Solutions for Applied Linear Regression Third Edition

Sanford Weisberg

2005, Revised February 1, 2011

Contents

Preface

Most of the solutions in this manual were computed using R. You can get a copy of the scripts that can be used to do the computations from `sandy@umn.edu`. The scripts were updated in January, 2011, to correspond to version 2.0.0 of the **alr3** package. The graphs produced by the latest version of **alr3** are much more esthetic than are the graphs shown in this solutions manual, so the scripts will not reproduce these graphs exactly.

If you use other programs, like SAS or JMP, then the solutions you get can to be different because different programs stress different approaches to regression. For example, when using stepwise regression in R, the default criterion is *AIC*; in SPSS, the default is a change in an F-statistic. Adding almost any smoother is fairly easy in R and S-Plus, but other programs aren't so flexible and may make only one particular type of smoother easy to use. I recommend that you use the methodology that is easy to use with your program, and you may need to adapt the solutions accordingly. I think the basic ideas are much more important than the implementation details. Few regression problems have a unique *correct* solution in any case.

Most of the homework problems require drawing graphs—there are 115 figures in this solutions manual, and some of the figures contain more than one graph. Drawing and interpreting graphs is a central theme of this book.

You may find that some problems simply can't be done with the software you have chosen to use. Identifying cases in a plot is easy in JMP, fairly easy with SAS (but only for graphs created with the insight procedure), and clumsy but possible with R. If you are teaching from this book for the first time, you will need to work through the material in the book with the program of your choice. The *computer primers* on the website for the book (www.stat.umn.edu/alr) should help you and your students with some of the common programs. The primers are free, and you should urge your students to get the primer that is relevant to their computer program.

If you choose to use a program that is not covered by a *primer*, you will probably need to provide handouts and other help for your students. If you do this, please consider combining them into a primer of your own. If you send your primer to sandy@stat.umn.edu, I can add it to the website and save future instructors from starting with no help.

Have mercy

I seem to write very challenging problems. The non-data problems that require manipulating formulas or easy proofs are probably inappropriate in service courses for non-statistics majors. The data problems are almost all based on real data, and often do not coöperate with textbook discussions. Some questions are vague, designed to simulate real-life data analysis problems that are almost always vague. Have mercy on your students: read the problem solutions before you assign a problem, or else you may assign too many problems. You can also use the problems as starting points for questions of your own. Since the data are genuine, they can stand up to anything you might want to try.

Scripts

Scripts for the computations shown in *Applied Linear Regression* are available on the book's website (www.stat.umn.edu/alr) for R, S-Plus and SAS. We have not written scripts using SPSS, but if you do so, please send them to me. Scripts for the homework problems are not on the website, but I'll be happy to send the R scripts to instructors, `sandy@stat.umn.edu`).

Install the package or library

As of January, 2011, the **alr3** package for R is no longer the same as the **alr3** library for S-Plus, and the R package has been updated by the S-Plus library, which is much harder to maintain, has not been updated.

The **alr3** package in R is now almost exclusively data sets. Almost all of the functions have been renamed, improved, and moved to a different package called **car**. You can install the **alr3** package, and **car**, with this command:

```
> install.packages("alr3", dependencies=TRUE)
```

When using **alr3**, whether or not a function is in **alr3** or **car** will be completely transparent to the user.

If you are using SAS, I have put all but one of the data files into a SAS transport file. The file and instructions for using it are also at www.stat.umn.edu/alr. The instructions seem very complex to me, and if you know an easier way to install a SAS library, let me know so I can simplify this.

Help

Please let me know about any errors you find in the solutions, the primers, the scripts or the book itself at the email address below.

SANFORD WEISBERG

sandy@stat.umn.edu
University of Minnesota
Minneapolis, Minnesota
September 2004 (light revision January 2011)

1

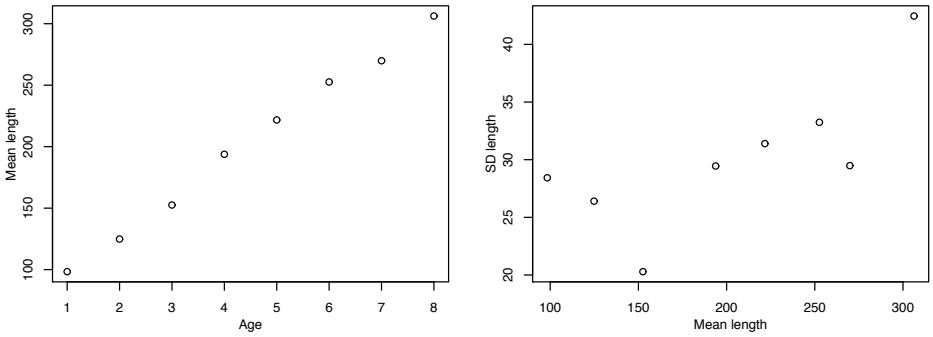
Scatterplots and Regression

Problems

1.1 Smallmouth bass data. Compute the means and the variances for each of the eight subpopulations in the smallmouth bass data. Draw a graph of average length versus *Age* and compare to Figure 1.5. Draw a graph of the standard deviations versus age. If the variance function is constant, then the plot of standard deviation versus *Age* should be a null plot. Summarize the information.

Solution:

```
       N  Mean   SD
Age 1 38  98.3 28.4
Age 2 72 124.8 26.4
Age 3 94 152.6 20.3
Age 4 15 193.8 29.5
Age 5 68 221.7 31.4
Age 6 87 252.6 33.2
Age 7 61 269.9 29.5
Age 8  4 306.3 42.5
```



The means appear to fall very close to a straight line, in agreement with the text. Apart from age 8, with only 4 fish, the SDs are mostly around 30 for all age classes, although there may be some evidence that SD increases with age. ∎

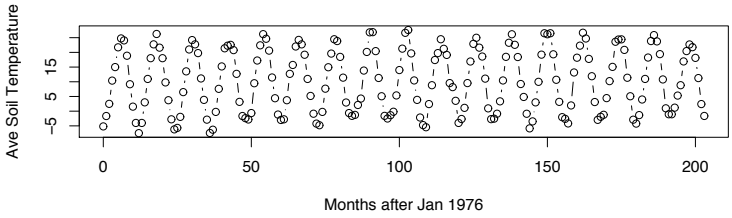
1.2 Mitchell data The data shown in Figure 1.12 give average soil temperature in degrees C at 20 cm depth in Mitchell, Nebraska for 17 years beginning January, 1976, plotted versus the month number. The data were collected by K. Hubbard and provided by O. Burnside.

1.2.1. Summarize the information in the graph about the dependence of soil temperature on month number.

Solution: This appears to be a null plot, with no particularly interesting characteristics. ∎

1.2.2. The data used to draw Figure 1.12 are in the file `Mitchell.txt`. Redraw the graph, but this time make the length of the horizontal axis at least four times the length of the vertical axis. Repeat question 1.2.1.

Solution:



Scaling matters! The points have also been joined with lines to emphasize the temporal pattern in the data: temperature is high in the summer and low in the winter. ∎

1.3 United Nations The data in the file `UN1.txt` contains *PPgdp*, the 2001 gross national product per person in U. S. dollars, and *Fertility*, the birth rate per 1000 females in the population in the year 2000. The data are for 193 localities, mostly UN member countries, but also other areas such

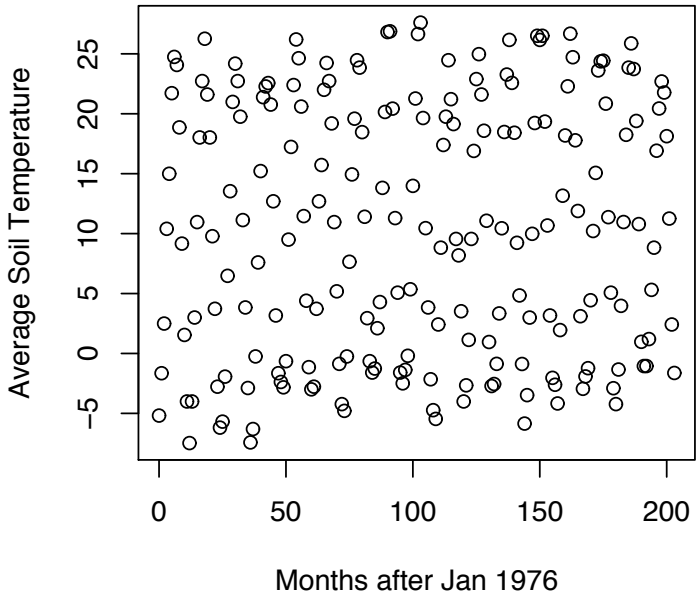


Fig. 1.12 Monthly soil temperature data.

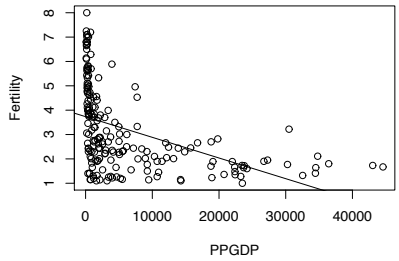
as Hong Kong that are not independent countries; the third variable on the file called *Locality* gives the name of the locality. The data were collected from unstats.un.org/unsd/demographic. In this problem, we will study the conditional distribution of *Fertility* given *PPgdp*.

1.3.1. Identify the predictor and the response.

Solution: The predictor is a function of *PPgdp* and the response is a function of *Fertility*. ∎

1.3.2. Draw the scatterplot of *Fertility* on the vertical axis versus *PPgdp* on the horizontal axis, and summarize the information in this graph. Does a straight line mean function seem to be a plausible for a summary of this graph?

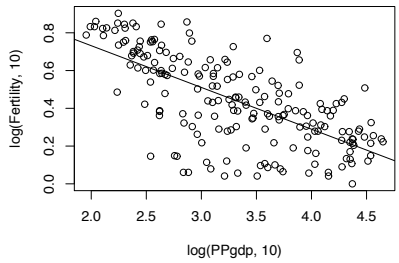
Solution:



Simple linear regression is not a good summary of this graph. The mean function does not appear to be linear, variance does not appear to be constant. ∎

1.3.3. Draw the scatterplot of log(*Fertility*) versus log(*PPgdp*), using logs to the base two. Does the simple linear regression model seem plausible for a summary of this graph?

Solution: In the figure below we actually used base-ten logarithms, but all the base of the logs do is change the labels for the tick marks in the graph, but not the shape of the graph.



Simple linear regression is much more appropriate in this scaling, as the mean function appears to be linear with fairly constant variance. The possible exception is for localities for which log(*PPGDP*) is very small, where the log(*Fertility*) is generally higher than would be predicted by simple linear regression. ∎

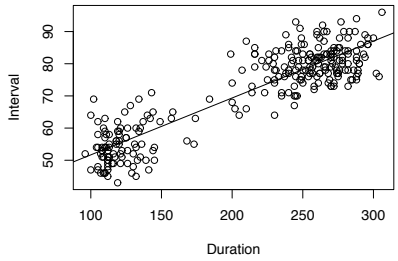
1.4 Old Faithful The data in the data file `oldfaith.txt` gives information about eruptions of Old Faithful Geyser during October 1980. Variables are the *Duration* in seconds of the current eruption, and the *Interval*, the time in minutes to the next eruption. The data were collected by volunteers and were provided by R. Hutchinson. Apart from missing data for the period from midnight to 6 AM, this is a complete record of eruptions for that month.

Old Faithful Geyser is an important tourist attraction, with up to several thousand people watching it erupt on pleasant summer days. The park service

uses data like these to obtain a prediction equation for the time to the next eruption.

Draw the relevant summary graph for predicting interval from duration, and summarize your results.

Solution:

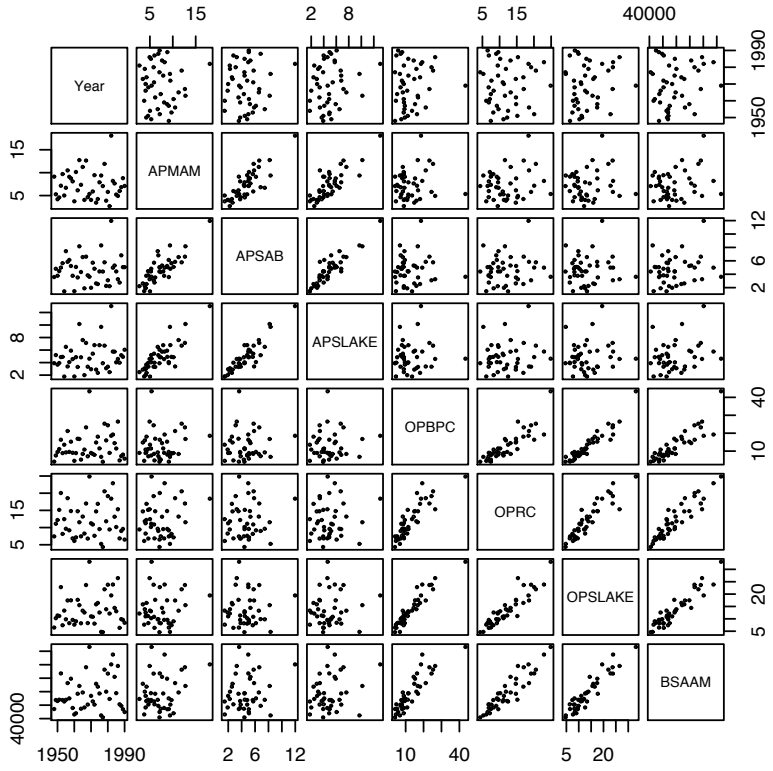


This is certainly not a null plot, as short durations are generally associated with shorter intervals. The points appear to form two clusters, and within clusters, the mean functions may be a bit different. ∎

1.5 Water runoff in the Sierras Can Southern California's water supply in future years be predicted from past data? One factor affecting water availability is stream runoff. If runoff could be predicted, engineers, planners and policy makers could do their jobs more efficiently. The data in the file `water.txt` contains 43 years worth of precipitation measurements taken at six sites in the Sierra Nevada mountains (labelled *APMAM*, *APSAB*, *APSLAKE*, *OPBPC*, *OPRC*, and *OPSLAKE*), and stream runoff volume at a site near Bishop, California, labelled *BSAAM*. The data are from the UCLA Statistics WWW server.

Draw the scatterplot matrix for these data, and summarize the information available from these plots.

Solution: (1) *Year* appears to be largely unrelated to each of the other variables; (2) the three variables starting with "O" seem to be correlated with each other, while the three variables starting with "A" also seems to be another correlated group; (3) *BSAAM* is more closely related to the "O" variables than the "A" variables; (4) there is at least one separated point with very high runoff. When we continue with this example in later chapters, we will end up talking logs of everything and combining the predictors into to terms.


Year
APMAM
APSAB
APSLAKE
OPBPC
OPRC
OPSLAKE
BSAAM
1950
1990
5
15
2
8
5
15
40000
2
6
12
10
40
5
20

2

Simple Linear Regression

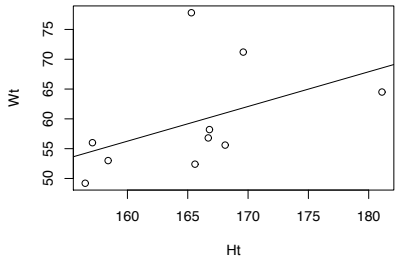
Problems

2.1 Height and weight data. The table below and in the data file `htwt.txt` gives Ht = height in centimeters and Wt = weight in kilograms for a sample of $n = 10$ 18-year-old girls. The data are taken from a larger study described in Problem 3.1. Interest is in predicting weight from height.

Ht	*Wt*
169.6	71.2
166.8	58.2
157.1	56.0
181.1	64.5
158.4	53.0
165.6	52.4
166.7	56.8
156.5	49.2
168.1	55.6
165.3	77.8

2.1.1. Draw a scatterplot of Wt on the vertical axis versus Ht on the horizontal axis. On the basis of this plot, does a simple linear regression model make sense for these data? Why or why not?

Solution:



With only 10 points, judging the adequacy of the model is hard, but it may be plausible here. ∎

2.1.2. Show that $\bar{x} = 165.52, \bar{y} = 59.47, SXX = 472.076, SYY = 731.961$, and $SXY = 274.786$. Compute estimates of the slope and the intercept for the regression of Y on X. Draw the fitted line on your scatterplot.

Solution: These computations are straightforward on a calculator, or using a computer language like R. Using a standard computer package, it is easiest to get means and the sample covariance matrix, and then use Table 2.1 to get the summary statistics. In R, the following will do the trick:

```
> library(alr3) # makes data available in R or S-Plus
> ave <- mean(htwt)     Computes the mean of each variable
> ave                   Display the means
    Ht     Wt
165.52  59.47
> cp <-(10-1)*cov(htwt)   Compute 9 times the covariance matrix
> cp
        Ht      Wt
Ht 472.076 274.786
Wt 274.786 731.961
```

so $SXX = 472.076, SYY = 274.786$ and $SYY = 731.961$. ∎

2.1.3. Obtain the estimate of σ^2 and find the estimated standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. Also find the estimated covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$. Compute the t-tests for the hypotheses that $\beta_0 = 0$ and that $\beta_1 = 0$, and find the appropriate p-values for these tests, using two-sided tests.

Solution: Using the computations from the last subproblem:

```
> bhat1 <- cp[1, 2]/cp[1, 1]
> bhat0 <- ave[2] - bhat1*ave[1]
> s2 <- (cp[2, 2] - cp[1, 2]^2/cp[1, 1])/8
> sebhat1 <- sqrt(s2 * (1/ cp[1, 1]))
> sebhat0 <- sqrt(s2 * (1/10 + ave[1]^2/cp[1, 1]))
> cov12 <- - s2 * ave[1]/cp[1, 1]
> t1 <- bhat1/sebhat1
> t0 <- bhat0/sebhat0
> c(bhat0, bhat1)
```

```
-36.87588   0.58208
> c(sebhat0, sebhat1, cov12)
 64.4728000   0.3891815 -25.0700325
> c(t0, t1)
-0.5719603  1.4956517
```

■

2.1.4. Obtain the analysis of variance table and F-test for regression. Show numerically that $F = t^2$, where t was computed in Problem 2.1.3 for testing $\beta_1 = 0$.

Solution:

```
> RSS <- cp[2,2] - cp[1,2]^2/cp[1,1]
> SSreg <- cp[2,2] - RSS
> F <- (SSreg/1) / (RSS/8)
> c(RSS,SSreg,F)
[1] 572.013564 159.947436   2.236974
> 1-pf(F,1,8)
[1] 0.1731089
```

These can be compared to the *Anova* table that will be obtained from R using the `lm` linear model fitting function:

```
> m <- lm(Wt ~ Ht, data=htwt)
> anova(m)
Analysis of Variance Table

Response: Wt
          Df Sum Sq Mean Sq F value Pr(>F)
d$Ht       1 159.95  159.95   2.237 0.1731
Residuals  8 572.01   71.50
```

■

2.2 More with Forbes' data An alternative approach to the analysis of Forbes' experiments comes from the Clausius–Clapeyron formula of classical thermodynamics, which dates to Clausius (1850). According to this theory, we should find that

$$\mathrm{E}(Lpres|\,Temp) = \beta_0 + \beta_1 \frac{1}{Ktemp} \tag{2.27}$$

where *Ktemp* is temperature in degrees Kelvin, which equals 255.37 plus $(5/9) \times$ *Temp*. If we were to graph this mean function on a plot of *Lpres* versus *Ktemp*, we would get a curve, not a straight line. However, we can estimate the parameters β_0 and β_1 using simple linear regression methods by defining u_1 to be the inverse of temperature in degrees Kelvin,

$$u_1 = \frac{1}{Ktemp} = \frac{1}{(5/9)\,Temp + 255.37}$$

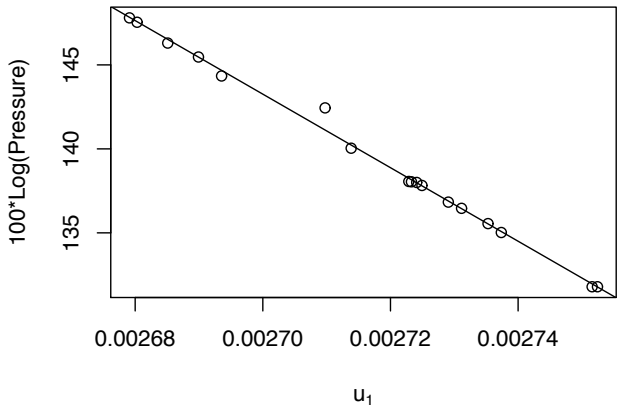
Then the mean function (2.27) can be rewritten as

$$\mathrm{E}(Lpres|\, Temp) = \beta_0 + \beta_1 u_1 \tag{2.28}$$

for which simple linear regression is suitable. The notation we have used in (2.28) is a little different, as the left side of the equation says we are conditioning on *Temp*, but the variable *Temp* does not appear explicitly on the right side of the equation.

2.2.1. Draw the plot of *Lpres* versus u_1, and verify that apart from case 12 the seventeen points in Forbes' data fall close to a straight line.

Solution: Thanks to Eric D. Kolaczyk for pointing out that the solution given in the solution manual used an incorrect definition of u_1. This has been corrected, as of May 16, 2007.



∎

2.2.2. Compute the linear regression implied by (2.28), and summarize your results.

Solution:

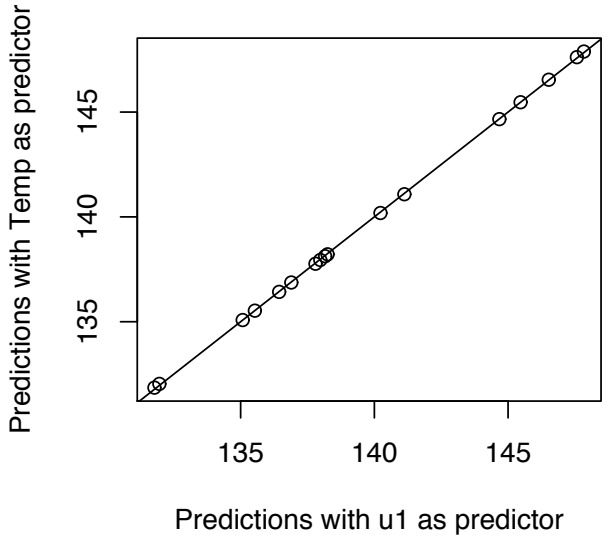
```
> m2 <- lm(Lpres ~ u1, data=forbes)
> summary(m2)

Call:
lm(formula = Lpres ~ u1, data = forbes)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     734.47      10.59   69.37   <2e-16
u1          -218968.41    3897.33  -56.18   <2e-16

Residual standard error: 0.3673 on 15 degrees of freedom
```

```
Multiple R-Squared: 0.9953
F-statistic:  3157 on 1 and 15 DF,  p-value: < 2.2e-16

> anova(m2)
Analysis of Variance Table

Response: Lpres
          Df Sum Sq Mean Sq F value    Pr(>F)
u1         1 425.77  425.77  3156.7 < 2.2e-16
Residuals 15   2.02    0.13
---
```

Apart from case 12, this mean function seems to match the data very well. ∎

2.2.3. We now have two possible models for the same data based on the regression of *Lpres* on *Temp* used by Forbes, and (2.28) based on the Clausius–Clapeyron formula. To compare these two, draw the plot of the fitted values from Forbes' mean function fit versus the fitted values from (2.28). Based on these and any other computations you think might help, is it possible to prefer one approach over the other? Why?

Solution:



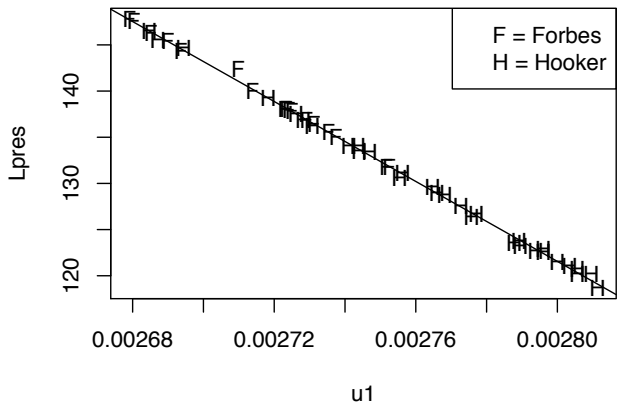
The line shown on the figure is the line $y = x$, indicating that both models give essentially the same predicted values and are therefore essentially indistinguishable. ∎

2.2.4. In his original paper, Forbes provided additional data collected by the botanist Dr. Joseph Hooker on temperatures and boiling points measured

often at higher altitudes in the Himalaya Mountains. The data for $n = 31$ locations is given in the file `hooker.txt`. Find the estimated mean function (2.28) for Hooker's data.

Solution: We begin by reading the data into R, combining the two data sets and drawing a graph:

```
> hooker$Lpres <- 100* logb(hooker$Pressure, 10)
> hooker$u1 <- 1/( (5/9)*hooker$Temp + 255.37)
> # create a combined data set for plotting
> combined.data <-data.frame(u1=c(forbes$u1,hooker$u1),
+                         Lpres=c(forbes$Lpres,hooker$Lpres),
+                 set=c(rep(c("F","H"),c(17,31))))
> attach(combined.data)
> plot(u1,Lpres,pch=as.character(set))
> legend("topright", c("F = Forbes", "H = Hooker"))
> abline(lm(Lpres~u1))
> detach(combined.data)
```



The variable `set` consists of "H" for Hooker and "F" for Forbes. R automatically converted this text variable to a factor, and so to use it to get plotting characters (the `pch=as.character(set)`), we need to convert `set` to a character vector. Both a key (using the `legend` function) and the OLS line have been added. From the graph, we see the two sets of data agree very closely, except perhaps at the very largest values of u_1, corresponding to the highest altitudes. Most of Hooker's data was collected at higher altitudes.

