorr`, `hoeffd`, `spearman2` (Harrell)
- Variable clustering: `varclus` (Harrell)

- `transcan`, `aregImpute` (Harrell)
- Correspondence analysis: see Web page
- Restricted cubic spline design matrix: `rcspline.eval` (Harrell)
- Re-state restricted spline in simpler form: `rcspline.restate`

6.3 The `Design` Library

- `datadist` function to compute predictor distribution summaries

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

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

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

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

Table 6.1: `Design` *Fitting Functions*

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

Table 6.2: `Design` *Transformation Functions*

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

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

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

```
library(Design, T)                  # make new functions available
ddist ← datadist(cholesterol, treat, num.diseases, age)
# Could have used ddist ← datadist(data.frame.name)
options(datadist="ddist")           # defines data dist. to Design
cholesterol ← impute(cholesterol)
fit ← lrm(y ∼ treat + scored(num.diseases) + rcs(age) +
                 log(cholesterol+10) + treat:log(cholesterol+10))
describe(y ∼ treat + scored(num.diseases) + rcs(age))
# or use describe(formula(fit)) for all variables used in fit
# describe function (in Hmisc) gets simple statistics on variables
# fit ← robcov(fit)                 # Would make all statistics that follow
                                    # use a robust covariance matrix
                                    # would need x=T, y=T in lrm()
specs(fit)                          # Describe the design characteristics
anova(fit)
anova(fit, treat, cholesterol)      # Test these 2 by themselves
plot(anova(fit))                    # Summarize anova graphically
summary(fit)                        # Estimate effects using default ranges
plot(summary(fit))                  # Graphical display of effects with C.I.
summary(fit, treat="b", age=60)     # Specify reference cell and adjustment val
```

```
summary(fit, age=c(50,70))          # Estimate effect of increasing age from
                                    # 50 to 70
summary(fit, age=c(50,60,70))       # Increase age from 50 to 70, adjust to
                                    # 60 when estimating effects of other
                                    # factors
# If had not defined datadist, would have to define ranges for all var.

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

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

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

ag ← 10:80
logit ← predict(fit, expand.grid(treat="a", num.dis=0, age=ag,
                 cholesterol=median(cholesterol)), type="terms")[,"age"]
# Note: if age interacted with anything, this would be the age
#       "main effect" ignoring interaction terms
# Could also use
#    logit ← plot(f, age=ag, ...)$x.xbeta[,2]
# which allows evaluation of the shape for any level of interacting
#  factors.  When age does not interact with anything, the result from
# predict(f, ..., type="terms") would equal the result from
# plot if all other terms were ignored

# Could also specify
```

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

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

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

# Compose S function to evaluate linear predictors analytically
g <- Function(fit)
g(treat='b', cholesterol=260, age=50)
# Letting num.diseases default to reference value
```

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

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

6.4 Other Functions

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

- `validate.tree` in `Design`

- `loess`: multi-dimensional scatterplot smoother

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

Chapter 9

Overview of Maximum Likelihood Estimation

- In ordinary least squares regression, main objective function (criterion for deriving $\hat{\beta}$) is SSE
- If residuals are normally distributed, the resulting least squares estimates are optimal (consistently estimate β as $n \rightarrow \infty$ and have lowest variances among unbiased estimates)
- Other fitting criteria such as minimizing sum of absolute errors are needed for non-normal residuals (or residuals not assumed to be symmetrically distributed)
- With binary Y a drastic change is needed
- Need a general way to write down a good fitting criterion for many different types of Y and for any distribution of $Y|X$
- *Maximum likelihood* (ML) is a general solution

- $\hat{\beta}$ is the vector of values of β making the data *most likely to have been observed* given $\beta = \hat{\beta}$

- Example: 1-sample binomial problem

- Single unknown P = probability of an event in a population unknown parameter, the probability of an event in a population.

- Occurrence of the event signaled by $Y = 1$, non-occurrence by $Y = 0$, for an individual

- $\text{Prob}\{Y = 1\} = P$

- Draw a random sample of size $n = 3$ from the population

- Observed occurrences of events $Y = 1, 0, 1$

- Assuming individuals in the sample act completely independently, probability of intersection of the 3 events is $P(1 - P)P = P^2(1 - P)$; this joint probability is called the *likelihood*

- P is unknown but the ML estimate (MLE) is ready to be computed by solving for P that makes the likelihood of the observed data the maximum

- In other words, the MLE of P is that value which makes the population parameter most consistent with the observed data (or the data most likely to have arisen from that population)

- Optimum value of P is the value giving the maximum likelihood, which is also the value where the slope of the likelihood vs. P is zero

- Slope of the likelihood is $P^2 - P^3$; first derivative is $2P - 3P^2 = P(2 - 3P)$

- Set to zero; $2 - 3P = 0 \rightarrow \hat{P} = 2/3$

- In general if Y is binary so that the sample is $Y_1, \ldots, Y_n$ and s is $\sum Y_i$, the likelihood is

$$\begin{aligned} L &= \prod_{i=1}^{n} P^{Y_i}(1-P)^{1-Y_i} \\ &= P^s(1-P)^{n-s} \end{aligned}$$

- For numerical and statistical reasons we work with the *log-likelihood function*

$$\log L = s\log(P) + (n-s)\log(1-P)$$

- Slope of this function is $\frac{s}{P} - \frac{n-s}{1-P}$

- Equating this function to zero requires that $s/P = (n-s)/(1-P)$. Multiplying both sides of the equation by $P(1-P)$ yields $s(1-P) = (n-s)P$ or that $s = (n-s)P + sP = nP$ $\hat{P} = p = s/n$

- Later in logistic regression we allow for differences in subject characteristics through Xs instead of just addressing the one-sample problem as above

- Example: $Y = 1\ 0\ 1$, $X = 18\ 16\ 28$ ($n = 3, p = 1$)

- If the model is $\mathrm{Prob}[Y = 1|X] = 1/[1 + \exp(-\beta_0 - \beta_1 X)]$, the likelihood is $\frac{1}{1+e^{-(\beta_0+\beta_1 18)}} \times [1 - \frac{1}{1+e^{-(\beta_0+\beta_1 16)}}] \times \frac{1}{1+e^{-(\beta_0+\beta_1 28)}}$

- Solve for β_0, β_1

- P is a function of X

- Normal one-sample problem: $\hat{\mu} = \bar{Y}$, MLE of σ^2 is s^2 but with n in the denominator instead of $n - 1$

- Normal regression problem: $\hat{\beta}$ = least squares estimates

9.1 Test Statistics

- With normal distribution, test statistics are t and F

- These make use of $\hat{\sigma}^2$

- Other models do not have σ^2 parameter and do not use a normal distribution

- For one-parameter test with ML we use a z test (estimate divided by standard error, is approximately normal)

- z^2 is (estimate)^2/(estimated variance) which has a χ^2 distribution with 1 d.f.

- For p parameters, a joint test statistic is χ^2_p

- $F_{p,n-p-1} \approx \frac{\chi^2_p}{p}$

- z and χ^2 statistics derived from estimates and standard errors are called *Wald statistics*

- Statistics that have even better agreement with the χ^2 distribution are *likelihood ratio* χ^2

- These are computed by subtracting the best log likelihood from the log likelihood evaluated at the null hypothesis and multiplying by -2

- LR χ^2 statistics do not assume that the log likelihood is quadratic like the normal distribution's

Chapter 10

Binary Logistic Regression

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

10.1 Model

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

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

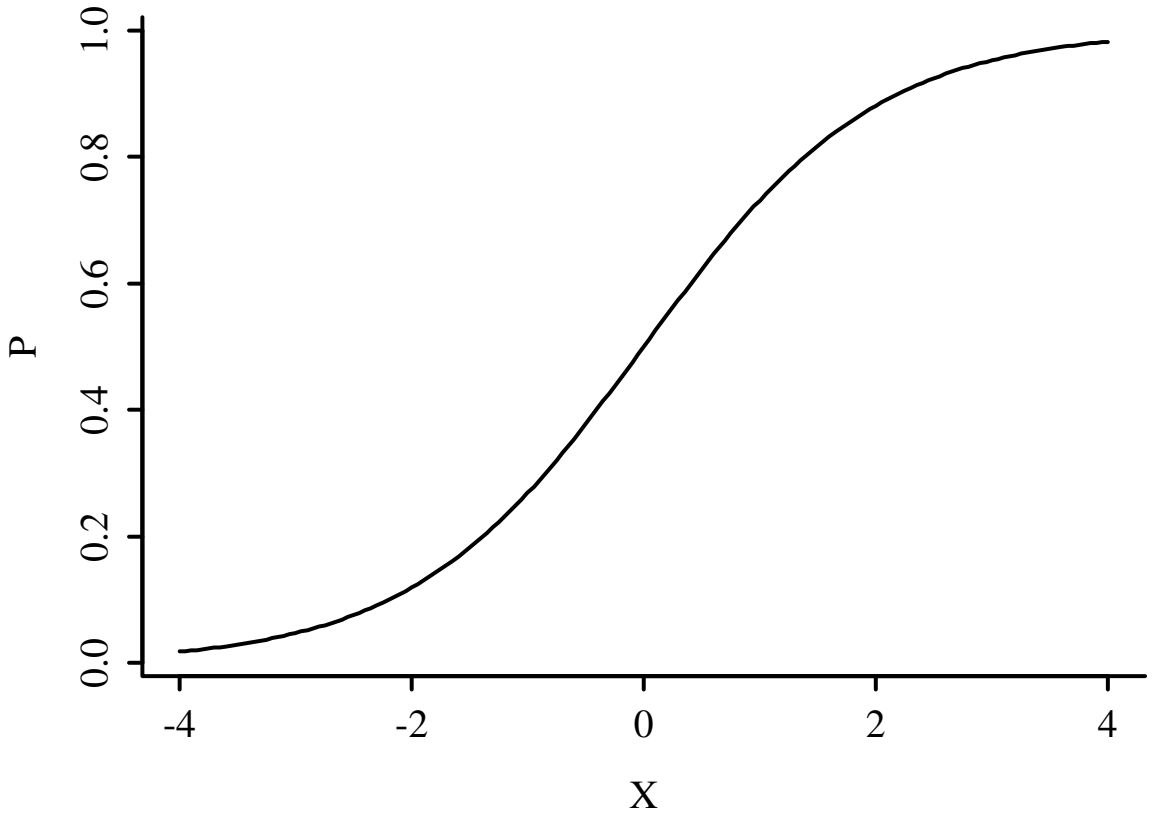

Figure 10.1: *Logistic function*

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

10.1.1 Model Assumptions and Interpretation of Parameters

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

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

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

- One continuous predictor:

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

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

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

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

10.1.2 Odds Ratio, Risk Ratio, and Risk Difference

- Odds ratio capable of being constant

- Ex: risk factor doubles odds of disease

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

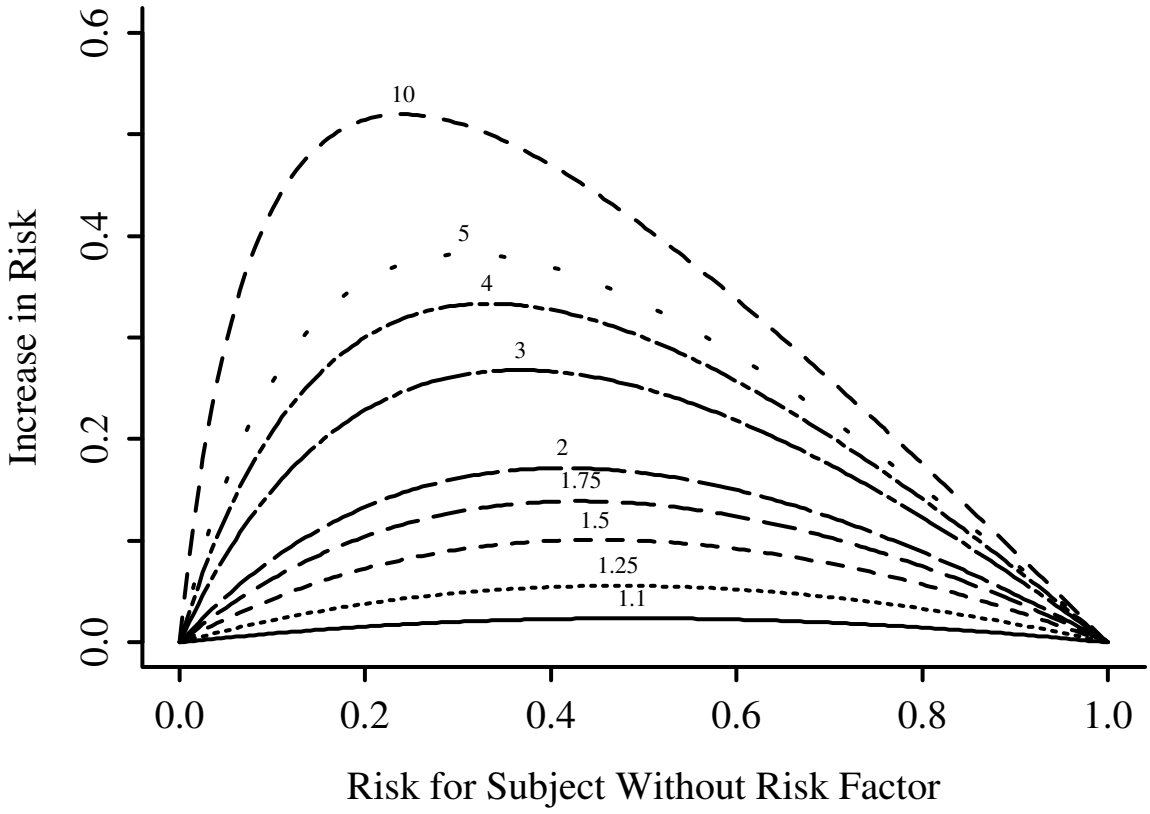

Figure 10.2: *Absolute benefit as a function of risk of the event in a control subject and the relative effect (odds ratio) of the risk factor. The odds ratios are given for each curve.*

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

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

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

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

- Risk ratio is $\frac{1+e^{-X_2\beta}}{1+e^{-X_1\beta}}$

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

10.1.3 Detailed Example

TABLE OF SEX BY RESPONSE

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

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

STATISTICS FOR TABLE OF SEX BY RESPONSE

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

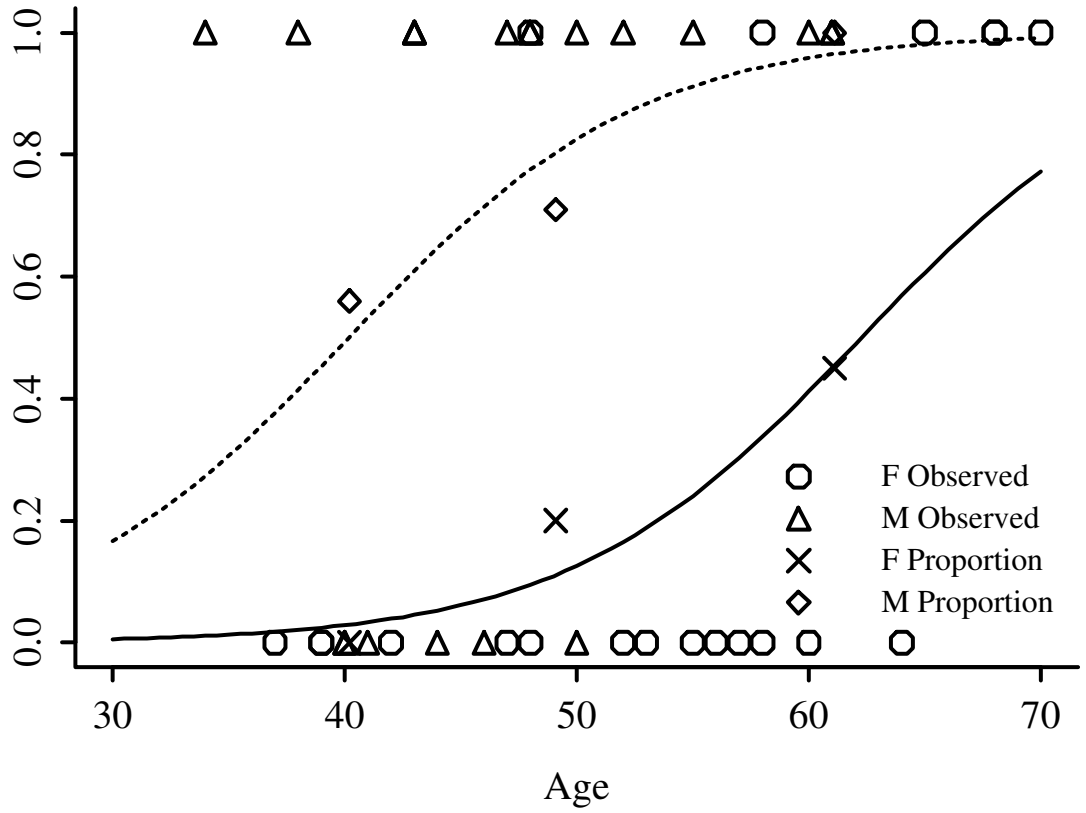

Figure 10.3: *Data, subgroup proportions, and fitted logistic model*

Fitted Logistic Model

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

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

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

TABLE OF AGE BY RESPONSE

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

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

Fitted Logistic Model

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

The estimate of β_1 is in rough agreement with that obtained from the frequency table. The 55+:$<$45 log odds ratio is .875, and since the respective mean ages in the 55+ and $<$45 age groups are 61.1 and 40.2, an estimate of the log odds ratio increase per year is .875/(61.1–40.2)=.875/20.9=.042.

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

Log likelihood ($\beta_1 = 0$)	:	-27.727
Log likelihood (max)	:	-26.511
LR $\chi^2(H_0 : \beta_1 = 0)$	:	-2(-27.727- -26.511) = 2.432

(Compare 2.432 with the Wald statistic 2.28.)

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

SEX=F

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

SEX=M

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

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

Fitted Logistic Model

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

Likelihood ratio tests are obtained from the information below.

Log likelihood ($\beta_1 = 0, \beta_2 = 0$)	:	-27.727
Log likelihood (max)	:	-19.458
Log likelihood ($\beta_1 = 0$)	:	-26.511
Log likelihood ($\beta_2 = 0$)	:	-24.435
LR χ^2 ($H_0 : \beta_1 = \beta_2 = 0$)	:	-2(-27.727- -19.458)= 16.538
LR χ^2 ($H_0 : \beta_1 = 0$) sex|age	:	-2(-26.511- -19.458) = 14.106
LR χ^2 ($H_0 : \beta_2 = 0$) age|sex	:	-2(-24.435- -19.458) = 9.954

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

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

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

10.1.4 Design Formulations

- Can do ANOVA using $k - 1$ dummies for a k-level predictor
- Can get same χ^2 statistics as from a contingency table
- Can go farther: covariable adjustment

- Simultaneous comparison of multiple variables between two groups: Turn problem backwards to predict group from all the *dependent* variables

- This is more robust than a parametric multivariate test

- Propensity scores for adjusting for nonrandom treatment selection: Predict treatment from all baseline variables

- Adjusting for the predicted probability of getting a treatment adjusts adequately for confounding from all of the variables

- In a randomized study, using logistic model to adjust for covariables, even with perfect balance, will improve the treatment effect estimate

10.2 Estimation

10.2.1 Maximum Likelihood Estimates

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

10.2.2 Estimation of Odds Ratios and Probabilities

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

10.3 Test Statistics

- Likelihood ratio test best

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

10.4 Residuals

Partial residuals (to check predictor transformations)

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

10.5 Assessment of Model Fit

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

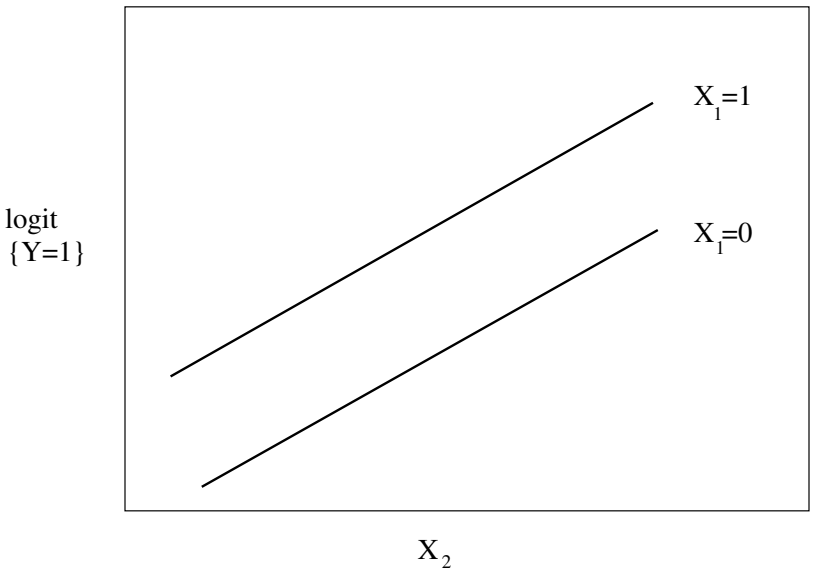

Figure 10.4: *Logistic regression assumptions for one binary and one continuous predictor*

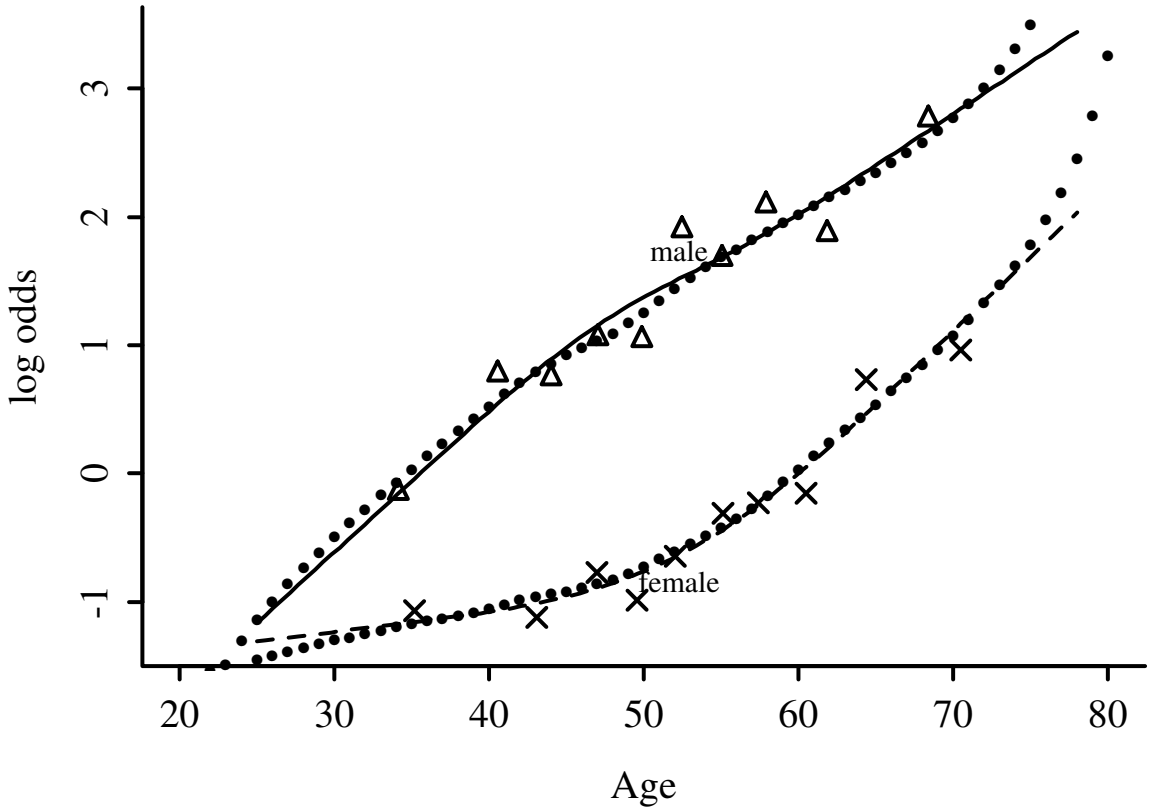

Figure 10.5: *Logit proportions of significant coronary artery disease by sex and deciles of age for n=3504 patients, with spline fits (smooth curves). Spline fits are for $k = 4$ knots at age=36, 48, 56, and 68 years, and interaction between age and sex is allowed. Smooth nonparametric estimates are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

- Can verify by plotting stratified proportions
- $\hat{P} =$ number of events divided by stratum size
- $\hat{O} = \frac{\hat{P}}{1-\hat{P}}$
- Plot $\log \hat{O}$ (scale on which linearity is assumed)
- Stratified estimates are noisy
- 1 or 2 Xs $\rightarrow$ nonparametric smoother
- `plsmo` function makes it easy to use loess to compute logits of nonparametric estimates (`fun=qlogis`)
- General: restricted cubic spline expansion of one or more predictors

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

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

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

- Example of finding transform. of a single continuous predictor
- Duration of symptoms vs. odds of severe coronary disease
- Look at AIC to find best # knots for the money

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

- Sample of 2258 pts
- Predict significant coronary disease
- For now stratify age into tertiles to examine interactions simply

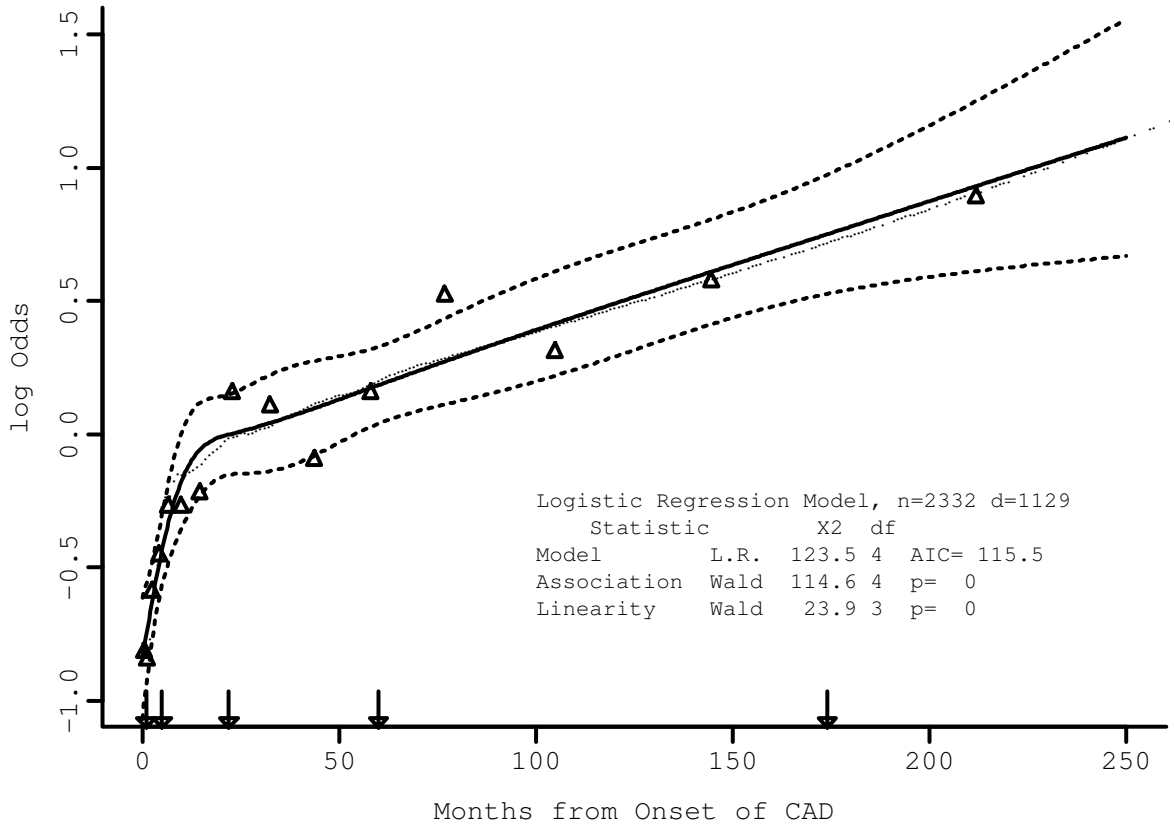

Figure 10.6: *Estimated relationship between duration of symptoms and the log odds of severe coronary artery disease for $k = 5$. Knots are marked with arrows. Solid line is spline fit; dotted line is a nonparametric "super-smoothed" estimate.*

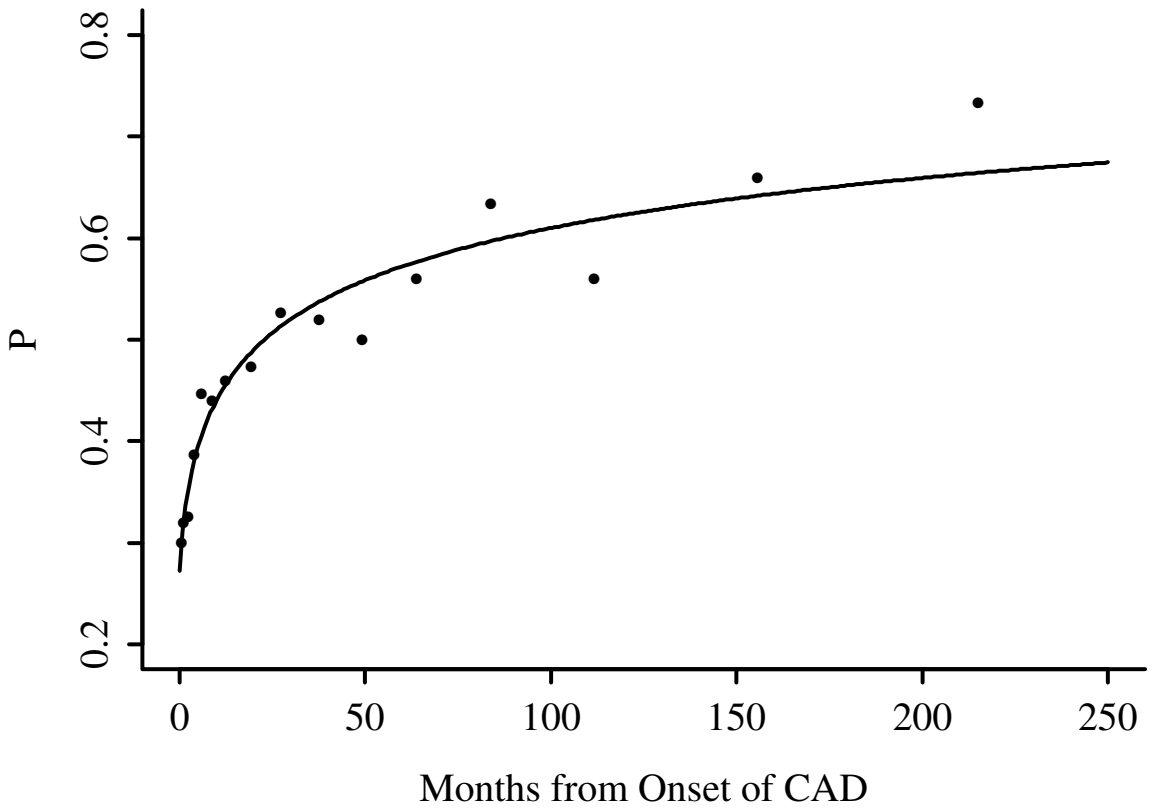

Figure 10.7: *Fitted linear logistic model in* $\log_{10}$*(*duration*+1), with subgroup estimates using groups of 150 patients. Fitted equation is logit(*`tvdlm`*)* $= -.9809 + .7122 \log_{10}(\text{months} + 1)$.

- Model has 2 dummies for age, sex, age $\times$ sex, 4-knot restricted cubic spline in cholesterol, age tertile $\times$ cholesterol

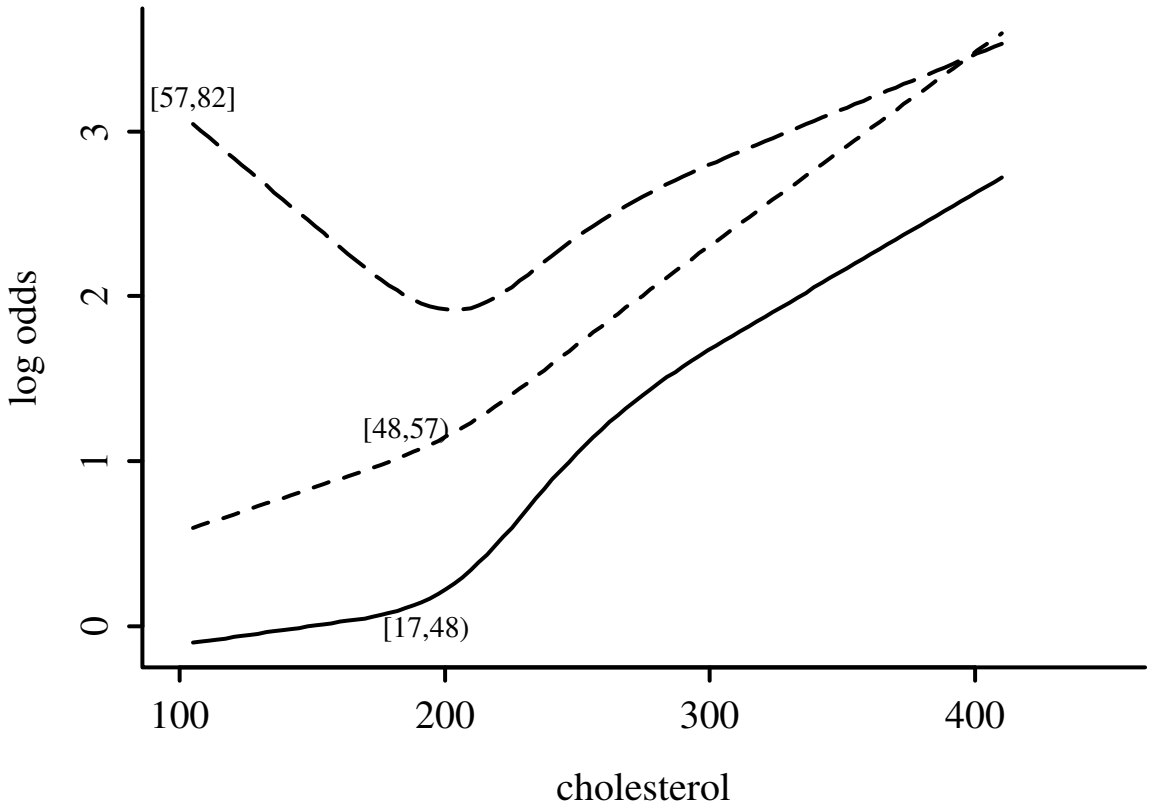

Figure 10.8: *Log odds of significant coronary artery disease modeling age with two dummy variables*

```
anova(fit)
```

Wald Statistics

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

- Now model age as continuous predictor
- Start with nonparametric surface using $Y = 0/1$
- Next try parametric fit using linear spline in age, chol. (3 knots each), all product terms
- Next try smooth spline surface, include all cross-products

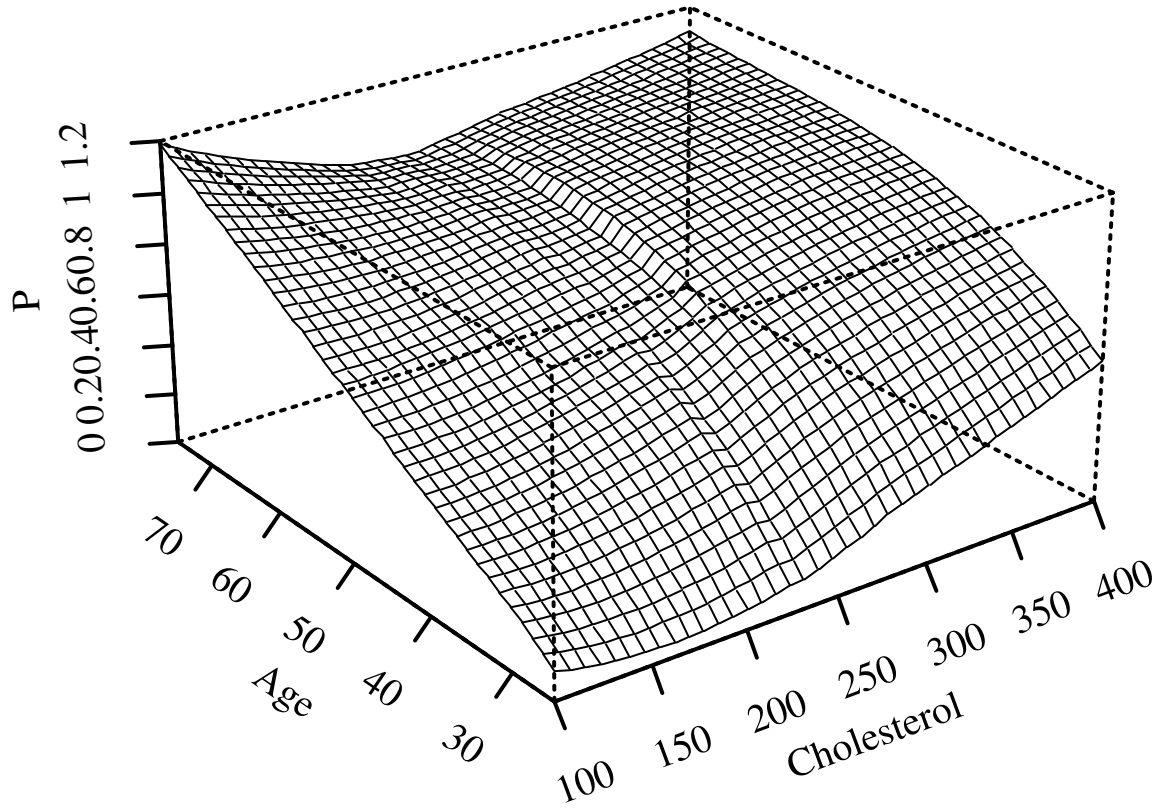

Figure 10.9: *Local regression fit for the probability of significant coronary disease vs. age and cholesterol for males, based on the S-P*LUS `loess` *function*

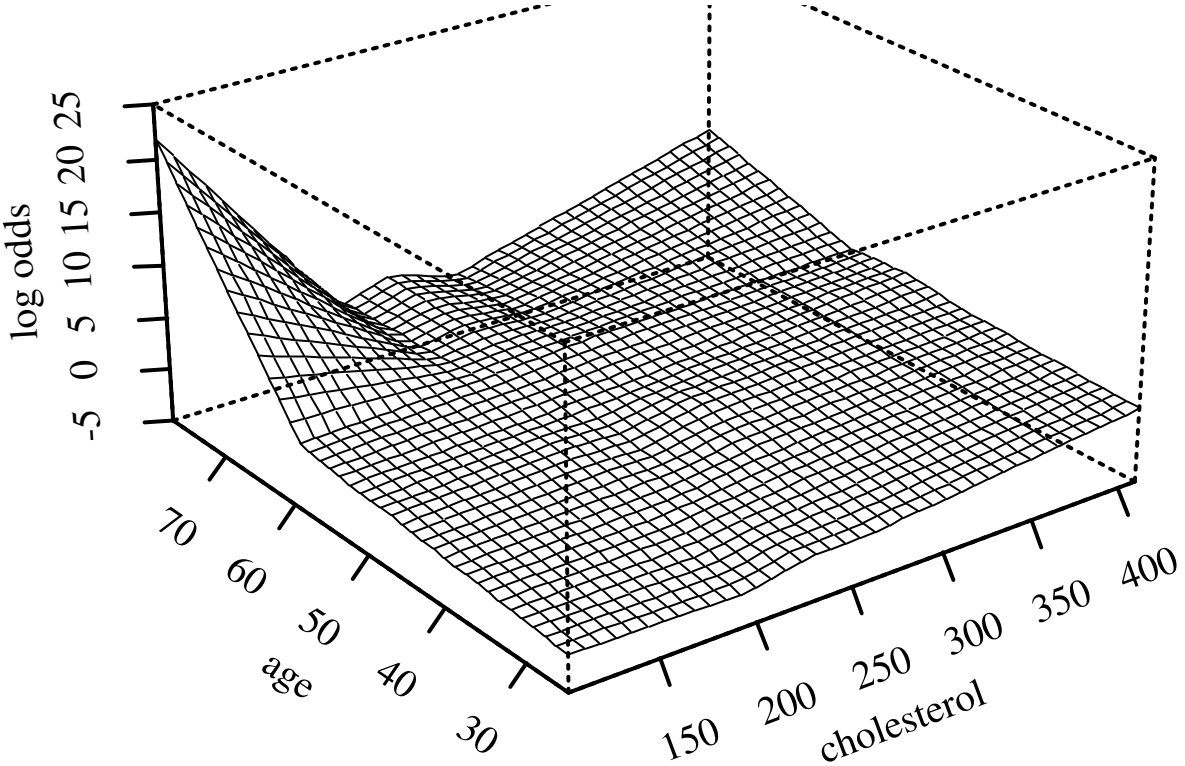

Figure 10.10: *Linear spline surface for males, with knots for age at 46, 52, 59 and knots for cholesterol at 196, 224, and 259 (quartiles)*

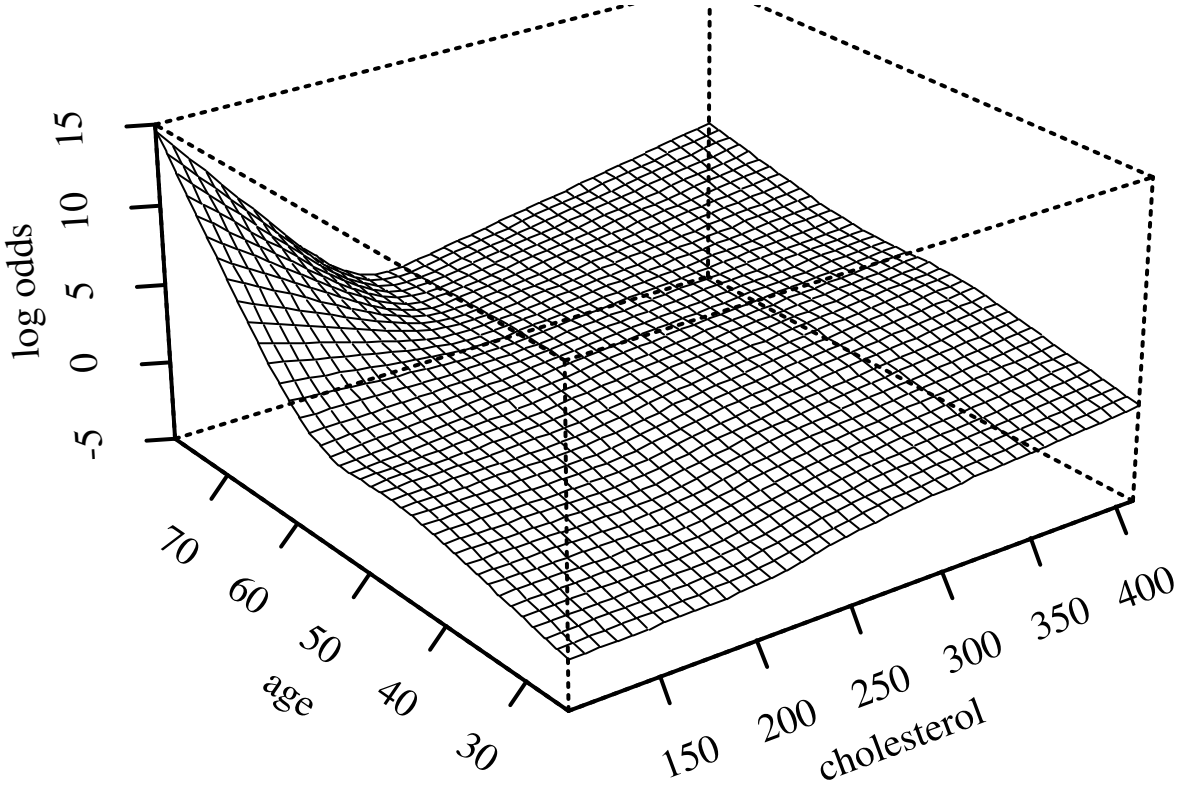

Figure 10.11: *Restricted cubic spline surface in two variables, each with $k = 4$ knots*

Wald Statistics

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

- Now restrict surface by excluding doubly nonlinear terms

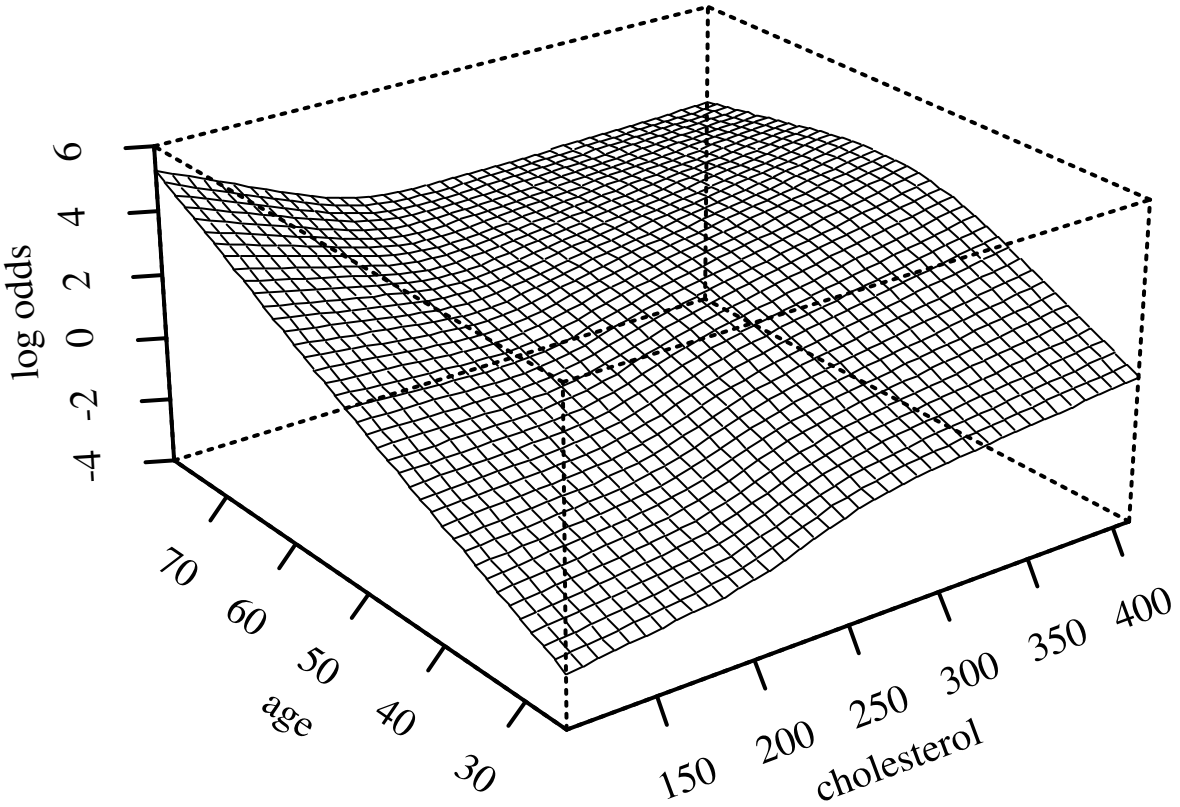

Figure 10.12: *Restricted cubic spline fit with age $\times$ spline(cholesterol) and cholesterol $\times$ spline(age)*

Wald Statistics

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

- Finally restrict the interaction to be a simple product The Wald test for age $\times$

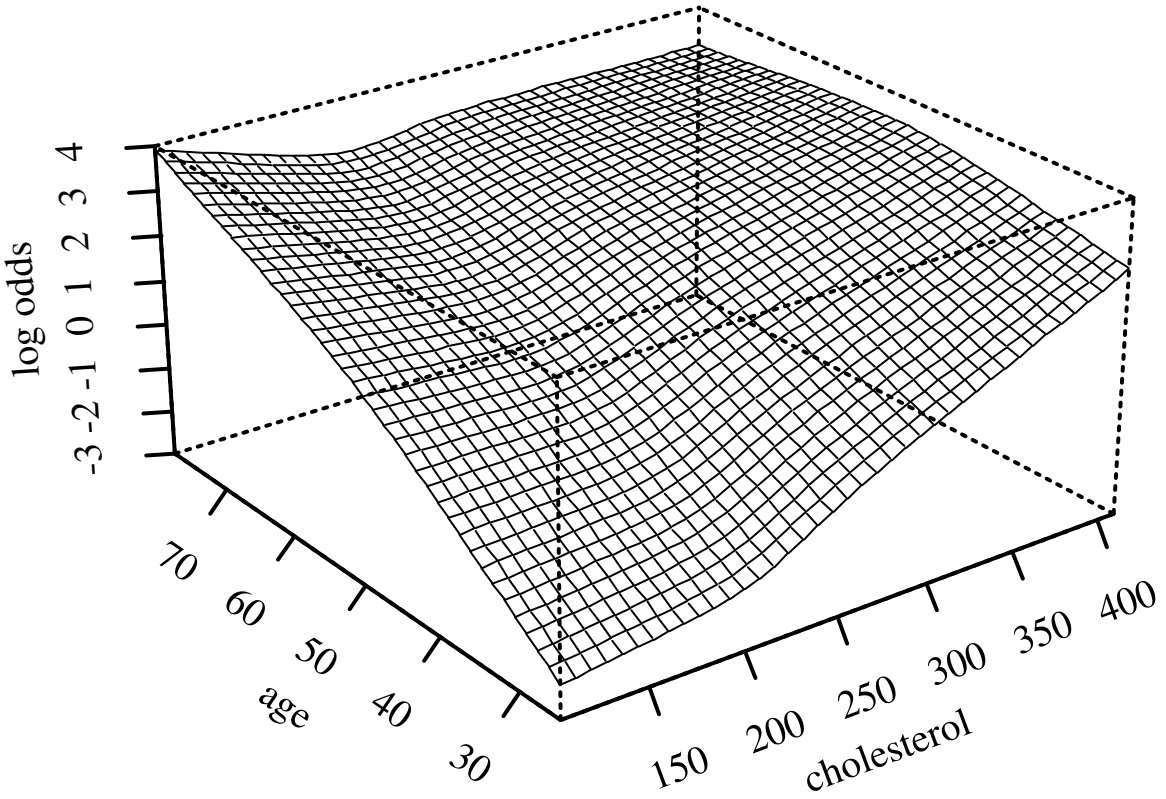

Figure 10.13: *Spline fit with nonlinear effects of cholesterol and age and a simple product interaction*

cholesterol interaction yields $\chi^2 = 7.99$ with 1 d.f., p=.005.

- See how well this simple interaction model compares with initial model using 2 dummies for age

- Request predictions to be made at mean age within tertiles

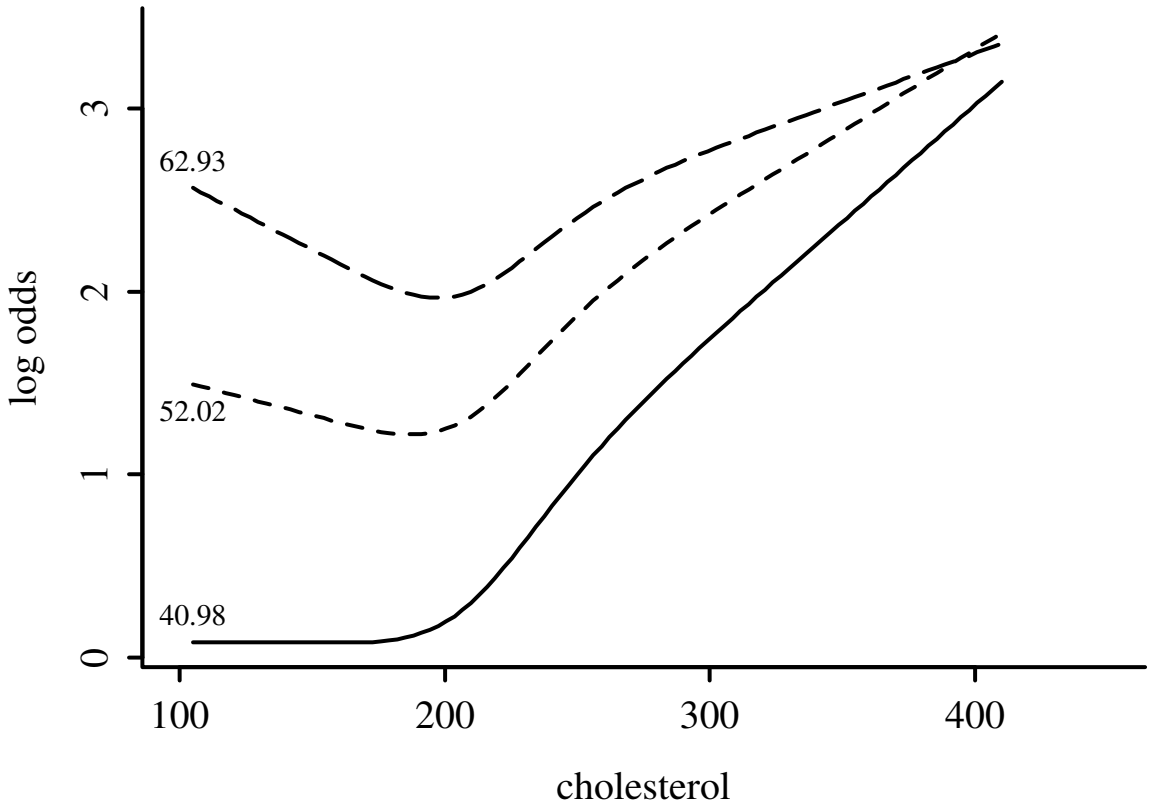

Figure 10.14: *Predictions from linear interaction model with mean age in tertiles indicated.*

- Using residuals for "duration of symptoms" example

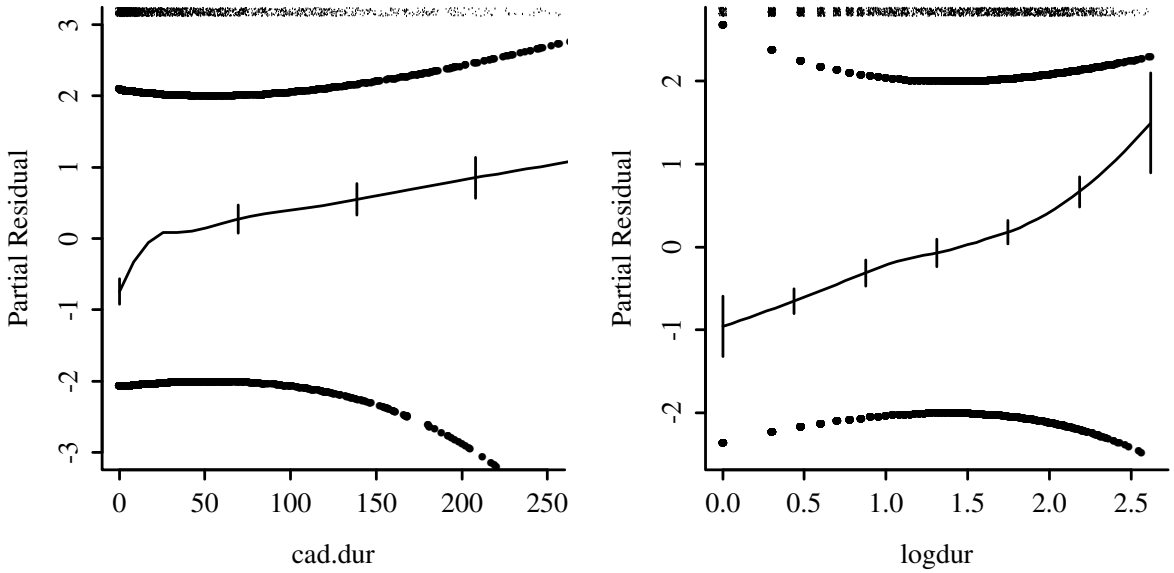

Figure 10.15: *Partial residuals for duration and* $\log_{10}$*(duration+1). Data density shown at top of each plot.*

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

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

- Hosmer-Lemeshow test is a commonly used test of goodness-of-fit of a binary logistic model
  Compares proportion of events with mean predicted probability within deciles of $\hat{P}$
  - Arbitrary (number of groups, how to form groups)
  - Low power (too many d.f.)

  - Does not reveal the culprits

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

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

10.6 Collinearity

10.7 Overly Influential Observations

10.8 Quantifying Predictive Ability

- Generalized R^2: equals ordinary R^2 in normal case:

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

- Brier score (calibration + discrimination):

$$B = \frac{1}{n}\sum_{i=1}^{n}(\hat{P}_i - Y_i)^2,$$

- c = "concordance probability" = ROC area

- Related to Wilcoxon-Mann-Whitney stat and Somers' D_{xy}

$$D_{xy} = 2(c - .5).$$

- "Percent classified correctly" has lots of problems

10.9 Validating the Fitted Model

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

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

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

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

Table 10.1: *Validation of 2-variable Logistic Model*

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

- Allow for step-down at each re-sample
- Use individual tests at $\alpha = 0.10$
- Both age and sex selected in 76 of 80, neither in 1 sample

Table 10.2: *Validation of 2-variable Stepwise Model*

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

- Try adding 5 noise candidate variables

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

- The first 15 patterns of factors selected are:

```
age sex x1 x2 x3 x4 x5


*   *
*   *           *
*


*   *
*   *   *       *  *


*   *           *

*   *       *
```

Table 10.3: *Validation of Model with 5 Noise Variables*

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

10.10 Describing the Fitted Model

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

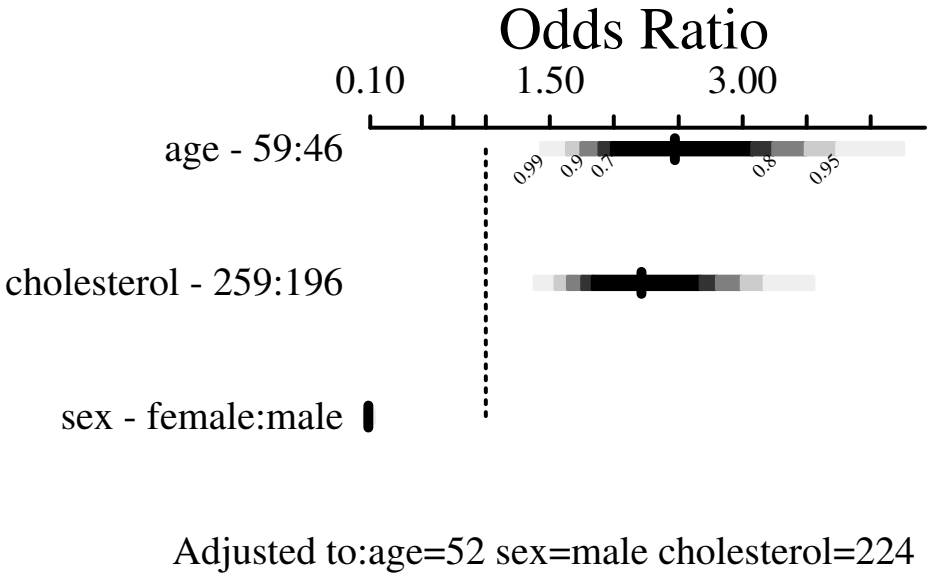

Figure 10.16: *Odds ratios and confidence bars, using quartiles of age and cholesterol for assessing their effects on the odds of coronary disease.*

10.11 S-PLUS Functions

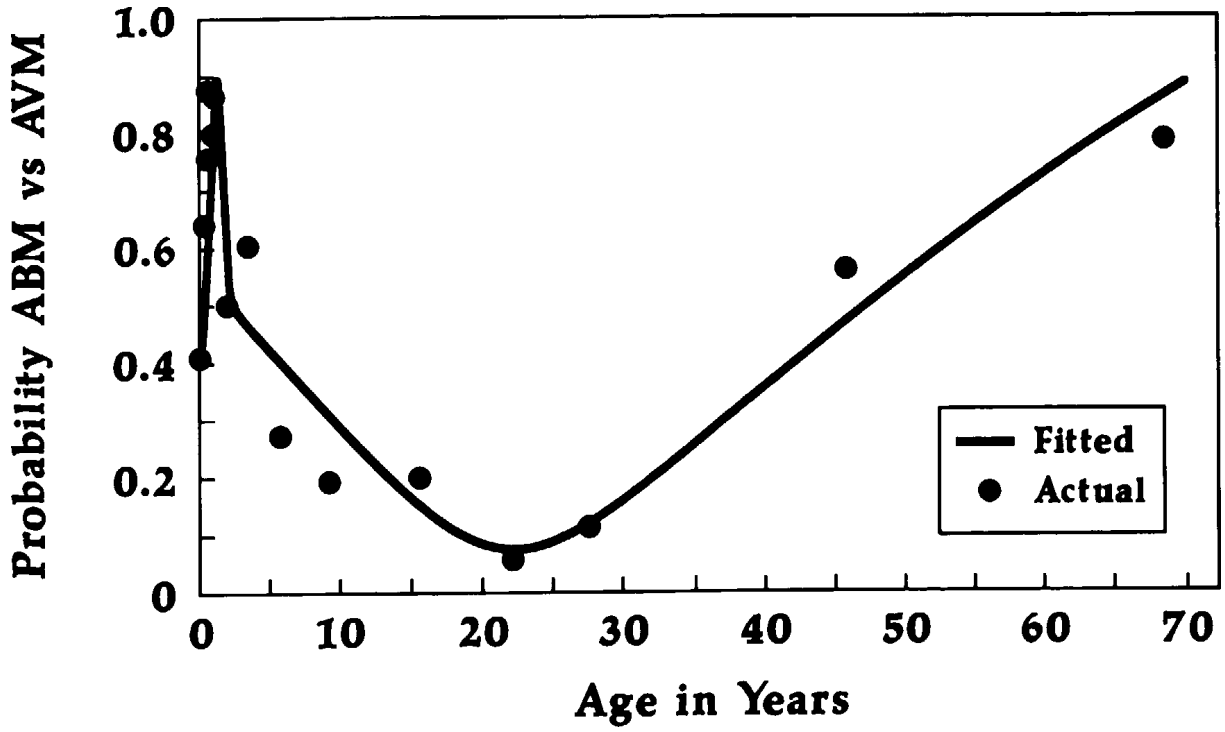

Figure 10.17: *Linear spline fit for probability of bacterial vs. viral meningitis as a function of age at onset . Copyrighted 1989, American Medical Associated. Reprinted by permission.*

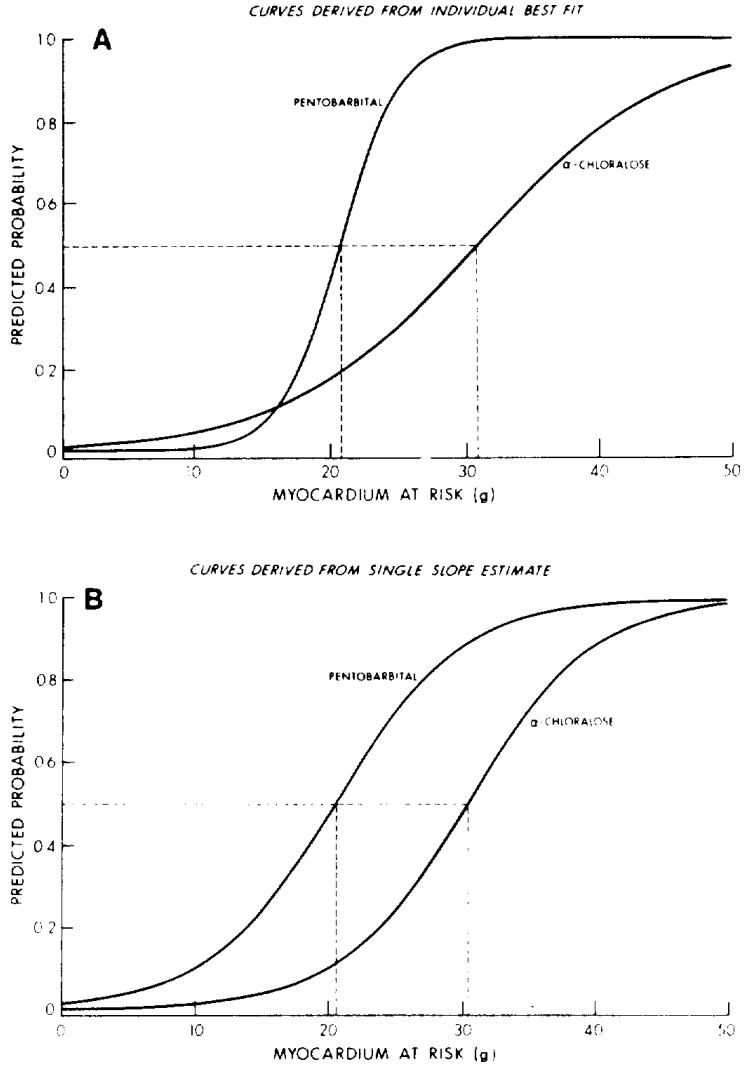

Figure 10.18: *(A) Relationship between myocardium at risk and ventricular fibrillation, based on the individual best fit equations for animals anesthetized with pentobarbital and α-chloralose. The amount of myocardium at risk at which 0.5 of the animals are expected to fibrillate (MAR_{50}) is shown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

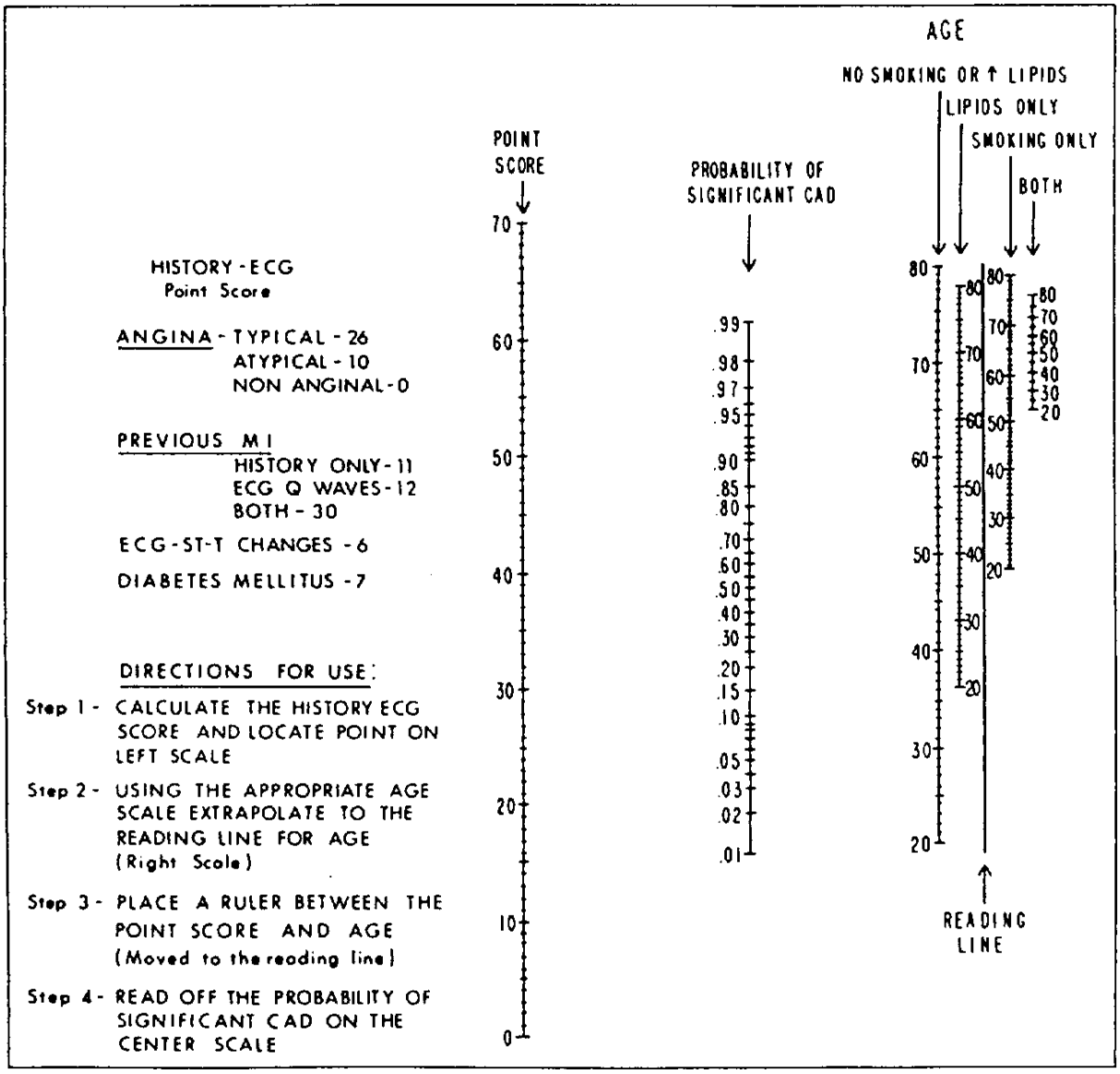

Figure 10.19: *A nomogram for estimating the likelihood of significant coronary artery disease (CAD) in women. ECG = electrocardiographic; MI = myocardial infarction . Reprinted from American Journal of Medicine, Vol 75, Pryor DB et al., "Estimating the likelihood of significant coronary artery disease",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

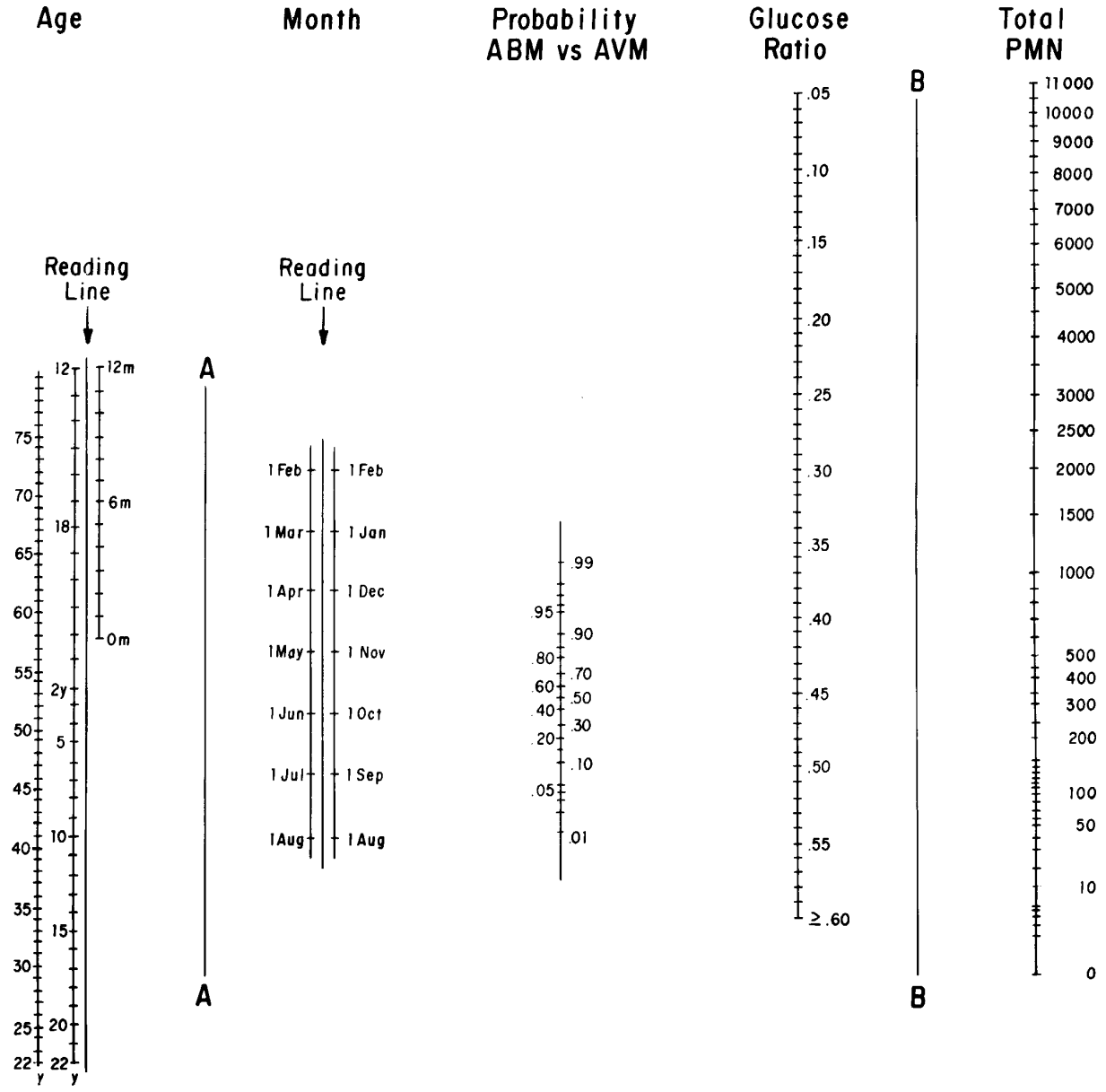

Figure 10.20: *Nomogram for estimating probability of bacterial (ABM) vs. viral (AVM) meningitis. Step 1, place ruler on reading lines for patient's age and month of presentation and mark intersection with line A; step 2, place ruler on values for glucose ratio and total polymorphonuclear leukocyte (PMN) count in cerbrospinal fluid and mark intersection with line B; step 3, use ruler to join marks on lines A and B, then read off the probability of ABM vs. AVM . Copyrighted 1989, American Medical Associated. Reprinted by permission.*

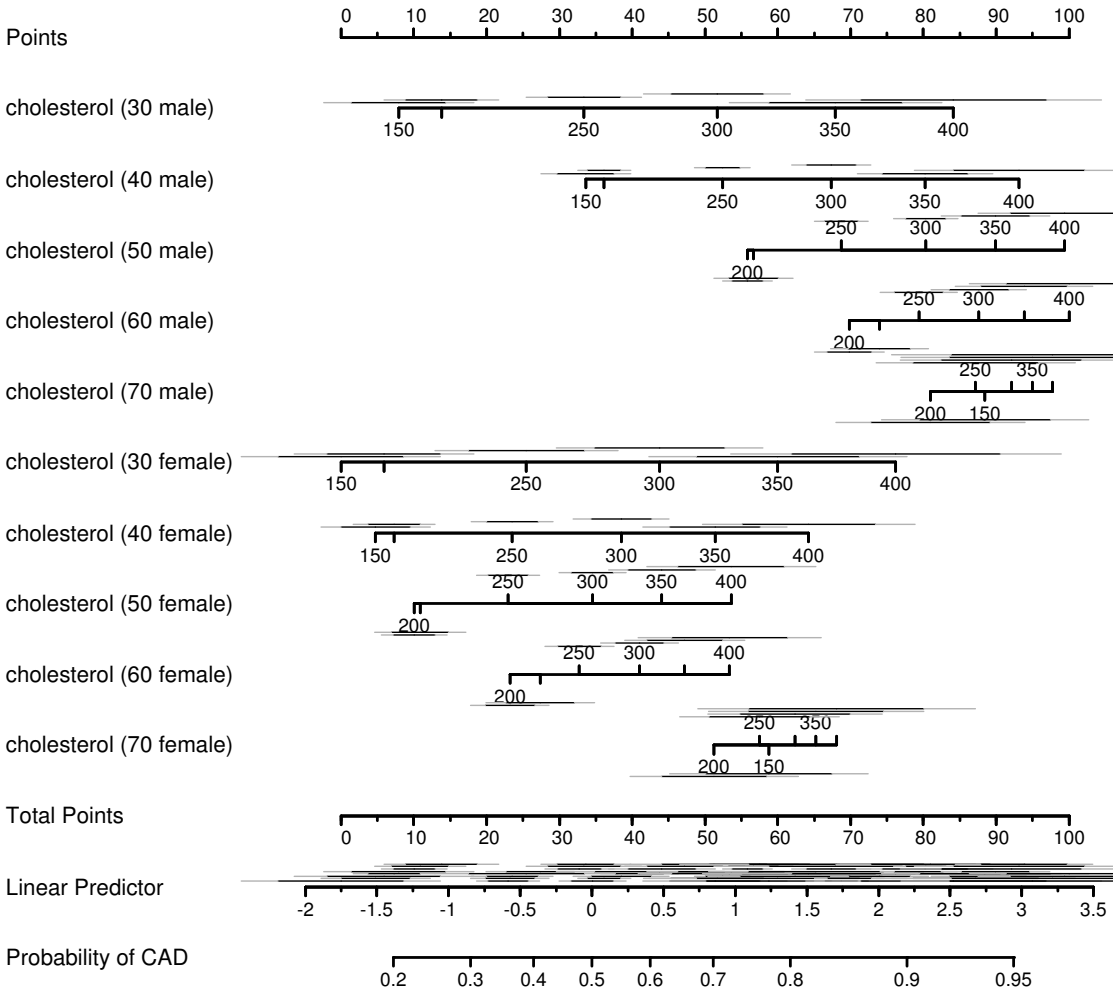

Figure 10.21: *Nomogram relating age, sex, and cholesterol to the log odds and to the probability of significant coronary artery disease. Select one axis corresponding to sex and to age* $\in \{30, 40, 50, 60, 70\}$*. There was linear interaction between age and sex and between age and cholesterol. 0.70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

Chapter 11

Ordinal Logistic Regression

11.1 Background

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

- Continuation ratio model: discrete proportional hazards model

11.2 Ordinality Assumption

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

11.3 Proportional Odds Model

11.3.1 Model

- Walker & Duncan — most popular ordinal response model
- For convenience $Y = 0, 1, 2, \ldots, k$

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

  where $j = 1, 2, \ldots, k$.

- α_j is the logit of Prob$[Y \geq j]$ when all Xs are zero
- Odds$[Y \geq j|X] = \exp(\alpha_j + X\beta)$
- Odds$[Y \geq j|X_m = a + 1]$ / Odds$[Y \geq j|X_m = a] = e^{\beta_m}$
- Same odds ratio e^{β_k} for any $j = 1, 2, \ldots, k$
- Odds$[Y \geq j|X]$ / Odds$[Y \geq v|X] = \frac{e^{\alpha_j + X\beta}}{e^{\alpha_v + X\beta}} = e^{\alpha_j - \alpha_v}$

- Odds$[Y \geq j|X] = constant \times$ Odds$[Y \geq v|X]$

- Assumes OR for 1 unit increase in age is the same when considering the probability of death as when considering the probability of death or heart attack

- PO model only uses ranks of Y; same $\hat{\beta}$s if transform Y; is robust to outliers

11.3.2 Assumptions and Interpretation of Parameters

11.3.3 Estimation

11.3.4 Residuals

- Construct binary events $Y \geq j, j = 1, 2, \ldots, k$ and use corresponding predicted probabilities

$$\hat{P}_{ij} = \frac{1}{1 + \exp[-(\hat{\alpha}_j + X_i\hat{\beta})]},$$

- Score residual for subject i predictor m:

$$U_{im} = X_{im}([Y_i \geq j] - \hat{P}_{ij}),$$

- For each column of U plot mean $\bar{U}_{\cdot m}$ and C.L. against Y

- Partial residuals are more useful as they can also estimate covariable transformations :

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

where

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

- Smooth r_{im} vs. X_{im} to estimate how X_m relates to the log relative odds that $Y = 1|X_m$

- For ordinal Y compute binary model partial res. for all cutoffs j:

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

11.3.5 Assessment of Model Fit

- Section 11.2

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

11.3.6 Quantifying Predictive Ability

11.3.7 Validating the Fitted Model

11.3.8 S-PLUS Functions

The `Design` library's `lrm` function fits the PO model directly, assuming that the levels of the response variable (e.g., the `levels` of a `factor` variable) are listed in the proper order.

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

The function `plot.xmean.ordinaly` in `Design` computes and graphs the quantities described in Section 11.2. It plots simple Y-stratified means overlaid with $\hat{E}(X|Y = j)$, with j on the x-axis. The $\hat{E}$s are computed for both PO and continuation ratio ordinal logistic models.

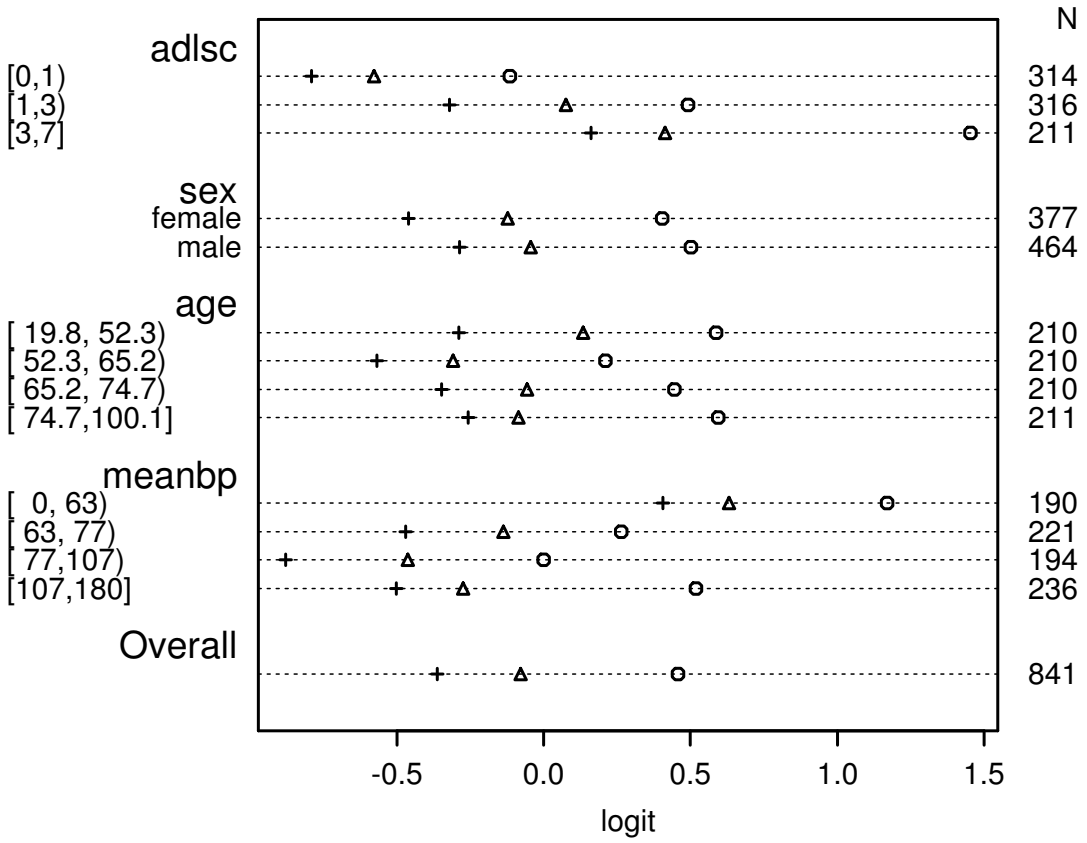

Figure 11.1: *Checking PO assumption separately for a series of predictors. The circle, triangle, and plus sign correspond to* $Y \geq 1, 2, 3$*, respectively. PO is checked by examining the vertical constancy of distances between any two of these three symbols. Response variable is the severe functional disability scale* `sfdm2` *from the 1000-patient SUPPORT dataset, with the last two categories combined because of low frequency of coma/intubation.*

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

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

Chapter 16

Introduction to Survival Analysis

16.1 Background

- Use when time to occurrence of event is important
- Don't just count events; event at 6m worse than event at 9y
- Response called *failure time, survival time, event time*
- Ex: time until CV death, light bulb failure, pregnancy, ECG abnormality during exercise
- Allow for censoring
- Ex: 5y f/u study; subject still alive at 5y has failure time $5+$
- Length of f/u can vary
- Even in a well-designed randomized clinical trial, survival modeling can allow

one to

1. Test for and describe interactions with treatment. Subgroup analyses can easily generate spurious results and they do not consider interacting factors in a dose-response manner. Once interactions are modeled, relative treatment benefits can be estimated (e.g., hazard ratios), and analyses can be done to determine if some patients are too sick or too well to have even a relative benefit.
2. Understand prognostic factors (strength and shape).
3. Model absolute clinical benefit. First, a model for the probability of surviving past time t is developed. Then differences in survival probabilities for patients on treatments A and B can be estimated. The differences will be due primarily to sickness (overall risk) of the patient and to treatment interactions.
4. Understand time course of treatment effect. The period of maximum effect or period of any substantial effect can be estimated from a plot of relative effects of treatment over time.
5. Gain power for testing treatment effects.
6. Adjust for imbalances in treatment allocation.

16.2 Censoring, Delayed Entry, and Truncation

- Left-censoring
- Interval censoring
- Left-truncation (unknown subset of subjects who failed before qualifying for the study)
- Delayed entry (exposure after varying periods of survival)
- Choice of time zero important

- Take into account *waiting time bias*

- Usually have random type I censoring (on duration, not # events)

- Must usually have *non-informative censoring*:
  censoring independent of impending failure

- Intention-to-treat is a preventative measure

16.3 Notation, Survival, and Hazard Functions

- $T =$ response variable

$$S(t) = \text{Prob}\{T > t\} = 1 - F(t),$$

- Hazard function (force of mortality; instantaneous event rate)

- $\approx$ Prob{event will occur in small interval around t given has not occurred before t}

- Very useful for learning about mechanisms and forces of risk over time

$$\lambda(t) = \lim_{u \to 0} \frac{\text{Prob}\{t < T \leq t + u | T > t\}}{u},$$

  which using the law of conditional probability becomes

$$\begin{aligned} \lambda(t) &= \lim_{u \to 0} \frac{\text{Prob}\{t < T \leq t + u\}/\text{Prob}\{T > t\}}{u} \\ &= \lim_{u \to 0} \frac{[F(t+u) - F(t)]/u}{S(t)} \\ &= \frac{\partial F(t)/\partial t}{S(t)} \end{aligned}$$

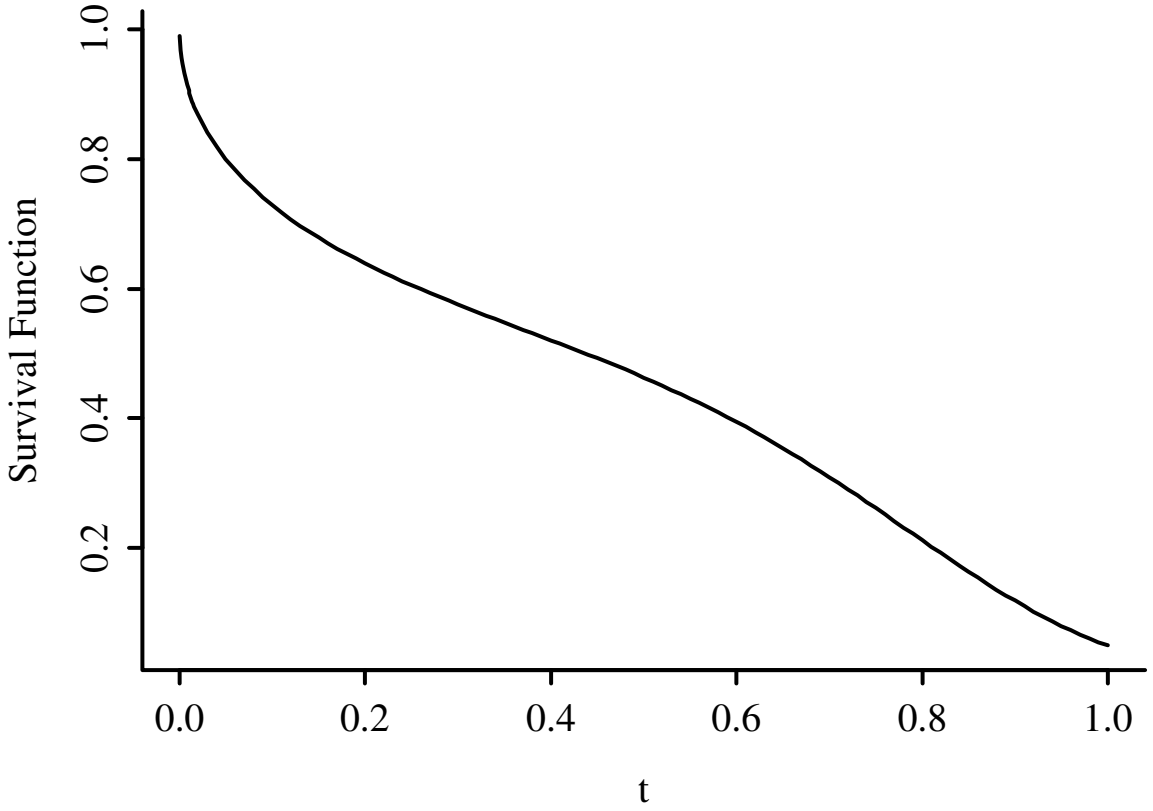

Figure 16.1: *Survival function*

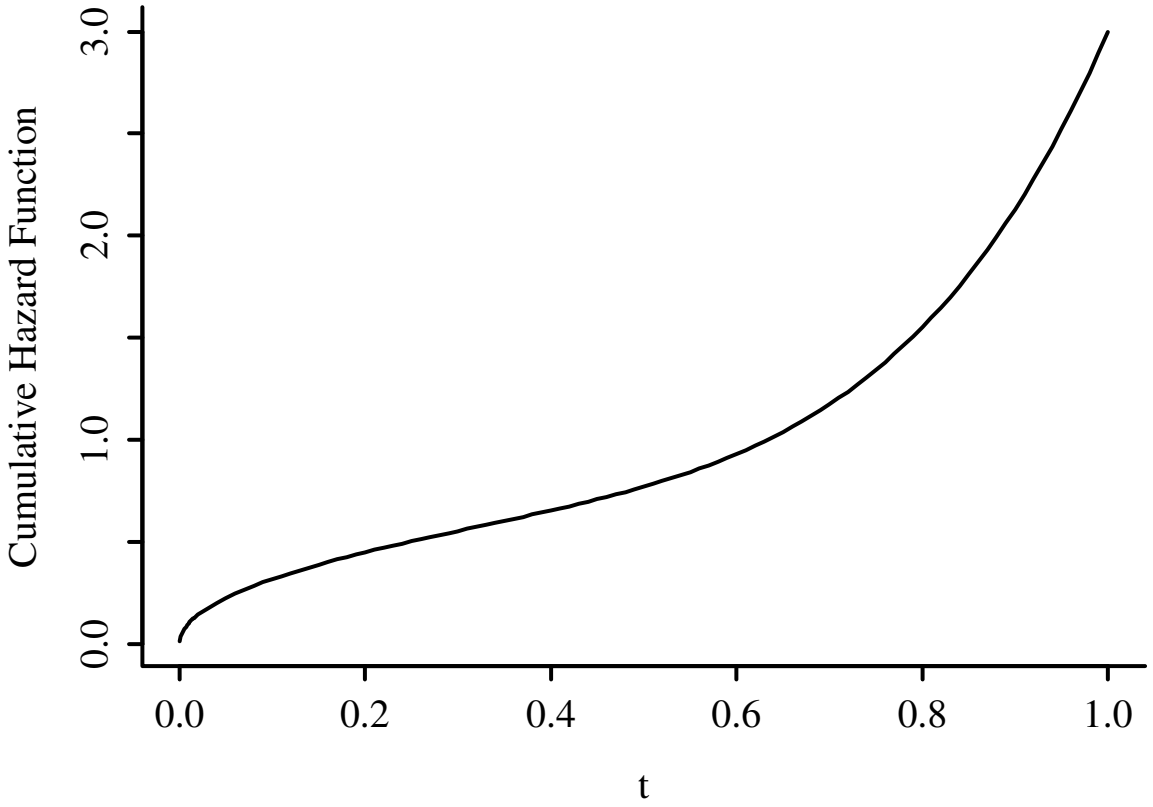

Figure 16.2: *Cumulative hazard function*

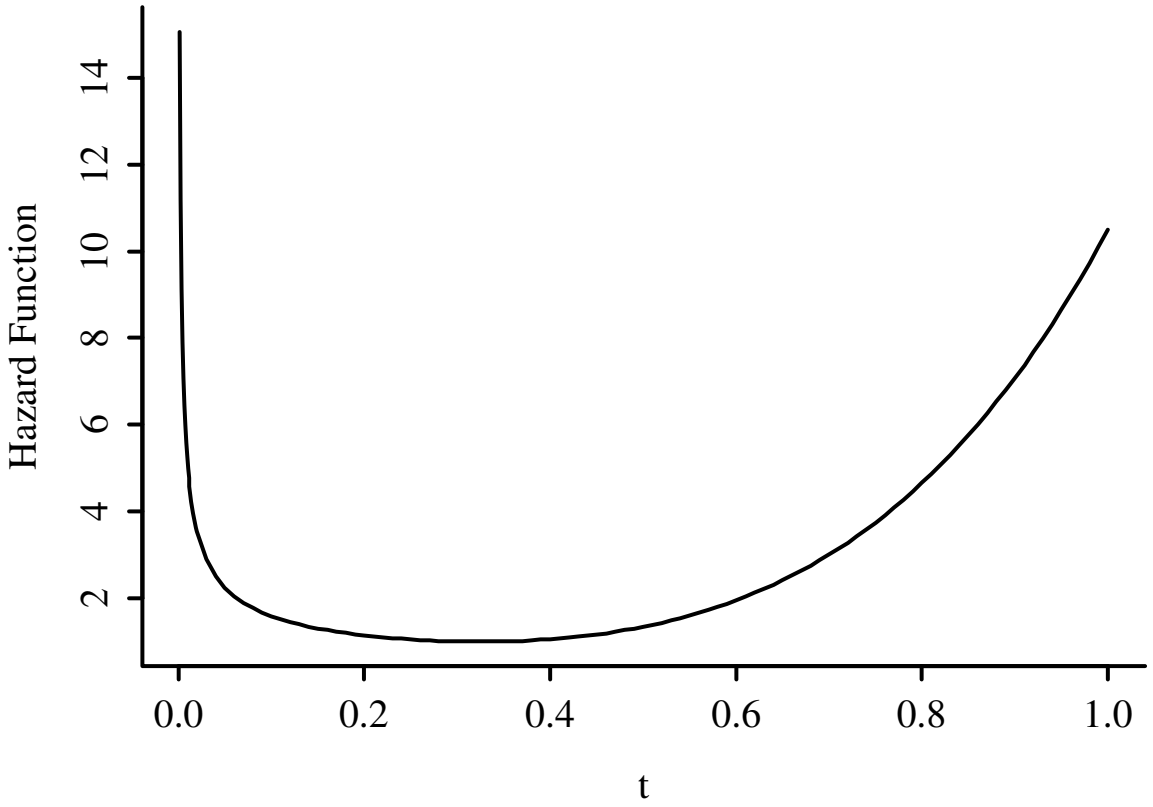

Figure 16.3: *Hazard function*

$$= \frac{f(t)}{S(t)},$$

$$\frac{\partial \log S(t)}{\partial t} = \frac{\partial S(t)/\partial t}{S(t)} = -\frac{f(t)}{S(t)},$$

$$\lambda(t) = -\frac{\partial \log S(t)}{\partial t},$$

$$\int_0^t \lambda(v)dv = -\log S(t).$$

$$\Lambda(t) = -\log S(t),$$

$$S(t) = \exp[-\Lambda(t)].$$

- Expected value of $\Lambda(T) = 1$

$$T_q = S^{-1}(1-q).$$

$$T_{0.5} = S^{-1}(0.5).$$

$$\begin{aligned} T_q &= \Lambda^{-1}[-\log(1-q)] \text{ and as a special case,} \\ T_{.5} &= \Lambda^{-1}(\log 2). \end{aligned}$$

$$\mu = \int_0^\infty S(v)dv.$$

- Event time for subject i: T_i

- Censoring time: C_i

- Event indicator:

$$\begin{aligned} e_i &= 1 \text{ if the event was observed } (T_i \leq C_i), \\ &= 0 \text{ if the response was censored } (T_i > C_i). \end{aligned}$$

- The observed response is

$$Y_i = \min(T_i, C_i),$$

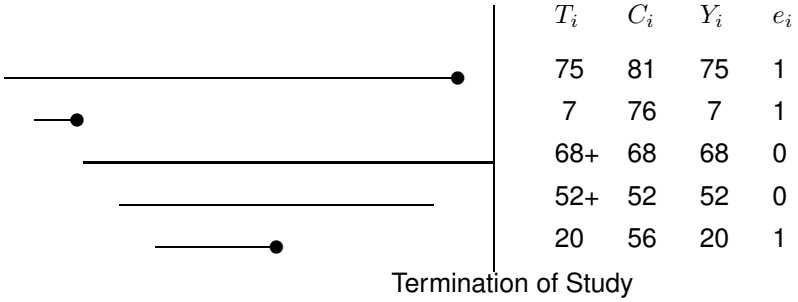

Figure 16.4: *Some censored data. Circles denote events.*

16.4 Homogeneous Failure Time Distributions

- Exponential distribution: constant hazard

$$\begin{aligned}\Lambda(t) &= \lambda t \text{ and} \\ S(t) &= \exp(-\Lambda(t)) = \exp(-\lambda t).\end{aligned}$$

$$T_{0.5} = \log(2)/\lambda.$$

- Weibull distribution

$$\begin{aligned}\lambda(t) &= \alpha\gamma t^{\gamma-1} \\ \Lambda(t) &= \alpha t^{\gamma} \\ S(t) &= \exp(-\alpha t^{\gamma}).\end{aligned}$$

$$T_{0.5} = [(\log 2)/\alpha]^{1/\gamma}.$$

- The restricted cubic spline hazard model with k knots is

$$\lambda_k(t) = a + bt + \sum_{j=1}^{k-2} \gamma_j w_j(t),$$

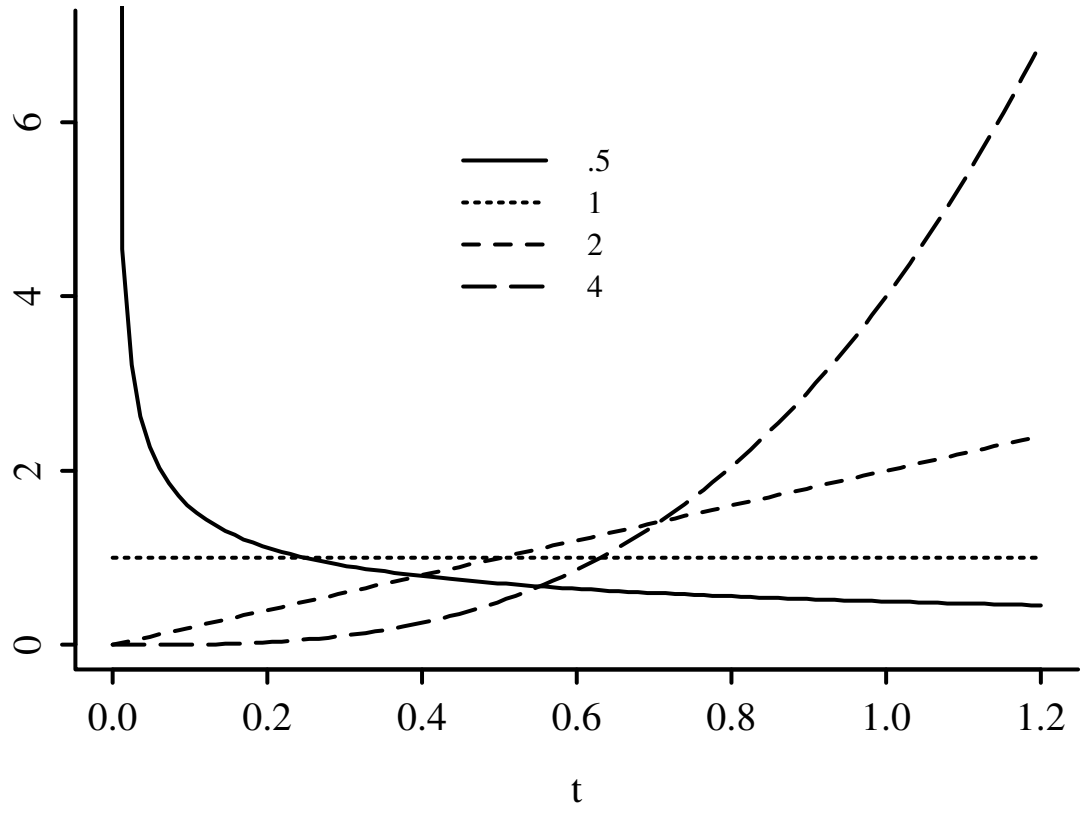

Figure 16.5: *Some Weibull hazard functions with $\alpha = 1$ and various values of γ*

16.5 Nonparametric Estimation of S and Λ

16.5.1 Kaplan–Meier Estimator

- No censoring $\rightarrow$

$$S_n(t) = [\text{number of } T_i > t]/n.$$

- Kaplan–Meier (product-limit) estimator

Day	No. Subjects At Risk	Deaths	Censored	Cumulative Survival
12	100	1	0	$99/100 = .99$
30	99	2	1	$97/99 \times 99/100 = .97$
60	96	0	3	$96/96 \times .97 = .97$
72	93	3	0	$90/93 \times .97 = .94$
.	.	.	.	.
.	.	.	.	.

$$S_{\text{KM}}(t) = \prod_{i:t_i<t} (1 - d_i/n_i).$$

- The Kaplan–Meier estimator of $\Lambda(t)$ is $\Lambda_{\text{KM}}(t) = -\log S_{\text{KM}}(t)$.

- Simple example

$$1 \;\; 3 \;\; 3 \;\; 6^+ \;\; 8^+ \;\; 9 \;\; 10^+.$$

i	t_i	n_i	d_i	$(n_i - d_i)/n_i$
1	1	7	1	6/7
2	3	6	2	4/6
3	9	2	1	1/2

$$
\begin{aligned}
S_{\mathrm{KM}}(t) &= 1, \quad 0 \le t < 1 \\
&= 6/7 = .85, \;\; 1 \le t < 3 \\
&= (6/7)(4/6) = .57, \;\; 3 \le t < 9 \\
&= (6/7)(4/6)(1/2) = .29, \;\; 9 \le t < 10.
\end{aligned}
$$

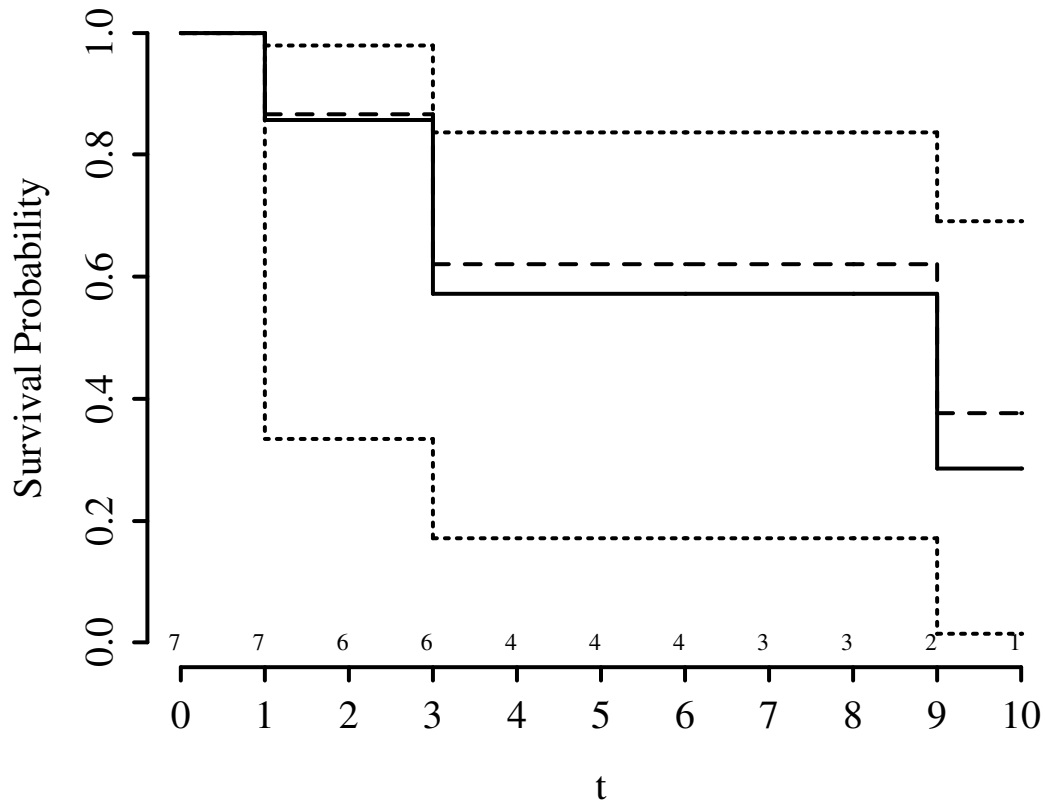

Figure 16.6: *Kaplan–Meier product-limit estimator with 0.95 confidence bands. The Altschuler–Nelson–Aalen–Fleming–Harrington estimator is depicted with the dashed lines.*

$$
\mathrm{Var}\{\log \Lambda_{\mathrm{KM}}(t)\} = \frac{\sum_{i:t_i<t} d_i/[n_i(n_i - d_i)]}{\{\sum_{i:t_i<t} \log[(n_i - d_i)/n_i]\}^2}
$$

$$S_{\mathrm{KM}}(t)^{\exp(\pm zs)}.$$

16.5.2 Altschuler–Nelson Estimator

$$\begin{aligned} \hat{\Lambda}(t) &= \sum_{i:t_i<t} \frac{d_i}{n_i} \\ S_{\Lambda}(t) &= \exp(-\hat{\Lambda}(t)) \end{aligned}$$

16.6 Analysis of Multiple Endpoints

- Cancer trial: recurrence of disease or death
- CV trial: nonfatal MI or death
- Analyze usual way but watch out for differing risk factors
- Analyzing multiple causes of terminating event →
  Cause-specific hazards, censor on cause not currently analyzed
  Not assume mechanism for cause removal or correlations of causes
  Problem if apply to a setting where causes are removed differently
- More complex if explicitly handle mixture of nonfatal outcomes with fatal outcome

16.6.1 Competing Risks

16.6.2 Competing Dependent Risks

16.6.3 State Transitions and Multiple Types of Nonfatal Events

16.6.4 Joint Analysis of Time and Severity of an Event

16.6.5 Analysis of Multiple Events

16.7 S-PLUS Functions

- `event.chart` in `Hmisc` draws a variety of charts for displaying raw survival time data, for both single and multiple events per subject (see also `event.history`)
- Analyses in this chapter can be done as special cases of the Cox model
- Particular functions for this chapter (no covariables) from Therneau:
- `Surv` function: Combines time to event variable and event/censoring indicator into a single survival time matrix object
- Right censoring: `Surv(y, event)`; `event` is event/censoring indicator, usually coded `T/F`, 0=censored 1=event or 1=censored 2=event. If the event status variable has other coding, e.g., 3 means death, use `Surv(y, s==3)`.
- `survfit`: Kaplan–Meier and other nonparametric survival curves