The above code will not work with S-Plus for several reasons. First, S-Plus does not allow adding new variables to a data set from a library. Second, you can't specify different plotting characters (or colors) on one call to plot. The following will work for S-Plus:

```
combined.data <- data.frame(Pres=c(forbes$Pres,hooker$Pres),
```

```
                          Temp=c(forbes$Temp,hooker$Temp))
combined.data$u1 <- 1/( (5/9)*combined.data$Temp + 255.37)
combined.data$Lpres <- logb(combined.data$Pres,10)
combined.data$set <- c(rep("H",31),rep("F",17))
attach(combined.data)
plot(u1,Lpres,type="n") # draws the axes only
points(u1[set=="H"],Lpres[set=="H"], pch="H") # hooker
points(u1[set=="F"],Lpres[set=="F"], pch="F") # forbes
legend(.00278, 145, c("F = Forbes", "H = Hooker"))
abline(lm(Lpres~u1))
```

The fitted regression is:

```
> h2 <- lm(Lpres ~ u1, data = hooker)
> summary(h2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.249e+02  4.844e+00   149.6   <2e-16
u1          -2.155e+05  1.753e+03  -122.9   <2e-16

Residual standard error: 0.353 on 29 degrees of freedom
Multiple R-Squared: 0.9981
F-statistic: 1.511e+04 on 1 and 29 DF,  p-value: < 2.2e-16

> anova(h2)
Analysis of Variance Table

Response: Lpres
          Df  Sum Sq Mean Sq F value    Pr(>F)
u1         1 1882.53 1882.53   15112 < 2.2e-16
Residuals 29    3.61    0.12
```

2.2.5. *This problem is not recommended unless you have access to a package with a programming language, like R, S-plus, Mathematica, or SAS.* For each of the cases in Hooker's data, compute the predicted values $\hat{y}$, and the standard error of prediction. Then compute $z = (Lpre - \hat{y})/\text{sepred}$. Each of the zs is a random variable, but if the model is correct each has mean zero and standard deviation close to one. Compute the sample mean and standard deviation of the zs, and summarize results.

Solution: To do the computing in R:

```
> fit.hooker <- predict(h2, newdata=hooker, se.fit=TRUE)
> # compute se prediction from se.fit:
> se.pred <- sqrt(fit.hooker$residual.scale^2 + fit.hooker$se.fit^2)
> # compute (observed - pred)/sepred
> options(width=60, digits=4) # for printing in this book
> zscores.hooker <- (hooker$Lpres - fit.hooker$fit)/se.pred
> zscores.hooker
```

```
      1       2       3       4       5       6
 0.34110 -0.86498  0.86367  0.25003  0.09698 -0.48894
      7       8       9      10      11      12
-0.76394 -0.66038  1.03607 -0.22953  1.25236  0.34015
     13      14      15      16      17      18
-1.76056  1.08116  0.19300  0.82277  0.36730 -1.16997
     19      20      21      22      23      24
 0.59080 -0.56299 -1.65749 -1.74009  1.40050 -0.43928
     25      26      27      28      29      30
 0.39479 -0.39375  0.57170  0.91007 -0.40823  1.83704
     31
-1.23229
> mean(zscores.hooker) ; sd(zscores.hooker)
[1] -0.00074
[1] 0.955
```

The `predict` function computes both the prediction and the standard error of the *fitted value* for all the points given by the argument `new.data`. This argument must be a data.frame, and the function gets the variables it needs, in this case `Temp`, from the data frame. The function returns a structure with relevant components `fit.hooker$fit` for the fitted values, `fit.hooker$se.fit` for the sefit, and `fit.hooker$residual.scale` for $\hat{\sigma}^2$. Since the z should have approximately mean zero and variance one, we see that the actual behavior of the z matches the theoretical behavior. ∎

2.2.6. Repeat Problem 2.2.5, but this time predict and compute the z-scores for the seventeen cases in Forbes data, again using the fitted mean function from Hooker's data. If the mean function for Hooker's data applies to Forbes' data then each of the z-scores should have zero mean and standard deviation close to one. Compute the z scores, compare them to those in the last problem, and comment on the results.

Solution:

```
> # predict from Hooker's data to Forbes' data
> fit.forbes <- predict(h2, newdata=forbes, se.fit=TRUE)
> # compute se prediction from se.fit:
> se.pred <- sqrt(fit.forbes$residual.scale^2 + fit.forbes$se.fit^2)
> # compute (observed - pred)/sepred
> options(width=60, digits=4) # for printing in this book
> zscores.forbes <- (100*log(forbes$Pressure, 10) - fit.forbes$fit)/se.pred
> zscores.forbes
      1       2       3       4       5       6
-0.62396 -0.11921 -0.13930  0.06034  0.10338 -0.09622
      7       8       9      10      11      12
 0.18563  0.19560 -0.38474 -0.19149 -0.29124  3.81591
     13      14      15      16      17
 0.44476 -0.46758 -0.09791  0.43133  0.43808
> mean(zscores.forbes) ; sd(zscores.forbes)
[1] 0.1920
[1] 0.9851
```

```
> mean(zscores.forbes[-12]) ; sd(zscores.forbes[-12])
[1] -0.03453
[1] 0.3239
```

The predictions from Hooker's data to Forbes' data are surprisingly accurate. The exception is case 12, which remains poorly fit. This exercise could be repeated, but using Forbes' original mean function.

```
> anova(h1)
Analysis of Variance Table

Response: Lpres
          Df Sum Sq Mean Sq F value Pr(>F)
Temp       1   1882    1882   14181 <2e-16
Residuals 29      4    0.13
> # predict from Hooker's data to Forbes' data
> fit.forbes1 <- predict(h1, newdata=forbes, se.fit=TRUE)
> # compute se prediction from se.fit:
> se.pred1 <- sqrt(fit.forbes$residual.scale^2 + fit.forbes1$se.fit^2)
> # compute (observed - pred)/sepred
> zscores.forbes1 <- (100*log(forbes$Pressure,10) - fit.forbes$fit)/se.pred
> zscores.forbes1
       1        2        3        4        5        6
-0.29983  0.18917  0.12482  0.30446  0.31313  0.10053
       7        8        9       10       11       12
 0.33004  0.33015 -0.24702 -0.05476 -0.28346  3.62679
      13       14       15       16       17
-0.05786 -0.85374 -0.70504 -0.32267 -0.34954
> mean(zscores.forbes1); sd(zscores.forbes1)
[1] 0.1262
[1] 0.9696
> mean(zscores.forbes1[-12]);sd(zscores.forbes1[-12])
[1] -0.0926
[1] 0.3672
```

Forbes' mean function appears be a bit less accurate. ∎

2.3 Deviations from the sample average Sometimes it is convenient to write the simple linear regression model in a different form that is a little easier to manipulate. Taking equation (2.1), and adding $\beta_1\bar{x} - \beta_1\bar{x}$, which equals zero, to the right-hand side, and combining terms, we can write

$$\begin{aligned} y_i &= \beta_0 + \beta_1\bar{x} + \beta_1 x_i - \beta_1\bar{x} + e_i \\ &= (\beta_0 + \beta_1\bar{x}) + \beta_1(x_i - \bar{x}) + e_i \\ &= \alpha + \beta_1(x_i - \bar{x}) + e_i \end{aligned} \tag{2.29}$$

where we have defined $\alpha = \beta_0 + \beta_1\bar{x}$. This is called the *deviations from the sample average form for simple regression.*

2.3.1. What is the meaning of the parameter α?

Solution: α is the value of $\mathrm{E}(Y|X = \bar{x})$. ∎

2.3.2. Show that the least squares estimates are

$$\hat{\alpha} = \bar{y} \qquad \hat{\beta}_1 \text{ as given by (2.5)}$$

Solution: The residual sum of squares function can be written as

$$\begin{aligned} RSS(\alpha, \beta_1) &= \sum (y_i - \alpha - \beta_1(x_i - \bar{x}))^2 \\ &= \sum (y_i - \alpha)^2 - 2\beta_1 \sum (y_i - \alpha)(x_i - \bar{x}) + \beta_1^2 \sum (x_i - \bar{x})^2 \end{aligned}$$

We can write

$$\begin{aligned} \beta_1 \sum (y_i - \alpha)(x_i - \bar{x}) &= \sum y_i(x_i - \bar{x}) + \alpha \sum (x_i - \bar{x}) \\ &= SXY + 0 \\ &= SXY \end{aligned}$$

Substituting into the last equation,

$$RSS(\alpha, \beta_1) = \sum (y_i - \alpha)^2 - 2\beta_1 SXY + \beta_1^2 SXX$$

Differentiating with respect to α and β_1 immediately gives the desired result. ∎

2.3.3. Find expressions for the variances of the estimates and the covariance between them.

Solution:

$$\mathrm{Var}(\hat{\alpha}) = \frac{\sigma^2}{n}, \mathrm{Var}(\hat{\beta}_1) = \sigma^2/SXX$$

The estimates $\hat{\beta}_1$ and $\hat{\alpha}$ are uncorrelated. ∎

2.4 Heights of Mothers and Daughters

2.4.1. For the heights data in the file `heights.txt`, compute the regression of *Dheight* on *Mheight*, and report the estimates, their standard errors, the value of the coefficient of determination, and the estimate of variance. Give the analysis of variance table the tests the hypothesis that $\mathrm{E}(Dheight|Mheight) = \beta_0$ versus the alternative that $\mathrm{E}(Dheight|Mheight) = \beta_0 + \beta_1 Mheight$. Write a sentence or two that summarizes the results of these computations.

Solution:

```
> mean(heights)
Mheight Dheight
  62.45   63.75    Daughters are a little taller
> var(heights)
         Mheight Dheight
Mheight    5.547   3.005    Daughters are a little more variable
Dheight    3.005   6.760
```

```
> m1 <- lm(Dheight ~ Mheight, data=heights)
> summary(m1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   29.917      1.623    18.4   <2e-16 ***
Mheight        0.542      0.026    20.9   <2e-16 ***
---
Residual standard error: 2.27 on 1373 degrees of freedom
Multiple R-Squared: 0.241,      Adjusted R-squared: 0.24
F-statistic:  435 on 1 and 1373 DF,  p-value: <2e-16
```

The F-statistic has a p-value very close to zero, suggesting strongly that $\beta_1 \neq 0$. The value of $R^2 = 0.241$, so only about one-forth of the variability in daughter's height is explained by mother's height. ∎

2.4.2. Write the mean function in the deviations from the mean form as in Problem 2.3. For this particular problem, give an interpretation for the value of β_1. In particular, discuss the three cases of $\beta_1 = 1$, $\beta_1 < 1$ and $\beta_1 > 1$. Obtain a 99% confidence interval for β_1 from the data.

Solution: If $\beta_1 = 1$, then on average *Dheight* is the same as *Mheight*. If $\beta_1 < 1$, then, while tall mothers tend to have tall daughters, on average they are shorter than themselves; this is the idea behind the word *regression*, in which extreme values from one generation tend to produce values not so extreme in the next generation. $\beta_1 > 1$ would imply that daughters tend to be taller than their mothers, suggesting that, eventually, we will all be giants.

The base R function `vcov` returns the covariance matrix of the estimated coefficients from a fitted model, so the diagonal elements of this matrix gives the squares of the standard errors of the coefficient estimates. The `alr3` library adds this function for S-Plus as well. In addition, the function `confint` in the `alr3` package can be used to get the confidence intervals:

```
> confint(m1, level=0.99)
                 0.5 %     99.5 %
(Intercept) 25.7324151 34.1024585
Mheight      0.4747836  0.6087104
```

∎

2.4.3. Obtain a prediction and 99% prediction interval for a daughter whose mother is 64 inches tall.

Solution: Using R,

```
> predict(m1,data.frame(Mheight=64),interval="prediction",level=.99)
       fit   lwr   upr
[1,] 64.59 58.74 70.44
```

∎

2.5 Small Mouth Bass

2.5.1. Using the West Bearskin Lake small mouth bass data in the file `wblake.txt`, obtain 95% intervals for the mean length at ages 2, 4 and 6 years.

Solution:

```
> m1 <- lm(Length~Age,smb)
> predict(m1,data.frame(Age=c(2,4,6)),interval="confidence")
     fit    lwr    upr
1 126.17 122.16 130.19
2 186.82 184.12 189.52
3 247.47 243.85 251.09
```

■

2.5.2. Obtain a 95% interval for the mean length at age 9. Explain why this interval is likely to be untrustworthy.

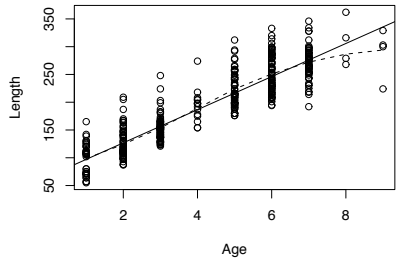
Solution:

```
> predict(m1,data.frame(Age=c(9)),interval="confidence")
        fit    lwr    upr
[1,] 338.44 331.42 345.46
```

This is an extrapolation outside the range of the data, as there were no nine year old fish in the sample. We don't know if the straight line mean function applies to these older fish. ■

2.5.3. The file `wblake2.txt` contains all the data for ages one to eight, and in addition includes a few older fishes. Using the methods we have learned in this chapter, show that the simple linear regression model is not appropriate for this larger data set.

Solution:



There are very few fish of age over seven, but comparison of the *loess* smooth and the straight line fit suggests that the straight line overestimates expected length for older fish. One could also look at residual plots to come to this conclusion. ■

2.6 United Nations data Refer to the UN data in Problem 1.3, page 2.

2.6.1. Using base-ten logarithms, use a software package to compute the simple linear regression model corresponding to the graph in Problem 1.3.3, and get the analysis of variance table.

Solution: Base-two logarithms were used in Problem 1.3.3, but here you are asked to use base-ten logarithms. The change of base has no material

effect on the solutions to this problem. If base-ten logs are used, then both the response and the predictor are multiplied by $\log_{10}(2) = 0.30103$, since both are in log scale.

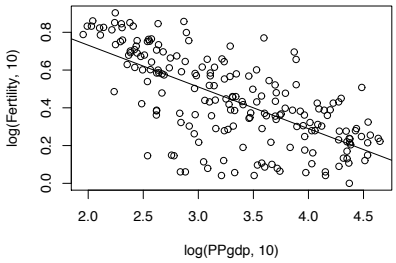
```
> m1 <- lm(log(Fertility,10)~log(PPgdp,10))
> summary(m1)
Call:
lm(formula = log(Fertility, 10) ~ log(PPgdp, 10))
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.1740     0.0588    20.0   <2e-16
log(PPgdp, 10)  -0.2212     0.0174   -12.7   <2e-16
Residual standard error: 0.172 on 191 degrees of freedom
Multiple R-Squared: 0.459
F-statistic:  162 on 1 and 191 DF,  p-value: <2e-16
> anova(m1)
Analysis of Variance Table
Response: log(Fertility, 10)
                Df Sum Sq Mean Sq F value Pr(>F)
log(PPgdp, 10)   1   4.80    4.80     162 <2e-16
Residuals      191   5.65    0.03
```

Unitless quantities like F and t tests and R^2 don't depend on the base of the logarithms. Other quantities are appropriately scaled quantities. ∎

2.6.2. Draw the summary graph, and add the fitted line to the graph.

```
> plot(log(PPgdp,10),log(Fertility,10))
> abline(lm(log(Fertility,10)~log(PPgdp,10)))
```

Solution:



∎

2.6.3. Test the hypothesis that the slope is zero versus the alternative that it is negative (a one-sided test). Give the significance level of the test, and a sentence that summarizes the result.

Solution: The t-test can be used, $t = -12.7$ with 191 df. The p-value given is essentially zero, so the one-sided p-value will also be near zero. We have strong evidence that $\beta_1 < 0$ suggesting that countries with higher $\log(PPgdp)$ have on average lower $\log(Fertility)$. ∎

2.6.4. Give the value of the coefficient of determination, and explain its meaning.

Solution: $R^2 = .4591$, so about 46% of the variability in log(*Fertility*) can be explained by conditioning on log(*PPgdp*). ∎

2.6.5. Increasing log(*PPgdp*) by one unit is the same as multiplying *PPgdp* by ten. If two localities differ in *PPgdp* by a factor of ten, give a 95% confidence interval on the difference in log(*Fertility*) for these two localities.

Solution: An increase in log(*PPgdp*) by one unit results in an increase in the mean of log(*Fertility*) by β_1 units, so this problem is asking for a 95% confidence interval for β_1. Using R,

```
> s1 <- coef(summary(m1)) # extract the summary table
> s1
               Estimate Std. Error t value   Pr(>|t|)
(Intercept)     1.17399   0.058795  19.968 1.2241e-48
log(PPgdp, 10) -0.22116   0.017368 -12.734 2.7310e-27
> s1[2,1] + c(1,-1)*qt(.975,191)*s1[2,2]
[1] -0.18690 -0.25542
> 10^(s1[2,1] + c(1,-1)*qt(.975,191)*s1[2,2])
[1] 0.65028 0.55537
> confint(m1)
                   2.5 %    97.5 %
(Intercept)     1.058022  1.289963
log(PPgdp, 10) -0.255418 -0.186902
> 10^confint(m1)
                    2.5 %     97.5 %
(Intercept)    11.4293649 19.4968018
log(PPgdp, 10)  0.5553694  0.6502764
```

which means that the fertility rate will be multiplied by a number between about 0.55 and 0.65, which amounts to a decrease of between 45% and 55%. ∎

2.6.6. For a locality not in the data with $PPgdp = 1000$, obtain a point prediction and a 95% prediction interval for log(*Fertility*). If the interval (a, b) is a 95% prediction interval for log(*Fertility*), then a 95% prediction interval for *Fertility* is given by $(10^a, 10^b)$. Use this result to get a 95% prediction interval for *Fertility*.

Solution: The prediction and its standard error can be obtained using the formulas in the chapter. To do the computation in R, we can use the `predict` function, as follows:

```
> new.data <- data.frame(PPgdp=1000)
> predict(m1,new.data,interval="prediction")
         fit     lwr     upr
[1,] 0.51051 0.17008 0.85094
> 10^predict(m1,new.data,interval="prediction")
        fit    lwr    upr
[1,] 3.2398 1.4794 7.0948
```

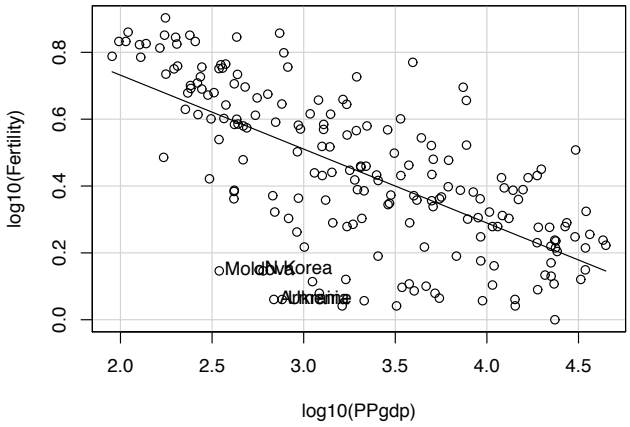
The first argument to the `predict` function is the name of a regression object. If no other arguments are given, then predictions are returned for each of the original data points. To get predictions for a different points, values must be supplied as the second argument. The function expects an object called a *data frame* to contain the values of the predictors for the new prediction. The variable `new.data` above is a data frame with just one value, `PPgdp=1000`. We do *not* need to take logarithms here because of the way that `m1` was defined, with the log in the definition of the mean function, so `m1` will take the log for us. If we wanted predictions at, say $PPgdp = 1000, 2000, 5000$, we would have defined `new.data` to be `data.frame(PPgdp=c(1000,2000,5000))`.

The `predict` function for R was used with the additional argument `interval="prediction"` to give the 95% prediction interval in log scale. Exponentiating the end points gives the interval for *Fertility* to be 1.48 to 7.09, a surprisingly wide interval. In S-Plus, the `predict` command has different arguments. ∎

2.6.7. Identify (1) the locality with the highest value of *Fertility*; (2) the locality with the lowest value of *Fertility*; and (3) the two localities with the largest positive residuals from the regression when both variables are in log scale, and the two countries with the largest negative residuals in log scales.

Solution: This problem should be solved using an interactive program for working with a graph. This is easily done, for example, in JMP. In R, the `identify` function, though clumsy, or if you use the scatterplot function in **car**, you can identify the odd points automatically:

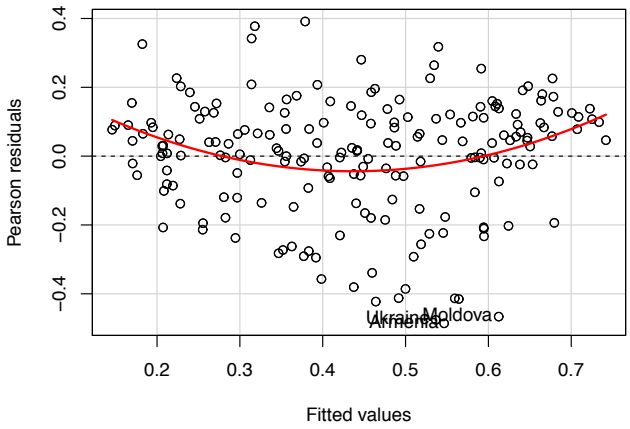
```
> scatterplot(log10(Fertility) ~ log10(PPgdp), UN1, id.n=4,
+   box=FALSE, smooth=FALSE)
[1] "Moldova" "Armenia" "Ukraine" "N.Korea"
```



This plot can be used to find Niger with the highest fertility rate and Hong Kong with the lowest. To find the residuals, it is convenient to plot

the residuals versus either the fitted values, and for this you can use the `residualPlots` function in **car**:

```
> residualPlots(m1, ~1, id.n=4, id.method="y")
           Test stat Pr(>|t|)
Tukey test     3.439    0.001
```



Equatorial Guinea and Oman have the largest positive residuals, and are therefore the two countries with fertility rates that are much larger than expected. Moldova, Armenia and Ukraine all have large negative residuals. The Tukey test, described later in the book, tests for lack of fit of the linear model; here significant lack of fit is found, due either to curvature or possibly outliers.

■

2.7 Regression through the origin Occasionally, a mean function in which the intercept is known *a priori* to be zero may be fit. This mean function is given by

$$\mathrm{E}(y|x) = \beta_1 x \tag{2.30}$$

The residual sum of squares for this model, assuming the errors are independent with common variance σ^2, is $RSS = \sum(y_i - \hat{\beta}_1 x_i)^2$.

2.7.1. Show that the least squares estimate of β_1 is $\hat{\beta}_1 = \sum x_i y_i / \sum x_i^2$. Show that $\hat{\beta}_1$ is unbiased and that $\mathrm{Var}(\hat{\beta}_1) = \sigma^2 / \sum x_i^2$. Find an expression for $\hat{\sigma}^2$. How many df does it have?

Solution: Differentiate the residual sum of squares function

$$RSS(\beta_1) = \sum(y_i - \beta_1 x_i)^2$$

and set the result to zero:

$$\frac{dRSS(\beta_1)}{d\beta_1} = -2\sum x_i(y_i - x_i\beta_1) = 0$$

or

$$\sum x_i y_i = \beta_1 \sum x_i^2$$

Solving for β_1 gives the desired result. To show unbiasedness,

$$\begin{aligned}
\mathrm{E}(\hat{\beta}_1) &= \mathrm{E}(\sum x_i y_i / \sum x_i^2) \\
&= \sum x_i \mathrm{E}(y_i|x_i) / \sum x_i^2 \\
&= \beta_1 \sum x_i^2 / \sum x_i^2 \\
&= \beta_1
\end{aligned}$$

as required. For the variance,

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}_1) &= \mathrm{Var}(\sum x_i y_i / \sum x_i^2) \\
&= \sum x_i^2 \mathrm{Var}(y_i|x_i) / (\sum x_i^2)^2 \\
&= \sigma^2 \sum x_i^2 / (\sum x_i^2)^2 \\
&= \sigma^2 / \sum x_i^2
\end{aligned}$$

To estimate variance, we need an expression for the residual sum of squares, which we will call RSS_0:

$$\begin{aligned}
RSS_0 &= \sum (y_i - \hat{\beta}_1 x_i) \\
&= \sum y_i^2 - 2\hat{\beta}_1 \sum x_i y_i + \hat{\beta}_1^2 \sum x_i^2 \\
&= \sum y_i^2 - 2(\sum x_i y_i)^2 / \sum x_i^2 + (\sum x_i y_i)^2 / \sum x_i^2 \\
&= \sum y_i^2 - (\sum x_i y_i)^2 / \sum x_i^2
\end{aligned}$$

which is the same as the simple regression formula for RSS except that uncorrected sums of squares and cross products replace corrected ones. Since the mean function has only one parameter, the estimate of σ^2 will have $(n-1)$ df, and $\hat{\sigma}^2 = RSS_0/(n-1)$. ∎

2.7.2. Derive the analysis of variance table with the larger model given by (2.16), but with the smaller model specified in (2.30). Show that the F-test derived from this table is numerically equivalent to the square of the t-test (2.23) with $\beta_0^* = 0$.

Solution: For (2.16), the residual sum of squares is $RSS = SYY - SXY^2/SXX$, and for (2.30) the residual sum of squares is $RSS_0 = \sum y_i^2 - (\sum x_i y_i)^2 / \sum x_i^2$. Thus, the regression sum of squares is $SSreg = RSS - RSS_0$. With these definitions, the *Anova* table is identical to Table 2.3, replacing the df for residual by $n-1$ rather than $n-2$, and replacing the total sum of squares SYY with $\sum y_i^2$.

The problem asked to show numerical equivalence between the F and t test. This can be done by fitting the two mean functions indicated to a set of data.

Table 2.6 Snake River data for Problem 2.7.

X	Y	X	Y
23.1	10.5	32.8	16.7
31.8	18.2	32.0	17.0
30.4	16.3	24.0	10.5
39.5	23.1	24.2	12.4
52.5	24.9	37.9	22.8
30.5	14.1	25.1	12.9
12.4	8.8	35.1	17.4
31.5	14.9	21.1	10.5
27.6	16.1		

It can also be shown mathematically by actually computing $F = SSreg/\hat{\sigma}^2$, and showing it is the same as $(\hat{\beta}_0/\text{se}(\hat{\beta}_0))^2$. ∎

2.7.3. The data in Table 2.6 and in the file `snake.txt` give X = water content of snow on April 1 and Y = water yield from April to July in inches in the Snake River watershed in Wyoming for $n = 17$ years from 1919 to 1935, from Wilm (1950).

Fit a regression through the origin and find $\hat{\beta}_1$ and σ^2. Obtain a 95% confidence interval for β_1. Test the hypothesis that the intercept is zero.

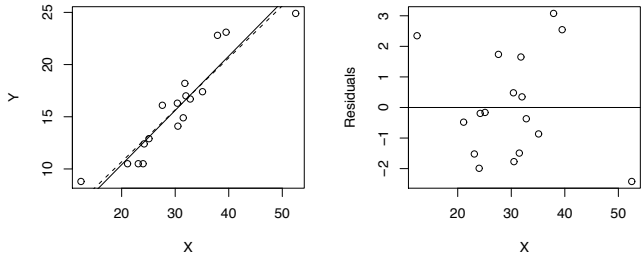
Solution:

```
> m0 <- lm(Y~X-1,data=snake)
> summary(m0)
Coefficients:
  Estimate Std. Error t value Pr(>|t|)
X  0.52039    0.01318   39.48   <2e-16 ***
---
Residual standard error: 1.7 on 16 degrees of freedom
Multiple R-Squared: 0.9898
F-statistic:  1559 on 1 and 16 DF,  p-value: < 2.2e-16
> m1 <- update(m0, ~ . + 1)
> anova(m0,m1)    A t-test is also possible giving same answer
Analysis of Variance Table

Model 1: Y ~ X - 1
Model 2: Y ~ X
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     16 46.226
2     15 45.560  1     0.666 0.2193 0.6463
```

∎

2.7.4. Plot the residuals versus the fitted values and comment on the adequacy of the mean function with zero intercept. In regression through the origin, $\sum \hat{e}_i \neq 0$.

Solution:



The plot at the left shows both the fit of the through-the-origin model (solid line) and the simple regression model (dashed line), suggesting little difference between them. The residual plot emphasizes the two points with the largest and smallest value of X as being somewhat separated from the other points, and fit somewhat less well. However, the through the origin model seems to be OK here. ∎

2.8 Scale invariance

2.8.1. In the simple regression model (2.1), suppose the value of the predictor X is replaced by cX, where c is some non-zero constant. How are $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$, R^2, and the t-test of NH: $\beta_1 = 0$ affected by this change?

Solution: Write

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X = \beta_0 + \frac{\beta_1}{c}(cX)$$

which suggests that the slope will change from β_1 to β_1/c, but no other summary statistics will change, and no tests will change. ∎

2.8.2. Suppose each value of the response Y is replaced by dY, for some $d \neq 0$. Repeat 2.8.1.

Solution: Write

$$\begin{aligned} \mathrm{E}(Y|X) &= \beta_0 + \beta_1 X \\ d\mathrm{E}(Y|X) &= d\beta_0 + d\beta_1 X \\ \mathrm{E}(dY|X) &= d\beta_0 + d\beta_1 X \end{aligned}$$

and so the slope and intercept and their estimates are all multiplied by d. The variance is also multiplied by d. Scale-free quantities like R^2 and test statistics are unchanged. ∎

2.9 Using Appendix A.3, verify equation (2.8).

Solution:

$$\begin{aligned} RSS &= \sum (y_i - \hat{y}_i)^2 \\ &= \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \end{aligned}$$

Table 2.7 The word count data.

Word	The word
Hamilton	Rate per 1000 words of this word in the writings of Alexander Hamilton
HamiltonRank	Rank of this word in Hamilton's wriings
Madison	Rate per 1000 words of this word in the writings of James Madison
MadisonRank	Rank of this word in Madison's wriings
Jay	Rate per 1000 words of this word in the writings of John Jay
JayRank	Rank of this word in Jay's wriings
Ulysses	Rate per 1000 words of this word in *Ulysses* by James Joyce
UlyssesRank	Rank of this word in *Ulysses*

$$
\begin{aligned}
&= \sum \left(y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i \right)^2 \\
&= \sum \left((y_i - \bar{y}) - \hat{\beta}_1 (x_i - \bar{x}) \right)^2 \\
&= \sum (y_i - \bar{y})^2 - 2\hat{\beta}_i \sum (y_i - \bar{y})(x_i - \bar{x}) + \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 \\
&= SYY - 2\hat{\beta}_1 SXY + \hat{\beta}_1^2 SXX
\end{aligned}
$$

Substituting $\hat{\beta}_1 = SXY/SXX$ and simplifying gives (2.8). ∎

2.10 Zipf's law Suppose we counted the number of times each word was used in the written works by Shakespeare, Alexander Hamilton, or some other author with a substantial written record. Can we say anything about the frequencies of the most common words?

Suppose we let f_i be the rate per 1000 words of text for the ith most frequent word used. The linguist George Zipf (1902–1950) observed a law-like relationship between rate and rank (Zipf, 1949),

$$\mathrm{E}(f_i|i) = a/i^b$$

and further observed that the exponent is close to $b = 1$. Taking logarithms of both sides, we get approximately

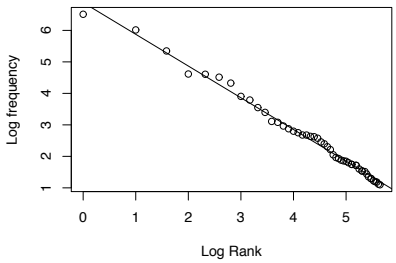
$$\mathrm{E}(\log(f_i)|\log(i)) = \log(a) - b\log(i) \tag{2.31}$$

Zipf's law has been applied to frequencies of may other classes of objects besides words, such as the frequency of visits to web pages on the internet, and the frequencies of species of insects in an ecosystem.

The data in `MWwords.txt` give the frequencies of words in works from four different sources: the political writings of eighteenth century American political figures Alexander Hamilton, James Madison, and John Jay, and the book *Ulysses* by twentieth century Irish writer James Joyce. The data are from Mosteller and Wallace (1964, Table 8.1-1), and give the frequencies of 165 very common words. Several missing values are occur in the data; these are really words that were used so infrequently that their count was not reported in Mosteller and Wallace's table.

2.10.1. Using only the fifty most frequent words in Hamilton's work (that is, using only rows in the data for which $HamiltonRank \leq 50$), draw the appropriate summary graph, estimate the mean function (2.31), and summarize your results.

Solution:



The scatterplot indicates that Zipf's law is remarkably accurate, as the points lie so close to the OLS line. The fitted regression is

```
> sel <- MWwords$HamiltonRank <= 50    select cases we want
> m1 <- lm(log2(Hamilton) ~ log2(HamiltonRank),
+         data= MWwords, subset=sel)
> summary(m1)    This is abbreviated here
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.88344    0.05696  120.84   <2e-16 ***
log2(HamiltonRank)   -1.00764    0.01275  -79.04   <2e-16 ***

Residual standard error: 0.1145 on 48 degrees of freedom
Multiple R-Squared: 0.9924
F-statistic:  6248 on 1 and 48 DF,  p-value: < 2.2e-16
```

The use of base-two logarithms is irrelevant here, as it only changes the intercept not the slope. ∎

2.10.2. Test the hypothesis that $b = 1$ against the two-sided alternative, and summarize.

Solution:

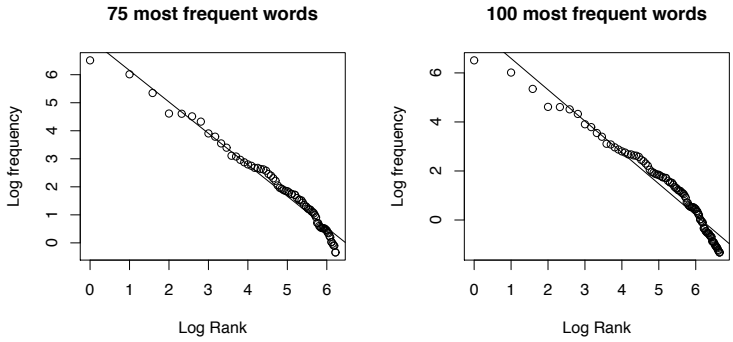
The test of $b = 1$ is equivalent to $\beta_1 = -1$ in the regression fit in the last subproblem. The test is $t = (-1.00764 - (-1.0))/.01275 = -0.5992157$, which can be compared to the $t(48)$ distribution:

```
> 2*pt(-0.5992157, 48)
[1] 0.551847
```

and the two-sided p-value is close to 0.95. There is no evidence against $b = 1$. ∎

2.10.3. Repeat Problem 2.10.1, but for words with rank of 75 or less, and with rank less than 100. For larger number of words, Zipf's law may break down. Does that seem to happen with these data?

Solution:



Zipf's law seems to work for 75 words, but does seem less adequate for 100 words. The frequencies of these less frequent words are lower than predicted by Zipf's Law.

Here is the R that generates the last two plots:

```
sel75 <- MWwords$HamiltonRank <= 75
sel100 <- MWwords$HamiltonRank <= 100
op <- par(mfrow=c(1, 2))
with(MWwords, plot(log(HamiltonRank[sel75], 2),
                   log(Hamilton[sel75], 2), main="75 words"))
abline(m1 <- lm(log2(Hamilton) ~ log2(HamiltonRank),
    data=MWwords, subset=sel75))
with(MWwords, plot(log(HamiltonRank[sel100], 2),
                   log(Hamilton[sel100], 2), main="100 words"))
abline(update(m1, subset=sel100))
```

■

2.11 For the Ft. Collins snow fall data, test the hypothesis that the slope is zero versus the alternative that it is not zero. Show that the t-test of this hypothesis is the same as the F-test; that is, $t^2 = F$.

Solution:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	28.6358	2.6149	10.95	0.0000
Early	0.2035	0.1310	1.55	0.1239

```
> anova(m1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Early	1	453.58	453.58	2.41	0.1239
Residuals	91	17118.83	188.12		

2.12 Old Faithful Use the data from Problem 1.4, page 4.

2.12.1. Use simple linear regression methodology to obtain a prediction equation for *interval* from *duration*. Summarize your results in a way that might be useful for the non-technical personnel who staff the Old Faithful Visitor's Center.

Solution:

```
> summary(m1 <- lm(Interval~Duration,oldfaith))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.98781    1.18122    28.8   <2e-16
Duration     0.17686    0.00535    33.0   <2e-16

Residual standard error: 6 on 268 degrees of freedom
Multiple R-Squared: 0.803,      Adjusted R-squared: 0.802
F-statistic: 1.09e+03 on 1 and 268 DF,  p-value: <2e-16

> predict(m1,data.frame(Duration=c(130,240,300)),interval="prediction")
     fit    lwr    upr
1 56.980 45.108 68.852
2 76.435 64.589 88.281
3 87.047 75.167 98.927
```

The prediction of time in minutes to next eruption is within about 11 minutes of $34 + .18 \times \text{Duration}$. The point prediction is the fitted value. The error bound is about two standard errors of prediction.

2.12.2. Construct a 95% confidence interval for

$$\mathrm{E}(\mathit{interval} | \mathit{duration} = 250)$$

Solution:

```
> predict(m1,data.frame(Duration=c(250)),interval="confidence")
         fit      lwr      upr
[1,] 78.20354 77.36915 79.03794
```

2.12.3. An individual has just arrived at the end of an eruption that lasted 250 seconds. Give a 95% confidence interval for the time the individual will have to wait for the next eruption.

Solution:

```
> predict(m1,data.frame(Duration=c(250)),interval="prediction")
         fit      lwr      upr
[1,] 78.20354 66.35401 90.05307
```

2.12.4. Estimate the 0.90 quantile of the conditional distribution of

$$interval|(duration = 250)$$

assuming that the population is normally distributed.

Solution:

```
> predict(m1,data.frame(Duration=c(250)),interval="confidence",
+   level=0.80)
          fit      lwr    upr
[1,] 78.20354 77.65908 78.748
```

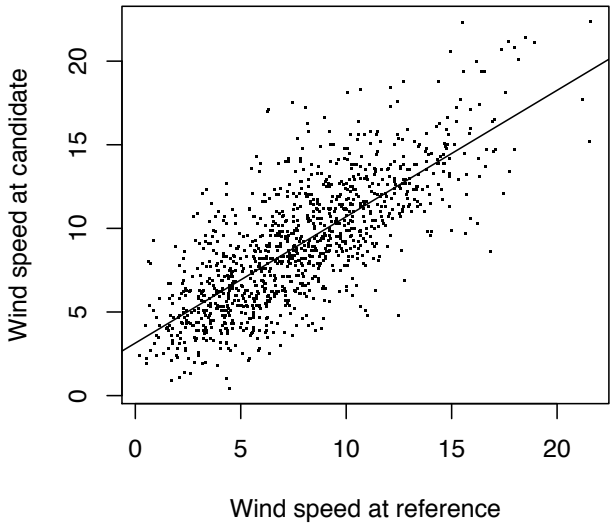
The upr value for the 80% interval is the 0.90 quantile. ∎

2.13 Windmills Energy can be produced from wind using windmills. Choosing a site for a *wind farm*, the location of the windmills, can be a multi-million dollar gamble. If wind is inadequate at the site, then the energy produced over the lifetime of the wind farm can be much less than the cost of building and operation. Prediction of long-term wind speed at a candidate site can be an important component in the decision to build or not to build. Since energy produced varies as the square of the wind speed, even small errors can have serious consequences.

The data in the file `wm1.txt` provides measurements that can be used to help in the prediction process. Data were collected every six hours for the year 2002, except that the month of May, 2002 is missing. The values *Cspd* are the calculated wind speeds in meters per second at a candidate site for building a wind farm. These values were collected at tower erected on the site. The values *RSpd* are wind speeds at a *reference site*, which is a nearby location for which wind speeds have been recorded over a very long time period. Airports sometimes serve as reference sites, but in this case the reference data comes from the National Center for Environmental Modeling; these data are described at http://dss.ucar.edu/datasets/ds090.0/. The reference is about 50 km south west of the candidate site. Both sites are in the northern part of South Dakota. The data were provided by Mark Ahlstrom and Rolf Miller of WindLogics.

2.13.1. Draw the scatterplot of the response *CSpd* versus the predictor *RSpd*. Is the simple linear regression model plausible for these data?

Solution:



A straight-line mean function with constant variance seems reasonable here, although there is clearly plenty of remaining variation. As with the heights data, the ranges of the data on the two axes are similar. Further analysis might look at the marginal distributions to see if they are similar as well. ∎

2.13.2. Fit the simple regression of the response on the predictor, and present the appropriate regression summaries.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.1412     0.1696    18.5   <2e-16
RSpd          0.7557     0.0196    38.5   <2e-16

Residual standard error: 2.47 on 1114 degrees of freedom
Multiple R-Squared: 0.571,
F-statistic: 1.48e+03 on 1 and 1114 DF,  p-value: <2e-16
```

The value of $R^2 = .57$ indicates that only about half the variation in *CSpd* is explained by *RSpd*. The large value of $\hat{\sigma}$ also suggests that predictions are likely to be of only modest quality. ∎

2.13.3. Obtain a 95% prediction interval for *CSpd* at a time when $RSpd = 7.4285$.

Solution: The prediction is

$$\widehat{CSpd} = 3.1412 + .7557 \times 6.4285 = 8.7552$$

with standard error given by the square root of $\hat{\sigma}^2 + \hat{\sigma}^2(1/1116 + (7.4285 - \overline{CSpd})^2/SXX) = (2.467)^2$. Since the df are so large, we can use the normal distribution to get the prediction interval to be from 3.914 to 13.596 meters per second. ∎

2.13.4. For this problem, we revert to generic notation, and let $x = CSpd$ and $y = CSpd$, and let n be the number of cases used in the regression ($n = 1116$ in the data we have used in this problem), and $\bar{x}$ and SXX defined on on the basis of these n observations. Suppose we want to make predictions at m time points with values of wind speed $x_{*1}, \ldots, x_{*m}$ that are different from the n cases used in constructing the prediction equation. Show that (1) the average of the m predictions is equal to the prediction taken at the average value $\bar{x}_*$ of the m values of the predictor, and (2), using the first result, the standard error of the average of m predictions is

$$\text{se of average prediction} = \sqrt{\frac{\hat{\sigma}^2}{m} + \hat{\sigma}^2\left(\frac{1}{n} + \frac{(\bar{x}_* - \bar{x})^2}{SXX}\right)} \tag{2.32}$$

If m is very large, then the first term in the square root is negligible and the standard error of average prediction is essentially the same as the standard error of a fitted value at $\bar{x}_*$.

Solution: For the first result,

$$\frac{1}{m}\sum_{i=1}^{m}\tilde{y}_{*i} = \frac{1}{m}\sum_{i=1}^{m}(\hat{\beta}_0 + \hat{\beta}_1 x_{*i}) = \hat{\beta}_0 + \hat{\beta}_1\frac{1}{m}\sum_{i=1}^{m}x_{*i} = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_*$$

so the average of the predictions is the same as the prediction at the average.

For the second result, we use the results of Appendix A.4. The variance of the average prediction will consist of two parts, the estimated error for estimating the coefficients, $\hat{\sigma}^2(1/n + (\bar{x}_* - \bar{x})^2/SXX)$, and the *average of the variance of the m independent errors attached to the m future predictions*, with estimated variance $\hat{\sigma}^2/m$. Adding these two and taking square roots gives (2.32). This standard error is *not* the average of the m standard errors for the m individual predictions, as all the predictions are correlated. ∎

2.13.5. For the period from January 1, 1948 to July 31, 2003, a total of $m = 62039$ wind speed measurements are available at the reference site, excluding the data from the year 2002. For these measurements, the average wind speed was $\bar{x}_* = 7.4285$. Give a 95% prediction interval on the long-term average wind speed at the candidate site. This long-term average of the past is then taken as an estimate of the long-term average of the future, and can be used to help decide whether the candidate is a suitable site for a wind farm.

Solution: The point estimate is the same as in Problem 2.13.3. We are now interested in the *average* of m predictions, so the standard error of this average prediction will be given by the square root of $\hat{\sigma}^2/m + \hat{\sigma}^2(1/1116 + (7.4285 - \overline{CSpd})^2/SXX) = (0.0748)^2$. If the year 2002 were a typical year, then this standard error would be close to $\hat{\sigma}/\sqrt{n}$, since the other terms will

all be relatively smaller. The 95% prediction interval for the mean wind speed over more than fifty years at the candidate site is from 8.609 to 8.902 meters per second. ∎

3

Multiple Regression

Problems

3.1 Berkeley Guidance Study The Berkeley Guidance Study enrolled children born in Berkeley, California, between January 1928 and June 1929, and then measured them periodically until age eighteen (Tuddenham and Snyder, 1954). The data we use is described in Table 3.6, and the data is given in the data files `BGSgirls.txt` for girls only, `BGSboys.txt` for boys only, and `BGSall.txt` for boys and girls combined. For this example use only the data on the girls.

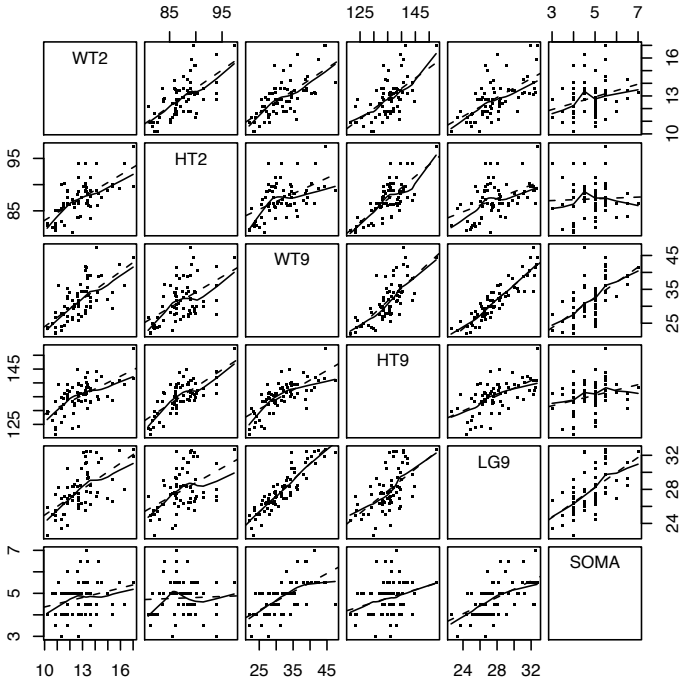
3.1.1. For the girls only, draw the scatterplot matrix of all the age two variables, all the age nine variables, and *Soma*. Write a summary of the information in this scatterplot matrix. Also obtain the matrix of sample correlations between the height variables.

Solution: The scatterplot matrix below is enhanced by adding two smoothers to each of the plots. In each frame, the solid line is the OLS fit for the simple regression of the vertical axis variable given the horizontal axis variable. The dashed line is the *loess* smooth with smoothing parameter 2/3. This plot is default behavior of the `scatterplotMatrix` function using the `car` package in R. Most other packages, such as S-Plus, SAS, JMP and SPSS, do not allow adding a smoother to a scatterplot matrix.

```
> scatterplotMatrix(BGSgirls[,c(2,3,4,5,6,12),], smooth=FALSE))
```

Table 3.6 Variable definitions for the Berkeley Guidance Study in the files `BGSgirls.txt`, `BGSboys.txt` and `BGSall.txt`.

Variable	Description
Sex	0 for males, 1 for females
WT2	Age 2 weight, kg
HT2	Age 2 height, cm
WT9	Age 9 weight, kg
HT9	Age 9 height, cm
LG9	Age 9 leg circumference, cm
ST9	Age 9 strength, kg
WT18	Age 18 weight, kg
HT18	Age 18 height, cm
LG18	Age 18 leg circumference, cm
ST18	Age 18 strength, kg
Soma	Somatotype, a scale from 1, very thin, to 7, obese, of body type



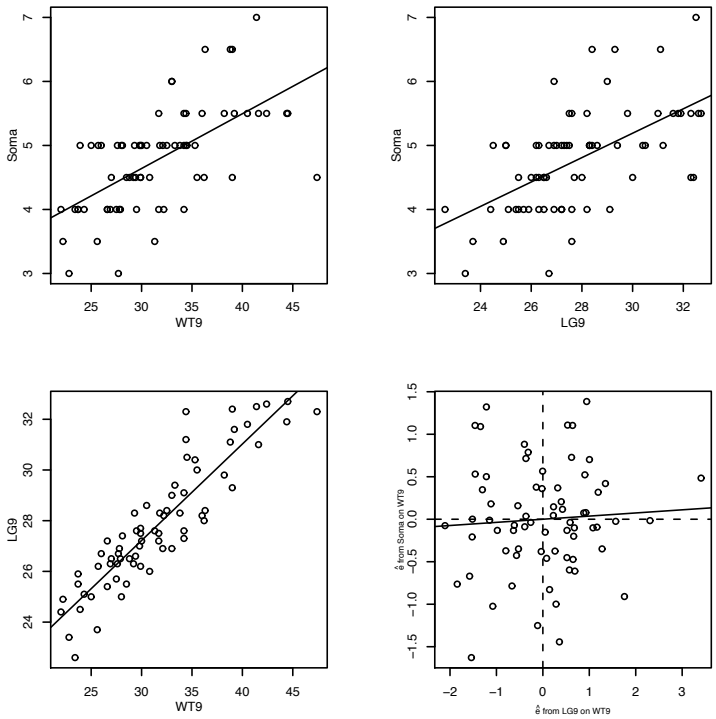
In virtually all of the frames, the regressions have nearly linear mean functions, which means that the OLS fit and the smoother agree. This is the ideal case for multiple linear regression. When we look at the regression of *Soma*

on the predictors on at a time, the last row of the scatterplot matrix suggests that only *WT9*, *LG9* and possibly *WT2* are predictive of *Soma*. We can't tell if these will be important predictors in multiple regression. Since the regressions are all linear, the correlation gives essentially the same information as the scatterplot matrix:

```
> print(cor(bgs.girls[,c(2,3,4,5,6,12),]),3)
       WT2    HT2   WT9   HT9   LG9   Soma
WT2  1.000 0.6445 0.693 0.607 0.616 0.2715
HT2  0.645 1.0000 0.523 0.738 0.469 0.0398
WT9  0.693 0.5229 1.000 0.728 0.904 0.6181
HT9  0.607 0.7384 0.728 1.000 0.598 0.2740
LG9  0.616 0.4688 0.904 0.598 1.000 0.5794
Soma 0.272 0.0398 0.618 0.274 0.579 1.0000
```

3.1.2. Starting with the mean function $\mathrm{E}(Soma|\,WT9) = \beta_0 + \beta_1\,WT9$, use added-variable plots to explore adding *LG9* to get the mean function $\mathrm{E}(Soma|\,WT9, LG9) = \beta_0 + \beta_1\,WT9 + \beta_2 LG9$. In particular, obtain the four plots equivalent to Figure 3.1, and summarize the information in the plots.

Solution: The four plots in Figure 3.1 can be drawn in essentially any computer package, since all that is required is two-dimensional scatterplots and saving residuals. Some programs (for example, JMP) draw added-variable plots whenever a multiple linear regression model is fit; others, like S-plus and R, have pre-written function in the `car` library for added variable plots.



While *Soma* and *LG9* positively correlated, as shown in the top-right figure, the added-variable plot for *LG9* after *WT9* shows that after adjustment *Soma* and *LG9* are essentially unrelated. This means that *LG9* and *WT9* are explaining essentially the same variation. ∎

3.1.3. Fit the multiple linear regression model with mean function

$$\mathrm{E}(Soma|X) = \beta_0 + \beta_1 HT2 + \beta_2 WT2 + \beta_3 HT9 + \beta_4 WT9 + \beta_5 ST9 \quad (3.25)$$

Find $\hat{\sigma}$, R^2, the overall analysis of variance table and overall F-test. Compute the t-statistics to be used to test each of the β_j to be zero against two-sided alternatives. Explicitly state the hypotheses tested and the conclusions.