```
units(y) <- "Month"
# Default is "Day" - used for axis labels, etc.
survfit(Surv(y, event) ~ svar1 + svar2 + ... , data, subset,
        na.action=na.delete,
        type=c("kaplan-meier","fleming-harrington"),
        error=c("greenwood","tsiatis"), se.fit=T,
        conf.int=.95,
```

```
          conf.type=c("log-log","log","plain","none"))
```

If there are no stratification variables (`svar1`, ...), omit them. To print a table of estimates, use

```
f ← survfit(. . .)
print(f)     # print brief summary of f
summary(f, times, censored=F, digits=3)
```

For failure times stored in days, use

```
f ← survfit(Surv(futime, event) ∼ sex)
summary(f, seq(30,180,by=30))
```

to print monthly estimates.

To plot the object returned by `survfit`, use

```
plot(f, conf.int=T, mark.time=T, mark=3, col=1, lty=1,
     lwd=1, cex=1, log=F, yscale=1, xscale=1,
     xlab="", ylab="", xaxs="i", ...)
```

This invokes `plot.survfit`. You can also use `survplot` in `Design` (here, actually `survplot.survfit`) for other options that include automatic curve labeling and showing the number of subjects at risk at selected times. Figure 16.6 was drawn with the statements

```
tt ← c(1,3,3,6,8,9,10)
stat ← c(1,1,1,0,0,1,0)
S ← Surv(tt, stat)
survplot(survfit(S),conf="bands",n.risk=T,xlab="t")
survplot(survfit(S, type="fleming-harrington", conf.int=F),
         add=T, lty=3)
```

Stratified estimates, with four treatments distinguished by line type and curve labels, could be drawn by

```
units(y) <- "Year"
f ← survfit(Surv(y, stat) ∼ treatment)
survplot(f, ylab="Fraction Pain-Free")
```

- `bootkm` function in `Hmisc` bootstraps Kaplan–Meier survival estimates or Kaplan–Meier estimates of quantiles of the survival time distribution. It is easy to use `bootkm` to compute for example a nonparametric confidence interval for the ratio of median survival times for two groups.

Chapter 19

Cox Proportional Hazards Regression Model

19.1 Model

19.1.1 Preliminaries

- Most popular survival model
- Semi-parametric (nonparametric hazard; parametric regression)
- Usually more interest in effects of X than on shape of $\lambda(t)$
- Uses only rank ordering of failures/censoring times $\rightarrow$ more robust, easier to write protocol
- Even if parametric PH assumptions true, Cox still fully efficient for β
- Model diagnostics are advanced

- Log-rank test is a special case with one binary X

19.1.2 Model Definition

$$\lambda(t|X) = \lambda(t)\exp(X\beta).$$

- No intercept parameter
- Shape of λ not important

19.1.3 Estimation of β

19.1.4 Model Assumptions and Interpretation of Parameters

19.1.5 Example

Model	Group Regression Coefficient	S.E.	Wald p Value	Group 2:1 Hazard Ratio
Cox (Exact)	-0.629	0.361	0.08	0.533
Cox (Efron)	-0.569	0.347	0.10	0.566
Cox (Breslow)	-0.596	0.348	0.09	0.551
Exponential	-0.093	0.334	0.78	0.911
Weibull (AFT)	0.132	0.061	0.03	
Weibull (PH)	-0.721			0.486

19.1.6 Design Formulations

- $k-1$ dummies for k treatments, one treatment $\rightarrow \lambda(t)$
- Only provides relative effects

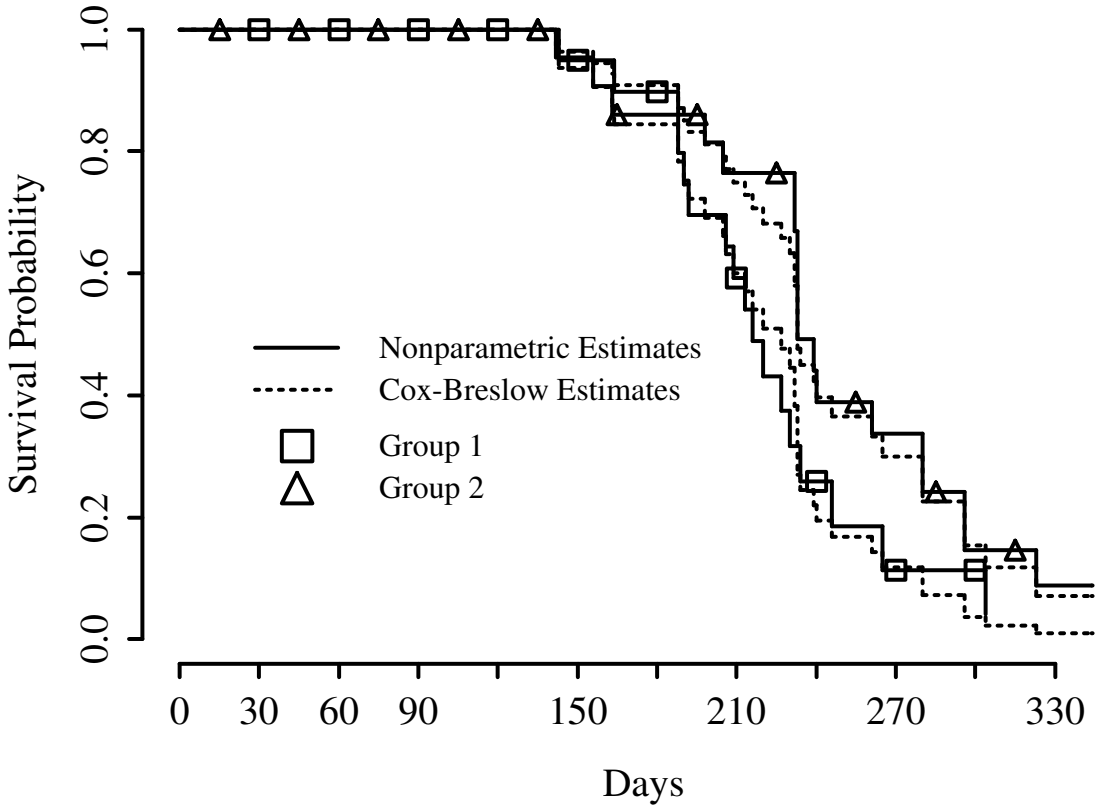

Figure 19.1: *Altschuler-Nelson-Fleming-Harrington nonparametric survival estimates and Cox-Breslow estimates for rat data*

19.1.7 Extending the Model by Stratification

- Is a unique feature of the Cox model
- Adjust for non-modeled factors
- Factors too difficult to model or fail PH assumption
- Commonly used in RCTs to adjust for site variation
- Allow form of λ to vary across strata
- Rank failure times *within* strata
- b strata, stratum ID is C

$$\begin{aligned} \lambda(t|X, C = j) &= \lambda_j(t)\exp(X\beta), \quad \text{or} \\ S(t|X, C = j) &= S_j(t)^{\exp(X\beta)}. \end{aligned}$$

- Not assume connection between shapes of λ_j
- By default, assume common β
- Ex: model age, stratify on sex
  Estimates common age slope pooling F and M
  No assumption about effect of sex except no age interact.
- Can stratify on multiple factors (cross-classify)
- Loss of efficiency not bad unless number of events in strata very small
- Stratum with no events is ignored

- Estimate β by getting separate log-likelihood for each stratum and adding up (independence)

- No inference about strat. factors

- Useful for checking PH and linearity assumptions: Model, then stratify on an X

- Can extend to strata $\times$ covariable interaction

$$\begin{aligned}\lambda(t|X_1, C=1) &= \lambda_1(t)\exp(\beta_1 X_1)\\ \lambda(t|X_1, C=2) &= \lambda_2(t)\exp(\beta_1 X_1 + \beta_2 X_1).\end{aligned}$$

$$\lambda(t|X_1, C=j) = \lambda_j(t)\exp(\beta_1 X_1 + \beta_2 X_2)$$

- X_2 is product interaction term (0 for F, X_1 for M)

- Are testing interaction with sex without modeling main effect!

19.2 Estimation of Survival Probability and Secondary Parameters

- Kalbfleisch-Prentice discrete hazard model method $\rightarrow$ K-M if $\hat{\beta} = 0$

- Breslow method $\rightarrow$ Nelson *et al.* if $\hat{\beta} = 0$

$$\hat{S}(t|X) = \hat{S}(t)^{\exp(X\hat{\beta})}.$$

- Stratified model $\rightarrow$ estimate underlying hazard parameters separately within strata

- "Adjusted K-M estimates"

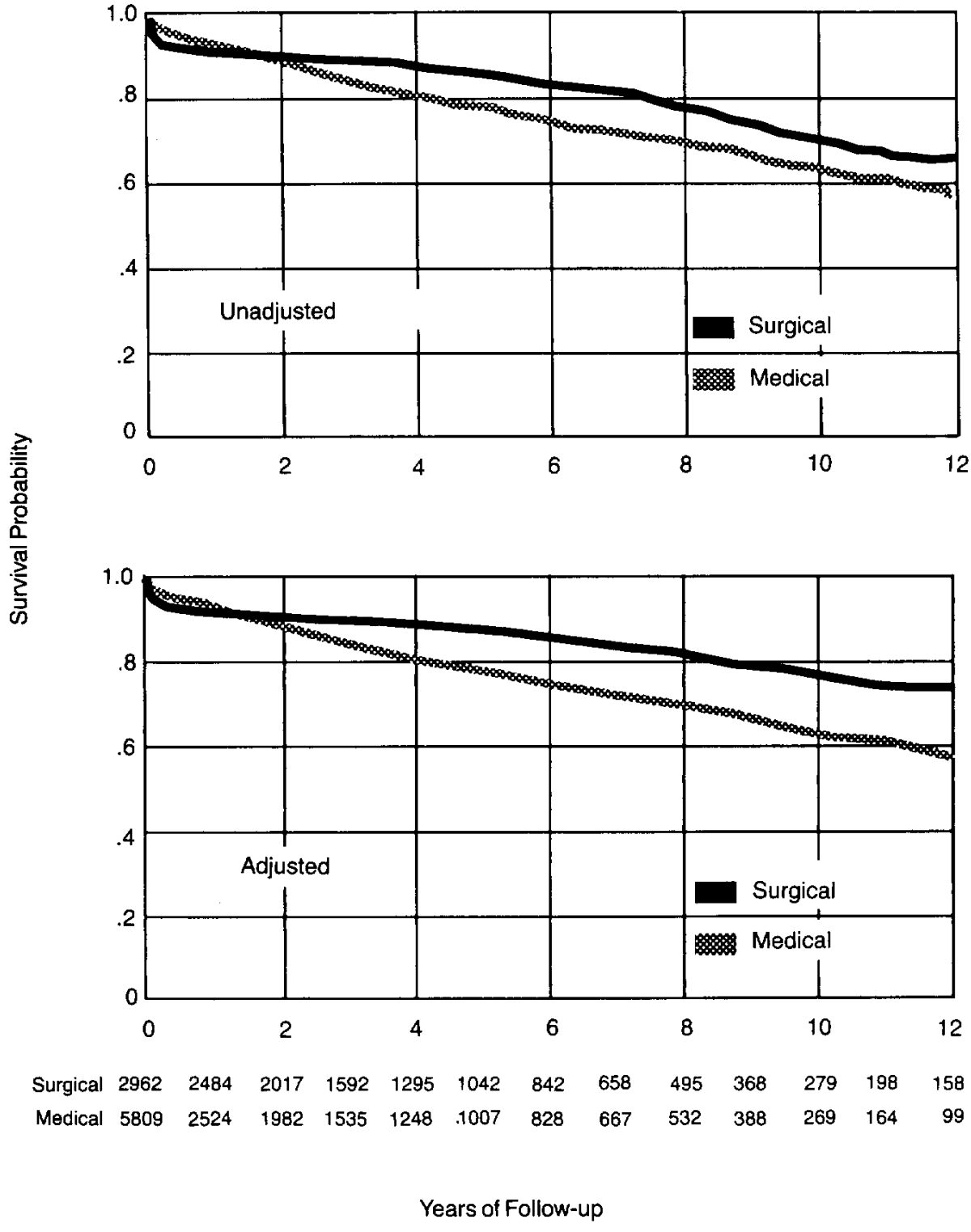

Figure 19.2: *Unadjusted (Kaplan-Meier) and adjusted (Cox-Kalbfleisch-Prentice) estimates of survival. Top, Kaplan-Meier estimates for patients treated medically and surgically at Duke University Medical Center from November 1969 through December 1984. These survival curves are not adjusted for baseline prognostic factors. Numbers of patients alive at each follow-up interval for each group are given at the bottom of the figure. Bottom, survival curves for patients treated medically or surgically after adjusting for all known important baseline prognostic characteristics . Reprinted by permission, American Medical Association.*

$$\hat{\Lambda}(t) = \sum_{i:t_i<t} \frac{d_i}{\sum_{Y_i \geq t_i} \exp(X_i \hat{\beta})}.$$

For any X, the estimates of Λ and S are

$$\begin{aligned} \hat{\Lambda}(t|X) &= \hat{\Lambda}(t) \exp(X\hat{\beta}) \\ \hat{S}(t|X) &= \exp[-\hat{\Lambda}(t) \exp(X\hat{\beta})]. \end{aligned}$$

19.3 Test Statistics

19.4 Residuals

Residual	Purposes
martingale	Assessing adequacy of a hypothesized predictor transformation; Graphing an estimate of a predictor transformation (Section 19.5.1)
score	Detecting overly influential observations
Schoenfeld	Testing PH assumption (Section 19.5.2) Graphing estimate of hazard ratio function (Section 19.5.2)

19.5 Assessment of Model Fit

19.5.1 Regression Assumptions

- Stratified KM estimates have problems
- 2000 simulated subject, $d = 368$, 1196 M, 804 F
- Exponential with known log hazard, linear in age, additive in sex

$$\lambda(t|X_1, X_2) = .02 \exp[.8X_1 + .04(X_2 - 50)],$$

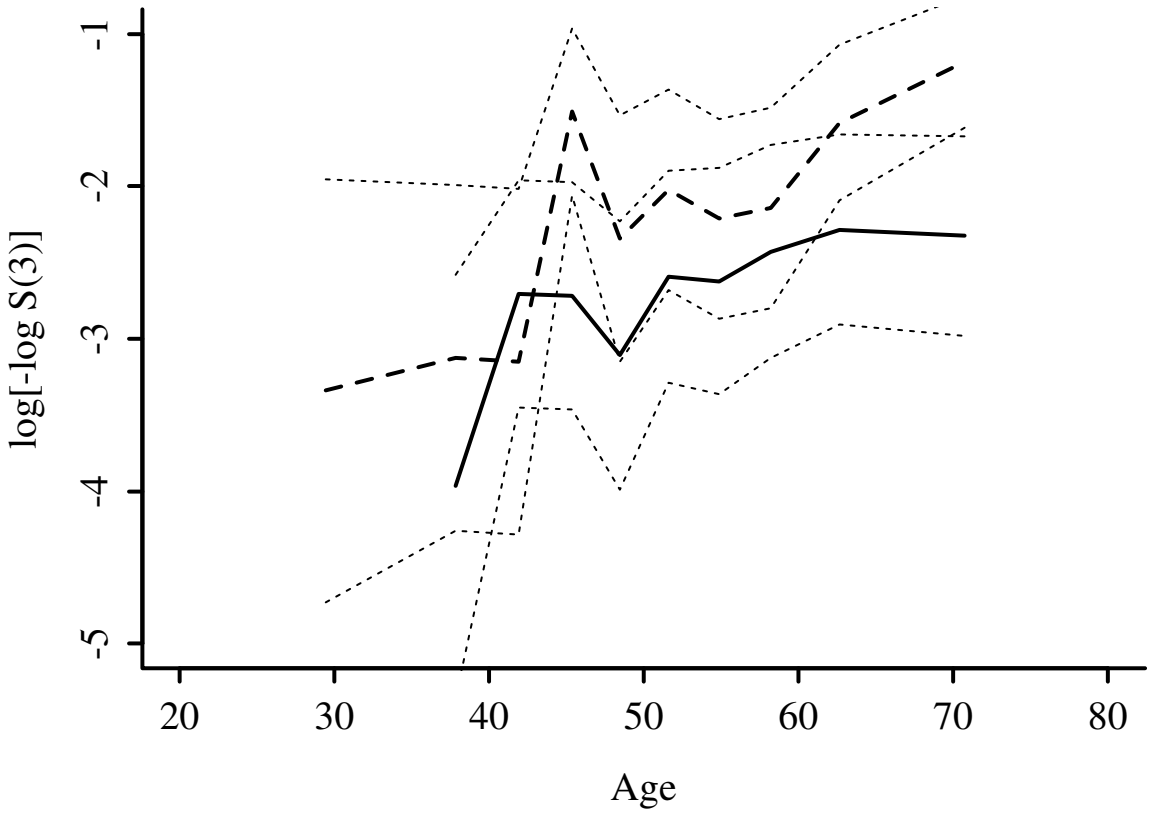

Figure 19.3: *Kaplan-Meier log Λ estimates by sex and deciles of age, with 0.95 confidence limits.*

Better: A 4-knot spline Cox PH model in two variables (X_1, X_2) which assumes linearity in X_1 and no $X_1 \times X_2$ interaction

$$\begin{aligned}\lambda(t|X) &= \lambda(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2' + \beta_4 X_2''), \\ &= \lambda(t) \exp(\beta_1 X_1 + f(X_2)),\end{aligned}$$

$$f(X_2) = \beta_2 X_2 + \beta_3 X_2' + \beta_4 X_2''.$$

$$\log \lambda(t|X) = \log \lambda(t) + \beta_1 X_1 + f(X_2).$$

To not assume PH in X_1, stratify on it:

$$\begin{aligned}\log \lambda(t|X_2, C = j) &= \log \lambda_j(t) + \beta_1 X_2 + \beta_2 X_2' + \beta_3 X_2'' \\ &= \log \lambda_j(t) + f(X_2).\end{aligned}$$

Formal test of linearity: $H_0 : \beta_2 = \beta_3 = 0, \chi^2 = 4.84$, 2 d.f., $P = 0.09$.

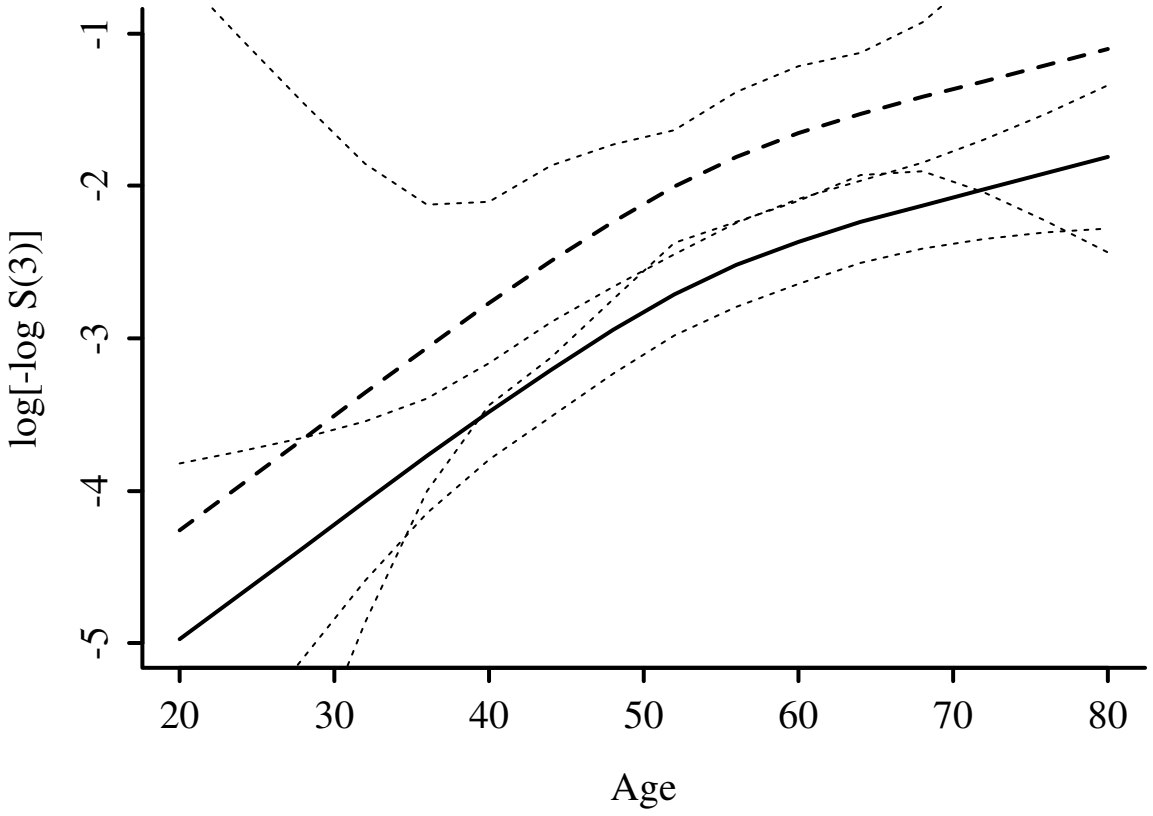

Figure 19.4: *Cox PH model stratified on sex, using spline function for age, no interaction. 0.95 confidence limits also shown.*

- Model allowing interaction with sex strata:

$$\begin{aligned}\log \lambda(t|X_2, C=j) &= \log \lambda_j(t) + \beta_1 X_2 \\ &+ \beta_2 X_2' + \beta_3 X_2'' \\ &+ \beta_4 X_1 X_2 + \beta_5 X_1 X_2' + \beta_6 X_1 X_2''.\end{aligned}$$

Test for interaction: $P = 0.33$.

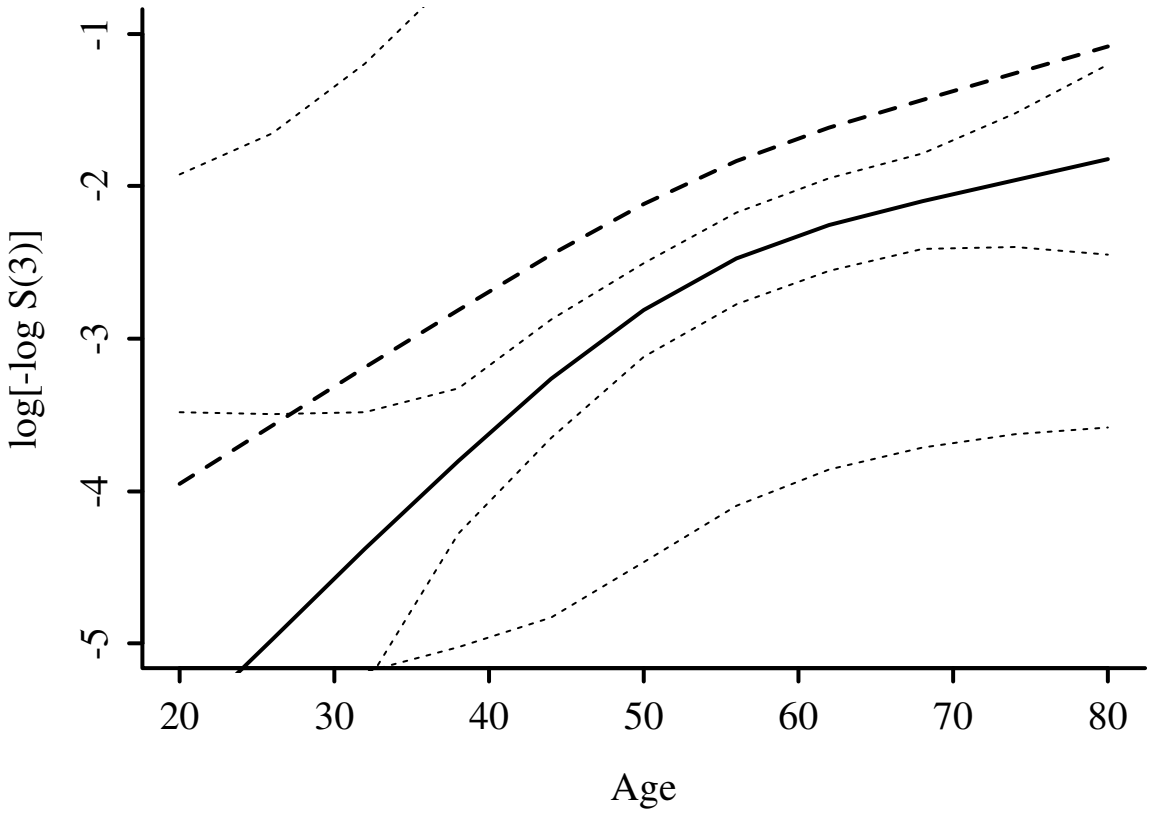

Figure 19.5: *Cox PH model stratified on sex, with interaction between age spline and sex. 0.95 confidence limits are also shown.*

- Example of modeling a single continuous variable (left ventricular ejection fraction), outcome = time to cardiovascular death

$$\begin{aligned}\text{LVEF}' &= \text{LVEF} \quad \text{if } \text{LVEF} \leq 0.5, \\ &= 0.5 \quad \text{if } \text{LVEF} > 0.5,\end{aligned}$$

The AICs for 3, 4, 5, and 6-knots spline fits were respectively 126, 124, 122, and 120. Smoothed residual plot: Martingale residuals, loess smoother

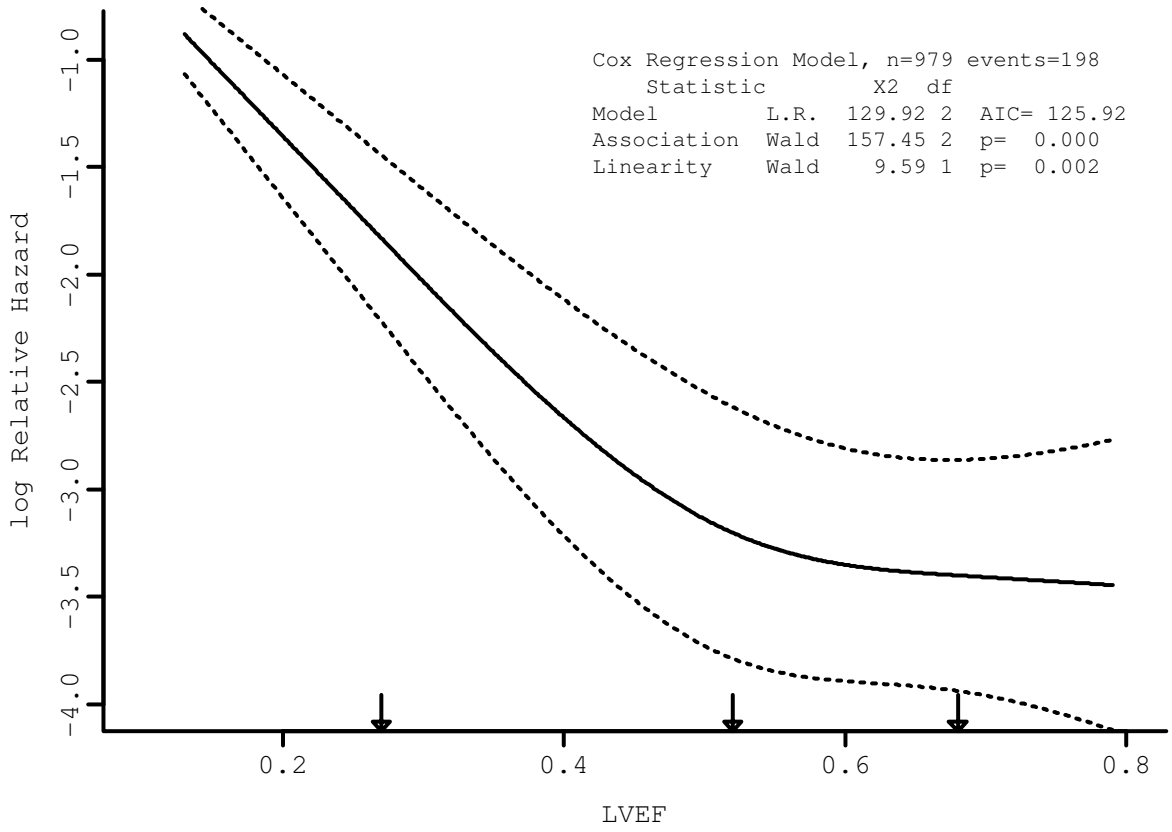

Figure 19.6: *Restricted cubic spline estimate of relationship between LVEF and relative log hazard from a sample of 979 patients and 198 cardiovascular deaths. Data from the Duke Cardiovascular Disease Databank.*

- One vector of residuals no matter how many covariables

- Unadjusted estimates of regression shape obtained by fixing $\hat{\beta} = 0$ for all Xs

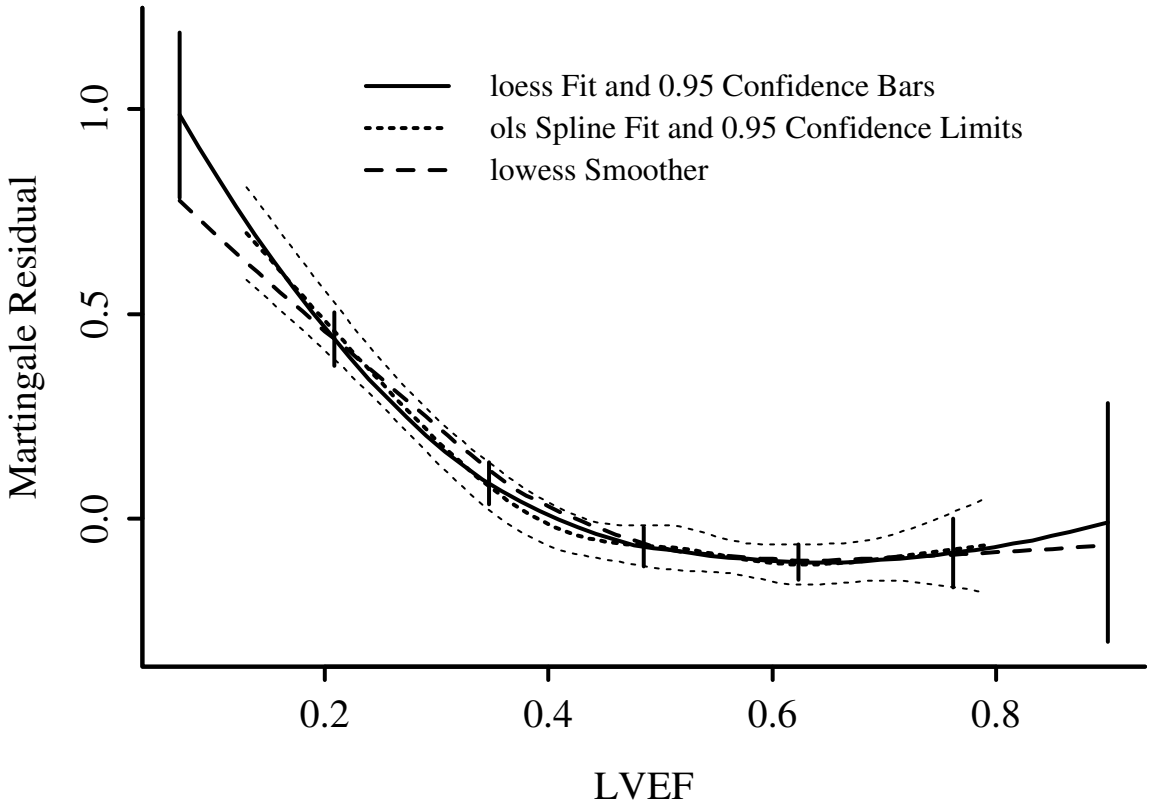

Figure 19.7: *Three smoothed estimates relating martingale residuals to LVEF.*

Purpose	Method
Estimate transformation for a single variable	Force $\hat{\beta}_1 = 0$ and compute residuals off of the null regression
Check linearity assumption for a single variable	Compute $\hat{\beta}_1$ and compute residuals off of the linear regression
Estimate marginal transformations for p variables	Force $\hat{\beta}_1, \ldots, \hat{\beta}_p = 0$ and compute residuals off the global null model
Estimate transformation for variable i adjusted for other $p-1$ variables	Estimate $p-1$ βs, forcing $\hat{\beta}_i = 0$ Compute residuals off of mixed global/null model

19.5.2 Proportional Hazards Assumption

- Parallelism of $\log \Lambda$ plots

- Comparison of stratified and modeled estimates of $S(t)$

- Plot actual ratio of estimated Λ, or get differences in $\log \Lambda$

- Plot $\hat{\Lambda}$ vs. cumulative number of events as $t \uparrow$

- Stratify time, get interval-specific Cox regression coefficients:
  In an interval, exclude all subjects with
  event/censoring time before start of interval
  Censor all events at end of interval

Time Interval	Observations	Deaths	Log Hazard Ratio	Standard Error
$[0, 209)$	40	12	-0.47	0.59
$[209, 234)$	27	12	-0.72	0.58
$234+$	14	12	-0.50	0.64

  Overall Cox $\hat{\beta} = -0.57$.

- VA Lung Cancer dataset, squamous vs. (small, adeno)

Time Interval	Observations	Deaths	Log Hazard Ratio	Standard Error
$[0, 21)$	110	26	-0.46	0.47
$[21, 52)$	84	26	-0.90	0.50
$[52, 118)$	59	26	-1.35	0.50
$118+$	28	26	-1.04	0.45

  Estimates for Karnofsky performance status weight over time:

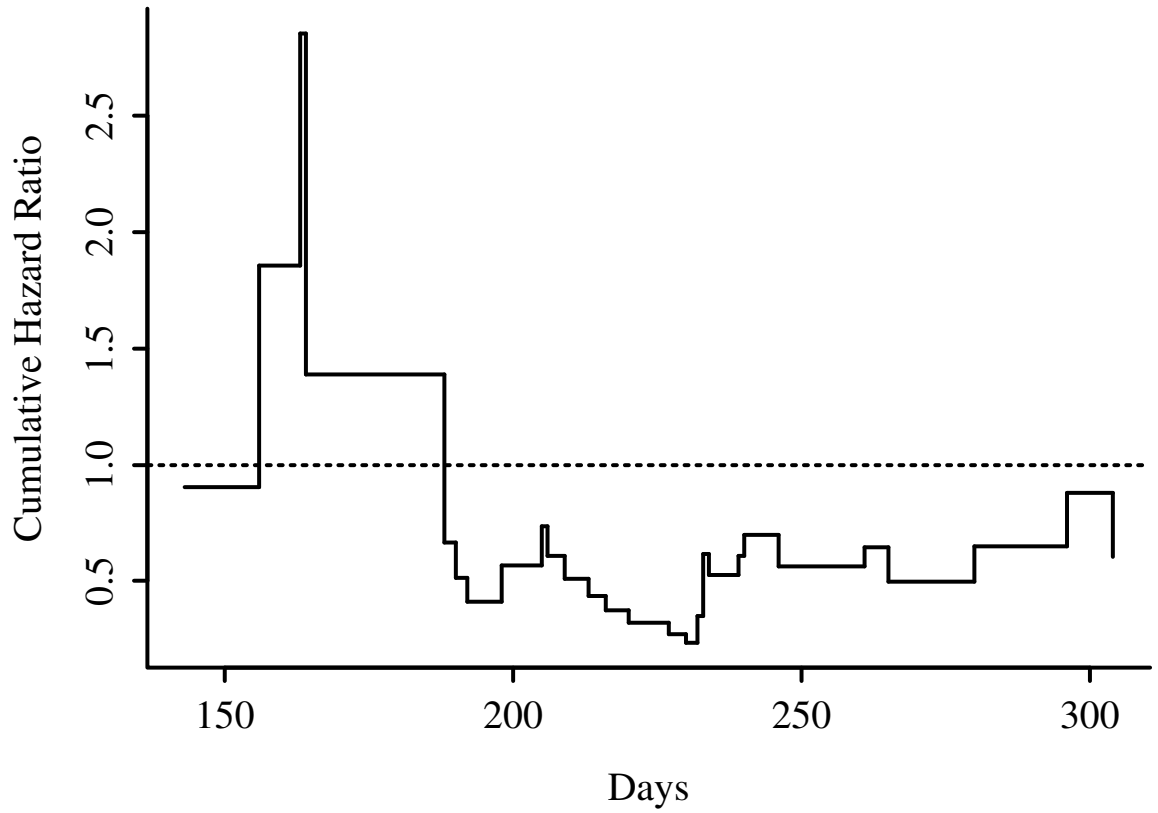

Figure 19.8: *Estimate of $\frac{\Lambda_2}{\Lambda_1}$ based on $-\log$ of Altschuler-Nelson-Fleming-Harrington nonparametric survival estimates*

Time Interval	Observations	Deaths	Log Hazard Ratio	Standard Error
$[0, 19]$	137	27	-0.053	0.010
$[19, 49)$	112	26	-0.047	0.009
$[49, 99]$	85	27	-0.036	0.012
$99+$	28	26	-0.012	0.014

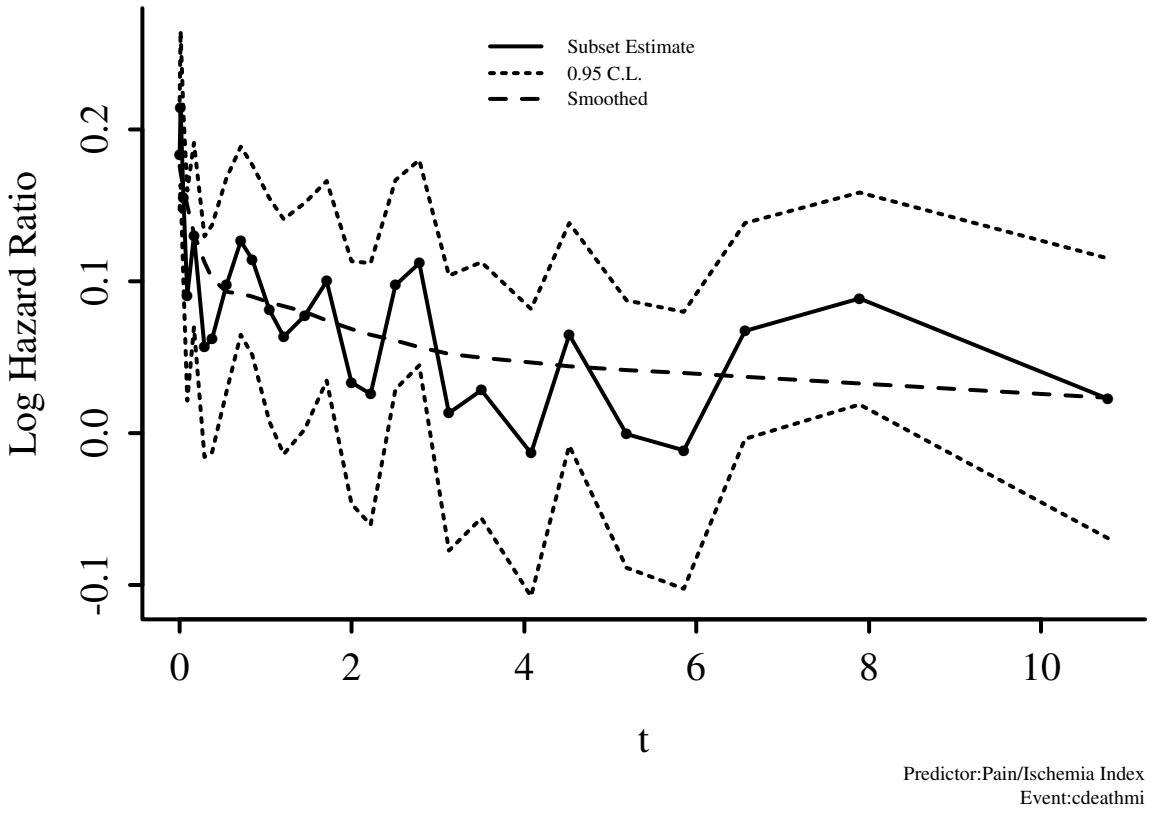

Figure 19.9: *Stratified hazard ratios for pain/ischemia index over time. Data from the Duke Cardiovascular Disease Databank.*

- Schoenfeld residuals computed at each unique failure time
- Partial derivative of $\log L$ with respect to each X in turn
- Grambsch and Therneau scale to yield estimates of $\beta(t)$

- Can form a powerful test of PH (Z:PH in old SAS PROC PHGLM)

$$\hat{\beta} + dR\hat{V},$$

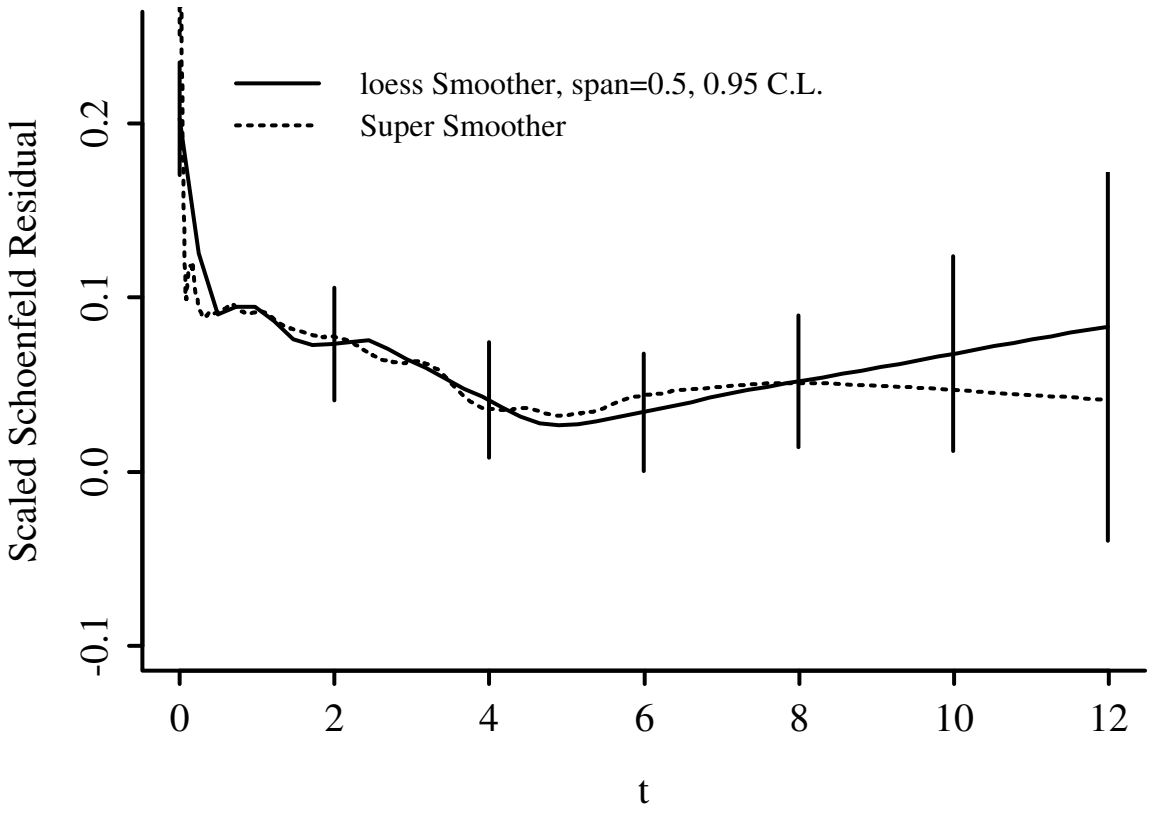

Figure 19.10: *Smoothed weighted Schoenfeld residuals for the same data in Figure 19.9. Test for PH based on the correlation (ρ) between the individual weighted Schoenfeld residuals and the rank of failure time yielded* $\rho = -0.23, z = -6.73, P = 2 \times 10^{-11}$.

- Can test PH by testing $t \times X$ interaction using time- dependent covariables

- Separate parametric fits, e.g. Weibull with differing γ; hazard ratio is

$$\frac{\alpha\gamma t^{\gamma-1}}{\delta\theta t^{\theta-1}} = \frac{\alpha\gamma}{\delta\theta} t^{\gamma-\theta}.$$

t	log Hazard Ratio
10	-0.36
36	-0.64
83.5	-0.83
200	-1.02

- Interaction between X and spline function of t:

$$\log \lambda(t|X) = \log \lambda(t) + \beta_1 X + \beta_2 Xt + \beta_3 Xt' + \beta_4 Xt'',$$

The $X + 1 : X$ log hazard ratio function is estimated by

$$\hat{\beta}_1 + \hat{\beta}_2 t + \hat{\beta}_3 t' + \hat{\beta}_4 t''.$$

Assumptions of the Proportional Hazards Model

$$\lambda(t|X) = \lambda(t)e^{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p}$$

Variables	Assumptions	Verification
Response Variable T Time Until Event	Shape of $\lambda(t\|X)$ for fixed X as $t \uparrow$ Cox: none Weibull: t^θ	Shape of $S_{\mathrm{KM}}(t)$
Interaction between X and T	Proportional hazards – effect of X does not depend on T. E.g. treatment effect is constant over time.	• Categorical X: check parallelism of stratified $\log[-\log S(t)]$ plots as $t \uparrow$ • Muenz cum. hazard ratio plots • Arjas cum. hazard plots • Check agreement of stratified and modeled estimates • Hazard ratio plots • Smoothed Schoenfeld residual plots and correlation test (time vs. residual) • Test time-dependent covariable such as $X \times \log(t+1)$ • Ratio of parametrically estimated $\lambda(t)$
Individual Predictors X	Shape of $\lambda(t\|X)$ for fixed t as $X \uparrow$ Linear: $\log \lambda(t\|X) = \log \lambda(t) + \beta X$ Nonlinear: $\log \lambda(t\|X) = \log \lambda(t) + f(X)$	• k-level ordinal X : linear term + $k-2$ dummy variables • Continuous X: Polynomials, spline functions, smoothed martingale residual plots
Interaction between X_1 and X_2	Additive effects: effect of X_1 on $\log \lambda$ is independent of X_2 and vice-versa	Test non-additive terms, e.g. products

19.6 What to Do When PH Fails

- Test of association not needed → stratify

Method	Requires Grouping X	Requires Grouping t	Computational Efficiency	Yields Formal Test	Yields Estimate of $\lambda_2(t)/\lambda_1(t)$	Requires Fitting 2 Models	Must Choose Smoothing Parameter
$\log[-\log]$, Muenz, Arjas plots	x		x			x	
Dabrowska $\log\hat{\Lambda}$ difference plots	x		x	x		x	
Stratified vs. Modeled Estimates	x		x			x	
Hazard ratio plot		x		?	x	x	?
Schoenfeld residual plot			x		x		x
Schoenfeld residual correlation test			x	x			
Fit time-dependent covariables				x	x		
Ratio of parametric estimates of $\lambda(t)$	x		x	x	x	x	

- P-value for testing variable may still be useful (conservative)
- Survival estimates wrong in certain time intervals
- Can model non-PH:

$$\lambda(t|X) = \lambda_0(t)\exp(\beta_1 X + \beta_2 X \times \log(t+1))$$

- Can also use time intervals:

$$\lambda(t|X) = \lambda_0(t)\exp[\beta_1 X + \beta_2 X \times I(t > c)],$$

- Or fit one model for early follow-up, one for late
- Try another model, e.g. log-normal, log-logistic can have effects of X changing constantly over time
- Differences in mean restricted life length can be useful in comparing therapies when PH fails

19.7 Collinearity

19.8 Overly Influential Observations

19.9 Quantifying Predictive Ability

- $$\begin{aligned} R^2_{\mathrm{LR}} &= 1 - \exp(-\mathrm{LR}/n) \\ &= 1 - \omega^{2/n}, \end{aligned}$$

  where ω is the null model likelihood divided by the fitted model likelihood. Divide by max attainable value to get R^2_{N}.

- $c =$ concordance probability (predicted vs. observed)

- All possible pairs of subjects whose ordering of failure times can be determined

- Fraction of these for which X ordered same as Y

- Somers' $D_{xy} = 2(c - 0.5)$

19.10 Validating the Fitted Model

Separate bootstrap validations for calibration and for discrimination.

19.10.1 Validation of Model Calibration

- Calibration at fixed t

- Get $\hat{S}(t|X)$ for all subjects
- Divide into intervals each containing say 50 subjects
- Compare mean predicted survival with K-M
- Bootstrap this process to add back optimism in difference of these 2, due to overfitting
- Ex: 20 random predictors, $n = 200$

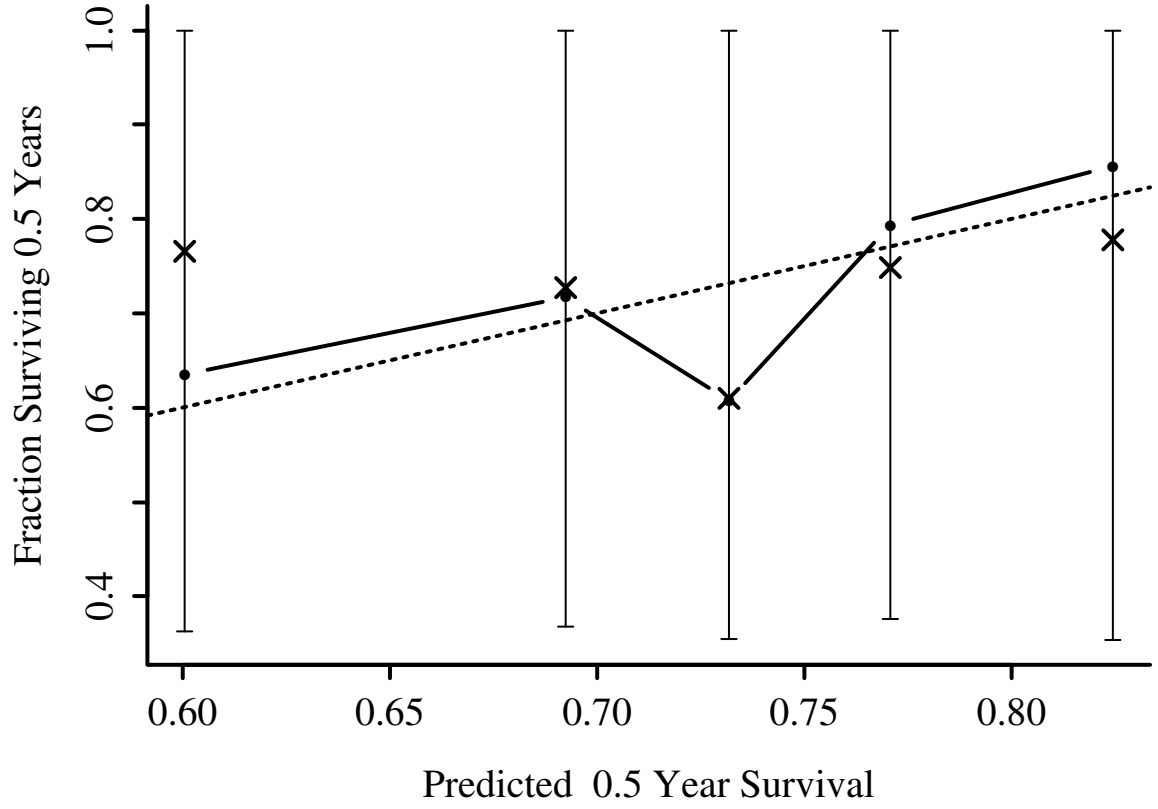

Figure 19.11: *Calibration of random predictions using Efron's bootstrap with B=50 re-samples and 40 patients per interval. Dataset has n=200, 100 uncensored observations, 20 random predictors,* $\chi^2_{20} = 9.87$. *•: apparent calibration; X: bias-corrected calibration.*

19.10.2 Validation of Discrimination and Other Statistical Indexes

Validate slope calibration (estimate shrinkage from overfitting):

$$\lambda(t|X) = \lambda(t)\exp(\gamma Xb).$$

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index
D_{xy}	-0.16	-0.31	-0.09	-0.22	0.06
R^2_{N}	0.05	0.15	0.00	0.15	-0.10
Slope	1.00	1.00	0.25	0.75	0.25
D	0.01	0.04	0.00	0.04	-0.02
U	0.00	0.00	0.00	0.00	0.00
Q	0.01	0.04	0.00	0.04	-0.02

19.11 Describing the Fitted Model

- Can use coefficients if linear and additive
- In general, use e.g. inter-quartile-range hazard ratios for various levels of interacting factors
- Nomogram to compute $X\hat{\beta}$
- Also $\hat{S}(t|X)$ for a few t
- Can have axis for median failure time if sample is high risk

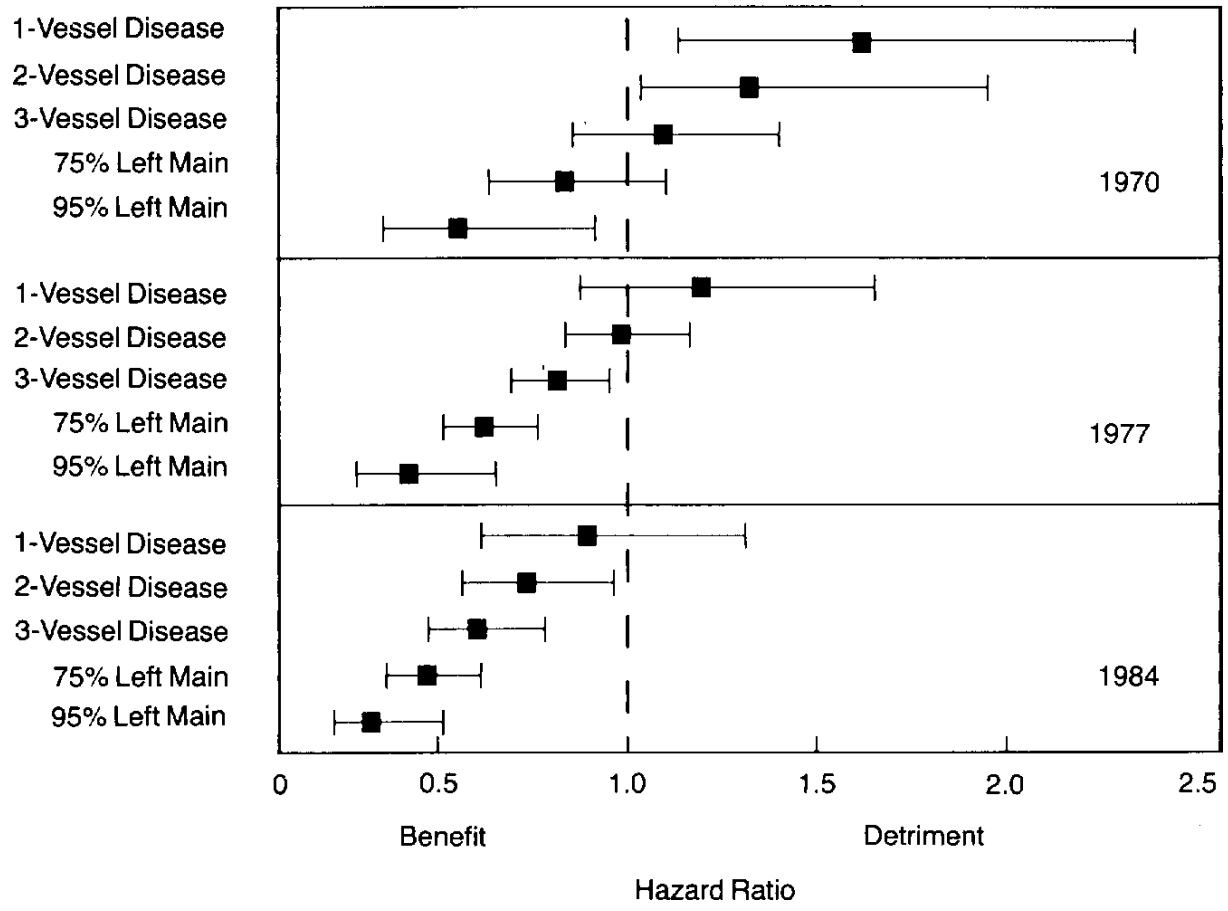

Figure 19.12: *A display of an interaction between treatment, extent of disease, and calendar year of start of treatment. Comparison of medical and surgical average hazard ratios for patients treated in 1970, 1977, and 1984 according to coronary anatomy. Closed squares represent point estimates; bars represent 0.95 confidence limits of average hazard ratios . Reprinted by permission, American Medical Association.*

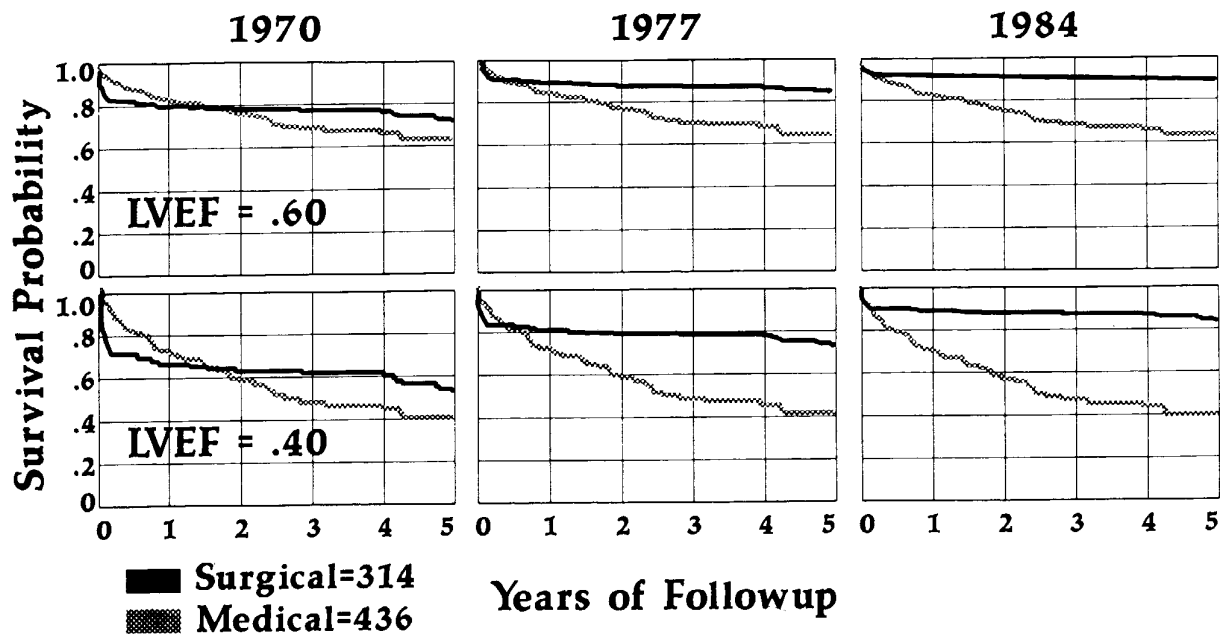

Figure 19.13: *Cox-Kalbfleisch-Prentice survival estimates stratifying on treatment and adjusting for several predictors. Estimates are for patients with left main disease and normal or impaired ventricular function . Reprinted by permission, Mosby, Inc. / Harcourt Health Sciences.*

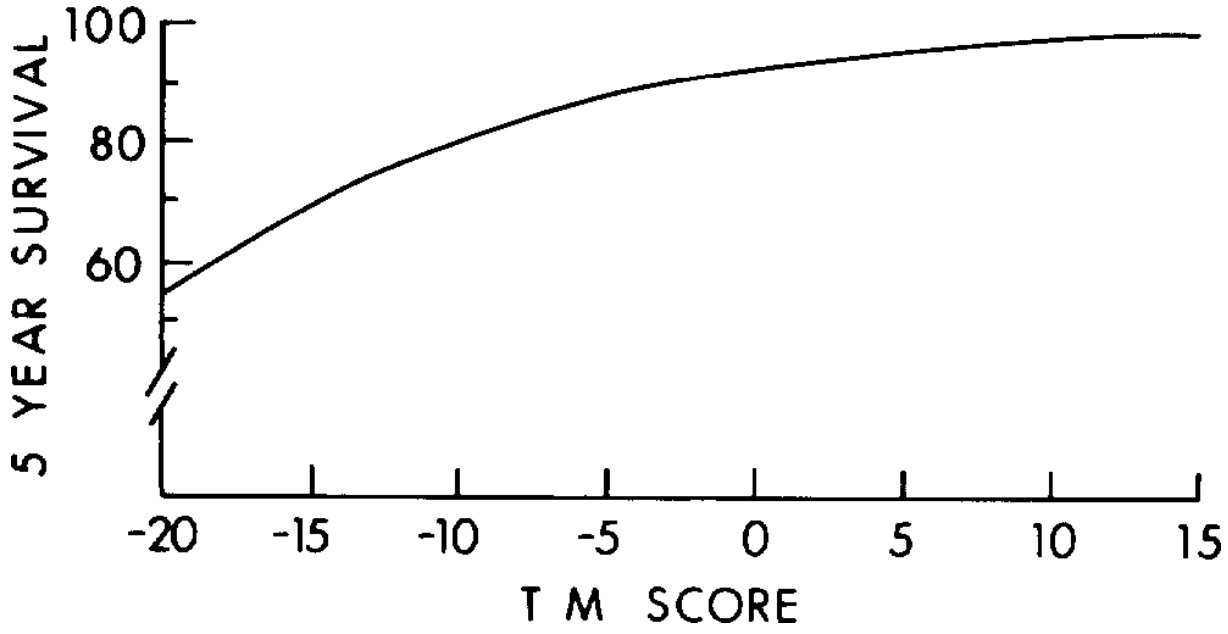

Figure 19.14: *Cox model predictions with respect to a continuous variable. X-axis shows the range of the treadmill score seen in clinical practice and Y-axis shows the corresponding 5-year survival probability predicted by the Cox regression model for the 2842 study patients . Reprinted by permission, American College of Physicians—American Society of Internal Medicine.*

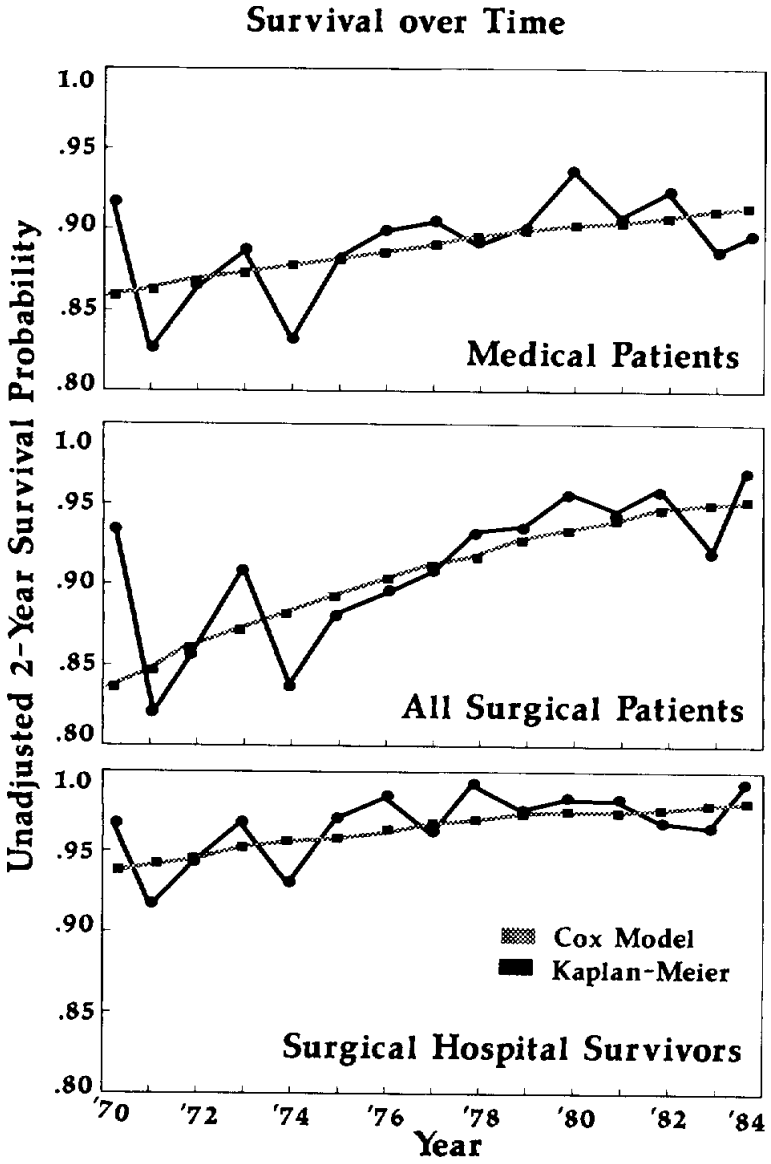

Figure 19.15: *Modeled and stratified nonparametric survival estimates. Kaplan-Meier observed 2 year survival (solid line) and estimated Cox model trends (dashed line) for each year of entry into the study for all medical patients, surgical patients, and surgical survivors . Reprinted by permission, Mosby, Inc. / Harcourt Health Sciences.*

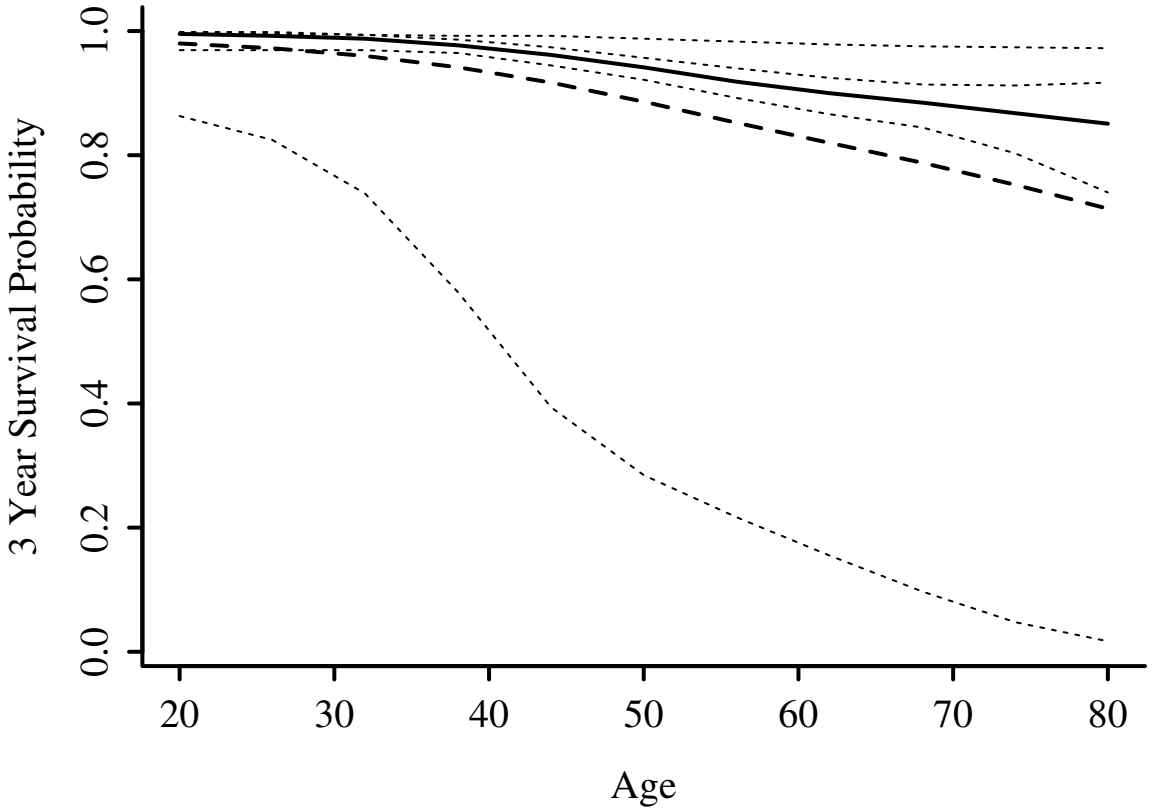

Figure 19.16: *Survival estimates for model stratified on sex, with interaction*

19.12 S-PLUS Functions

19.12.1 Power and Sample Size Calculations, Hmisc Library

- `cpower`: computes power for a two-sample Cox test with random patient entry over a fixed duration and a given length of minimum follow-up, using exponential distribution with handling of dropout and drop-in

- `ciapower`: computes power of the Cox interaction test in a 2×2 setup using the method of Peterson and George

- `spower`: simulates power for 2-sample tests (the log-rank test by default) allowing for very complex conditions such as continuously varying treatment effect and non-compliance probabilities.

19.12.2 Cox Model using Design Library

- `cph`: slight modification of Therneau's `survival` library `coxph` function

- `print` method prints the Nagelkerke index R^2_{N} (Section 19.9)

- `cph` works with generic functions such as `specs, predict, summary, anova, fastbw, which.influence, latex, residuals, coef, nomogram`, and `plot`,

- `plot` has an additional argument `time` for plotting `cph` fits. It also has an argument `loglog` which if `T` causes instead log -log survival to be plotted on the y-axis.

- `Survival.cph, Quantile.cph, Mean.cph` create other S functions to evaluate survival probabilities, survival time quantiles, and mean and mean restricted lifetimes, based on a `cph` fit with `surv=T`

- `Quantile` and `Mean` are especially useful with `plot` and `nomogram`. `Survival` is useful with `nomogram`