Solution:

```
> m1 <- lm(Soma~HT2+WT2+HT9+WT9+ST9,bgs.girls)
> summary(m1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.8590417  2.3764431   3.728 0.000411
HT2         -0.0792535  0.0354034  -2.239 0.028668
WT2         -0.0409358  0.0754343  -0.543 0.589244
HT9         -0.0009613  0.0260735  -0.037 0.970704
WT9          0.1280506  0.0203544   6.291  3.2e-08
ST9         -0.0092629  0.0060130  -1.540 0.128373
```

```
---
Residual standard error: 0.5791 on 64 degrees of freedom
Multiple R-Squared: 0.5211,
F-statistic: 13.93 on 5 and 64 DF,  p-value: 3.309e-09
```

The hypothesis that all the βs apart from the intercept are zero against a general alternative has p-value of 3×10^{-9}, so there is strong evidence against the null hypothesis. The hypotheses tested by the t-values are that each of the $\beta_j = 0$ with the other βs arbitrary versus $\beta_j \neq 0$ with all the other βs arbitrary. For this test, only *WT9* and *HT2* have t-values with p-values smaller than 0.05. This seems to conflict with the information from the scatterplot matrix, but the scatterplot matrix contains information about *marginal tests* ignoring other variables, while the t-tests are *conditional*, and correspond to added-variable plots. ∎

3.1.4. Obtain the sequential analysis of variance table for fitting the variables in the order they are given in (3.25). State the hypotheses tested, and the conclusions for each of the tests.

Solution:

```
> anova(m1)
Analysis of Variance Table

Response: Soma
          Df  Sum Sq Mean Sq F value    Pr(>F)
HT2        1  0.0710  0.0710  0.2116 0.6470887
WT2        1  4.6349  4.6349 13.8212 0.0004252
HT9        1  3.7792  3.7792 11.2695 0.0013299
WT9        1 14.0746 14.0746 41.9700 1.516e-08
ST9        1  0.7958  0.7958  2.3731 0.1283728
Residuals 64 21.4623  0.3353
```

The test concerning *HT2* is for the null hypothesis $\mathrm{E}(Soma|X) = \beta_0$ versus the alternative $\mathrm{E}(Soma|X) = \beta_0 + \beta_1 HT2$. The F for *WT2* compares the null hypothesis $\mathrm{E}(Soma|X) = \beta_0 + \beta_1 HT2$ to the alternative $\mathrm{E}(Soma|X) = \beta_0 + \beta_1 HT2 + \beta_2 WT2$. Thus, each F compares the mean function with all preceding terms to the mean function that adds the current term to the mean function. For this order, *WT2*, *HT9* and *W9* all have small p-values. ∎

3.1.5. Obtain analysis of variance again, this time fitting with the five terms in the order given from right to left in (3.25). Explain the differences with the table you obtained in Problem 3.1.4. What graphs could help understand the issues?

Solution:

```
Analysis of Variance Table
Response: Soma
          Df  Sum Sq Mean Sq F value    Pr(>F)
ST9        1  0.3524  0.3524  1.0509   0.30916
WT9        1 18.8328 18.8328 56.1587 2.516e-10
```

```
HT9        1  1.4375  1.4375  4.2867   0.04245
WT2        1  1.0523  1.0523  3.1379   0.08125
HT2        1  1.6805  1.6805  5.0112   0.02867
Residuals 64 21.4623  0.3353
```

Order matters! *HT2* has a small p-value adjusted for the other predictors, but is unimportant ignoring them. *WT2* and *HT9* have very small p-values ignoring *WT9*, but much smaller p-values adjusted for *WT9*. Added-variable plots can be helpful to understand the effects of a variable adjusted for others. ∎

3.2 Added-variable plots This problem uses the United Nations example in Section 3.1 to demonstrate many of the properties of added-variable plots. This problem is based on the mean function

$$\mathrm{E}(\log(\mathit{Fertility})|\log(\mathit{PPgdp}) = x_1, \mathit{Purban} = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

There is nothing special about the two-predictor regression mean function, but we are using this case for simplicity.

3.2.1. Show that the estimated coefficient for $\log(\mathit{PPgdp})$ is the same as the estimated slope in the added-variable plot for $\log(\mathit{PPgdp})$ after *Purban*. This correctly suggests that *all the estimates in a multiple linear regression model are adjusted for all the other terms in the mean function.* Also, show that the residuals in the added-variable plot are identical to the residuals from the mean function with both predictors.

Solution:

```
> attach(UN)
> m1 <- lm(logFertility~Purban)    ignore log(PPgdp)
> m2 <- lm(logPPgdp~Purban)        second regression
> m3 <- lm(residuals(m1) ~ residuals(m2))  regression for the avp
> m4 <- lm(logFertility~Purban+logPPgdp)   regression with both terms
> summary(m3)
Coefficients:
               Estimate Std. Error   t value Pr(>|t|)
(Intercept)   -1.487e-17  2.826e-02 -5.26e-16        1
residuals(m2) -1.255e-01  1.904e-02    -6.588 4.21e-10 ***
---
Residual standard error: 0.3926 on 191 degrees of freedom
Multiple R-Squared: 0.1852
F-statistic: 43.41 on 1 and 191 DF,  p-value: 4.208e-10

> summary(m4)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.592996   0.146864  17.656  < 2e-16
Purban      -0.003522   0.001884  -1.869   0.0631
logPPgdp    -0.125475   0.019095  -6.571 4.67e-10
---
```

```
Residual standard error: 0.3936 on 190 degrees of freedom
Multiple R-Squared: 0.4689
F-statistic: 83.88 on 2 and 190 DF,  p-value: < 2.2e-16
```

The coefficients for log(*PPgdp*) are identical in the two regressions, although one is printed in scientific notation and the other in standard notation. The residuals can be shown to be the same by either plotting one set against the other, or by subtracting them. ∎

3.2.2. Show that the t-test for the coefficient for log(*PPgdp*) is not quite the same from the added-variable plot and from the regression with both terms, and explain why they are slightly different.

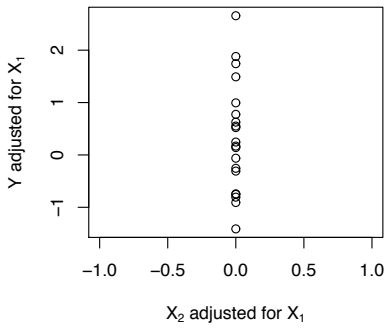
Solution: The added-variable plot computation has the df wrong, with one extra df. After correcting the df, the computations are identical. ∎

3.3 The following questions all refer to the mean function

$$\mathrm{E}(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{3.26}$$

3.3.1. Suppose we fit (3.26) to data for which $x_1 = 2.2x_2$, with no error. For example, x_1 could be a weight in pounds, and x_2 the weight of the same object in kg. Describe the appearance of the added-variable plot for X_2 after X_1.

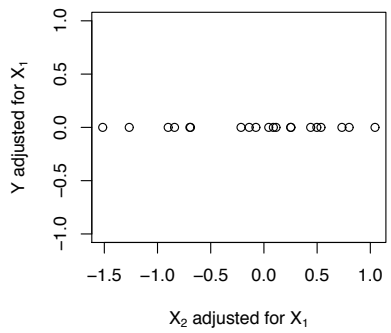
Solution: Since X_2 is an exact linear function of X_1, the residuals from the regression of X_2 on X_1 will all be zero, and so the plot will look like



In general, if X_1 and X_2 are highly correlated, the variability on the horizontal axis of an added-variable plot will be very small compared to the variability of the original variable. The coefficient for such a variable will be very poorly estimated. ∎

3.3.2. Again referring to (3.26), suppose now that Y and X_1 are perfectly correlated, so $Y = 3X_1$, without any error. Describe the appearance of the added-variable plot for X_2 after X_1.

Solution: Since $Y = 3X_1$ the residuals from the regression of Y on X_1 will all be zero, and so the plot will look like



In general, if Y and X_1 are highly correlated, the variability on the vertical axis of an added-variable plot will be very small compared to the variability of the original variable, and we will get an approximately null plot. ∎

3.3.3. Under what conditions will the added-variable plot for X_2 after X_1 have exactly the same shape as the scatterplot of Y versus X_2?

Solution: If X_1 is uncorrelated with both X_2 *and* Y, then these two plots will be the same. ∎

3.3.4. True or false: The vertical variation in an added-variable plot for X_2 after X_1 is always less than or equal to the vertical variation in a plot of Y versus X_2. Explain.

Solution: Since the vertical variable is the residuals from the regression of Y on X_1, the vertical variation in the added-variable plot is never larger than the vertical variation in the plot of Y versus X_2. ∎

3.4 Suppose we have a regression in which we want to fit the mean function (3.1). Following the outline in Section 3.1, suppose that the two terms X_1 and X_2 have sample correlation zero. This means that, if $x_{ij}, i = 1, \ldots, n$ and $j = 1, 2$ are the observed values of these two terms for the n cases in the data, $\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = 0$.

3.4.1. Give the formula for the slope of the regression for Y on X_1, and for Y on X_2. Give the value of the slope of the regression for X_2 on X_1.

Solution: (1) $\hat{\beta}_1 = SX_1Y/SX_1X_1$; (2) $\hat{\beta}_2 = SX_2Y/SX_2X_2$; (3) $\hat{\beta}_3 = 0$. ∎

3.4.2. Give formulas for the residuals for the regressions of Y on X_1 and for X_2 on X_1. The plot of these two sets of residuals corresponds to the added-variable plot in Figure 3.1d.

Solution: (1) $\hat{e}_{1i} = y_i - \bar{y} - \hat{\beta}_1(x_{i1} - \bar{x}_1)$; (2) $\hat{e}_{3i} = x_{i2} - \bar{x}_2$. ∎

3.4.3. Compute the slope of the regression corresponding to the added-variable plot for the regression of Y on X_2 after X_1, and show that this slope is exactly the same as the slope for the simple regression of Y on X_2 ignoring X_1. Also find the intercept for the added variable plot.

Solution: Because $\sum \hat{e}_{3i} = 0$,

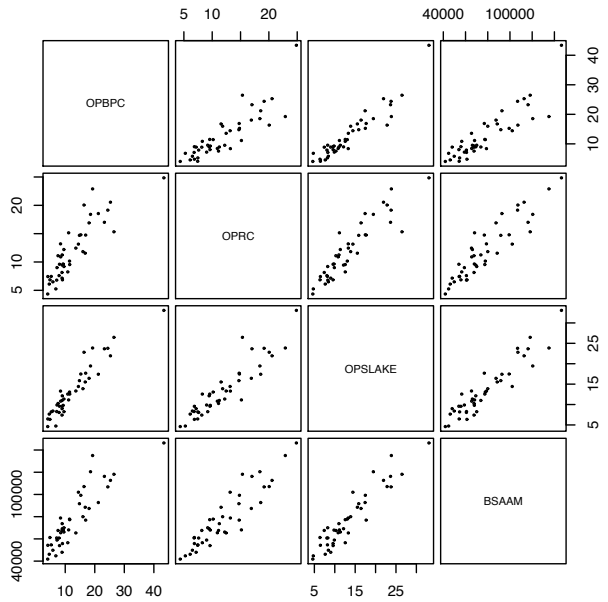
$$\begin{aligned}
\text{Slope} &= \sum \hat{e}_{3i}\hat{e}_{1i} / \sum \hat{e}_{3i}^2 \\
&= \sum (x_{i2} - \bar{x}_2)(y_i - \bar{y} - \hat{\beta}_1(x_{i1} - \bar{x}_1)) / \sum (x_{i2} - \bar{x}_2)^2 \\
&= \left(SX_2\,Y - \hat{\beta}_1 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) \right) / SX_2X_2 \\
&= SX_2\,Y / SX_2X_2 \\
&= \hat{\beta}_2
\end{aligned}$$

The estimated intercept is exactly zero, and the R^2 from this regression is exactly the same as the R^2 from the regression of Y on X_2. ∎

3.5 Refer to the data described in Problem 1.5, page 5. For this problem, consider the regression problem with response *BSAAM*, and three predictors as terms given by *OPBPC*, *OPRC* and *OPSLAKE*.

3.5.1. Examine the scatterplot matrix drawn for these three terms and the response. What should the correlation matrix look like (that is, which correlations are large and positive, which large and negative, and which are small)? Compute the correlation matrix to verify your results. Get the regression summary for the regression of *BSAAM* on these three terms. Explain what the "t values" column of your output means.

Solution: The scatterplot matrix is



All the variables are strongly, positively, related, which can lead to problems in understanding coefficients, since each of the three predictors is nearly the same variable. The correlation matrix and regression output are:

```
> cor(water[sel])
          OPBPC    OPRC OPSLAKE   BSAAM
OPBPC   1.00000 0.86471 0.94335 0.88575
OPRC    0.86471 1.00000 0.91914 0.91963
OPSLAKE 0.94335 0.91914 1.00000 0.93844
BSAAM   0.88575 0.91963 0.93844 1.00000

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22991.9     3545.3    6.49  1.1e-07
OPBPC           40.6      502.4    0.08   0.9360
OPRC          1867.5      647.0    2.89   0.0063
OPSLAKE       2354.0      771.7    3.05   0.0041

Residual standard error: 8300 on 39 degrees of freedom
Multiple R-Squared: 0.902,
F-statistic:  119 on 3 and 39 DF,  p-value: <2e-16
```

The variable *OPBPC* is unimportant after the others because of its large p-value, in spite of its high correlation with the response of more than 0.86. ∎

3.5.2. Obtain the overall test if the hypothesis that *BSAAM* is independent of the three terms versus the alternative that it is not independent of them, and summarize your results.

Solution:

```
Analysis of Variance Table

Model 1: BSAAM ~ 1
Model 2: BSAAM ~ OPBPC + OPRC + OPSLAKE
  Res.Df         RSS Df   Sum of Sq         F     Pr(>F)
1     42 27351018334
2     39  2689509185  3 24661509149 119.20376 < 2.22e-16
```

The tiny p-value suggests very strong evidence against the null hypothesis. ∎

3.5.3. Obtain three analysis of variance tables fitting in the order (*OPBPC*, *OPRC* and *OPSLAKE*), then (*OPBPC*, *OPSLAKE* and *OPRC*), and finally (*OPSLAKE*, *OPRC* and *OPBPC*). Explain the resulting tables, and discuss in particular any apparent inconsistencies. Which F-tests in the Anova tables are equivalent to t-tests in the regression output?

Solution:

```
Response: BSAAM
          Df   Sum Sq  Mean Sq F value  Pr(>F)
OPBPC      1 2.15e+10 2.15e+10   311.2 < 2e-16
```

```
OPRC       1 2.56e+09 2.56e+09    37.1 3.8e-07
OPSLAKE    1 6.42e+08 6.42e+08     9.3  0.0041
Residuals 39 2.69e+09 6.90e+07
> anova(m2)
Analysis of Variance Table

Response: BSAAM
          Df   Sum Sq  Mean Sq F value Pr(>F)
OPSLAKE    1 2.41e+10 2.41e+10  349.28 <2e-16
OPRC       1 5.74e+08 5.74e+08    8.32 0.0063
OPBPC      1 4.51e+05 4.51e+05    0.01 0.9360
Residuals 39 2.69e+09 6.90e+07
> anova(m3)
Analysis of Variance Table

Response: BSAAM
          Df   Sum Sq  Mean Sq F value Pr(>F)
OPSLAKE    1 2.41e+10 2.41e+10  349.28 <2e-16
OPBPC      1 5.64e+04 5.64e+04 0.00082 0.9773
OPRC       1 5.74e+08 5.74e+08    8.33 0.0063
Residuals 39 2.69e+09 6.90e+07
```

The key difference is that *OPBPC* is unimportant adjusted for the others, but significant ignoring the others. The F for each term fit last is equivalent to the t for that term in the regression output. ∎

3.5.4. Using the output from the last problem, test the hypothesis that the coefficients for both *OPRC* and *OPBPC* are both zero against the alternative that they are not both zero.

Solution:

```
Analysis of Variance Table

Model 1: BSAAM ~ OPSLAKE
Model 2: BSAAM ~ OPBPC + OPRC + OPSLAKE
  Res.Df      RSS Df Sum of Sq    F Pr(>F)
1     41 3.26e+09
2     39 2.69e+09  2  5.75e+08 4.17  0.023
```

The p-value of about 0.02 suggest modest evidence against the null hypothesis. At least one of these two terms is likely to have a nonzero coefficient. ∎

4
Drawing conclusions

Problems

4.1 Fit the regression of *Soma* on *AVE*, *LIN* and *QUAD* as defined in Section 4.1 for the girls in the Berkeley Guidance Study data, and compare to the results in Section 4.1.

Solution:

```
> summary(m1)   Mean function 1 from Table 4.1
Call:
lm(formula = Soma ~ WT2 + WT9 + WT18)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5921     0.6742    2.36   0.0212
WT2          -0.1156     0.0617   -1.87   0.0653
WT9           0.0562     0.0201    2.80   0.0068
WT18          0.0483     0.0106    4.56  2.3e-05

Residual standard error: 0.543 on 66 degrees of freedom
Multiple R-Squared: 0.566
F-statistic: 28.7 on 3 and 66 DF,  p-value: 5.5e-12

> summary(m2)    Mean function with transformed terms

Call:
lm(formula = Soma ~ AVE + LIN + QUAD)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.4030 -0.2608 -0.0318  0.3801  1.4409

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.5921     0.6742    2.36   0.0212
AVE          -0.0111     0.0519   -0.21   0.8321
LIN          -0.0820     0.0304   -2.70   0.0089
QUAD         -0.0300     0.0162   -1.85   0.0688

Residual standard error: 0.543 on 66 degrees of freedom
Multiple R-Squared: 0.566,      Adjusted R-squared: 0.546
F-statistic: 28.7 on 3 and 66 DF,  p-value: 5.5e-12
```

(1) All summary statistics are identical. (2) All residuals are identical. (3) Intercepts are the same. The mean function for the first model is

$$\mathrm{E}(Soma|W) = \beta_0 + \beta_1\, WT2 + \beta_2\, WT9 + \beta_3\, WT18$$

Substituting the definitions of AVE, LIN and $QUAD$, the mean function for the second model is

$$\begin{aligned}
\mathrm{E}(Soma|W) &= \eta_0 + \eta_1 AVE + \eta_2 LIN + \eta_3 QUAD \\
&= \eta_0 + \eta_1(WT2 + WT9 + WT18)/3 \\
&\quad +\eta_2(WT2 - WT18) + \eta_3(WT2 - 2\,WT9 + WT18) \\
&= \eta_0 + (\eta_1/3 + \eta_2 + \eta_3)\, WT2 + (\eta_1/3 - 2\eta_3)\, WT9 \\
&\quad +(\eta_1/3 - \eta_2 + \eta_3)\, WT18
\end{aligned}$$

which shows the relationships between the βs and the ηs (for example, $\hat{\beta}_1 = \hat{\eta}_1/3 + \hat{\eta}_2 + \hat{\eta}_3$). The interpretation in the transformed scale may be a bit easier, as only the linear trend has a small p-value, so we might be willing to describe the change in $Soma$ over time as increasing by the same amount each year. ∎

4.2

4.2.1. Starting with (4.10), we can write

$$y_i = \mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x_i - \mu_x) + \varepsilon_i$$

Ignoring the error term ε_i, solve this equation for x_i as a function of y_i and the parameters.

Solution:

$$x_i = \mu_x + \frac{1}{\rho_{xy}}\frac{\sigma_x}{\sigma_y}(y_i - \mu_y)$$

This is undefined if $\rho_{xy} = 0$. ∎

4.2.2. Find the conditional distribution of $x_i|y_i$. Under what conditions is the equation you obtained in Problem 4.2.1, which is computed by inverting the regression of y on x, is the same as the regression of x on y?

Solution: Simply reverse the role of x and y in (4.10) to get

$$x_i|y_i \sim \mathrm{N}\left(\mu_x + \rho_{xy}\frac{\sigma_x}{\sigma_y}(y_i - \mu_y), \sigma_y^2(1-\rho_{xy}^2)\right)$$

These two are the same if and only if the correlation is equal to plus or minus one. In general there are two regressions. ∎

4.3 For the transactions data described in Section 4.6.1, define $A = (T_1 + T_2)/2$ to be the average transaction time, and $D = T_1 - T_2$, and fit the following four mean functions

$$\begin{aligned}
M1 : \mathrm{E}(Y|T_1, T_2) &= \beta_{01} + \beta_{11}T_1 + \beta_{21}T_2 \\
M2 : \mathrm{E}(Y|T_1, T_2) &= \beta_{02} + \beta_{32}A + \beta_{42}D \\
M3 : \mathrm{E}(Y|T_1, T_2) &= \beta_{03} + \beta_{23}T_2 + \beta_{43}D \\
M4 : \mathrm{E}(Y|T_1, T_2) &= \beta_{04} + \beta_{14}T_1 + \beta_{24}T_2 + \beta_{34}A + \beta_{44}D
\end{aligned}$$

4.3.1. In the fit of M4, some of the coefficients estimates are labelled as either "aliased" or as missing. Explain what this means.

Solution: Since A and D are exact linear combinations of T_1 and T_2, only two of the four terms added after the intercept can be estimated. ∎

4.3.2. What aspects of the fitted regressions are the same? What is different?

Solution:

	Mean function for equation			
Term	(M1)	(M2)	(M3)	(M4)
Constant	144.37	144.37	144.37	144.37
T_1	5.46			5.46
T_2	2.03		7.50	2.03
A		7.50		aliased
D		1.71	5.46	aliased
$\hat{\sigma} = 1142.56$, $R^2 = 0.909$				

The intercept, $\hat{\sigma}$ and R^2 are the same for each fit. The estimates for T_1 and T_2 is the same in M1 and M4, since after deleting the aliased variables, the two are really the identical fit.∎

4.3.3. Why is the estimate for T_2 different in M1 and M3?

Solution: In M1, the estimate is the change in the response for a unit change in T_2 with T_1 fixed. In M3, the estimate is the change in Y for unit change in T_2 when $D = T_1 - T_2$ is fixed. The only way that T_1 can be increased by one unit with D fixed is to increase T_2 by one unit as well, so the coefficient for T_2 in M3 is the sum of the coefficients for T_1 and T_2 in M1. ∎

4.4 Interpreting coefficients with logarithms

4.4.1. For the simple regression with mean function $\mathrm{E}(\log(Y)|X = x) = \beta_0 + \beta_1 \log(x)$, provide an interpretation for β_1 as a rate of change in Y for a small change in x.

Solution: Write the approximate mean function

$$\mathrm{E}(Y|X = x) \approx e^{\beta_0} x^{\beta_1}$$

and we get

$$\frac{d\mathrm{E}(Y|X = x)/dx}{\mathrm{E}(Y|X = x)} = \beta_1/x$$

so the rate of change per unit of Y decreases inversely with x. ∎

4.4.2. Show that the results of Section 4.1.7 do not depend on the base of the logarithms.

Solution: Changing the base of logs would multiply the equations shown by a constant, but the value of β_1 will be divided by the same constant, resulting in no effect on the results. ∎

4.5 Use the bootstrap to estimate confidence intervals of the coefficients in the fuel data.

Solution: Here is output using the `bootCase` command in the `alr3` library for R:

```
> m1 <- lm(Fuel~Tax+Dlic+Income+logMiles,f)
> ans <- bootCase(m1,f=coef,B=999)
> # print percentile confidence intervals:
> print(results <-t(apply(ans,2,function(x)
                    c(mean(x),quantile(x, c(.025,.975))))))
                          2.5%     97.5%
(Intercept) 200.35788 -124.452052 730.65110
Tax          -4.55119  -10.711211   0.57699
Dlic          0.45132    0.098412   0.77125
Income       -6.19183  -10.050968  -2.78302
logMiles     17.35175   -3.396241  32.50824
> # compare to normal theory
> confint(m1)
             Coef est      Lower     Upper
(Intercept) 154.19284 -238.13291 546.51860
Tax          -4.22798   -8.31441  -0.14156
Dlic          0.47187    0.21319   0.73056
Income       -6.13533  -10.55089  -1.71978
logMiles     18.54527    5.51746  31.57309
```

The is some disagreement between normal theory in the bootstrap for most of the coefficients, particularly for the intercept. The effect of *Tax* is apparent from normal theory, but not from the bootstrap. ∎

4.6 Windmill data For the windmill data in the data file `wm1.txt` discussed in Problem 2.13, page 30, use $B = 999$ replications of the bootstrap to estimate a 95% confidence interval for the long-term average wind speed at the

candidate site and compare this to the prediction interval in Problem 2.13.5. See the comment at the end of Problem 2.13.4 to justify using a bootstrap confidence interval for the mean as a prediction interval for the long-term mean.

Solution: This requires a straightforward application of the bootstrap as outlined in the text. Here is an R program that will carry out the bootstrap for this problem.

```
> m1 <- lm(CSpd ~ RSpd, wm1)
> f <- function(m) predict(m, data.frame(RSpd=7.4285))
> results <- bootCase(m1, f=f, B=999)
> quantile(results, c(.025, .975))
    2.5%    97.5%
8.613350 8.898304
> data.frame(Mean=mean(results), SD=sd(results))
      Mean         SD
1 8.750454 0.07293487
```

The command `bootCase` is used a little differently here. On each of the B bootstraps, the function `f` will be applied to the regression using the bootstrap sample. In this case, `f` simply returns the fitted wind speed at the long term reference wind speed of 7.4285. The remainder of the code shown is similar to the code used previously.

For one realization of this bootstrap, we got the interval from 8.608 to 8.895, with average prediction 8.755. This compares to the interval 8.609 to 8.902 from normal theory, with average prediction 8.755. Normal theory and the bootstrap agree almost perfectly. ∎

4.7 Suppose we fit a regression with the true mean function

$$\mathrm{E}(Y|X_1 = x_1, X_2 = x_2) = 3 + 4x_1 + 2x_2$$

Provide conditions under which the mean function for $\mathrm{E}(Y|X_1 = x_1)$ is linear but has a negative coefficient for x_1.

Solution: Using (4.4),

$$\mathrm{E}(Y|X_1 = x_1) = 3 + 4x_1 + 2\mathrm{E}(X_2|X_1 = x_1)$$

This mean function will be linear if $\mathrm{E}(X_2|X_1 = x_1) = \gamma_0 + \gamma_1 x_1$, and then

$$\begin{aligned}\mathrm{E}(Y|X_1 = x_1) &= 3 + 4x_1 + 2(\gamma_0 + \gamma_1 x_1)\\ &= (3 + \gamma_0) + (4 + 2\gamma_1)x_1\end{aligned}$$

and the coefficient for $\mathbf{x}_1$ will be negative if $4 + 2\gamma_1 < 0$ or if $\gamma_1 < -2$. ∎

4.8 In a study of faculty salaries in a small college in the midwest, a linear regression model was fit, giving, the fitted mean function

$$\widehat{\mathrm{E}(Salary|Sex)} = 24697 - 3340Sex \tag{4.18}$$

where *Sex* equals one if the faculty member was female and zero if male. The response *Salary* is measured in dollars (the data are from the 1970s).