```
f ← cph(..., surv=T)
med ← Quantile(f)
nomogram(f, fun=function(x) med(lp=x),
         funlabel='Median Survival Time')
# fun tranforms the linear predictors
srv ← Survival(f)
rmean ← Mean(f, tmax=3, method='approx')
nomogram(f, fun=list(function(x) srv(3, x), rmean),
         funlabel=c('3-Year Survival Prob.','Restricted Mean'))
# med, srv, expected are more complicated if strata are present
```

Figures 19.3, 19.4, 19.5 and 19.16 were produced by

```
n ← 2000
.Random.seed ← c(49,39,17,36,23,0,43,51,6,54,50,1)
# to be able to re-generate same data
age ← 50 + 12*rnorm(n)
label(age) ← "Age"
sex <- sample(c('Male','Female'), n, rep=T, prob=c(.6, .4))
cens ← 15*runif(n)
h ← .02*exp(.04*(age-50)+.8*(sex=='Female'))
t ← -log(runif(n))/h
e ← ifelse(t<=cens,1,0)
t ← pmin(t, cens)
units(t) ← "Year"
age.dec ← cut2(age, g=10, levels.mean=T)
Srv ← Surv(t,e)
f ← cph(Srv ~ strat(age.dec)+strat(sex), surv=T)
# surv=T speeds up computations, and confidence limits
# when there are no covariables are still accurate.
plot(f, age.dec=NA, sex=NA, time=3, loglog=T,
     val.lev=T, ylim=c(-5,-1))

f ← cph(Srv ~ rcs(age,4)+strat(sex), x=T, y=T)
# Get accurate C.L. for any age
# Note: for evaluating shape of regression, we would not
#  ordinarily bother to get 3-year survival probabilities -
#  would just use X * beta. We do so here to use same scale
#  as nonparametric estimates
f
anova(f)
ages ← seq(20, 80, by=4)
# Evaluate at fewer points. Default is 100
# Take much RAM if we use the exact C.L. formula with n=100
plot(f, age=ages, sex=NA, time=3, loglog=T, ylim=c(-5,-1))

f ← cph(Srv ~ rcs(age,4)*strat(sex), x=T, y=T)
anova(f)
ages ← seq(20, 80, by=6)
# Still fewer points - more parameters in model
plot(f, age=ages, sex=NA, time=3, loglog=T, ylim=c(-5,-1))
plot(f, age=ages, sex=NA, time=3)
```

```
# Having x=T, y=T in fit also allows computation of
#  influence statistics
resid(f, "dfbetas")
which.influence(f)
```

The S-PLUS program below demonstrates how several `cph`-related functions work well with the `nomogram` function to display this last fit. Here predicted 3-year survival probabilities and median survival time (when defined) are displayed against age and sex. The fact that a nonlinear effect interacts with a stratified factor is taken into account.

```
srv ← Survival(f)  # use an f that used surv=T
# Define functions to compute 3-year estimates as a function
#  of the linear predictors (X*Beta)
surv.f ← function(lp) srv(3, lp, stratum="sex=Female")
surv.m ← function(lp) srv(3, lp, stratum="sex=Male")
quant ← Quantile(f)
# Define functions to compute median survival time
med.f ← function(lp) quant(.5, lp, stratum="sex=Female")
med.m ← function(lp) quant(.5, lp, stratum="sex=Male")
nomogram(f, fun=list(surv.m, surv.f, med.m, med.f),
         funlabel=c("S(3 | Male)","S(3 | Female)",
                    "Median (Male)","Median (Female)"),
         fun.at=list(c(.8,.9,.95,.98,.99),
                     c(.1,.3,.5,.7,.8,.9,.95,.98),
                     c(8,12),c(1,2,4,8,12)))
```

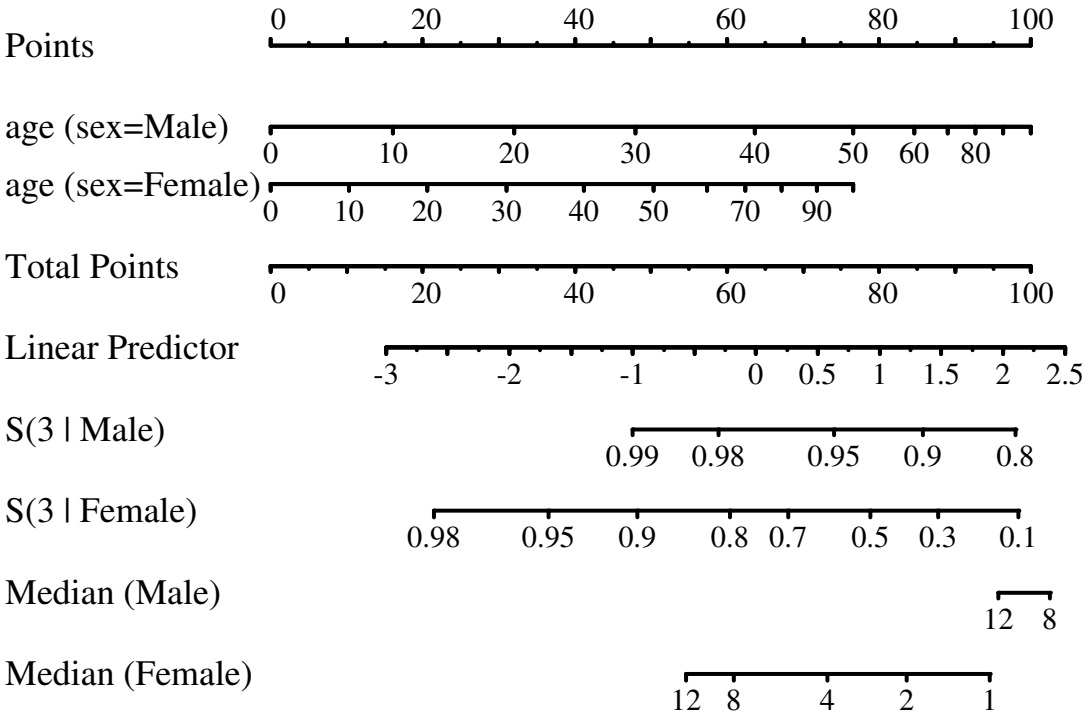

Figure 19.17: *Nomogram from a fitted stratified Cox model that allowed for interaction between age and sex, and nonlinearity in age. The axis for median survival time is truncated on the left where the median is beyond the last follow-up time.*

Chapter 20

Modeling Longitudinal Responses using Generalized Least Squares

20.1 Notation

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

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

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

20.2.1 Common Basis Functions

- k dummy variables for $k+1$ unique times (assumes no functional form for time but may spend many d.f.)

- $k = 1$ for linear time trend, $g_1(t) = t$

- k–order polynomial in t

- $k+1$–knot restricted cubic spline (one linear term, $k-1$ nonlinear terms)

20.2.2 Model for Mean Profile

- A model for mean time-response profile without interactions between time and any X:
$E[Y_{it}|X_i] = X_i\beta + \gamma_1 g_1(t) + \gamma_2 g_2(t) + \ldots + \gamma_k g_k(t)$

- Model with interactions between time and some X's: add product terms for desired interaction effects

- Example: To allow the mean time trend for subjects in group 1 (reference group) to be arbitrarily different from time trend for subjects in group 2, have a dummy variable for group 2, a time "main effect" curve with k d.f. and all k products of these time components with the dummy variable for group 2

20.2.3 Model Specification for Treatment Comparisons

- In studies comparing two or more treatments, a response is often measured at baseline (pre-randomization)

- Analyst has the option to use this measurement as Y_{i0} or as part of X_i

- Jim Rochon (Dept. of Biostatistics & Bioinformatics, Duke University) has the following comments about this:

> For RCTs, I draw a sharp line at the point when the intervention begins. The LHS 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.

20.3 Modeling Within-Subject Dependence

- Random effects and mixed effects models have become very popular

- Disadvantages:
  - Induced correlation structure for Y may be unrealistic
  - Numerically demanding

  - 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)

- Pinheiro and Bates (Section 5.1.2) state that "in some applications, one may wish to avoid incorporating random effects in the model to account for dependence among observations, choosing to use the within-group component Λ_i to directly model variance-covariance structure of the response."

- We will assume that $Y_{it}|X_i$ has a multivariate normal distribution with mean given above and with variance-covariance matrix V_i, an $n_i \times n_i$ matrix that is a function of $t_{i1}, \ldots, t_{in_i}$

- We further assume that the diagonals of V_i are all equal

- Procedure can be generalized to allow for heteroscedasticity over time or with respect to X (e.g., males may be allowed to have a different variance than females)

- This *extended linear model* has the following assumptions:
  - all the assumptions of OLS at a single time point including correct modeling of predictor effects and univariate normality of responses conditional on X
  - the distribution of two responses at two different times for the same subject, conditional on X, is bivariate normal with a specified correlation coefficient
  - the joint distribution of all n_i responses for the i^{th} subject is multivariate normal with the given correlation pattern (which implies the previous two distributional assumptions)

  - responses from any times for any two different subjects are uncorrelated

20.4 Parameter Estimation Procedure

- Generalized least squares

- Like weighted least squares but uses a covariance matrix that is not diagonal

- Each subject can have her own shape of V_i due to each subject being measured at a different set of times

- Maximum likelihood

- Newton-Raphson or other trial-and-error methods used for estimating parameters

- For small number of subjects, advantages in using REML (restricted maximum likelihood) instead of ordinary MLE , (esp. to get more unbiased estimate of the covariance matrix)

- When imbalances are not severe, OLS fitted ignoring subject identifiers may be efficient
  - But OLS standard errors will be too small as they don't take intra-cluster correlation into account

  - May be rectified by substituting covariance matrix estimated from Huber-White cluster sandwich estimator or from cluster bootstrap

- When imbalances are severe and intra-subject correlations are strong, OLS is not expected to be efficient because it gives equal weight to each observation

  - a subject contributing two distant observations receives $\frac{1}{5}$ the weight of a subject having 10 tightly-spaced observations

20.5 Common Correlation Structures

- Usually restrict ourselves to *isotropic* correlation structures — correlation between responses within subject at two times depends only on a measure of distance between the two times, not the individual times

- We simplify further and assume depends on $|t_1 - t_2|$

- Can speak interchangeably of correlations of residuals within subjects or correlations between responses measured at different times on the same subject, conditional on covariates X

- Assume that the correlation coefficient for Y_{it_1} vs. Y_{it_2} conditional on baseline covariates X_i for subject i is $h(|t_1 - t_2|, \rho)$, where ρ is a vector (usually a scalar) set of fundamental correlation parameters

- Some commonly used structures when times are continuous and are not equally spaced :

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

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

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

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

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

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

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

The last 5 structures use ρ as a scaling parameter, not as something restricted to be in $[0, 1]$

20.6 Checking Model Fit

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

20.7 S Software

- Nonlinear mixed effects model library of Pinheiro & Bates in S-PLUS and R
- For linear models, fitting functions are
  - `lme` for mixed effects models
  - `gls` for generalized least squares without random effects
- R has a new version of `gls`

- For this version the Design library has `glsD` so that many features of Design can be used:

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

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

  `plot` : continuous effect plots

  `nomogram` : nomogram

  `Function` : generate S function code for fitted model

  `latex` : LaTeX representation of fitted model

  In addition, `glsD` has a bootstrap option (hence you do not use Design's `bootcov` for `glsD` 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

20.8 Case Study

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

- Randomized to placebo ($N = 36$), 5000 units of BotB ($N = 36$), 10,000 units of BotB ($N = 37$)

- Response variable: total score on Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS), measuring severity, pain, and disability of cervical dystonia (high scores mean more impairment)

- TWSTRS measured at baseline (week 0) and weeks 2, 4, 8, 12, 16 after treatment began

- Dataset `cdystonia` from web site

20.8.1 Graphical Exploration of Data

```
library(Hmisc)
getHdata(cdystonia)
# Or load('cdystonia.sav') if using R, data.restore('cdystonia.sdd') if using S-Plus
attach(cdystonia)

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

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

            0         0 2 4   0 2 4 12 16       0 2 4 8    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
xYplot(twstrs ~ week | site*treat, groups=uid,
       type='b', label.curves=FALSE)                                  # Fig. 20.1

# Show quartiles
xYplot(twstrs ~ week | treat, method='quantile', nx=0)      # Fig. 20.2

# Show means with bootstrap nonparametric CLs
xYplot(twstrs ~ week | treat, method=smean.cl.boot, nx=0)  # Fig. 20.3
```

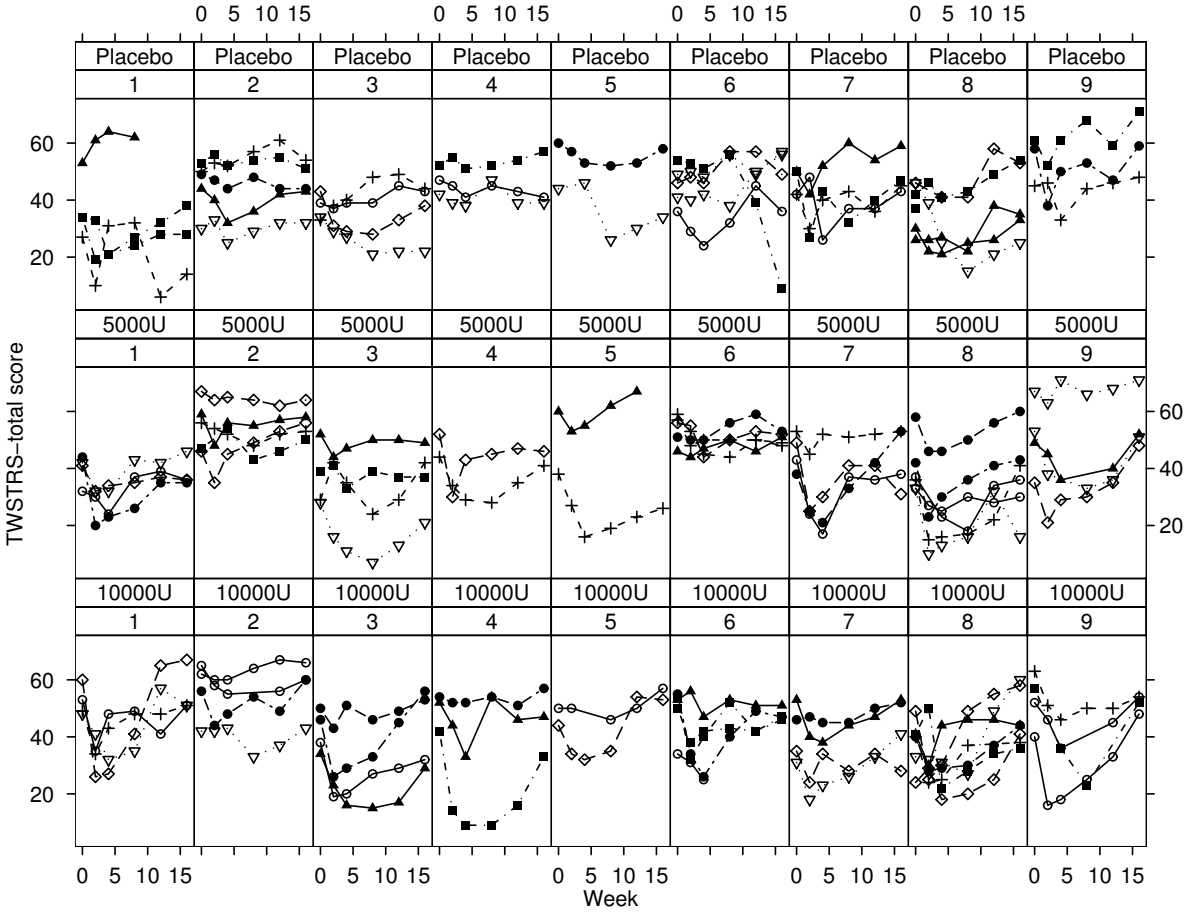

Figure 20.1: *Time profiles for individual subjects, stratified by study site and dose*

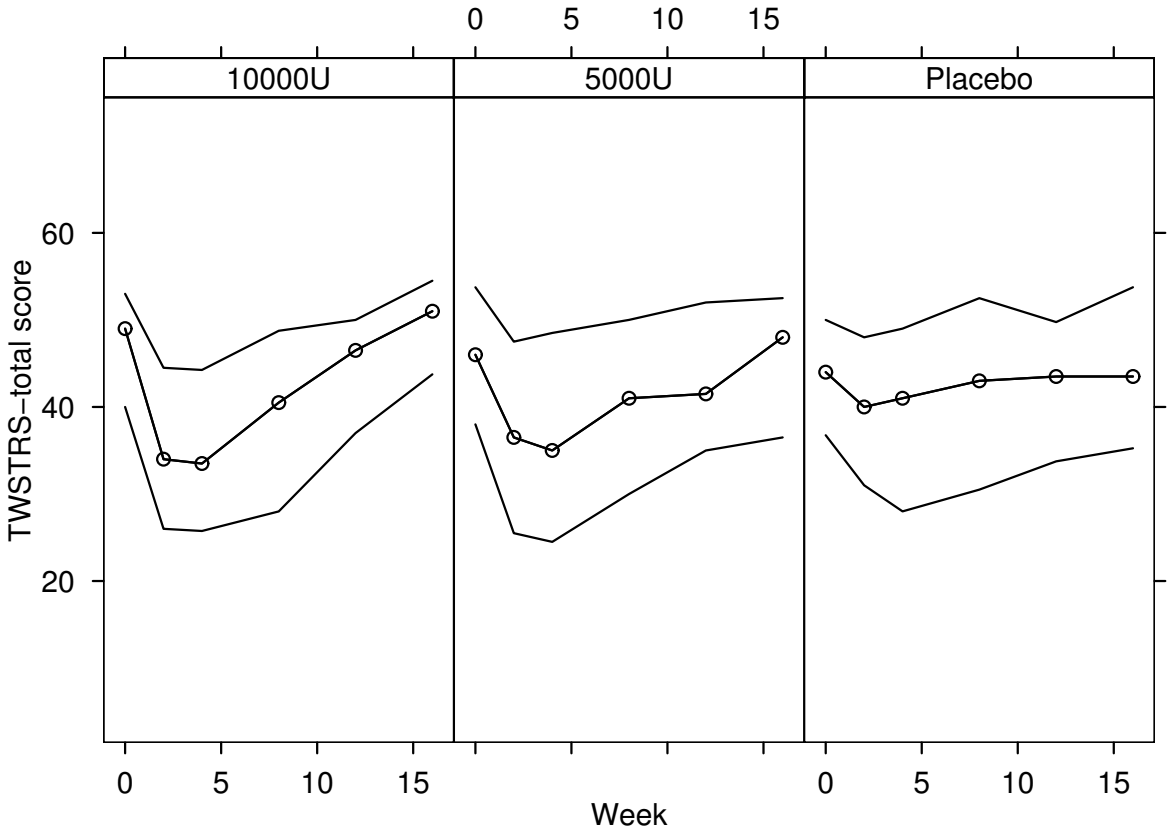

Figure 20.2: *Quartiles of* `TWSTRS` *stratified by dose*

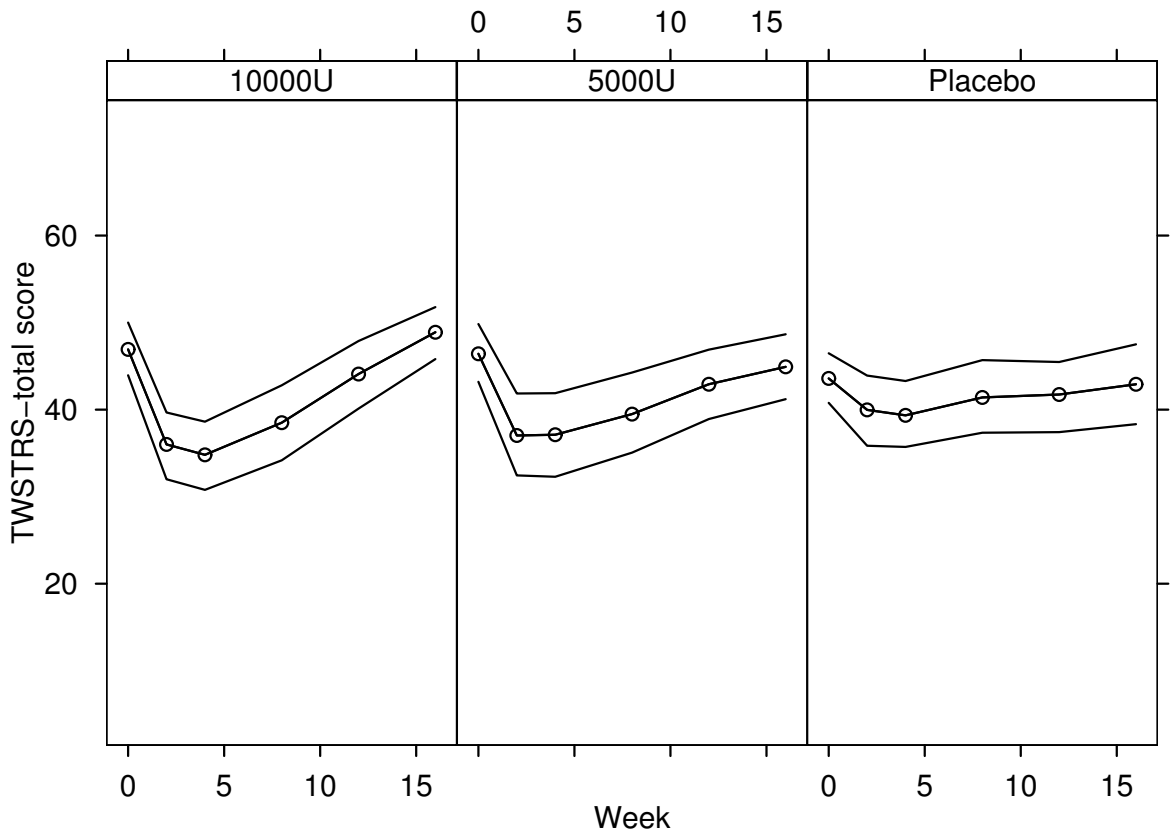

Figure 20.3: *Mean responses and nonparametric bootstrap 0.95 confidence limits for population means, stratified by dose*

20.8.2 Using OLS and Correcting Variances for Intra-Subject Correlation

Model with Y_{i0}

```
library(Design,T)
f ← ols(twstrs ~ treat*rcs(week,4) + rcs(age,4)*sex, x=T,y=T)
options(digits=4)
anova(f)
```

```
                Analysis of Variance          Response: twstrs

 Factor                                          d.f. Partial SS MS      F    P
 treat  (Factor+Higher Order Factors)              8   1662.07    207.76 1.40 0.1940
  All Interactions                                 6   1650.93    275.15 1.85 0.0869
 week  (Factor+Higher Order Factors)               9   9028.40   1003.16 6.75 <.0001
  All Interactions                                 6   1650.93    275.15 1.85 0.0869
  Nonlinear (Factor+Higher Order Factors)          6   7416.36   1236.06 8.32 <.0001
 age  (Factor+Higher Order Factors)                6   1262.29    210.38 1.42 0.2060
  All Interactions                                 3    364.77    121.59 0.82 0.4840
  Nonlinear (Factor+Higher Order Factors)          4    888.42    222.11 1.49 0.2022
 sex  (Factor+Higher Order Factors)                4   1044.90    261.22 1.76 0.1357
  All Interactions                                 3    364.77    121.59 0.82 0.4840
 treat * week  (Factor+Higher Order Factors)       6   1650.93    275.15 1.85 0.0869
  Nonlinear                                        4   1163.81    290.95 1.96 0.0994
  Nonlinear Interaction : f(A,B) vs. AB            4   1163.81    290.95 1.96 0.0994
 age * sex  (Factor+Higher Order Factors)          3    364.77    121.59 0.82 0.4840
  Nonlinear                                        2     62.22     31.11 0.21 0.8112
  Nonlinear Interaction : f(A,B) vs. AB            2     62.22     31.11 0.21 0.8112
 TOTAL NONLINEAR                                 10   8299.94    829.99 5.59 <.0001
 TOTAL INTERACTION                                9   2018.73    224.30 1.51 0.1407
 TOTAL NONLINEAR + INTERACTION                   13   9143.44    703.34 4.73 <.0001
 REGRESSION                                      18  10943.68    607.98 4.09 <.0001
 ERROR                                          612  90937.78    148.59
```

Adjust variance-covariance matrix for intra-subject correlation
No assumption about correlation pattern
First use Huber cluster sandwich covariance estimator

```
g ← robcov(f, uid)
anova(g)
```

```
 Factor                                          d.f. Partial SS MS      F     P
 treat  (Factor+Higher Order Factors)              8   4653.19    581.65  3.91 0.0002
  All Interactions                                 6   4138.29    689.72  4.64 0.0001
 week  (Factor+Higher Order Factors)               9  23306.57   2589.62 17.43 <.0001
  All Interactions                                 6   4138.29    689.72  4.64 0.0001
```

```
 Nonlinear (Factor+Higher Order Factors)      6  22363.81   3727.30 25.08 <.0001
age  (Factor+Higher Order Factors)            6    279.84     46.64  0.31 0.9298
 All Interactions                            3    105.20     35.07  0.24 0.8713
 Nonlinear (Factor+Higher Order Factors)     4    223.34     55.84  0.38 0.8260
sex  (Factor+Higher Order Factors)           4    215.98     54.00  0.36 0.8347
 All Interactions                            3    105.20     35.07  0.24 0.8713
treat * week  (Factor+Higher Order Factors)  6   4138.29    689.72  4.64 0.0001
 Nonlinear                                   4   3803.58    950.90  6.40 <.0001
 Nonlinear Interaction : f(A,B) vs. AB       4   3803.58    950.90  6.40 <.0001
age * sex  (Factor+Higher Order Factors)     3    105.20     35.07  0.24 0.8713
 Nonlinear                                   2     28.07     14.04  0.09 0.9099
 Nonlinear Interaction : f(A,B) vs. AB       2     28.07     14.04  0.09 0.9099
TOTAL NONLINEAR                             10  23672.75   2367.28 15.93 <.0001
TOTAL INTERACTION                            9   4199.22    466.58  3.14 0.0010
TOTAL NONLINEAR + INTERACTION               13  24087.86   1852.91 12.47 <.0001
REGRESSION                                  18  27906.36   1550.35 10.43 <.0001
ERROR                                      612  90937.78    148.59
```

Now use cluster bootstrap covariance estimator

```
h ← bootcov(f, uid, B=100)
anova(h)
```

```
Factor                                      d.f. Partial SS MS      F     P
treat  (Factor+Higher Order Factors)          8   4879.90    609.99  4.11 0.0001
 All Interactions                             6   3846.39    641.06  4.31 0.0003
week  (Factor+Higher Order Factors)           9  23583.06   2620.34 17.63 <.0001
 All Interactions                             6   3846.39    641.06  4.31 0.0003
 Nonlinear (Factor+Higher Order Factors)      6  21775.79   3629.30 24.42 <.0001
age  (Factor+Higher Order Factors)            6    335.65     55.94  0.38 0.8941
 All Interactions                             3    114.87     38.29  0.26 0.8559
 Nonlinear (Factor+Higher Order Factors)      4    211.89     52.97  0.36 0.8396
sex  (Factor+Higher Order Factors)           4    214.18     53.55  0.36 0.8369
 All Interactions                             3    114.87     38.29  0.26 0.8559
treat * week  (Factor+Higher Order Factors)  6   3846.39    641.06  4.31 0.0003
 Nonlinear                                   4   3252.66    813.17  5.47 0.0002
 Nonlinear Interaction : f(A,B) vs. AB       4   3252.66    813.17  5.47 0.0002
age * sex  (Factor+Higher Order Factors)     3    114.87     38.29  0.26 0.8559
 Nonlinear                                   2     20.21     10.10  0.07 0.9343
 Nonlinear Interaction : f(A,B) vs. AB       2     20.21     10.10  0.07 0.9343
TOTAL NONLINEAR                             10  24748.34   2474.83 16.66 <.0001
TOTAL INTERACTION                            9   4103.71    455.97  3.07 0.0013
TOTAL NONLINEAR + INTERACTION               13  26263.10   2020.24 13.60 <.0001
REGRESSION                                  18  31638.78   1757.71 11.83 <.0001
ERROR                                      612  90937.78    148.59
```

```
# Compare variances estimates
cbind(OLS=diag(Varcov(f)),Huber=diag(Varcov(g)),Bootstrap=diag(Varcov(h)))
```

```
                              OLS    Huber Bootstrap
Intercept                 69.88935 164.7059  237.9581
treat=5000U                7.40136   5.9373    5.7783
treat=Placebo              7.39587   4.7386    5.3305
week                       0.90621   0.2750    0.3292
week'                     21.42611   6.3595    7.1285
week''                   132.07963  36.9766   39.9628
age                        0.03694   0.1029    0.1474
age'                       0.28017   1.1665    1.6329
age''                      4.36057  19.4210   27.9583
sex=M                    233.59791 437.5595  857.5275
treat=5000U * week         1.83488   0.5159    0.5453
treat=Placebo * week       1.83838   0.4084    0.5059
treat=5000U * week'       43.18567  12.0905   11.8689
treat=Placebo * week'     43.44157   8.8561   10.0041
treat=5000U * week''     265.70345  71.9817   68.3556
treat=Placebo * week'' 267.60677  51.4669   55.1328
age * sex=M                0.13387   0.2952    0.5099
age' * sex=M               1.04418   3.6837    5.5063
age'' * sex=M             15.73105  61.1878   95.9781
```

Model with Y_{i0} as Baseline Covariate

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

dd       ← datadist(both)

attach(both)

dd ← datadist(dd, twstrs0, week)

f2 ← ols(twstrs ~ treat*rcs(week,4) + rcs(twstrs0,4) +
          rcs(age,4)*sex, x=T,y=T)

f2$stats
         n Model L.R.       d.f.          R2      Sigma
  522.0000   494.6362    21.0000     0.6123     8.3256
```

Compare R^2 with original R^2

```
f$stats
         n Model L.R.       d.f.          R2      Sigma
  631.0000    71.7033    18.0000     0.1074    12.1898
```

```
anova(f2)
```

Factor	d.f.	Partial SS	MS	F	P
treat (Factor+Higher Order Factors)	8	2878.5	359.82	5.19	<.0001
All Interactions	6	1236.6	206.10	2.97	0.0073
week (Factor+Higher Order Factors)	9	6771.3	752.37	10.85	<.0001
All Interactions	6	1236.6	206.10	2.97	0.0073
Nonlinear (Factor+Higher Order Factors)	6	289.5	48.25	0.70	0.6529
twstrs0	3	45899.4	15299.80	220.73	<.0001
Nonlinear	2	465.5	232.75	3.36	0.0356
age (Factor+Higher Order Factors)	6	1336.6	222.76	3.21	0.0042
All Interactions	3	816.9	272.30	3.93	0.0086
Nonlinear (Factor+Higher Order Factors)	4	1012.4	253.11	3.65	0.0061
sex (Factor+Higher Order Factors)	4	1080.5	270.13	3.90	0.0040
All Interactions	3	816.9	272.30	3.93	0.0086
treat * week (Factor+Higher Order Factors)	6	1236.6	206.10	2.97	0.0073
Nonlinear	4	104.8	26.20	0.38	0.8244
Nonlinear Interaction : f(A,B) vs. AB	4	104.8	26.20	0.38	0.8244
age * sex (Factor+Higher Order Factors)	3	816.9	272.30	3.93	0.0086
Nonlinear	2	645.1	322.56	4.65	0.0099
Nonlinear Interaction : f(A,B) vs. AB	2	645.1	322.56	4.65	0.0099
TOTAL NONLINEAR	12	1753.1	146.09	2.11	0.0152
TOTAL INTERACTION	9	2064.8	229.43	3.31	0.0006
TOTAL NONLINEAR + INTERACTION	15	3028.5	201.90	2.91	0.0002
REGRESSION	21	54740.6	2606.70	37.61	<.0001
ERROR	500	34657.9	69.32		

Huber cluster sandwich covariance estimator

```
g2 ← robcov(f2, both$uid)
anova(g2)
```

Factor	d.f.	Partial SS	MS	F	P
treat (Factor+Higher Order Factors)	8	2027.3	253.42	3.66	0.0004
All Interactions	6	1098.6	183.09	2.64	0.0157
week (Factor+Higher Order Factors)	9	8159.4	906.60	13.08	<.0001
All Interactions	6	1098.6	183.09	2.64	0.0157
Nonlinear (Factor+Higher Order Factors)	6	445.9	74.31	1.07	0.3781
twstrs0	3	18368.5	6122.83	88.33	<.0001
Nonlinear	2	176.7	88.35	1.27	0.2805

```
 age  (Factor+Higher Order Factors)            6    777.2     129.54  1.87 0.0843
  All Interactions                            3    453.4     151.14  2.18 0.0894
  Nonlinear (Factor+Higher Order Factors)     4    539.8     134.94  1.95 0.1015
 sex  (Factor+Higher Order Factors)           4    667.0     166.74  2.41 0.0487
  All Interactions                            3    453.4     151.14  2.18 0.0894
 treat * week  (Factor+Higher Order Factors)  6   1098.6     183.09  2.64 0.0157
  Nonlinear                                   4    156.4      39.11  0.56 0.6887
  Nonlinear Interaction : f(A,B) vs. AB       4    156.4      39.11  0.56 0.6887
 age * sex  (Factor+Higher Order Factors)     3    453.4     151.14  2.18 0.0894
  Nonlinear                                   2    316.7     158.36  2.28 0.1029
  Nonlinear Interaction : f(A,B) vs. AB       2    316.7     158.36  2.28 0.1029
 TOTAL NONLINEAR                             12   1013.5      84.46  1.22 0.2667
 TOTAL INTERACTION                            9   1372.7     152.52  2.20 0.0209
 TOTAL NONLINEAR + INTERACTION               15   1907.3     127.15  1.83 0.0277
 REGRESSION                                  21  38377.1    1827.48 26.36 <.0001
 ERROR                                      500  34657.9      69.32
```

Cluster bootstrap covariance estimator

```
h2 ← bootcov(f2, both$uid, B=100)
anova(h2)
```

```
 Factor                                      d.f. Partial SS MS      F     P
 treat  (Factor+Higher Order Factors)          8   2745.2     343.15  4.95 <.0001
  All Interactions                             6   1650.8     275.13  3.97 0.0007
 week  (Factor+Higher Order Factors)           9   9561.3    1062.37 15.33 <.0001
  All Interactions                             6   1650.8     275.13  3.97 0.0007
  Nonlinear (Factor+Higher Order Factors)      6    614.2     102.37  1.48 0.1840
 twstrs0                                       3  13845.2    4615.06 66.58 <.0001
  Nonlinear                                    2    175.7      87.85  1.27 0.2825
 age  (Factor+Higher Order Factors)            6    394.4      65.73  0.95 0.4599
  All Interactions                             3    193.6      64.53  0.93 0.4255
  Nonlinear (Factor+Higher Order Factors)      4    310.9      77.72  1.12 0.3457
 sex  (Factor+Higher Order Factors)            4    274.9      68.73  0.99 0.4117
  All Interactions                             3    193.6      64.53  0.93 0.4255
 treat * week  (Factor+Higher Order Factors)   6   1650.8     275.13  3.97 0.0007
  Nonlinear                                    4    165.7      41.41  0.60 0.6646
  Nonlinear Interaction : f(A,B) vs. AB        4    165.7      41.41  0.60 0.6646
 age * sex  (Factor+Higher Order Factors)      3    193.6      64.53  0.93 0.4255
  Nonlinear                                    2    165.5      82.77  1.19 0.3038
  Nonlinear Interaction : f(A,B) vs. AB        2    165.5      82.77  1.19 0.3038
 TOTAL NONLINEAR                             12   1139.4      94.95  1.37 0.1764
 TOTAL INTERACTION                            9   1795.5     199.50  2.88 0.0025
 TOTAL NONLINEAR + INTERACTION               15   3290.1     219.34  3.16 0.0001
 REGRESSION                                  21  38994.3    1856.87 26.79 <.0001
 ERROR                                      500  34657.9      69.32
```

20.8.3 Using Generalized Least Squares

We stay with baseline adjustment and use a variety of correlation structures, with constant variance.

```
rm(uid)

library(nlme)         # R
library(nlme3)        # S-Plus 6.1

cp ← Cs(corCAR1,corExp,corCompSymm,corLin,corGaus,corSpher)
fits ← vector('list',length(cp))
names(fits) ← cp
for(k in 1:length(cp))
  z ← gls(twstrs ∼ treat*rcs(week,4) + rcs(twstrs0,4) + age*sex,
          correlation= get(cp[k])(form=∼week | uid))
  fits[[k]] ← z

eval(parse(text=paste('anova(',
        paste('fits[[',1:length(cp),']]',collapse=','),')')))
# Same as saying anova(fits[[1]],fits[[2]],...)

          Model df  AIC  BIC logLik
fits[[1]]     1 24 3539 3640  -1746
fits[[2]]     2 24 3591 3692  -1771
fits[[3]]     3 24 3574 3675  -1763
fits[[4]]     4 24 3560 3661  -1756
fits[[5]]     5 24 3591 3692  -1771
fits[[6]]     6 24 3591 3692  -1771
```

AIC computed above is set up so that smaller values are best. From this the continuous-time AR1 structure fits the best, followed by linear structure. For the remainder of the analysis use `corCAR1`, using `glsD`.