4.8.1. Give a sentence that describes the meaning of the two estimated coefficients.

Solution: The intercept is \$24697, which is the estimated salary for a male faculty members. Female faculty members have expected salaries that are \$3340 lower. ∎

4.8.2. An alternative mean function fit to these data with an additional term, *Years*, the number of years employed at this college, gives the estimated mean function

$$\widehat{\mathrm{E}(Salary|Sex}, Years) = 18065 + 201\,Sex + 759\,Years \tag{4.19}$$

The important difference between these two mean functions is that the coefficient for *Sex* has changed signs. Using the results of this chapter, explain how this could happen. (Data consistent with these equations are presented in Problem 6.13).

Solution: Using Section 4.1.6, given (4.2), we get to (4.1) by replacing *Years* by the conditional expectation of *Years* given the other three terms,

$$\widehat{\mathrm{E}(Salary|Sex}) = 18065 + 201\,Sex + 759\mathrm{E}(\,Years|Sex)$$

Equating the right side of this last equation with the right side of (4.2), we can solve for $\mathrm{E}(\,Years|Sex)$,

$$\begin{aligned}\mathrm{E}(\,Years|Sex) &= \frac{24697-18065}{759} - \frac{3340+201}{759}Sex \\ &\approx 8.7 - 4.7\,Sex\end{aligned}$$

The two mean functions are consistent if the average male has about 8.7 years of experience but the average female has only about $8.7 - 4.7 = 4.0$ years of experience. ∎

4.9 Sleep data

4.9.1. For the sleep data described in Section 4.5, describe conditions under which the missing at random assumption is reasonable. In this case, deleting the partially observed species and analyzing the complete data can make sense.

Solution: MAR will be reasonable if the chance of a value being missing does not depend on the value that would be observed. For example, if *SWS* is not observed because the experimenters who collected the data were not interested in it, then the value is MAR. ∎

4.9.2. Describe conditions under which the missing at random assumption for the sleep data is not reasonable. In this case, deleting partially observed species can change the inferences by changing the definition of the sampled population.

Solution: If it is not observed because the value is very short and hard to measure, the MAR fails. ∎

4.9.3. Suppose that the sleep data were fully observed, meaning that values for all the variables were available for all 62 species. Assuming that there are more than 62 species of mammals, provide a situation where examining the missing at random assumption could still be important.

Solution: If the goal is inference to the population of mammal species, then if we failed to observe species because they had unusual sleep patterns, then the MAR assumption fails and inference to the population of species from the data is questionable. ∎

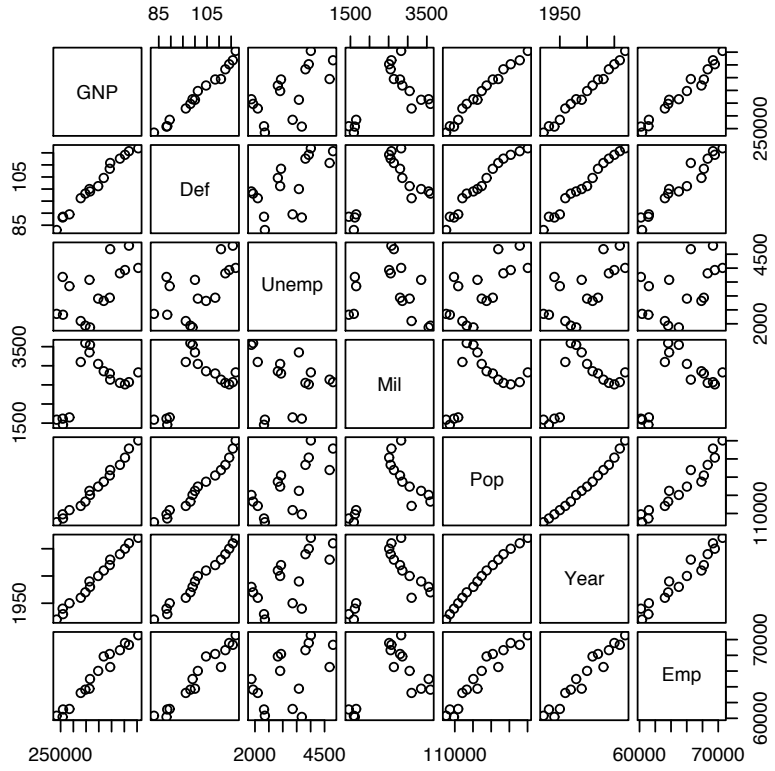
4.10 The data given in `longley.txt` were first given by Longley (1967) to demonstrate inadequacies of regression computer programs then available. The variables are:

GNP.deflator	=	GNP price deflator, in percent
GNP	=	GNP, in millions of dollars
Unemployed	=	Unemployment, in thousands of persons
Armed.Forces	=	Size of armed forces, in thousands
Population	=	Population 14 years of age and over, in thousands
Employed	=	Total derived employment in thousands the response
Year	=	Year

(The variable names are incorrect in the text book; the names above are correct.)

4.10.1. Draw the scatterplot matrix for these data excluding *Year*, and explain from the plot why this might be a good example to illustrate numerical problems of regression programs. (Hint: Numerical problems arise through rounding errors, and these are most likely to occur when terms in the regression model are very highly correlated.)

Solution: Almost all the variables are almost perfectly and linearly related with *Year*. The exceptions are *Unemployed*, for which the linear increase is more variable, and *Armed.Forces*, which was low in the period immediately after World War II, and quickly increased during the Korean War and stayed at a high level during the succeeding years. The very high correlations between most of the predictors suggest that all coefficients will be poorly eliminated, as we will essentially be explaining the same variability over and over again.



4.10.2. Fit the regression of *Employed* on the others excluding *Year*.
Solution:

```
> summary(m1 <- lm(Employed ~ ., longley))

Call:
lm(formula = Employed ~ ., data = longley)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.482e+03  8.904e+02  -3.911 0.003560 **
GNP.deflator  1.506e-02  8.492e-02   0.177 0.863141
GNP          -3.582e-02  3.349e-02  -1.070 0.312681
Unemployed   -2.020e-02  4.884e-03  -4.136 0.002535 **
Armed.Forces -1.033e-02  2.143e-03  -4.822 0.000944 ***
Population   -5.110e-02  2.261e-01  -0.226 0.826212
Year          1.829e+00  4.555e-01   4.016 0.003037 **
---
```

```
Residual standard error: 0.3049 on 9 degrees of freedom
Multiple R-squared: 0.9955,     Adjusted R-squared: 0.9925
F-statistic: 330.3 on 6 and 9 DF,  p-value: 4.984e-10
```

R^2 is nearly one. ∎

4.10.3. Suppose that the values given in this example were only accurate to three significant figures (two figures for *Def*). The effects of measurement errors can be assessed using a simulation study in which we add uniform random values to the observed values, and recompute estimates for each simulation. For example, *Unemployed* for 1947 is given as 2356, which corresponds to 2,356,000. If we assume only three significant figures, we only believe the first three digits. In the simulation we would replace 2356 by $2356 + u$, where u is a uniform random number between -5 and $+5$. Repeat the simulation 1000 times, and on each simulation compute the coefficient estimates. Compare the standard deviation of the coefficient estimates from the simulation to the coefficient standard errors from the regression on the unperturbed data. If the standard deviations in the simulation are as large or larger than the standard errors, we would have evidence that rounding would have important impact on results.

Solution:

```
> #longley simulation experiment, assuming uniform rounding
> #error on the last digit
> dim(l)  # get the number of rows and columns in the data.
[1] 16  7
> # write a function that will add random rounding error to the
> # observed data.  Don't add error to the response or to Year
> # The function creates  a matrix of uniform (-.5,.5) random numbers,
> # and then multiplies by a diagonal matrix of scale factors to scale
> # random numbers to the right size for each predictor.
> perturb.data <- function(data)
+    data + matrix( runif(16*7)-.5, nrow=16) %*%
+                                        diag(c(1000,1,10,10,100,0,0))
> # do the simulation
> simulate <- function(m=m1,data=l,B=999)
+   ans <- NULL
+   for (j in 1:B)
+     ans <- rbind(ans,coef(update(m1,data=perturb.data(data))))
+   ans
> # set the seed, so results can be reproduced exactly
> set.seed(1044)
> ans <- simulate()
> apply(ans,2,mean)   Simulation means
(Intercept)         Def          GNP       Unemp          Mil          Pop
 9.1553e+04 -4.4969e+01  7.1147e-02 -4.1427e-01 -5.6169e-01 -3.9566e-01
> apply(ans,2,sd)    Simulation sd's
(Intercept)         Def          GNP       Unemp          Mil          Pop
```

```
 7.5224e+03  3.4361e+01  7.3671e-03  8.8592e-02  2.7969e-02  6.2434e-02
> apply(ans,2,function(a) quantile(a, c(.025,.975)))
      (Intercept)      Def      GNP    Unemp      Mil      Pop
2.5%        77311 -115.460 0.057436 -0.57952 -0.61146 -0.52648
97.5%      106789   20.589 0.086156 -0.23068 -0.49983 -0.27668
> apply(ans,2,sd)/sqrt(diag(vcov(m1)))   Ratios
(Intercept)         Def         GNP       Unemp         Mil         Pop
   0.213891    0.259820    0.232154    0.202019    0.098547    0.189043
```

All the ratios except for *Mil* are close to .2, suggesting that the variation due to the rounding as about 20% of the unexplained variation. If the digits beyond the third can't be believed, then neither can the regression coefficients. ∎

5

Weights, Lack of Fit, and More

Problems

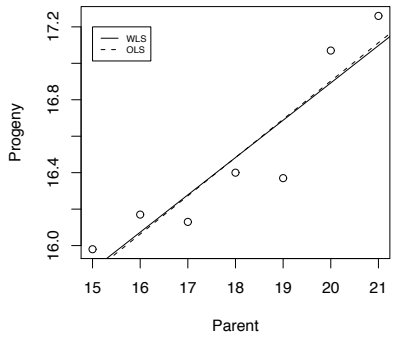
5.1 Galton's sweet peas Many of the ideas of regression first appeared in the work of Sir Francis Galton on the inheritance of characteristics from one generation to the next. In a paper on "Typical Laws of Heredity," delivered to the Royal Institution on February 9, 1877, Galton discussed some experiments on sweet peas. By comparing the sweet peas produced by parent plants to those produced by offspring plants, he could observe inheritance from one generation to the next. Galton categorized parent plants according to the typical diameter of the peas they produced. For seven size classes from 0.15 to 0.21 inches he arranged for each of nine of his friends to grow ten plants from seed in each size class; however, two of the crops were total failures. A summary of Galton's data was later published by Karl Pearson (1930) (see Table 5.8 and the data file `galtonpeas.txt`). Only average diameter and standard deviation of the offspring peas are given by Pearson; sample sizes are unknown.

5.1.1. Draw the scatterplot of *Progeny* versus *Parent*.

Solution:

Table 5.8 Galton's peas data.

Parent diameter (.01 in)	*Progeny* diameter (.01 in)	SD
21	17.26	1.988
20	17.07	1.938
19	16.37	1.896
18	16.40	2.037
17	16.13	1.654
16	16.17	1.594
15	15.98	1.763



■

5.1.2. Assuming that the standard deviations given are population values, compute the weighted regression of *Progeny* on *Parent*. Draw the fitted mean function on your scatterplot.

Solution:

```
> summary(m1)
lm(formula = Progeny ~ Parent, weights = 1/SD^2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.7964     0.6811   18.79  7.9e-06
Parent        0.2048     0.0382    5.37    0.003

Residual standard error: 0.11 on 5 degrees of freedom
Multiple R-Squared: 0.852
F-statistic: 28.8 on 1 and 5 DF,  p-value: 0.00302

Analysis of Variance Table
Response: Progeny
          Df Sum Sq Mean Sq F value Pr(>F)
```

```
Parent     1  0.349   0.349    28.8  0.003
Residuals  5  0.061   0.012
```

In addition, the OLS line is virtually identical to the WLS line. ∎

5.1.3. Galton wanted to know if characteristics of the parent plant such as size were passed on to the offspring plants. In fitting the regression, a parameter value of $\beta_1 = 1$ would correspond to perfect inheritance, while $\beta_1 < 1$ would suggest that the offspring are "reverting" toward "what may be roughly and perhaps fairly described as the average ancestral type." (The substitution of "regression" for "reversion" was probably due to Galton in 1885.) Test the hypothesis that $\beta_1 = 1$ versus the alternative that $\beta_1 < 1$.

Solution:

```
> (.2084-1)/.0382     t statistic
[1] -20.723
> pt(-20.723,5)
[1] 2.4233e-06      significance level, one sided
```

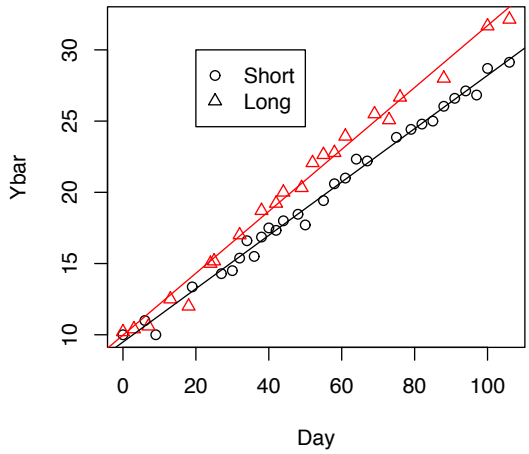
∎

5.1.4. In his experiments, Galton took the average size of all peas produced by a plant to determine the size class of the parental plant. Yet for seeds to represent that plant and produce offspring, Galton chose seeds that were as close to the overall average size as possible. Thus for a small plant, the exceptional large seed was chosen as a representative, while larger more robust plants were represented by relatively smaller seeds. What effects would you expect these experimental biases to have on (1) estimation of the intercept and slope and (2) estimates of error?

Solution: This should decrease the slope, and it could increase variances, making differences more difficult to detect. ∎

5.2 Apple shoots Apply the analysis of Section 5.3 to the data on short shoots in Table 5.6.

Solution:




```
Call:
lm(formula = ybar ~ Day, subset = Type == 1, weights = n)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.97375    0.31427    31.7   <2e-16
Day          0.21733    0.00534    40.7   <2e-16

Residual standard error: 1.93 on 20 degrees of freedom
Multiple R-Squared: 0.988
F-statistic: 1.66e+03 on 1 and 20 DF,  p-value: <2e-16
```

The visual impression is that the two groups of shoots have the same intercept (start the same place on *Day* zero), but different slopes, with short shoots increasing more slowly.

However, both groups show slight lack-of-fit:

```
            Short      Long
SSpe    246.73920 255.12150
dfpe    292.00000 167.00000
MSpe      0.84500   1.52767
Flof      2.54527   2.43482
pvalues   0.00004   0.00112
```

■

5.3 Nonparametric lack of fit The lack of fit tests in Sections 5.2–5.3 require either a known value for σ^2 or repeated observations for a given value of the predictor that can be used obtain a model-free, or pure-error, estimate of

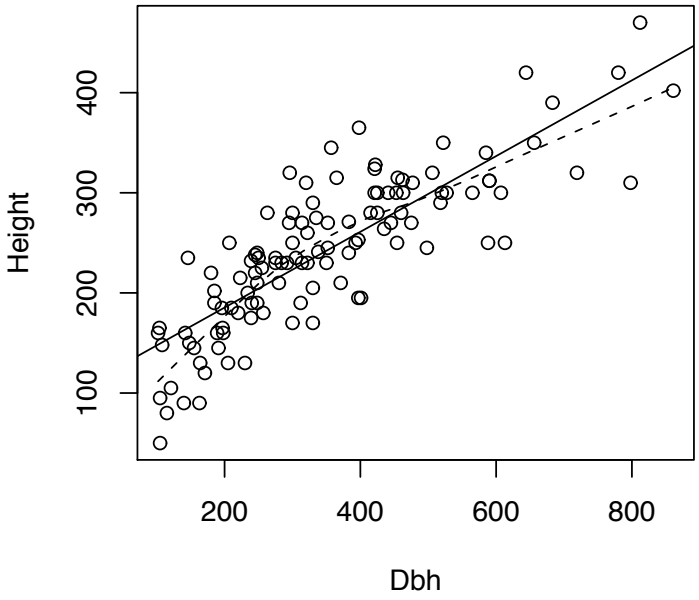


Fig. 5.4 *Height* versus *Dbh* for the Upper Flat Creek grand fir data. The solid line is the OLS fit. The dashed line is the *loess* fit with smoothing parameter 2/3, using one iteration and using local linear fitting.

σ^2. Loader (2004, Sec. 4.3) describes a lack of fit test that can be used without repeated observations or prior knowledge of σ^2 based on comparing the fit of the parametric model to the fit of a smoother. For illustration, consider Figure 5.4, which uses data that will be described later in this problem. For each data point, we can find the fitted value $\hat{y}_i$ from the parametric fit, which is just a point on the solid line, and $\tilde{y}_i$, the fitted value from the smoother, which is a point on the dashed line. If the parametric model is appropriate for the data, then the differences $(\hat{y}_i - \tilde{y}_i)$ should all be relatively small. A suggested test statistic is based on looking at the squared differences, and then dividing by an estimate of σ^2,

$$G = \frac{\sum_{i=1}^{n}(\hat{y}_i - \tilde{y}_i)^2}{\hat{\sigma}^2} \tag{5.23}$$

where $\hat{\sigma}^2$ is the estimate of variance from the parametric fit. Large values of G provide evidence against the NH that the parametric mean function matches

the data. Loader (2004) provides an approximation to the distribution of G, and also a bootstrap for computing an approximate significance level for a test based on G. In this problem, we will present the bootstrap.

5.3.1. The appropriate bootstrap algorithm is a little different from what we have seen before, and uses a *parametric bootstrap*. It works as follows:

1. Fit the parametric and smooth regression to the data, and compute G from (5.23). Save the residuals, $\hat{e}_i = y_i - \hat{y}_i$ from the parametric fit.

2. Obtain a bootstrap sample $\hat{e}_1^*, \ldots, \hat{e}_n^*$ by sampling with replacement from $\hat{e}_1, \ldots, \hat{e}_n$. Some residuals will appear in the sample many times, some not at all.

3. Given the bootstrap residuals, compute a bootstrap response $\mathbf{Y}^*$ with elements $y_i^* = \hat{y}_i + \hat{e}_i^*$. Use the original predictors unchanged in every bootstrap sample. Obtain the parametric and nonparametric fitted values with the response $\mathbf{Y}^*$, and then compute G from (5.23).

4. Repeat steps 2–3 B times, perhaps $B = 999$.

5. The significance level of the test is estimated to be the fraction of bootstrap samples that give a value of (5.23) that exceed the observed G.

The important problem of selecting a smoothing parameter for the smoother has been ignored. If the *loess* smoother is used, selecting the smoothing parameter to be 2/3 is a reasonable default, and statistical packages may include methods to choose a smoothing parameter. See Simonoff (1996), Bowman and Azzalini (1997), and Loader (2004) for more discursion of this issue.

Write a computer program that implements this algorithm for regression with one predictor.

Solution: Here is a program that works in R/S-plus:

```
nplof <- function(x,y,B=999,...){
   compute.G <- function(yhat,ytilde){
      sum( (yhat-ytilde)^2)/ (sum(yhat^2)/(length(yhat)-2))}
   smooth.fit <- function(x,y,span=2/3,degree=1,...){
      predict(loess(y~x,span=span,degree=degree,...))}
   m <- lm(y ~ x, ...)
   r <- residuals(m)
   fit <- predict(m)
   ans <- compute.G(fit,smooth.fit(x,y,...))
   n <- length(r)
   for (j in 1:B){
    sel <- sample(n,replace=TRUE)  # sample with replacement
    ystar <- fit + r[sel]
    ans <- c(ans,compute.G(predict(lm(ystar~x,...)),
                           smooth.fit(x,ystar,...)))}
   ans}
```

The function `nplof` has two required arguments x and y. The three dots "..." means that other arguments can be added to the function, and these will be passed to `lm`, which computes the simple linear regression and to `loess`, which computes the smoother. The local function `compute.G` computes G given by (5.23), and `smooth.fit` uses `loess` as a smoother to get the $\tilde{y}$. Default values for the span and for the degree are set in the definition of this function that are different from the "factory" defaults. In R/S-plus there are many other options for smoothers, and for selecting a smoothing parameter, and any of these could be substituted here. The function `nplof` first fits the parametric simple regression model using `lm`, and saves the residuals and the fitted values. G is computed for the original data. The `for` loop computes the bootstrap. `sel` samples the case numbers 1 to n with replacement, and in the next line y_i^* is computed. The function returns all $B + 1$ values of G, and these can be plotted or otherwise summarized, as illustrated in the solution to the next subproblem. ∎

5.3.2. The data file `ufcgf.txt` gives the diameter *Dbh* in millimeters at 137 cm perpendicular to the bole, and the *Height* of the tree in decimeters for a sample of Grand fir trees at Upper Flat Creek, Idaho, in 1991, courtesy of Andrew Robinson. Also included in the file are the *Plot* number, the *Tree* number in a plot, and the *Species*, which is always "GF" for these data. Use the computer program you wrote in the last subproblem to test for lack of fit of the simple linear regression mean function for the regression of *Height* on *Dbh*.

Solution:

```
> attach(ufcgf)
> set.seed(10131985)  # this allows reproducing this output
> ans <- nplof(Dbh,Height)
> print(paste("Statistic =", round(ans[1],3), "Significance level =",
+         round( (1+length(which(ans>ans[1])))/length(ans),3)))
[1] "Statistic = 0.378 Significance level = 0.001"
```

We used `set.seed` to make the results given here reproducible. The value of the statistic is .378 and the significance level is 0.001, suggesting that the straight-line model is clearly inadequate for a growth model for these trees. ∎

5.4 An F-test In simple regression derive an explicit formula for the F-test of

$$\begin{array}{ll} \text{NH:} & \mathrm{E}(Y|X=x) = x \qquad\qquad (\beta_0 = 0, \beta_1 = 1) \\ \text{AH:} & \mathrm{E}(Y|X=x) = \beta_0 + \beta_1 x \end{array}$$

Solution: Under the null hypothesis, the i-th fitted value is just x_i, and so $RSS_{NH} = \sum(y_i - x_i)^2$, with n df. The alternative hypothesis is the usual simple linear regression model, so the F test is

$$F = \frac{(\sum(y_i - x_i)^2 - RSS)/2}{\hat{\sigma}^2}$$

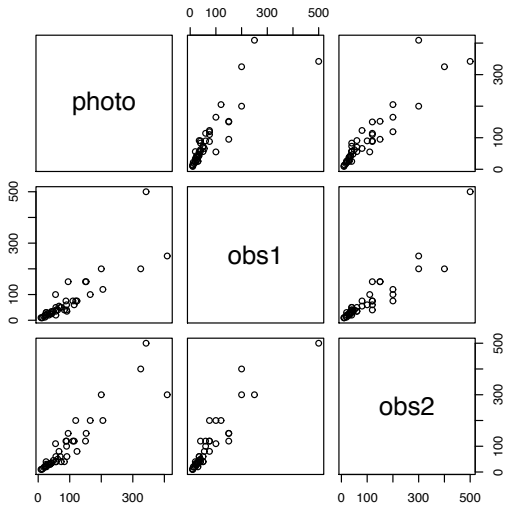
which is distributed as $F(2, n-2)$ under the null hypothesis. ∎

5.5 Snow geese Aerial surveys sometimes rely on visual methods to estimate the number of animals in an area. For example, to study snow geese in their summer range areas west of Hudson Bay in Canada, small aircraft were used to fly over the range and, when a flock of geese was spotted, an experienced person estimated the number of geese in the flock.

To investigate the reliability of this method of counting, an experiment was conducted in which an airplane carrying two observers flew over $n = 45$ flocks, and each observer made an independent estimate of the number of birds in each flock. Also, a photograph of the flock was taken so that a more or less exact count of the number of birds in the flock could be made. The resulting data are given in the data file `snowgeese.txt` (Cook and Jacobson, 1978). The three variables in the data set are *Photo* = photo count, *Obs1* = aerial count by observer one and *Obs2* = aerial count by observer 2.

5.5.1. Draw scatterplot matrix of three variables. Do these graphs suggest that a linear regression model might be appropriate for the regression of *Photo* on either of the observer counts, or on both of the observer counts? Why or why not? For the simple regression model of *Photo* on *Obs1*, what do the error terms measure? Why is it appropriate to fit the regression of *Photo* on *Obs1* rather than the regression of *Obs1* on *Photo*?

Solution:



A straight-line mean function seems plausible, but variance is clearly not constant, but rather is (much) larger for large flocks than for small ones. ∎

5.5.2. Compute the regression of *Photo* on *Obs1* using OLS, and test the hypothesis of Problem 5.4. State in words the meaning of this hypothesis, and the result of the test. Is the observer reliable (you must define reliable)? Summarize your results.