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

Generalized least squares fit by REML
  Model: twstrs ∼ treat * rcs(week, 4) + rcs(twstrs0, 4) + rcs(age, 4) * sex
  Data: NULL
  Log-restricted-likelihood: -1746

                        Value Std.Error  t-value  p-value
```

```
Intercept                6.4257   13.3777  0.48033 0.631202
treat=5000U             -0.1956    2.9432 -0.06645 0.947046
treat=Placebo            6.4337    2.9601  2.17346 0.030214
week                    -0.4409    0.5184 -0.85045 0.395481
week'                    7.2195    3.4841  2.07211 0.038767
week''                  -9.4496    4.9287 -1.91726 0.055774
twstrs0                  0.6410    0.2423  2.64535 0.008417
twstrs0'                 0.7331    0.6628  1.10609 0.269221
twstrs0''               -2.2254    2.5898 -0.85930 0.390589
age                     -0.1113    0.2350 -0.47346 0.636093
age'                     0.6574    0.6509  1.00997 0.312999
age''                   -3.2184    2.5727 -1.25097 0.211529
sex=M                   24.4994   18.6423  1.31418 0.189388
treat=5000U * week       0.3718    0.7394  0.50276 0.615352
treat=Placebo * week     0.2067    0.7423  0.27844 0.780791
treat=5000U * week'     -3.0258    4.9355 -0.61307 0.540112
treat=Placebo * week'   -4.0042    4.9629 -0.80683 0.420145
treat=5000U * week''     3.7655    6.9748  0.53987 0.589528
treat=Placebo * week''   4.9961    7.0157  0.71214 0.476713
age * sex=M             -0.5891    0.4452 -1.32319 0.186377
age' * sex=M             1.4969    1.2406  1.20655 0.228175
age'' * sex=M           -4.1604    4.8194 -0.86327 0.388402

Correlation Structure: Continuous AR(1)
 Formula: ~week | uid
 Parameter estimate(s):
   Phi
0.8672
Degrees of freedom: 522 total; 500 residual
Residual standard error: 8.59
Clusters: 108
```

$\hat{\rho} = 0.8672$, 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 $0.8672^{10} = 0.24$.

Plot Variogram with assumed pattern superimposed[a]

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

Check constant variance and normality assumptions:

[a] `nlme` in R currently has a bug in `Variogram` that prevents the plot from being drawn.

```
p1 ← xYplot(resid(a) ~ fitted(a) | treat, abline=list(h=0,lty=2))
p2 ← xYplot(resid(a) ~ twstrs0, abline=list(h=0,lty=2))
p3 ← xYplot(resid(a) ~ week, method=smean.sdl, nx=0,
            abline=list(h=0,lty=2), ylim=c(-20,20))
print(p1, more=TRUE, split=c(1,1,2,2)) # Figure 20.4
print(p2, more=TRUE, split=c(1,2,2,2))
print(p3, more=FALSE,split=c(2,1,2,2))

qqnorm(a, ~(resid(., type='n')))        # Invokes qqnorm.gls; Figure 20.5
```

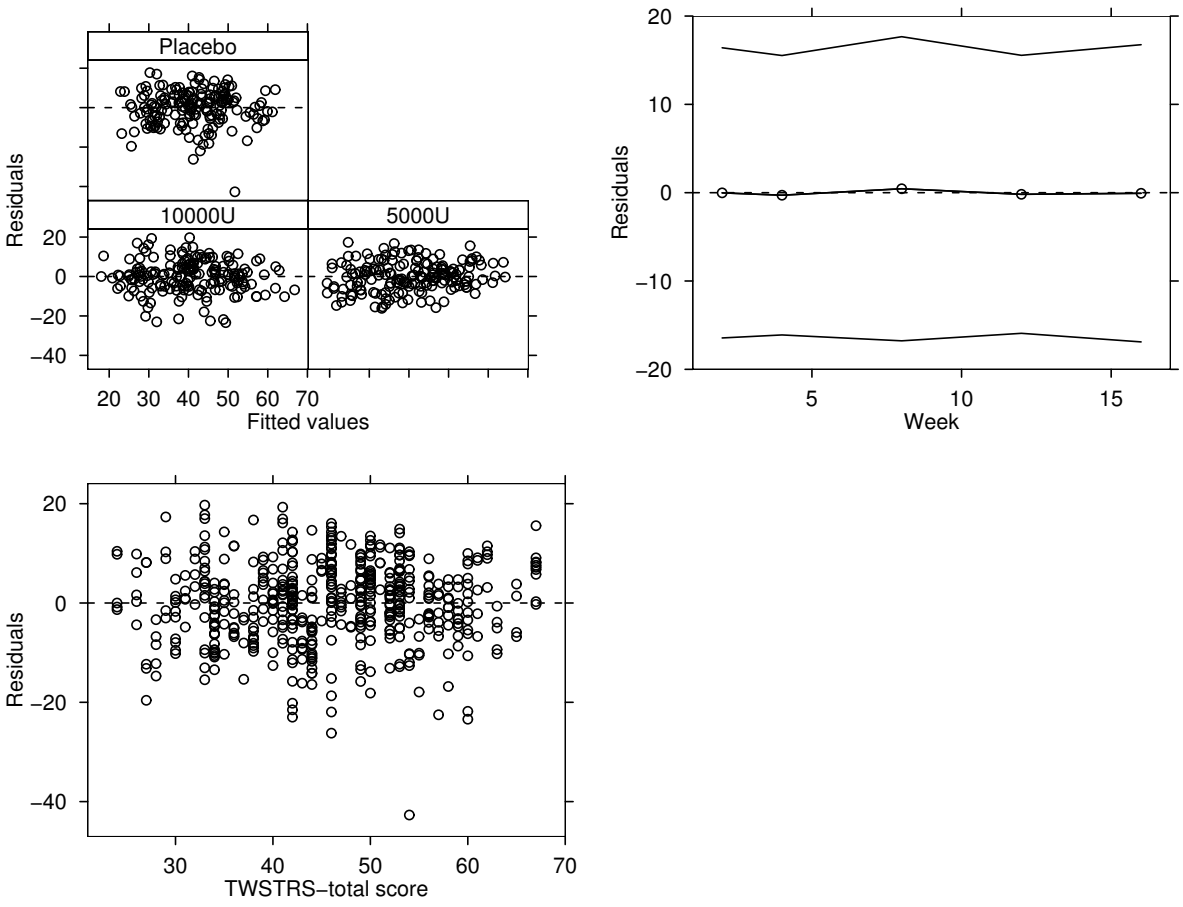

Figure 20.4: *Three residual plots to check for absence of trends in central tendency and in variability. Upper right panel shows the mean* $\pm 2 \times$ *SD.*

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

```
anova(a)

 Factor                                          Chi-Square d.f. P
 treat  (Factor+Higher Order Factors)             22.16      8   0.0046
```

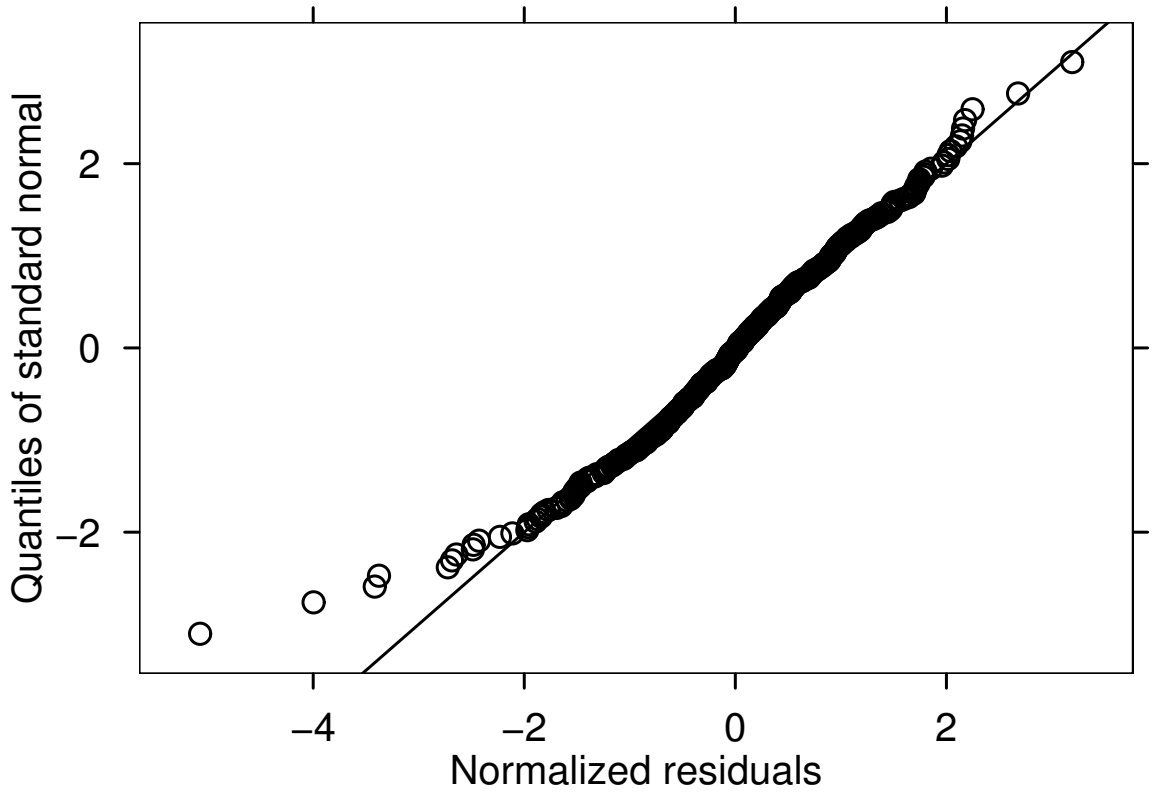

Figure 20.5: *QQ plot to check normality of residuals from GLS fit*

```
 All Interactions                                15.18      6   0.0189
week  (Factor+Higher Order Factors)              81.84      9   <.0001
 All Interactions                                15.18      6   0.0189
 Nonlinear (Factor+Higher Order Factors)         11.11      6   0.0849
twstrs0                                         234.10      3   <.0001
 Nonlinear                                        2.14      2   0.3436
age  (Factor+Higher Order Factors)                9.57      6   0.1441
 All Interactions                                 5.00      3   0.1720
 Nonlinear (Factor+Higher Order Factors)          7.36      4   0.1181
sex  (Factor+Higher Order Factors)                5.97      4   0.2017
 All Interactions                                 5.00      3   0.1720
treat * week  (Factor+Higher Order Factors)      15.18      6   0.0189
 Nonlinear                                        2.47      4   0.6501
 Nonlinear Interaction : f(A,B) vs. AB            2.47      4   0.6501
age * sex  (Factor+Higher Order Factors)          5.00      3   0.1720
 Nonlinear                                        3.80      2   0.1498
 Nonlinear Interaction : f(A,B) vs. AB            3.80      2   0.1498
TOTAL NONLINEAR                                  20.27     12   0.0622
TOTAL INTERACTION                                20.14      9   0.0171
TOTAL NONLINEAR + INTERACTION                    33.79     15   0.0036
TOTAL                                           327.84     21   <.0001
```

Compare coefficient estimates with OLS

```
cbind(OLS=coef(f2), GLS=coef(a))
```

```
                             OLS     GLS
Intercept                 2.64194  6.4257
treat=5000U              -0.66272 -0.1956
treat=Placebo             5.73742  6.4337
week                     -0.50795 -0.4409
week'                     8.12557  7.2195
week''                  -10.82480 -9.4496
twstrs0                   0.62806  0.6410
twstrs0'                  0.90792  0.7331
twstrs0''                -2.98243 -2.2254
age                      -0.02499 -0.1113
age'                      0.48903  0.6574
age''                    -2.78547 -3.2184
sex=M                    18.95866 24.4994
treat=5000U * week        0.59192  0.3718
treat=Placebo * week      0.48642  0.2067
treat=5000U * week'      -4.82604 -3.0258
treat=Placebo * week'    -6.18602 -4.0042
treat=5000U * week''      6.33322  3.7655
treat=Placebo * week''    8.10882  4.9961
age * sex=M              -0.43593 -0.5891
```

```
age' * sex=M              0.91918  1.4969

plot(anova(a))      # Figure 20.6
par(mfrow=c(1,2))   # Figure 20.7
plot(a, week=NA, treat=NA, conf.int=FALSE)
plot(a, twstrs0=NA)
```

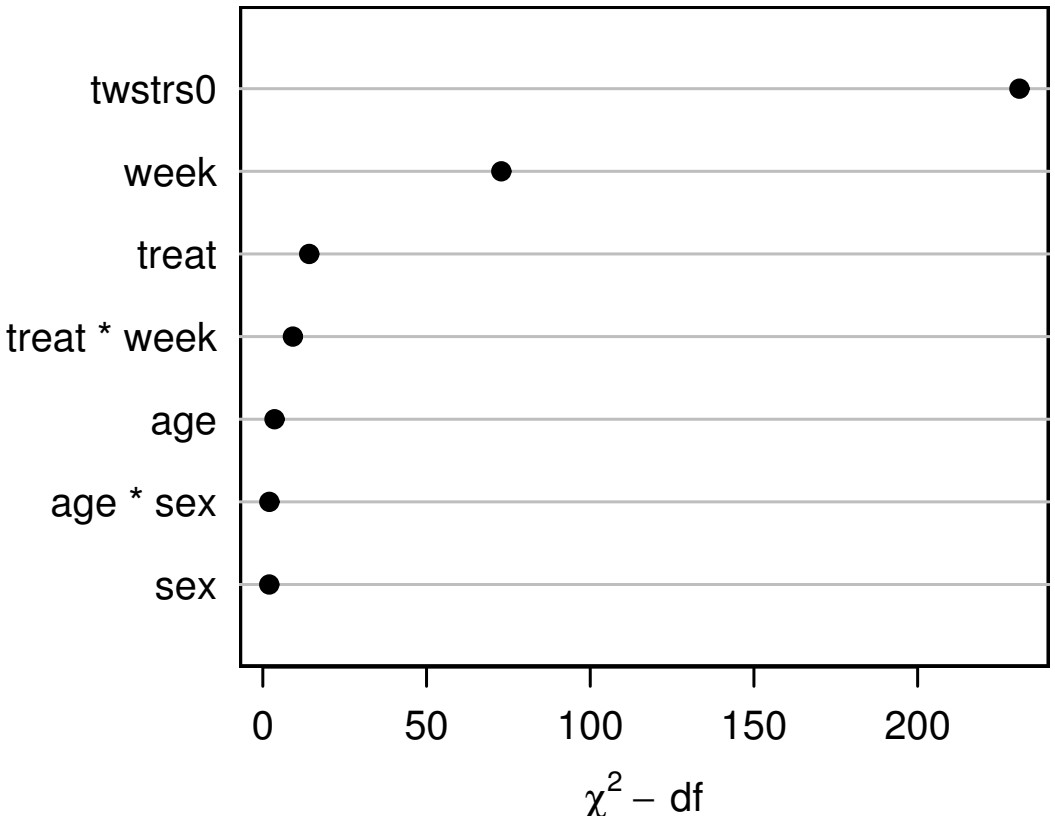

Figure 20.6: *Results of* `anova.Design` *from generalized least squares fit with continuous time AR1 correlation structure*

```
summary(a)      # Shows for week 8

            Effects              Response : twstrs

 Factor                 Low High Diff. Effect S.E. Lower 0.95 Upper 0.95
 week                    4  12    8     8.33  1.69  5.02      11.63
 twstrs0                39  53   14    14.67  1.69 11.35      17.99
 age                    46  65   19     2.34  2.06 -1.70       6.39
 treat - 5000U:10000U    1   2   NA     0.67  1.92 -3.09       4.43
 treat - Placebo:10000U  1   3   NA     5.31  1.92  1.54       9.07
```

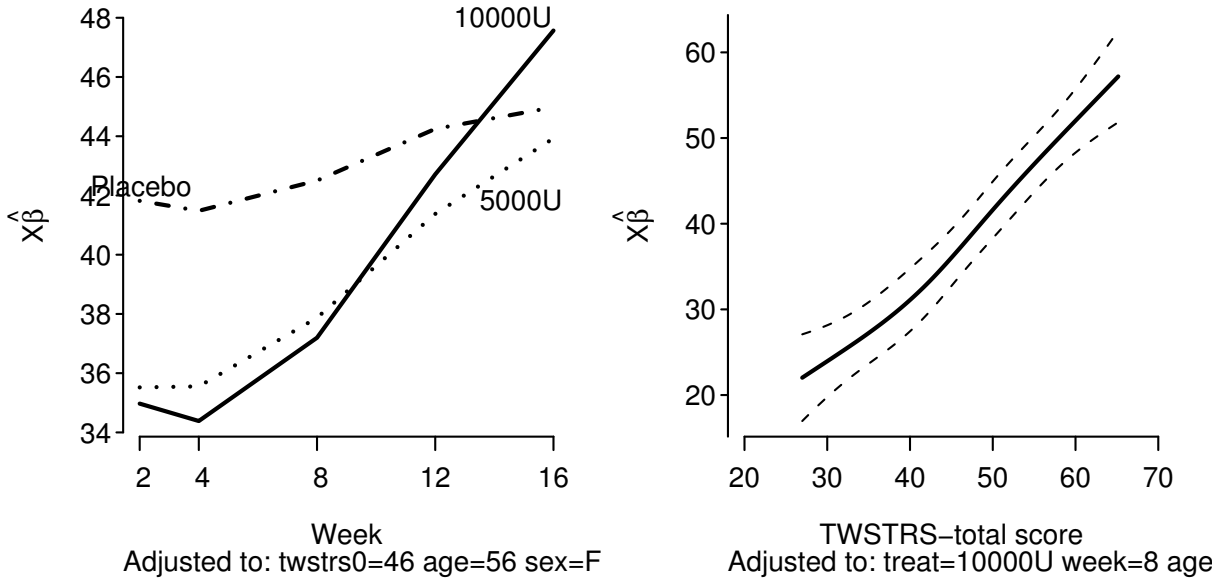

Figure 20.7: *Estimated effects of time and baseline TWSTRS*

```
 sex - M:F                1   2   NA    -0.96  1.78 -4.45       2.53

# 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'))
k1

 week twstrs0 age sex Contrast  S.E.     Lower      Upper      Z     Pr(>|z|)
  2   46      56  F   -6.299202 2.112422 -10.439474 -2.1589311 -2.98 0.0029
  4   46      56  F   -5.929143 2.000672  -9.850388 -2.0078972 -2.96 0.0030
  8   46      56  F   -4.632395 1.937588  -8.429999 -0.8347918 -2.39 0.0168
 12   46      56  F   -2.871302 2.047790  -6.884897  1.1422926 -1.40 0.1609
 16   46      56  F   -1.052674 2.110843  -5.189850  3.0845012 -0.50 0.6180

# 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'))
k2

 week twstrs0 age sex Contrast  S.E.     Lower      Upper     Z     Pr(>|z|)
  2   46      56  F   -6.847132 2.080123 -10.924099 -2.770165 -3.29 0.0010
  4   46      56  F   -7.097080 1.977315 -10.972546 -3.221614 -3.59 0.0003
  8   46      56  F   -5.305853 1.921916  -9.072738 -1.538967 -2.76 0.0058
 12   46      56  F   -1.535402 2.049264  -5.551886  2.481082 -0.75 0.4537
```

```
 16   46      56  F    2.588880 2.095188  -1.517613  6.695372  1.24 0.2166

p1 ← xYplot(Cbind(Contrast, Lower, Upper) ~ week, data=k1,
            ylab='Low Dose - Placebo', type='b',
            abline=list(h=0, lty=2), ylim=c(-15,10))

p2 ← xYplot(Cbind(Contrast, Lower, Upper) ~ week, data=k2,
            ylab='High Dose - Placebo', type='b',
            abline=list(h=0, lty=2), ylim=c(-15,10))

print(p1, more=T, split=c(1,1,1,2))     # Figure 20.8
print(p2, more=F, split=c(1,2,1,2))

nomogram(a, cex.axis=.7)                # Figure 20.9
```

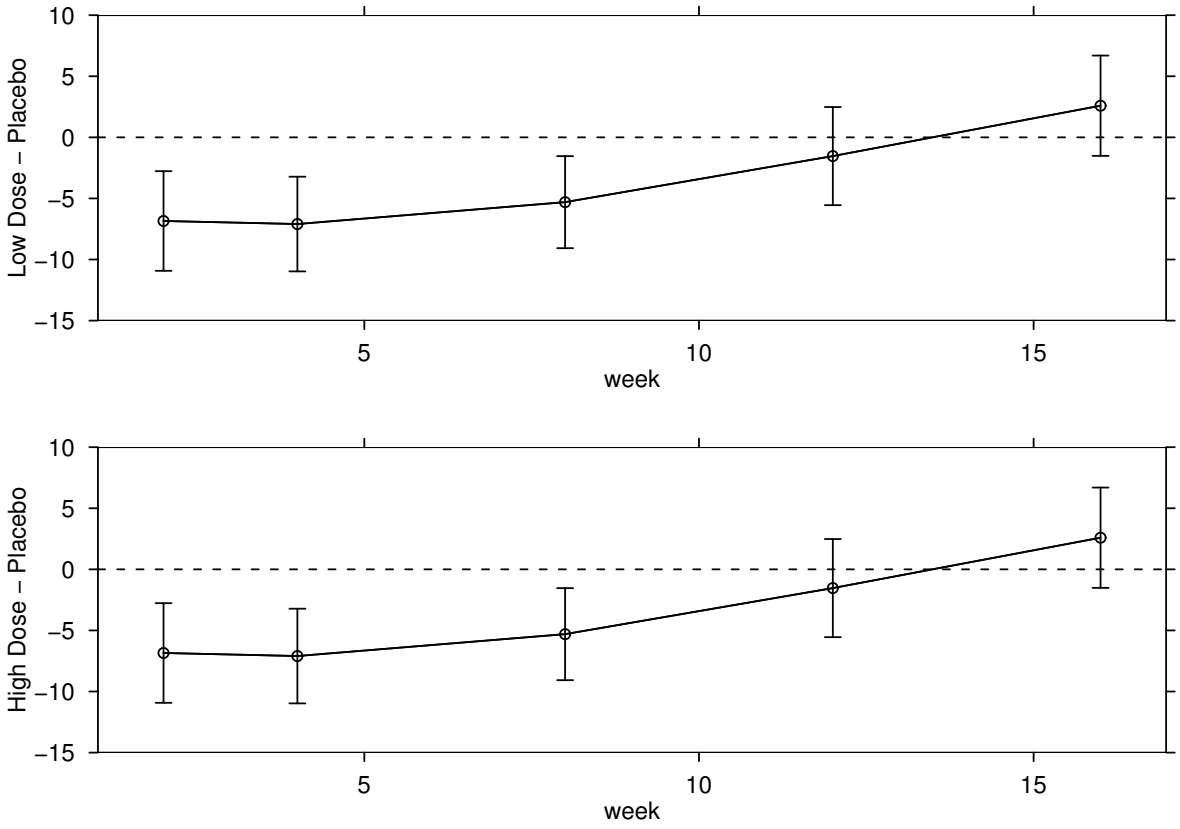

Figure 20.8: *Contrasts and 0.95 confidence limits from GLS fit*

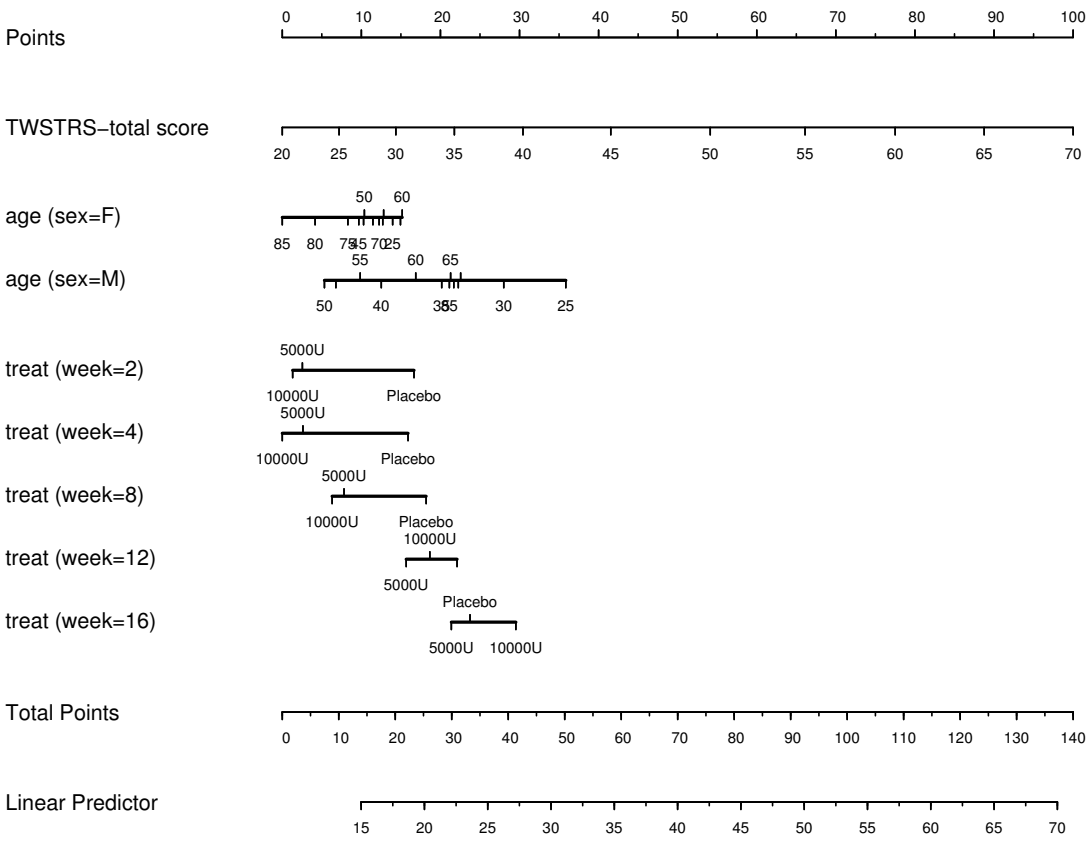

Figure 20.9: *Nomogram from GLS fit*

Bibliography

[1] C. S. Davis. *Statistical Methods for the Analysis of Repeated Measurements*. Springer, New York, 2002.

[2] P. J. Diggle, K.-Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Clarendon Press, Oxford UK, 1994.

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

[4] J. K. Lindsey. *Models for Repeated Measurements*. Clarendon Press, 1997.

[5] J. C. Pinheiro and D. M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, 2000.

[6] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer-Verlag, New York, fourth edition, 2002.

[7] G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, New York, 2000.