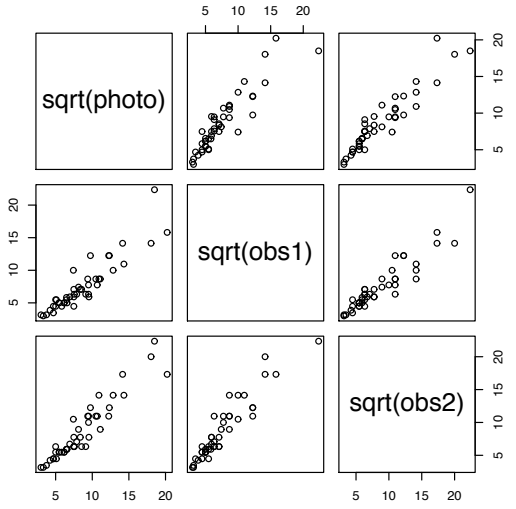
Solution:

```
> m1 <- lm(photo ~ obs1, data = snow)
> print(RSS.m1 <- sum (residuals(m1,type="pearson")^2))
[1] 84790  # RSS from the simple linear regression model
> print(RSS.5.4 <- sum ( (photo-obs1)^2))
[1] 104390  # Model of Problem 5.4, RSS = sum ( (y-x)^2)
> print(F <- ((RSS.5.4 - RSS.m1)/2)/sigmaHat(m1)^2)
[1] 4.9699
> print(pvalue <- 1 - pf(F,2,m1$df))
[1] 0.011436
```

The significance level of the test is 0.01, and so we would have evidence against the mean function in Problem 5.4. ∎

5.5.3. Repeat Problem 5.5.2, except fit the regression of $Photo^{1/2}$ on $Obs1^{1/2}$. The square-root scale is used to stabilize the error variance.

Solution: We begin by drawing the scatterplot matrix, with all the variables in square root scale:




```
> m2 <- update(m1, sqrt(photo) ~ sqrt(obs1))
> print(RSS.m2 <- sum (residuals(m2,type="pearson")^2))
[1] 114.47
> # Model of Problem 5.4, RSS = sum ( (y-x)^2)
> print(RSS.5.4 <- sum ( (photo-obs1)^2))
[1] 104390
> print(F <- ((RSS.5.4 - RSS.m2)/2)/sigmaHat(m2)^2)
[1] 19586
> print(pvalue <- 1 - pf(F,2,m2$df))
[1] 0
```

While a straight-line model is visually more appealing in the square root scale, there is strong evidence in this scale against the hypothesis that the intercept is zero and the slope is one. ∎

5.5.4. Repeat Problem 5.5.2, except assume that the variance of an error is $obs1 \times \sigma^2$.

Solution: We need to compute the residual SS under both hypotheses assuming $obs1 \times \sigma^2$.

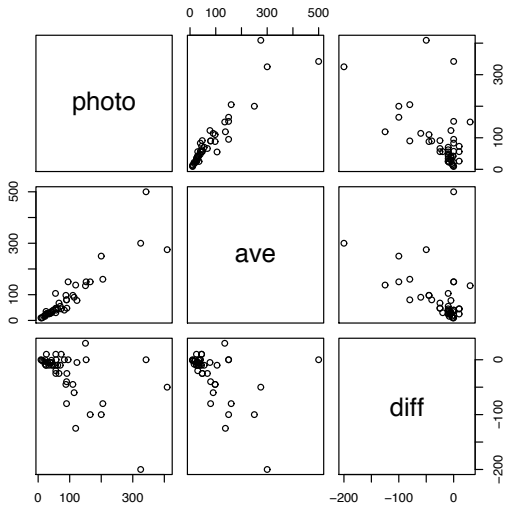
```
> m3 <- update(m1, weights = 1/obs1)
# gets the weights right
> print(RSS.m3 <- sum (residuals(m3,type="pearson")^2))
[1] 612.95
> # Model of Problem 5.4, RSS = sum ( (y-x)^2)
# include the weights here, too
> print(RSS.5.4 <- sum ((photo-obs1)^2/obs1))
[1] 891.03
> print(F <- ((RSS.5.4 - RSS.m3)/2)/sigmaHat(m3)^2)
[1] 9.7543
> print(pvalue <- 1 - pf(F,2,m3$df))
[1] 0.00032126
```

and so once again, the mean function is not acceptable. ∎

5.5.5. Do both observers combined do a better job at predicting *Photo* than either predictor separately? To answer this question, you may wish to look at the regression of *Photo* on both *Obs1* and *Obs2*. Since from the scatterplot matrix the two terms are highly correlated, interpretation of results might be a bit hard. An alternative is to replace *Obs1* and *Obs2* by $Average = (Obs1 + Obs2)/2$ and $Diff = Obs1 - Obs2$. The new terms have the same information as the observer counts, but they are much less correlated. You might also need to consider using WLS.

As a result of this experiment, the practice of using visual counts of flock size to determine population estimates was discontinued in favor of using photographs.

Solution: We again draw the scatterplot matrix:



Most of the differences are negative, and all of the large differences are negative and correspond to larger flocks. We learn immediately from the graph that the two observers are more likely to disagree with larger flocks than smaller ones, with one of the observers consistently higher than the other. Ignoring the few very large differences, there is little information in the two predictors beyond their average; a better way to look at this would be an added-variable plot.

It is clear from the graph that weights will be a good idea, and we use the average of *obs1* and *obs2* as weights.

```
> snow$ave <- (obs1 + obs2)/2
> snow$diff <- (obs1 - obs2)
> pairs(photo~ave+diff,data=snow)
> m4 <- lm(photo~ave+diff, data=snow, weights = 1/ave)
> anova(m4)
Analysis of Variance Table

Response: photo
          Df Sum Sq Mean Sq F value Pr(>F)
ave        1   2030    2030  254.36 <2e-16
diff       1      8       8    0.97   0.33
Residuals 42    335       8
```

This confirms that there is little to be gained beyond averaging the estimates by the two observers. ∎

5.6 Jevons' gold coins The data in this example are deduced from a diagram in a paper written by W. Stanley Jevons (1868), and provided by Stephen M. Stigler. In a study of coinage, Jevons weighed 274 gold sovereigns

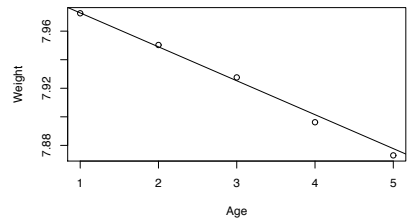
Table 5.9 Jevons gold coinage data

Age, decades	Sample size n	Average *Weight*	SD	Minimum *Weight*	Maximum *Weight*
1	123	7.9725	0.01409	7.900	7.999
2	78	7.9503	0.02272	7.892	7.993
3	32	7.9276	0.03426	7.848	7.984
4	17	7.8962	0.04057	7.827	7.965
5	24	7.873	0.05353	7.757	7.961

that he had collected from circulation in Manchester, England. For each coin, he recorded the weight after cleaning to the nearest .001 gram, and the date of issue. Table 5.9 lists the average, minimum and maximum weight for each age class. The age classes are coded 1 to 5, roughly corresponding to the age of the coin in decades. The standard weight of a gold sovereign was supposed to be 7.9876 grams; the minimum legal weight was 7.9379 grams. The data are given the file `jevons.txt`.

5.6.1. Draw a scatterplot of *Weight* versus *Age*, and comment on the applicability of the usual assumptions of the linear regression model. Also draw a scatterplot of SD versus *Age*, and summarize the information in this plot.

Solution:



The wear appears to be remarkably linear over time. The line is the WLS line with Weights n/SD^2. ∎

5.6.2. Since the numbers of coins n in each age class are all fairly large, it is reasonable to pretend that the variance of coin weight for each *Age* is well approximated by SD^2, and hence $\mathrm{Var}(\textit{Weight})$ is given by SD^2/n. Compute the implied WLS regression.

Solution:

```
> summary(m1)
Call:
lm(formula = Weight ~ Age, weights = n/SD^2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.99652    0.00132    6049    1e-11
Age         -0.02376    0.00088     -27  0.00011
```

```
Residual standard error: 0.555 on 3 degrees of freedom
Multiple R-Squared: 0.996
F-statistic:  729 on 1 and 3 DF,  p-value: 0.000111

> anova(m1)
Analysis of Variance Table

Response: Weight
          Df Sum Sq Mean Sq F value  Pr(>F)
Age        1  224.5   224.5     729 0.00011
Residuals  3    0.9     0.3
```

5.6.3. Compute a lack of fit test for the linear regression model, and summarize results.

Solution: Compare the RSS to the $\chi^2(3)$ distribution, to get a significance level of about 0.82. We have no evidence against the straight line mean function.

5.6.4. Is the fitted regression consistent with the known standard weight for a new coin?

Solution: This question is asking about the fitted value at $Age = 0$, so we need a confidence interval for the intercept:

```
> confint(m1)
             Coef est     Lower     Upper
(Intercept)  7.996522  7.992315  8.000729
Age         -0.023756 -0.026556 -0.020956
```

Since 7.9876 is not included in the 95% confidence interval for the mean at $Age = 0$, these results are a bit too high, and not consistent with the known standard weight. The computation of a fitted value for WLS is the same as the computation of the fitted value for OLS.

5.6.5. For previously unsampled coins of $Age = 1, 2, 3, 4, 5$, estimate the probability that the weight of the coin is less than the legal minimum. Hints: The standard error of prediction is a sum of two terms, the known variance of an unsampled coin of known Age, and the estimated variance of the fitted value for that Age. You should use the normal distribution rather than a t to get the probabilities.

Solution: The predictions are just the point on the line. We can compute the standard error of prediction as

$$\text{sepred}(\mathit{Weight}|\mathit{Age} = j) = \sqrt{\text{SD}_j^2 + \text{sefit}(\mathit{Weight}|\mathit{Age} = j)^2}$$

Here is the computation, using R:

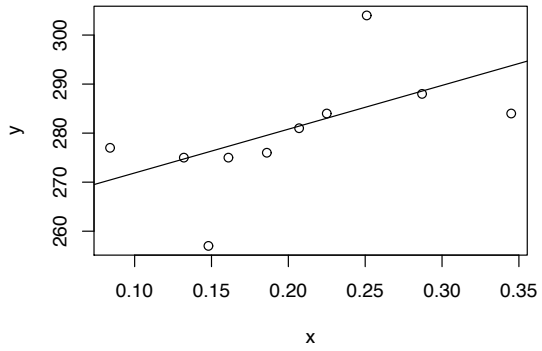
```
> ans <- predict(m1,data.frame(Age=1:5),se.fit=TRUE)
> se.pred <- sqrt(sd^2 + ans$se.fit^2)
```

```
> z <- (ans$fit - 7.9379) / se.pred
> prob <- 1-pnorm(z)
> ans1<- data.frame(Age,ans$fit,se.pred,z,pvalue)
> ans1
  Age ans.fit  se.pred        z      prob
1   1  7.9728 0.014106  2.47162 0.0067251
2   2  7.9490 0.022736  0.48863 0.3125519
3   3  7.9253 0.034297 -0.36874 0.6438400
4   4  7.9015 0.040643 -0.89568 0.8147894
5   5  7.8777 0.053631 -1.12173 0.8690108
```

5.7 The data file `physics1.txt` gives the results of the experiment described in Section 5.1.1, except in this case the input is the π^- meson as before, but the output is the π^+ meson.

Analyze these data following the analysis done in the text, and summarize your results.

Solution: As usual, begin with a graph:



Unlike the data in text, there are two (or more) points that fail to match the overall trend in the plot, although these values are not inconsistent given the size the measurement error. Here are the computations:

```
> m1 <- lm(y~x, weights=1/SD^2)
> summary(m1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   262.94       8.47   31.04  1.3e-09
x              89.30      43.08    2.07    0.072

Residual standard error: 1.11 on 8 degrees of freedom
Multiple R-Squared: 0.349
F-statistic:  4.3 on 1 and 8 DF,  p-value: 0.072
```

```
> anova(m1)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value Pr(>F)
x          1   5.26    5.26     4.3  0.072
Residuals  8   9.79    1.22
```

The lack-of-fit test is $X^2 = 9.79$ with 8 df, for a significance level near .32. ∎

6

Polynomials and Factors

Problems

6.1 Cake data The data for this example are in the data file `cakes.txt`.

6.1.1. Fit (6.4) and verify that the significance levels are all less than 0.005.

Solution:

```
> summary(m1 <- lm(Y ~ X1+X2+I(X1^2)+I(X2^2)+X1:X2, data=cake))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.20e+03   2.42e+02   -9.13  1.7e-05
X1           2.59e+01   4.66e+00    5.56  0.00053
X2           9.92e+00   1.17e+00    8.50  2.8e-05
I(X1^2)     -1.57e-01   3.94e-02   -3.98  0.00408
I(X2^2)     -1.20e-02   1.58e-03   -7.57  6.5e-05
X1:X2       -4.16e-02   1.07e-02   -3.88  0.00465

Residual standard error: 0.429 on 8 degrees of freedom
Multiple R-Squared: 0.949,      Adjusted R-squared: 0.917
F-statistic: 29.6 on 5 and 8 DF,  p-value: 5.86e-05
```

■

6.1.2. Estimate the optimal (X_1, X_2) combination $(\tilde{X}_1, \tilde{X}_2)$, and the standard errors of $\tilde{X}_1$ and $\tilde{X}_2$.

Solution: This is likely to be a very difficult problem for most students. Write the fitted mean function as

$$\mathrm{E}(Y|X) = b_0 + b_1X_1 + b_2X_2 + b_3X_1^2 + b_4X_2^2 + b_5X_1X_2$$

so the b's are the estimates from the table in the last sub-problem. Differentiate with respect to both X_1 and X_2:

$$\begin{aligned}\frac{d\mathrm{E}(Y|X)}{dX_1} &= b_1 + 2b_3X_1 + b_5X_2\\ \frac{d\mathrm{E}(Y|X)}{dX_2} &= b_2 + 2b_4X_2 + b_5X_1\end{aligned}$$

Set the two derivatives equal to zero, and then solve for X_1 and X_2,

$$\begin{aligned}\tilde{X}_1 &= \frac{b_2b_5 - 2b_1b_4}{4b_3b_4 - b_5^2}\\ \tilde{X}_2 &= \frac{b_1b_5 - 2b_2b_3}{4b_3b_4 - b_5^2}\end{aligned}$$

We can now use the deltaMethod to get estimates and standard errors (using R):

```
> summary(m1 <- lm(Y  ~  X1 +X2 +I(X1^2) +I(X2^2) +X1:X2, data=cakes))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.204e+03  2.416e+02  -9.125 1.67e-05 ***
X1           2.592e+01  4.659e+00   5.563 0.000533 ***
X2           9.918e+00  1.167e+00   8.502 2.81e-05 ***
I(X1^2)     -1.569e-01  3.945e-02  -3.977 0.004079 **
I(X2^2)     -1.195e-02  1.578e-03  -7.574 6.46e-05 ***
X1:X2       -4.163e-02  1.072e-02  -3.883 0.004654 **
---
Residual standard error: 0.4288 on 8 degrees of freedom
Multiple R-squared: 0.9487,     Adjusted R-squared: 0.9167
F-statistic:  29.6 on 5 and 8 DF,  p-value: 5.864e-05

> x1.max <- "(b2*b5 - 2*b1*b4)/(4*b3*b4 - b5^2)"
> x2.max <- "(b1*b5 - 2*b2*b3)/(4*b3*b4 - b5^2)"
> deltaMethod(m1, x1.max)
                                    Estimate        SE
(b2*b5 - 2*b1*b4)/(4*b3*b4 - b5^2) 35.82766 0.4330974
> deltaMethod(m1, x2.max)
                                    Estimate       SE
(b1*b5 - 2*b2*b3)/(4*b3*b4 - b5^2) 352.5917 1.203092
```

6.1.3. The cake experiment was carried out in two blocks of seven observations each. It is possible that the response might differ by block. For example, if the blocks were different days, then differences in air temperature or humidity when the cakes were mixed might have some effect on Y. We can allow for block effects by adding a factor for Block to the mean function, and possibly allowing for Block by term interactions. Add block effects to the mean function fit in Section 6.1.1 and summarize results. The blocking is indicated by the variable *Block* in the data file.

Solution:

```
> m2 <- update(m1, ~factor(block)+.)
> anova(m2)
Analysis of Variance Table

Response: Y
               Df Sum Sq Mean Sq F value  Pr(>F)
factor(block)   1   0.05    0.05    0.22 0.65001
X1              1   4.32    4.32   21.24 0.00246
X2              1   7.43    7.43   36.51 0.00052
I(X1^2)         1   2.13    2.13   10.47 0.01435
I(X2^2)         1  10.55   10.55   51.80 0.00018
X1:X2           1   2.77    2.77   13.62 0.00775
Residuals       7   1.43    0.20
```

We refit, with blocks fit *first*; the F test for blocks in the sequential anova suggests little effect due to blocks. All the other significance levels remain small, so there is unlikely to be much difference in an analysis that accounts for blocks. ∎

6.2 The data in the file `lathe1.txt` are the results of an experiment on characterizing the life of a drill bit in cutting steel on a lathe. Two factors were varied in the experiment, *Speed* and *Feed* rate. The response is *Life*, the total time until the drill bit fails, in minutes. The values of *Speed* in the data have been coded by computing

$$\begin{aligned} \mathit{Speed} &= \frac{(\text{Actual speed in feet per minute} - 900)}{300} \\ \mathit{Feed} &= \frac{(\text{Actual feed rate in thousandths of an inch per revolution } - 13)}{6} \end{aligned}$$

The coded variables are centered at zero. Coding has no material effect on the analysis, but can be convenient in interpreting coefficient estimates.

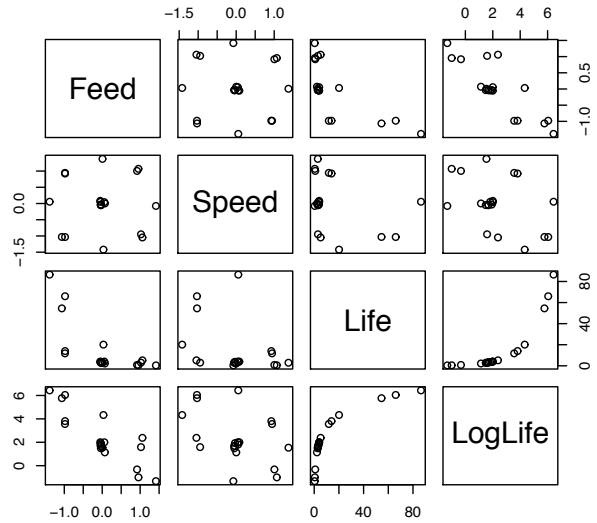
Solution:

∎

6.2.1. Draw a scatterplot matrix of *Speed*, *Feed*, *Life*, and log(*Life*), the base-two logarithm of tool life. Add a little jittering to *Speed* and *Feed* to reveal over-plotting. The plot of *Speed* versus *Feed* gives a picture of the experimental design, which is called a *central composite design*. It is useful

when we are trying to find a value of the factors that maximizes or minimizes the response. Also, several of the experimental conditions were replicated, allowing for a pure-error estimate of variance and lack of fit testing. Comment on the scatterplot matrix.

Solution:



We see from the third row of the scatterplot matrix that *Life* is highly variable, but generally decreasing with *Speed*; the role of *Feed* is less clear. When *Life* is replace by log(*Life*) as in the last row of the scatterplot matrix, the relationships with *Speed* and *Feed* appear more linear and variability appears to be more nearly constant, and so we will use log(*Life*) as the response variable. ∎

6.2.2. For experiments in which the response is a time to failure or time to event, the response often needs to be transformed to a more useful scale, typically by taking the log of the response, or sometimes by taking the inverse. For this experiment, log scale can be shown to be appropriate (Problem 9.7).

Fit the full second-order mean function (6.4) to these data using log(*Life*) as the response, and summarize results.

Solution:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.714      0.152   11.31  2.0e-08
Speed         -2.292      0.124  -18.52  3.0e-11
Speed2         0.416      0.145    2.86  0.01253
Feed          -1.140      0.124   -9.21  2.6e-07
Feed2          0.604      0.145    4.16  0.00096
Speed:Feed    -0.105      0.152   -0.69  0.49943
```

```
Residual standard error: 0.429 on 14 degrees of freedom
Multiple R-Squared: 0.97
F-statistic: 91.2 on 5 and 14 DF,  p-value: 3.55e-10

> m3 <- lm(LogLife ~ Speed + Speed2 +Feed + Feed2 + I(Speed*Feed))
> pureErrorAnova(m3)
Analysis of Variance Table

Response: LogLife
                Df Sum Sq Mean Sq F value   Pr(>F)
Speed            1  63.06   63.06 569.618 7.96e-11
Speed2           1   1.95    1.95  17.599 0.001498
Feed             1  15.60   15.60 140.874 1.30e-07
Feed2            1   3.18    3.18  28.727 0.000230
I(Speed * Feed)  1   0.09    0.09   0.798 0.390697
Lack.of.Fit      3   1.36    0.45   4.084 0.035581
Residuals       11   1.22    0.11
```

The analysis of variance output shown above is directly from R using the `pureErrorAnova` command. It mislabels "pure error" as "Residuals." In addition, the `pureErrorAnova` command works incorrectly with models that include interactions unless you surround the interactions with an `I()` as shown above. The F-values in the table use Pure error as the denomiator for F-tests.

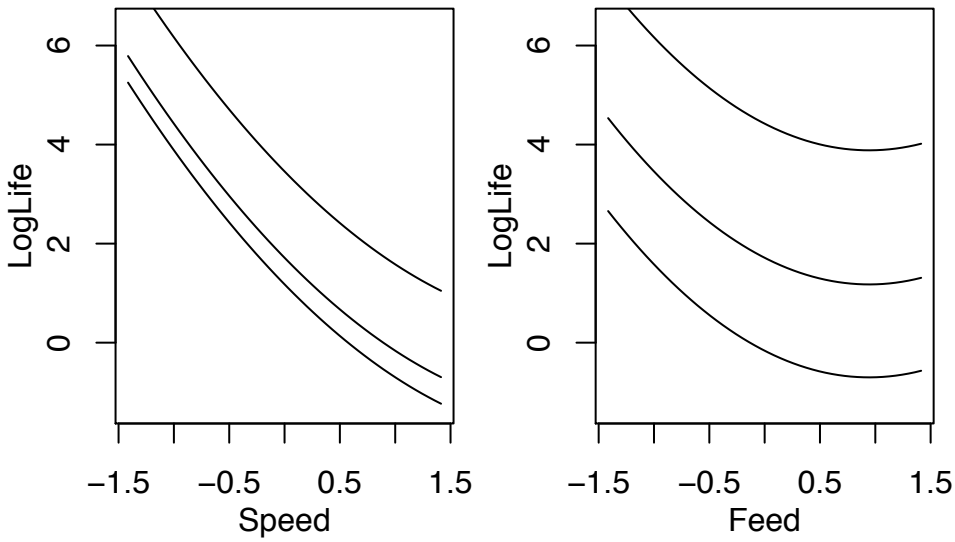
All the coefficients have fairly large t-values, except for the *Speed* $\times$ *Feed* interaction. The F-test for lack of fit is $F = 4.08$ with $(3, 11)$ df, for a p-value of about 0.04; if SF is dropped from the mean function, we get $F = 3.26$ with $(4, 11)$ df, and p-value $= 0.05$. ∎

6.2.3. Test for the necessity of the *Speed*$\times$ *Feed* interaction, and summarize your results. Draw appropriate summary graphs equivalent to Figure 6.3 or Figure 6.4, depending on the outcome of your test.

Solution: The p-value for this test is about 0.499, suggesting that the interaction is not needed. We can summarize with the simpler mean function,

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.714      0.149   11.51  7.6e-09
Speed         -2.292      0.122  -18.85  7.4e-12
Speed2         0.416      0.143    2.91  0.01069
Feed          -1.140      0.122   -9.37  1.2e-07
Feed2          0.604      0.143    4.23  0.00072

Residual standard error: 0.421 on 15 degrees of freedom
Multiple R-Squared: 0.969
F-statistic:  118 on 4 and 15 DF,  p-value: 3.81e-11
```

Because the interaction is not needed, a graph like Figure 6.4 can summarize the results of the experiment. Tool life is apparently minimized for values of *Speed* beyond the range of the data.



■

6.2.4. For $Speed = 0.5$, estimate the value of $Feed$ that minimizes $\log(Life)$, and obtain a 95% confidence interval for this value using the deltaMethod.

Solution: The results from the deltaMethod are:

```
Functions of parameters:  expression(-b4/(2 * b5))
Estimate = 0.944099 with se = 0.244724
```

and the confidence interval is about $.94 \pm 1.96(.25)$. Because the minimum occurs so close to the edge of the sampled region, the confidence interval is likely to be inaccurate, and a bootstrap is likely to provide a more reasonable interval. ■

6.3 In the sleep data, do a lack of fit test for D linear against the one way anova model, with response TS. Summarize results.

Solution:

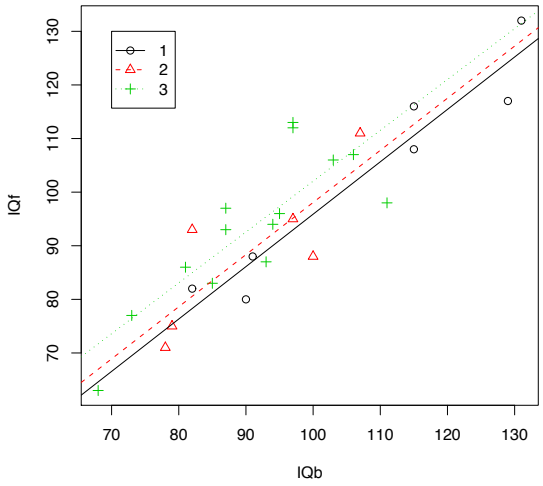
```
> m1 <- lm(TS ~ D, data = sleep1)
> pureErrorAnova(m1)
Analysis of Variance Table

Response: TS
            Df Sum Sq Mean Sq F value  Pr(>F)
D            1    418     418   29.43 1.5e-06
Lack.of.Fit  3     39      13    0.92    0.44
Residuals   53    752      14
```

There is no evidence of lack-of-fit, since the p-value is about 0.44. As a result, we would replace the factor D by a continuous predictor D without any particular loss of information. ■

6.4 The data in the file `twins.txt` give the IQ scores of identical twins, one raised in a foster home, *IQf*, and the other raised by birth parents, *IQb*. The data were published by Burt (1966), and their authenticity has been questioned. For purposes of this example, the twin pairs can be divided into three social classes C, low, middle or high, coded 1, 2, and 3, respectively, in the data file, according to the social class of the birth parents. Treat *IQf* as the response and *IQb* as the predictor, with C as a factor.

Perform an appropriate analysis of these data. Be sure to draw and discuss a relevant graph. Are the within-class mean functions straight lines? Are there class differences? If there are differences, what are they?

Solution:



Given the variation in the points, it is unlikely that there is any notable difference between levels of C:

```
> twin$C <- factor(twin$C)   make C a factor
> m1 <- lm(IQf ~ IQb, data=twin)   model ignoring C
> m2 <- update(m1, ~.+ C)     Separate intercepts
> m3 <- update(m1, ~ C:IQb)   Common intercept, separate slopes
> m4 <- update(m1, ~ C*IQb)   Separate intercepts and slope
> anova(m1,m2,m4)
Analysis of Variance Table

Model 1: IQf ~ IQb
Model 2: IQf ~ IQb + C
Model 3: IQf ~ C + IQb + C:IQb
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     25 1494
2     23 1318  2       175 1.40   0.27
3     21 1317  2         1 0.01   0.99
```

```
> anova(m1,m3,m4)
Analysis of Variance Table

Model 1: IQf ~ IQb
Model 2: IQf ~ C:IQb
Model 3: IQf ~ C + IQb + C:IQb
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     25 1494
2     23 1326  2       167 1.33   0.29
3     21 1317  2         9 0.07   0.93
```

All the p-values are large, so the simplest model, of no class differences, is supported. ∎

6.5 Referring to the data in Problem 2.2, compare the regression lines for Forbes' data and Hooker's data, for the mean function $\mathrm{E}(\log(\mathit{Pressure})|\,\mathit{Temp}) = \beta_0 + \beta_1\,\mathit{Temp}$.

Solution:

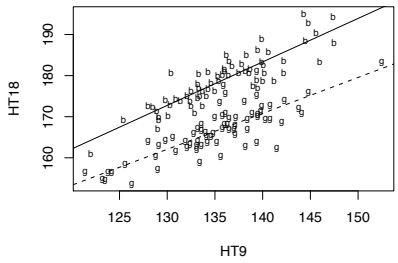
```
> m4 <- lm(100*log(Pressure) ~ Temp, data=d)
> m3 <- update(m4,~.+Temp:Source)
> m2 <- update(m4,~.+Source)
> m1 <- update(m4,~.+Source+Temp:Source)
> anova(m4,m3,m1)
Analysis of Variance Table

Model 1: 100 * log(Pressure) ~ Temp
Model 2: 100 * log(Pressure) ~ Temp + Temp:Source
Model 3: 100 * log(Pressure) ~ Temp + Source + Temp:Source
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     46 32.3
2     45 32.1  1       0.2 0.23   0.63
3     44 31.8  1       0.3 0.39   0.54
> anova(m4,m2,m1)
Analysis of Variance Table

Model 1: 100 * log(Pressure) ~ Temp
Model 2: 100 * log(Pressure) ~ Temp + Source
Model 3: 100 * log(Pressure) ~ Temp + Source + Temp:Source
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     46 32.3
2     45 32.1  1       0.2 0.25   0.62
3     44 31.8  1       0.3 0.36   0.55
```

The smallest model, model 1 of common regressions fits as well as any other the others, and so the same mean function can be used for each set of data. ∎

6.6 Refer to the Berkeley Guidance study described in Problem 3.1. Using the data file `BGSall.txt`, consider the regression of *HT18* on *HT9* and the grouping factor *Sex*.

6.6.1. Draw the scatterplot of *HT18* versus *HT9*, using a different symbol for males and females. Comment on the information in the graph about an appropriate mean function for these data.

Solution:



The lines shown on the graph are the OLS lines fit separately for each *Sex*. From the graph, a straight line mean function appears appropriate for each group. The parallel regressions mean function is plausible from the graph, as is the concurrent regressions mean function. ∎

6.6.2. Fit the four mean function suggested in Section 6.2.2, perform the appropriate tests, and summarize your findings.

Solution:

```
data(BGSall)
attach(BGSall)
# fit the four mean functions
model1 <- lm(HT18 ~ Sex + HT9 + Sex:HT9)
model2 <- lm(HT18 ~ Sex + HT9)
model3 <- lm(HT18 ~ HT9 + Sex:HT9)
model4 <- lm(HT18 ~ HT9)
anova(model4,model2,model1)
Analysis of Variance Table

Model 1: HT18 ~ HT9
Model 2: HT18 ~ Sex + HT9
Model 3: HT18 ~ Sex + HT9 + Sex:HT9
  Res.Df  RSS  Df Sum of Sq      F Pr(>F)
1    134 6191
2    133 1567   1      4624 398.29 <2e-16
3    132 1532   1        34   2.96  0.087
> anova(model4,model3,model1)
Analysis of Variance Table

Model 1: HT18 ~ HT9
Model 2: HT18 ~ HT9 + Sex:HT9
Model 3: HT18 ~ Sex + HT9 + Sex:HT9
  Res.Df  RSS  Df Sum of Sq      F Pr(>F)
1    134 6191
```

```
2    133 1542   1      4649 400.41 <2e-16
3    132 1532   1        10   0.84   0.36
```

The common regression mean function is firmly rejected, and the most general mean function, model 4, is probably not needed. We can't tell between the parallel model and the concurrent model; both provide an equivalent description of the data, although the *RSS* for the parallel model is somewhat smaller (1542 versus 1567).∎

6.7 In the Berkeley Guidance Study data, Problem 6.6, consider the response *HT18* and predictors *HT2* and *HT9*.

6.7.1. Model 1 in Section 6.2.2 allows each level of the grouping variable, in this example the variable *Sex*, to have its own mean function. Write down at least two generalization of this model for this problem with two continuous predictors rather than one.

Solution: Using the computer notation,

$$\begin{aligned} HT18 &\sim 1 + HT2 + HT9 + Sex + Sex{:}HT2 + Sex{:}HT9 \\ HT18 &\sim 1 + HT2 + HT9 + +HT2{:}HT18 + \\ &\quad Sex + Sex{:}HT2 + Sex{:}HT9 + Sex{:}HT2{:}HT18 \end{aligned}$$

Other generalizations are obtained from the this last mean function by deleting terms. There is no real requirement that the *same* terms be deleted for each *Sex*. ∎

6.8 In the Berkeley Guidance Study data, assuming no interaction between *HT2* and *HT9*, obtain a test for the null hypothesis that the regression planes are parallel for boys and girls versus the alternative that separate planes are required for each sex.

Solution:

```
> m2 <- lm(HT18 ~ Sex+ HT2+HT9) # parallel regressions
> m1 <- update(m2,~.+Sex:(HT9+HT2)) # general regressions
> anova(m2,m1)
Analysis of Variance Table

Model 1: HT18 ~ Sex + HT2 + HT9
Model 2: HT18 ~ Sex + HT2 + HT9 + Sex:HT9 + Sex:HT2
  Res.Df  RSS  Df Sum of Sq    F Pr(>F)
1    132 1566
2    130 1497   2        69 2.98  0.054
```

The p-value is about 0.054, suggesting some evidence against parallel regressions. ∎

6.9 Refer to the apple shoot data, Section 5.3, using the data file `allshoots.txt`, giving information on both long and short shoots.

6.9.1. Compute a mean square for pure error separately for long and short shoots, and show that the pure error estimate of variance for long shoots is

about twice the size of the estimate for short shoots. Since these two estimates are based on completely different observations, they are independent, and so their ratio will have an F distribution under the null hypothesis that the variance is the same for the two types of shoots. Obtain the appropriate test, and summarize results. (Hint: the alternative hypothesis is that the two variances are unequal, meaning that you need to compute a two-tailed significance level, not one-tailed as is usually done with F-tests.) Under the assumption that the variance for short shoots is σ^2 and the variance for long shoots is $2\sigma^2$ obtain a pooled pure error estimate of σ^2.

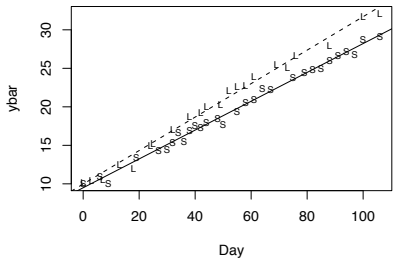
Solution:

```
> sel <- allshoots$Type == 1
> pure.error <- with(allshoots, data.frame(
+      df = c(sum(n[sel]-1), sum(n[!sel]-1), sum(n-1)),
+      SS = c( sum((n[sel]-1)*SD[sel]^2), sum((n[!sel]-1)*SD[!sel]^2),
+              sum((n[sel]-1)*SD[sel]^2)/2  + sum((n[!sel]-1)*SD[!sel]^2))))
> pure.error$pe <- pure.error$SS/pure.error$df
> row.names(pure.error) <- c("Long shoots", "Short shoots", "Pooled")
> pure.error
              df       SS        pe
Long shoots  167 255.1215 1.5276737
Short shoots 292 246.7392 0.8449973
Pooled       459 374.3000 0.8154683
> data.frame(F=pure.error$pe[1]/pure.error$pe[2],
+            pvalue=2*(1-pf(F, pure.error$df[1], pure.error$df[2])))
         F pvalue
1 1.807904      2
```

The significance level is zero to four decimals, so the variance in long shoots is not equal to the variance in short shoots. The pooled estimate only requires dividing the SS for long shoots by 2 to get the scaling right. ∎

6.9.2. Draw the scatterplot of *ybar* versus *Day*, with a separate symbol for each of the two types of shoots, and comment on the graph. Are straight line mean functions plausible? Are the two types of shoots different?

Solution:



The types are very likely different, and the points for the two groups do not overlap. From the regression lines shown, the concurrent mean functions with concurrence at day zero seems plausible.∎

6.9.3. Fit models 1, 3 and 4 from Section 6.2.2 to these data. You will need to use weighted least squares, since each of the responses is an average of n values. Also, in light of Problem 6.9.1, assume that the variance for short shoots is σ^2, but the variance for long shoots is $2\sigma^2$.

Solution: First, compute the weights. For short shoots, the variance of *ybar* is σ^2/n, while for long shoots it is $2\sigma^2/n$. The weights are for equal to n for short shoots and $n/2$ for long shoots. The three models are then

```
> anova(model4,model3,model1)
Analysis of Variance Table

Model 1: ybar ~ Day
Model 2: ybar ~ Day + Type:Day
Model 3: ybar ~ Day + Type + Type:Day
  Res.Df RSS Df Sum of Sq      F Pr(>F)
1     50 469
2     49 101  1       368 181.41 <2e-16
3     48  97  1         3   1.58   0.21
```

From this we conclude that the concurrent regression model is as good as the most general model, and better than the common regression model. However, the F test for lack of fit based on pure error,

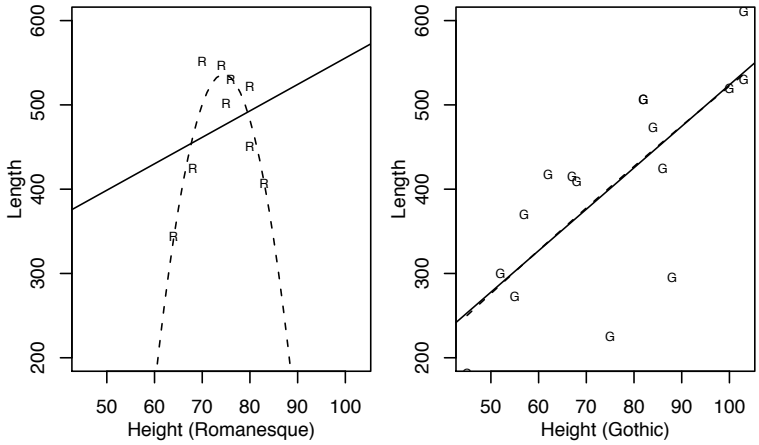
```
> data.frame(Flof=Flof <- sigmaHat(model3)^2/pure.error$pe[3] ,
+        pvalue=1-pf(Flof, model3$df, pure.error$df[3]))
      Flof       pvalue
1 2.518305 3.436992e-07
```

suggests that the straight-line models are not adequate for the data. As discussed in the text, this is probably an example of finding a relatively unimportant deviation from the straight line models because of the very large sample sizes giving very high power.∎

6.10 Gothic and Romanesque Cathedrals The data in the data file `cathedral.txt` gives *Height* = nave height and *Length* = total length, both in feet, for medieval English cathedrals. The cathedrals can be classified according to their architectural style, either Romanesque or, later, Gothic. Some cathedrals have both a Gothic and a Romanesque part, each of differing height; these cathedrals are included twice. Names of the cathedrals are also provided in the file.

6.10.1. For these data, it is useful to draw *separate* plots of *Length* versus *Height* for each architectural style. Summarize the differences apparent in the graphs in the regressions of *Length* on *Height* for the two styles.

Solution:



Fitted to each figure are the OLS simple and quadratic polynomials. For the earlier Romanesque style, a quadratic regression is apparent. Evidently, building taller cathedrals required smaller cathedrals. The flying buttress, characteristic of the later Gothic style, allowed taller cathedrals to be larger, as indicated by the similarity between the straight line and quadratic fits.

6.10.2. Use the data and the plots to fit regression models that summarize the relationship between the response *Length* and the predictor *Height* for the two architectural styles.

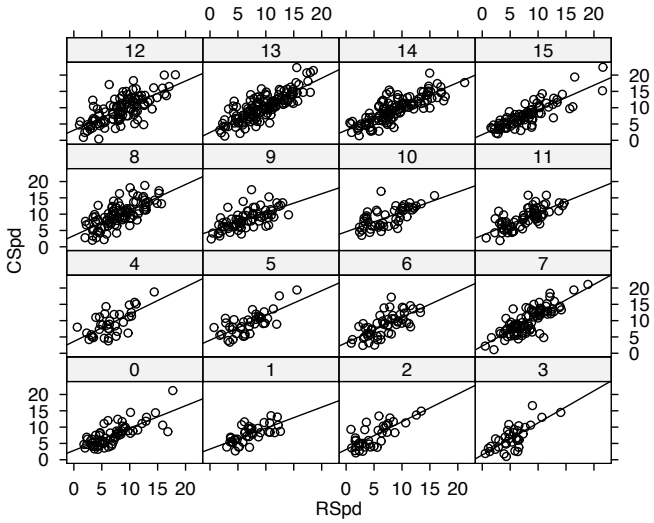
Solution: From the graph in the last subproblem, it is clear that a different mean function should be fit for each style, a quadratic for the earlier Romanesque and a linear mean function for the later Gothic styles. The quadratic fits to each of the two styles are

```
GOTHIC
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.28680  400.69553    0.01     0.99
Height        5.69289   10.98696    0.52     0.61
I(Height^2)  -0.00513    0.07227   -0.07     0.94
Residual standard error: 86.6 on 13 degrees of freedom
Multiple R-Squared: 0.558,
F-statistic: 8.22 on 2 and 13 DF,  p-value: 0.00492

ROMANESQUE
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9311.958   1951.640   -4.77   0.0031
Height        264.445     53.255    4.97   0.0025
I(Height^2)    -1.775      0.362   -4.91   0.0027
Residual standard error: 35.9 on 6 degrees of freedom
Multiple R-Squared: 0.815,
F-statistic: 13.2 on 2 and 6 DF,  p-value: 0.00631
```

6.11 Windmill data In Problem 2.13, we considered data to predict wind speed *CSpd* at a candidate site based on wind speed *RSpd* at a nearby reference site where long-term data is available. In addition to *RSpd*, we also have available the wind direction, *RDir*, measured in degrees. A standard method to include the direction data in the prediction is to divide the directions into several bins, and then fit a separate mean function for *CSpd* on *RSpdp* in each bin. In the wind farm literature, this is called the *measure, correlate, predict* method, Derrick (1992). The data file `wm2.txt` contains values of *CSpd*, *RSpd*, *RDir*, and *Bin* for 2002 for the same candidate and reference sites considered in Problem 2.13. Sixteen bins are used, the first bin for cases with *RDir* between 0 and 22.5 degrees, the second for cases with *RDir* between 22.5 and 45 degrees, ..., and the last bin between 337.5 and 360 degrees. Both the number of bins and their starting points are arbitrary.

6.11.1. Obtain tests that compare fitting the four mean functions discussed in Section 6.2.2 to the sixteen bins. How many parameters are in each of the mean functions?

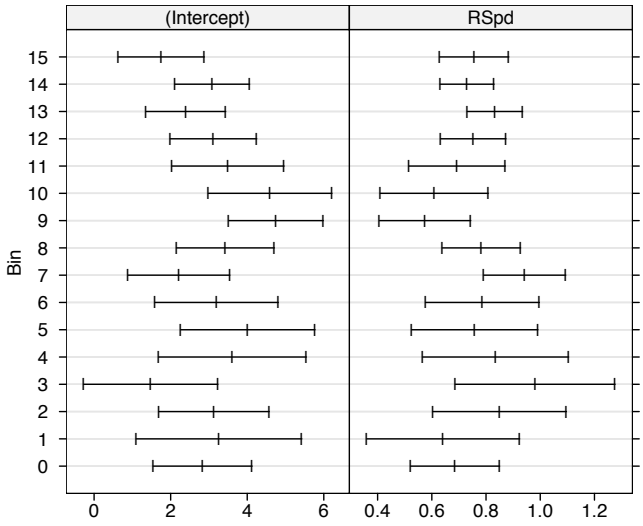
Solution:



This figure suggests that there is a linear regression within each bin, but with substantial variation remaining. The OLS lines shown appear to be very similar. Here are the F-tests, all using model 1 as the alternative hypothesis:

	df	RSS	F	P(>F)
Model 1, most general	1084	6272		
Model 2, parallel	1099	6388	1.33	0.176
Model 3, common intercept	1099	6414	1.63	0.059
Model 4, all the same	1114	6776	2.90	0.000

While model 4 is firmly rejected, there is little to decide between the other three models. This conclusion is echoed by looking at the confidence intervals for the slope and the intercept in each bin:



There is a substantial price to pay for estimating 32 parameters in the most general model, as compared to only 2 parameters in model 4. Some of the bins have as few as 35 observations, so the estimates in that bin are relatively poor. ∎

6.11.2. Do not attempt this problem unless your computer package has a programming language.

Table 6.5 gives the number of observations in each of the sixteen bins along with the average wind speed in that bin for the reference site for the period January 1, 1948 to July 31, 2003, excluding the year 2002; the table is also given in the data file `wm3.txt`. Assuming the most general model of a separate regression in each bin is appropriate, predict the average wind speed at the candidate site for each of the sixteen bins, and find the standard error. This will give you sixteen predictions and sixteen independent standard errors. Finally, combine these sixteen estimates into one overall estimate (you should weight according to the number of cases in a bin), and then compare your answer to the prediction and standard error from Problem 4.6.

Solution: In this solution, I have used the R function `lmList` which is part of the `nlme` library (also available in S-Plus in the library `nlme3`, although the name may have changed when you read this). This function assumes that the variance is different in each bin, and so it estimates 16 variances. Fitting with a common variance might have been preferred.

```
Bin      bin.count    pred  se.pred
0             2676  7.1701 0.287167
1             2073  6.9102 0.371992
2             1710  7.7746 0.410904
3             1851  6.8171 0.436893
4             2194  8.5228 0.489882
5             3427  9.0618 0.373026
```

Table 6.5 Bin counts and means for the windmill data. These data are also given in the file `wm3.txt`.

Bin	Bin.count	*RSpd*	Bin	Bin.count	*RSpd*
0	2676	6.3185	8	4522	7.7517
1	2073	5.6808	9	32077	6.4943
2	1710	5.4584	10	2694	6.1619
3	1851	5.4385	11	2945	6.5947
4	2194	5.8763	12	4580	7.6865
5	3427	6.6539	13	6528	8.8078
6	5201	7.8756	14	6705	8.5664
7	6392	8.4281	15	4218	7.5656

```
6             5201  9.3994 0.368846
7             6392 10.1622 0.249985
8             4522  9.5013 0.277303
9             3207  8.4871 0.283539
10            2694  8.3533 0.363055
11            2945  8.0730 0.275519
12            4580  8.9089 0.292846
13            6528  9.7357 0.205110
14            6705  9.3424 0.207876
15            4218  7.4782 0.220460
Combined     60923  8.8312 0.076584
```

In Problem 2.13, the prediction was 8.7552 with standard error of prediction equal to 0.0748. The prediction from the much more complicated mean function seems no better, and possibly a little worse, than the prediction from the one-bin mean function. ∎

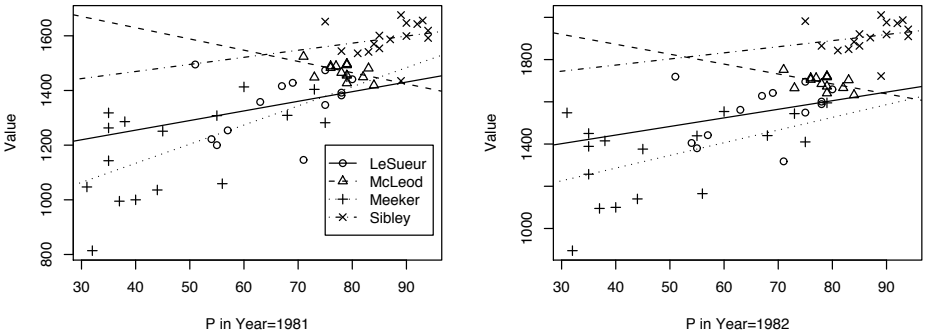
6.12 Land valuation Taxes on farmland enrolled in a "Green Acres" program in metropolitan Minneapolis–St. Paul are valued only with respect to the land's value as productive farmland; the fact that a shopping center or industrial park has been built nearby cannot enter into the valuation. This creates difficulties because almost all sales, which are the basis for setting assessed values, are priced according to the development potential of the land, not the land's value as farmland. A method of equalizing valuation of land of comparable quality was needed.

One method of equalization is based on a soil productivity score P, a number between 1 for very poor land, and 100, for the highest quality agricultural land. The data in the file `prodscore.txt`, provided by Doug Tiffany, gives P along with *Value*, the average assessed value, the *Year*, either 1981 or 1982 and the *County* name for four counties in Minnesota, Le Sueur, Meeker, McLeod, and Sibley, where development pressures had little effect on assessed value of land in 1981-82. The unit of analysis is a township, roughly six miles square.

The goal of analysis is to decide if soil productivity score is a good predictor of assessed value of farm land. Be sure to examine county and year differences,

and write a short summary that would be of use to decision makers who need to determine if this method can be used to set property taxes.

Solution:



The figure shows plots of *Value* versus P separately for each year, with a separate symbol and regression line for each county. Ignoring counties, the mean functions appear to be straight for each year, with similar scatter for each year. The range of P is very different in each county; for example in McLeod county where P is mostly in the 70s. As a result, the within county regressions are relatively poorly estimated. Thus, we suspect, but are not certain, that the variation between the fitted lines in the graph may be due to very small range in P within county.

Given this preliminary, we turn to models for help. We begin by comparing the model with parallel mean functions within each Year by County group to the most general model of a separate mean function for each group:

```
> m0 <- lm(Value ~ P+Year+County, data=prodscore)
> m1 <- lm(Value ~ P*Year*County, data=prodscore)
> anova(m0,m1)
Analysis of Variance Table

Model 1: Value ~ P + Year + County
Model 2: Value ~ P * Year * County
  Res.Df     RSS  Df Sum of Sq    F Pr(>F)
1    114 1423587
2    104 1235843  10    187744 1.58   0.12
```

This anova suggests that the parallel model may be appropriate. The regression summary is

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.37e+05   4.04e+04  -10.80  < 2e-16
P             5.38e+00   1.00e+00    5.36  4.5e-07
Year          2.21e+02   2.04e+01   10.83  < 2e-16
CountyMcLeod  7.16e+01   3.24e+01    2.21    0.029
```

Table 6.6 The salary data.

Variable	Description
Sex	Sex, 1 for female and 0 for male
Rank	Rank, 1 for Assistant Professor, 2 for Associate Professor and 3 for Full Professor
Year	Number of years in current rank
Degree	Highest degree, 1 if Doctorate, 0 if Masters
YSdeg	Number of years since highest degree was earned
Salary	Academic year salary in dollars

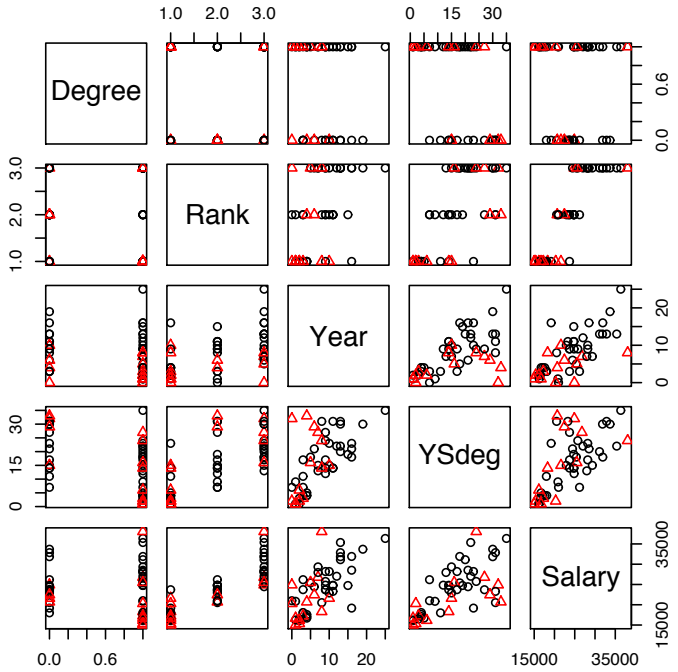
```
CountyMeeker -8.53e+01   3.42e+01   -2.50    0.014
CountySibley  1.93e+02   3.55e+01    5.43  3.2e-07

Residual standard error: 112 on 114 degrees of freedom
Multiple R-Squared: 0.805,      Adjusted R-squared: 0.796
F-statistic: 93.8 on 5 and 114 DF,  p-value: <2e-16
```

so each increase in P of one point is associated with a \$5.38 increase in assessed value; the increase from 1981 to 1982 was \$221, and counties differ by up to \$270 or so. ∎

6.13 Sex discrimination The data in the file `salary.txt` concern salary and other characteristics of all faculty in a small Midwestern college collected in the early 1980s for presentation in legal proceedings for which discrimination against women in salary was at issue. All persons in the data hold tenured or tenure track positions; temporary faculty are not included. The data were collected from personnel files, and consist of the quantities described in Table 6.6.

6.13.1. Draw an appropriate graphical summary of the data, and comment of the graph.

Solution:



This scatterplot matrix uses the *Sex* indicator to mark points; females are the red triangles. A scatterplot matrix is less helpful with categorical predictors, and a sequence of plots might have been preferable here. Nevertheless, we see: (1) females are concentrated in the lowest rank; (2) females generally have lower *Years* of service; (3) the mean function for the regression of *Salary* on *YSdeg* will probably have a different slope for males and females. ∎

6.13.2. Test the hypothesis that the mean salary for men and women is the same. What alternative hypothesis do you think is appropriate?

Solution: This is simply a two-sample t-test, which can be computed using regression software by fitting an intercept and a dummy variable for *Sex*:

```
> summary(m0 <- lm(Salary ~ Sex, salary))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    24697        938   26.33   <2e-16
Sex            -3340       1808   -1.85     0.07

Residual standard error: 5780 on 50 degrees of freedom
Multiple R-Squared: 0.0639,
F-statistic: 3.41 on 1 and 50 DF,  p-value: 0.0706
```

The significance level is 0.07 two-sided, and about 0.035 for the one-sided test that women are paid less. The point estimate of the *Sex* effect is $3340 in favor of men. ∎

6.13.3. Obtain a test of the hypothesis that salary adjusted for years in current rank, highest degree, and years since highest degree is the same for each of the three ranks, versus the alternative that the salaries are not the same. Test to see if the sex differential in salary is the same in each rank.

Solution: This problem asks for two hypothesis tests. The first test is ambiguous, and is either asking to test that the main effect of *Rank* is zero, meaning that rank has no effect on (adjusted) salary, or a test that all the *Rank* by other term interactions are zero, meaning that the regressions are parallel. We do both tests:

```
> m1 <- lm(Salary ~ Year +YSdeg + Degree, salary)
> m2 <- update(m1, ~.+ factor(Rank))
> m3 <- update(m2, ~.+ factor(Rank):(Year+YSdeg+Degree))
> anova(m1,m2,m3)
Analysis of Variance Table

Model 1: Salary ~ Year + YSdeg + Degree
Model 2: Salary ~ Year + YSdeg + Degree + factor(Rank)
Model 3: Salary ~ Year + YSdeg + Degree + factor(Rank) + Year:factor(Rank) +
    YSdeg:factor(Rank) + Degree:factor(Rank)
  Res.Df      RSS Df Sum of Sq     F  Pr(>F)
1     48 6.72e+08
2     46 2.68e+08  2  4.04e+08 35.84 1.2e-09
3     40 2.25e+08  6  4.25e+07  1.26     0.3
```

The small p-value for comparing models 1 and 2 suggests that there is indeed a rank effect (as those of us at higher ranks would hope...). The small p-value for comparing model 2 to model 3 suggest that the effects of the other variables are the same in each rank, meaning that the effect of rank is to add an amount to salary for any values of the other terms.

The second test asks specifically about a *Sex* by *Rank* interaction.

```
> m4 <- update(m1, ~.+Sex)
> m5 <- update(m4, ~.+Sex:factor(Rank))
> anova(m1,m4,m5)
Analysis of Variance Table

Model 1: Salary ~ Year + YSdeg + Degree
Model 2: Salary ~ Year + YSdeg + Degree + Sex
Model 3: Salary ~ Year + YSdeg + Degree + Sex + Sex:factor(Rank)
  Res.Df      RSS Df Sum of Sq    F Pr(>F)
1     48 6.72e+08
2     47 6.59e+08  1  1.35e+07 1.07  0.306
3     45 5.65e+08  2  9.36e+07 3.73  0.032
```

These tests should be examined from bottom to top, so we first compare model 2, including a *Sex* effect, to model 3, which includes a *Sex* by *Rank* interaction.

There is some evidence ($p = .032$) that the *Sex* differential depends on rank. The other test of no *Sex* effect is made irrelevant by the significance of the first test: given an interaction, a test for a main effect is not meaningful. Model 2 seems most appropriate, we examine it in a non-standard parameterization.

```
> summary(
lm(formula = Salary ~ -1 + Year + YSdeg + Degree + factor(Rank) +
    Sex:factor(Rank), data = salary))

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
Year                 522.1      105.5    4.95  1.2e-05
YSdeg               -148.6       86.8   -1.71    0.094
Degree             -1501.5     1029.8   -1.46    0.152
factor(Rank)1      17504.7     1285.0   13.62  < 2e-16
factor(Rank)2      22623.7     1580.9   14.31  < 2e-16
factor(Rank)3      28044.0     2103.1   13.33  < 2e-16
factor(Rank)1:Sex    444.3     1153.5    0.39    0.702
factor(Rank)2:Sex    942.6     2194.9    0.43    0.670
factor(Rank)3:Sex   2954.5     1609.3    1.84    0.073

Residual standard error: 2400 on 43 degrees of freedom
```

The coefficients for the three *Rank* terms correspond to intercept for the three ranks for males. The *Rank* by *Sex* terms give the *Sex* differentials in each of the three ranks; in each rank the differential for females is *positive*, although relatively small, meaning that adjusting for *Rank*, *Year*, *Degree* and *YSdeg*, the women are better paid than the men by a small amount. ∎

6.13.4. Finkelstein (1980), in a discussion of the use of regression in discrimination cases, wrote, "... [a] variable may reflect a position or status bestowed by the employer, in which case if there is discrimination in the award of the position or status, the variable may be 'tainted'." Thus, for example, if discrimination is at work in promotion of faculty to higher ranks, using rank to adjust salaries before comparing the sexes may not be acceptable to the courts.

Fit two mean functions, one including *Sex*, *Year*, *YSdeg* and *Degree*, and the second adding *Rank*. Summarize and compare the results of leaving out rank effects on inferences concerning differential in pay by sex.

Solution:

```
> summary(m7 <- update(m3, ~Sex+Year+YSdeg+Degree))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  13884.2     1639.8    8.47  5.2e-11
Sex          -1286.5     1313.1   -0.98  0.33221
Year           352.0      142.5    2.47  0.01719
YSdeg          339.4       80.6    4.21  0.00011
Degree        3299.3     1302.5    2.53  0.01470
```

```
Residual standard error: 3740 on 47 degrees of freedom
Multiple R-Squared: 0.631,
F-statistic: 20.1 on 4 and 47 DF,  p-value: 1.05e-09
```

If we ignore *Rank*, then the coefficient for *Sex* is again negative, indicating an advantage for males, but the p-value is .33 (or .165 for a one-sided test), indicating that the difference is not significant.

One could argue that other variables in this data set are tainted as well, so using data like these to resolve issues of discrimination will never satisfy everyone. ∎

6.14 Using the salary data in Problem 6.13, one fitted mean function is:

$$\mathrm{E}(\mathit{Salary}|\mathit{Sex}, \mathit{Year}) = 18223 - 571\mathit{Sex} + 741\,\mathit{Year} + 169\mathit{Sex} \times \mathit{Year}$$

6.14.1. Give the coefficients in the estimated mean function if *Sex* were coded so males had the value 2 and females had the value 1 (the coding given in the data file is 0 for males and 1 for females).

Solution: Changing the coding for the *Sex* indicator will change only the coefficient for *Sex* and the coefficient for the intercept. Suppose $\hat{\beta}_0$ and $\hat{\beta}_1$ are the intercept and estimate for *Sex* in the original parameterization, and let $\hat{\eta}_0$ and $\hat{\eta}_1$ be the corresponding estimates in the new coding for *Sex*. Then we must have:

$$\begin{aligned} \text{For males: } \hat{\beta}_0 + \hat{\beta}_1 \times 0 &= \hat{\eta}_0 + \hat{\eta}_1 \times 2 \\ \text{For females: } \hat{\beta}_0 + \hat{\beta}_1 \times 1 &= \hat{\eta}_0 + \hat{\eta}_1 \times 1 \end{aligned}$$

Substituting for $\hat{\beta}_0$ and $\hat{\beta}_1$,

$$\begin{aligned} 18223 &= \hat{\eta}_0 + 2\hat{\eta}_1 \\ 18823 - 571 &= \hat{\eta}_0 + \hat{\eta}_1 \end{aligned}$$

These two equations in two unknowns are easily solved to give $\hat{\eta}_0 = 17681$, and $\hat{\eta}_1 = +571$. ∎

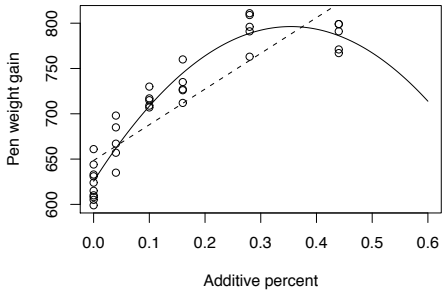
6.14.2. Give the coefficients if *Sex* is coded as -1 for males and $+1$ for females.

Solution: The intercept will change to $18223 + 571/2 = 18508.5$. The *Sex* coefficient will become $-571/2 = -285.5$. ∎

6.15 Pens of turkeys were grown with an identical diet, except that each pen was supplemented with an amount A of an amino acid methionine as a percentage of the total diet of the birds. The data in the file `turk0.txt` gives the response average weight *Gain* in grams of all the turkeys in the pen for 35 pens of turkeys receiving various levels of A.

6.15.1. Draw the scatterplot of *Gain* versus A and summarize. In particular, does simple linear regression appear plausible?

Solution:



For larger values of A, the response appears to level off, or possibly decrease. Variability appears constant across the plot. The lines on the plot refer to Problem 6.15.3. ∎

6.15.2. Obtain a lack of fit test for the simple linear regression mean function, and summarize results. Repeat for the quadratic regression mean function.

Solution:

```
Response: Gain
            Df Sum Sq Mean Sq F value  Pr(>F)
A            1 124689  124689   368.1 < 2e-16
Lack of fit  4  25353    6338    18.7 1.1e-07
Pure error  29   9824     339
```

Quadratic mean function:

```
Response: Gain
            Df Sum Sq Mean Sq F value Pr(>F)
A            1 124689  124689  368.09 <2e-16
I(A^2)       1  23836   23836   70.37  3e-09
Lack of fit  3   1516     505    1.49   0.24
Pure error  29   9824     339
```

There is lack of fit for the simple linear regression model, but the quadratic model is adequate. ∎

6.15.3. To the graph drawn in Problem 6.15.1 add the fitted mean functions based on both the simple linear regression mean function and the quadratic mean function, for values of A in the range from 0 to 0.60, and comment.

Solution: The straight line mean function does not match the data, and leads to the unlikely results that (1) *Gain* could be increased indefinitely as A is increased, and (2) the rate of increase is constant. The quadratic mean function is reasonable for the range of A observed in the data, but it implies that *Gain* actually decreases for $A > .4$ or so. This is probably also quite unrealistic. The conclusion is that the polynomial model is useful for interpolation here, but certainly not for extrapolation outside the range of the data. ∎

6.16 For the quadratic regression mean function for the turkey data discussed in Problem 6.15, use the bootstrap to estimate the standard error of the value of D that maximizes gain. Compare this estimated standard error with the answer obtained using the deltaMethod.

Solution: Using the `bootCase` command in the `alr3` library for R,

```
> deltaMethod(m2, "-b1/(2*b2)")
            Estimate         SE
-b1/(2*b2) 0.3540464 0.01925134
> ans <- bootCase(m2, coef, B=999)
> xmax <- -ans[, 2]/(2*ans[, 3])
> data.frame(mean=mean(xmax), sd=sd(xmax))
       mean         sd
1 0.3563093 0.01769811
```

The point estimates agree within 0.001, and the standard errors agree within about 5%. ∎

6.17 Refer to Jevons' coin data, Problem 5.6. Determine the *age* at which the predicted weight of coins is equal to the legal minimum, and use the deltaMethod to get a standard error for the estimated age. This problem is called *inverse regression*, and is discussed by Brown (1994).

Solution: A point estimate for this value of *Age* can be obtained by setting $Weight = 7.9379$, and solving the estimated regression equation for *Age*,

$$\begin{aligned} 7.9379 &= \hat{\beta}_0 + \hat{\beta}_1 Age \\ \tilde{Age} &= (7.9379 - \hat{\beta}_0)/\hat{\beta}_1 \end{aligned}$$

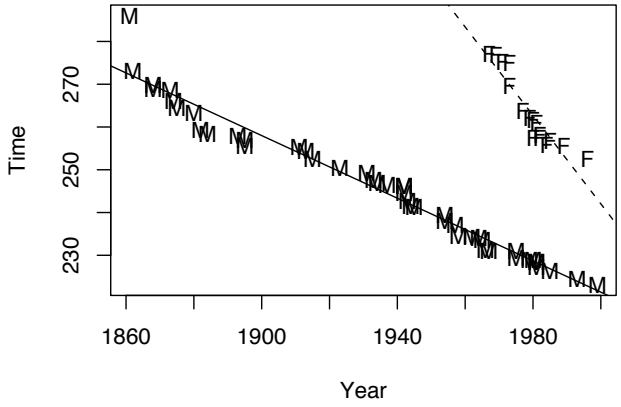
```
> m1 <- lm(Weight  ~  Age, weights = n/SD^2, data=jevons)
> deltaMethod(m1, "(7.9379-b0)/b1")
               Estimate         SE
(7.9379-b0)/b1 2.467645 0.04940154
```

The age at which the weight will on average achieve the legal minimum is 2.47 decades with a standard error of about 0.05. ∎

6.18 The data in the file `mile.txt` gives the world record times for the one mile run. For males, the records are for the period from 1861 to 2003, and for females for the period 1967–2003. The variables in the file are *Year*, year of the record, *Time*, the record time, in seconds, *Name*, the name of the runner, *Country*, the runner's home country, *Place*, the place where the record was run (missing for many of the early records), and *Gender*, either Male or Female. The data were taken from http://www.saunalahti.fi/~ sut/eng/.

6.18.1. Draw a scatterplot of *Time* versus *Year*, using a different symbol for men and women. Comment on the graph.

Solution:



For both genders, the mean function is remarkably straight. Women's records started much later than did men's, but the slope for women is clearly steeper; they are catching up. ∎

6.18.2. Fit separate simple linear regression mean functions to each sex, and show that separate slopes and intercepts are required. Provide an interpretation of the slope parameters for each sex.

Solution:

```
> a1 <- lm(Time~Year,mile)
> a2 <- update(a1,~.+Gender)
> a3 <- update(a2,~.+Gender:Year)
> anova(a1,a2,a3)
Analysis of Variance Table

Model 1: Time ~ Year
Model 2: Time ~ Year + Gender
Model 3: Time ~ Year + Gender + Year:Gender
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     60 11789
2     59   896  1     10893 1219.6 <2e-16
3     58   518  1       378   42.3  2e-08
```

The separate regressions model is appropriate here. ∎

6.18.3. Find the year in which the female record is expected to be 240 seconds, or four minutes. This will require inverting the fitted regression equation. Use the deltaMethod to estimated the standard error of this estimate.

Solution:

```
> m1 <- lm(Time~ -1 + Gender + Gender:Year)
```

```
> deltaMethod(m1, "(240-b1)/b3")
             Estimate       SE
(240-b1)/b3 2001.966 2.357027
```

The model `m1` is equivalent to `a3`, but this reparameterization gives the two intercepts and slopes directly. This computation uses the pooled estimate of error from both men and women. No woman has yet run a four-minute mile, but according to this regression, it is likely to happen within about $2 \times 2.357 \approx 5$ years of 2002; as I write in 2004, this could happen any time. If you are using `deltaMethod`, the first parameter in the mean function is called `b0` even when there is no intercept in the mean function. ∎

6.18.4. Using the model fit in Problem 6.18.2, estimate the year in which the female record will match the male record, and use the deltaMethod to estimate the standard error of the year in which they will agree. Comment on whether you think using the point at which the fitted regression lines cross as a reasonable estimator of the crossing time.

Solution: If males have estimated intercept and slope $\hat{\beta}_0$ and $\hat{\beta}_1$, and females have estimates $\hat{\gamma}_0$ and $\hat{\gamma}_1$, then the two lines cross when $\hat{\beta}_0 + \hat{\beta}_1 t = \hat{\gamma}_0 + \hat{\gamma}_1 t$, and, solving for t ,we get the year $\hat{t} = (\hat{\beta}_0 - \hat{\gamma}_0)/(\hat{\gamma}_1 - \hat{\beta}_1)$.

```
> deltaMethod(m1, "(b1-b2)/(b4-b3)")
                Estimate      SE
(b1-b2)/(b4-b3) 2030.950 8.16785
```

The crossing time is estimated to be in about 2031 with standard error of about 8 years.

It is easy to argue that this computation is without much merit. First, it is an extrapolation of twenty-five or so years. Second, it assumes that whatever athletes do to improve the world records will continue as it has in the past. Third, the larger (negative) slope for females may not be sustainable in the long run; it could be due to taking advantage of "easy" improvements to training and conditioning, and that when female speeds approach those of men in the last few years, the yearly increments will match men's increments more closely. ∎

6.19 Use the deltaMethod to get a 95% confidence interval for the ratio β_1/β_2 for the transactions data, and compare to the bootstrap interval obtained at the end of Section 4.6.1.

Solution:

```
> m1 <- lm(Time  ~  T1  + T2, transact)
> deltaMethod(m1, "b1/b2")
      Estimate        SE
b1/b2 2.684653 0.3189858
> b1 <- bootCase(m1, coef, B=999)
> data.frame(mean=mean(b1[, 2]/b1[, 3]), sd=sd(b1[, 2]/b1[, 3]))
      mean        sd
```

```
1 2.754168 0.5468505
```

While the means agree reasonably closely, the standard deviation computed by the deltaMethod is about 40% too small, so confidence intervals computed from the deltaMethod will be too short.∎

6.20 Refer to the wool data discussed in Section 6.3.

6.20.1. Write out in full the main-effects and the second-order mean functions, assuming that the three predictors will be turned into factors, each with three levels. This will require you to define appropriate dummy variables and parameters.

Solution: Using the parameterization used by default by R, for $i \in (Len, Amp, Load)$, let U_{ij} be the dummy variable for level j for variable i, $j = 2, 3$. This parameterization has a dummy variable for the middle and high level of each factor, dropping the low level. The two mean functions in R/S-Plus notation are

$$\begin{aligned}
\mathrm{E}(\log(Cycles)|\text{First order}) &= \beta_0 + \sum_{i=1}^{3}\sum_{j=2}^{3}\beta_{ij}U_{ij} \\
\mathrm{E}(\log(Cycles)|\text{Second order}) &= \beta_0 + \sum_{i=1}^{3}\sum_{j=2}^{3}\beta_{ij}U_{ij} + \\
&\quad \sum_{i=1}^{2}\sum_{k=i+1}^{3}\sum_{j=2}^{3}\beta_{ikj}U_{ij}U_{kj}
\end{aligned}$$

Most computer programs have a simple way of writing these mean functions. First, declare *Len*, *Amp*, and *Load* to be factors. The two mean functions are then just:

$$\begin{aligned}
\log(Cycles) &\sim Len + Amp + Load \\
\log(Cycles) &\sim (Len + Amp + Load)^2
\end{aligned}$$

The computer program is responsible for creating the correct dummy variables and products. ∎

6.20.2. For the two mean function in Problem 6.20.1, write out the expected change in the response when *Len* and *Amp* are fixed at their middle levels, but *Load* is increased from its middle level to its high level.

Solution: For the first-order model using the R parameterization, the change is $\beta_{33} - \beta_{32}$. Using the second-order mean function, the change is $\beta_{33} - \beta_{32} + \beta_{133} - \beta_{132} + \beta_{233} - \beta_{232}$. ∎

6.21 A partial one-dimensional or POD model for a problem with p predictors $X = (X_1 \ldots, X_p)$ and a factor F with d levels is specified, for the j-th

level of F, by

$$\mathrm{E}(Y|X=\mathbf{x}, F=j) = \eta_{0j} + \eta_{1j}(\mathbf{x}'\boldsymbol{\beta}^*) \tag{6.1}$$

This is a nonlinear model because η_{ij} multiplies the parameter $\boldsymbol{\beta}^*$. Estimation of parameters can use the following two-step algorithm:

1. Assume that the η_{1j}, $j = 1, \ldots, d$ are known. At the first step of the algorithm, set $\eta_{1j} = 1, j = 1, \ldots, d$. Define a new term $\mathbf{z}_j = \eta_{1j}\mathbf{x}$, and substituting into (6.1),

   $$\mathrm{E}(Y|X=\mathbf{x}, F=j) = \eta_{0j} + \mathbf{z}_j'\boldsymbol{\beta}^*$$

   We recognize this as a mean function for *parallel regressions* with common slopes $\boldsymbol{\beta}^*$ and a separate intercept for each level of F. This mean function can be fit using standard OLS linear regression software. Save the estimate $\hat{\boldsymbol{\beta}}^*$ of $\boldsymbol{\beta}^*$.

2. Let $v = \mathbf{x}'\hat{\boldsymbol{\beta}}^*$, where $\hat{\boldsymbol{\beta}}^*$ was computed in step 1. Substitute v for $\mathbf{x}'\boldsymbol{\beta}^*$ in (6.1) to get

   $$\mathrm{E}(Y|X=\mathbf{x}, F=j) = \eta_{0j} + \eta_{1j}v$$

   which we recognize as a mean function with a separate intercept and slope for each level of F. This mean function can also be fit using OLS linear regression software. Save the estimates of η_{1j} and use them in the next iteration of step 1.

Repeat this algorithm until the residual sum of squares obtained at the two steps is essentially the same. The estimates obtained at the last step will be the OLS estimates for the original mean function, and the residual sum of squares will be the residual sum of squares that would be obtained by fitting using nonlinear least squares. Estimated standard errors of the coefficients will be too small, so t-tests should not be used, but F-tests can be used to compare models.

Write a computer program that implements this algorithm.

6.22 Using the computer program written in the last problem or using any other convenient software, verify the results obtained in the text for the Australian Athletes data. Also, obtain tests for the general POD mean function versus the POD mean function with parallel mean functions.

Solution: Using the `alr3` package for R and S-Plus, POD models can be fit without writing the special purpose program:

```
> m1 <- pod(LBM~Ht+Wt+RCC, data= ais, group = Sex)
> summary(m1)

Formula: LBM ~ eta0 + eta1 * Ht + eta2 * Wt + eta3 * RCC + Sex1 * (th02 +
    th12 * (eta1 * Ht + eta2 * Wt + eta3 * RCC))

Parameters:
```

```
     Estimate Std. Error t value Pr(>|t|)
eta0 -14.6565     6.4645   -2.27  0.02447
eta1   0.1463     0.0342    4.27  3.0e-05
eta2   0.7093     0.0242   29.36  < 2e-16
eta3   0.7248     0.5854    1.24  0.21717
th02  12.8472     3.7634    3.41  0.00078
th12  -0.2587     0.0345   -7.51  2.1e-12

Residual standard error: 2.46 on 196 degrees of freedom

> anova(m1)
POD Analysis of Variance Table for LBM, grouped by Sex

1: LBM ~ Ht + Wt + RCC
2: LBM ~ Ht + Wt + RCC + factor(Sex)
3: LBM ~ eta0 + eta1 * Ht + eta2 * Wt + eta3 * RCC + Sex1 * (th02 +
3:     th12 * (eta1 * Ht + eta2 * Wt + eta3 * RCC))
4: LBM ~ (Ht + Wt + RCC) * factor(Sex)

             Res.Df  RSS  Df Sum of Sq      F  Pr(>F)
1: common       198 2937
2: parallel     197 1457   1      1479 245.65 < 2e-16
3: pod          196 1186   1       272  45.09 2.0e-10
4: pod + 2fi    194 1168   2        18   1.47    0.23
```


6.23 The Minnesota Twins professional baseball team plays its games in the Metrodome, an indoor stadium with a fabric roof. In addition to the large air fans required to keep to roof from collapsing, the baseball field is surrounded by ventilation fans that blow heated or cooled air into the stadium. Air is normally blown into the center of the field equally from all directions.

According to a retired supervisor in the Metrodome, in the late innings of some games the fans would be modified so that the ventilation air would blow out from home plate toward the outfield. The idea is that the air flow might increase the length of a fly ball. For example, if this were done in the middle of the eighth inning, then the air-flow advantage would be in favor of the home team for six outs, three in each of the eighth and ninth innings, and in favor of the visitor for three outs in the ninth inning, resulting in a slight advantage for the home team.

To see if manipulating the fans could possibly make any difference, a group of students at the University of Minnesota and their professor built a "cannon" that used compressed air to shoot baseballs. They then did the following experiment in the Metrodome in March, 2003:

1. A fixed angle of 50 degrees and velocity of 150 feet per second was selected. In the actual experiment, neither the velocity nor the angle

could be controlled exactly, so the actual angle and velocity varied from shot to shot.

2. The ventilation fans were set so that to the extent possible all the air was blowing in from the outfield towards home plate, providing a headwind. After waiting about 20 minutes for the air flows to stabilize, twenty balls were shot into the outfield, and their distances were recorded. Additional variables recorded on each shot include the weight (in grams) and diameter (in cm) of the ball used on that shot, and the actual velocity and angle.

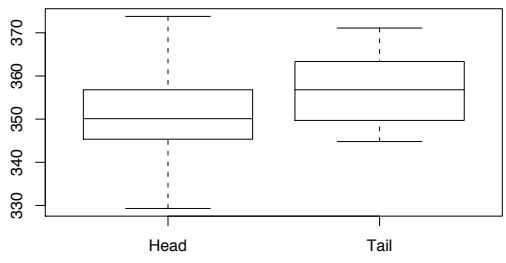
3. The ventilation fans were then reversed, so as much as possible air was blowing out towards the outfield, giving a tailwind. After waiting twenty minutes for air currents to stabilize, fifteen balls were shot into the outfield, again measuring the ball weight and diameter, and the actual velocity and angle on each shot.

The data from this experiment is available in the file `domedata.txt`, courtesy of Ivan Marusic. The variable names are: *Cond*, the condition, head or tail wind; *Velocity*, the actual velocity in feet per second; *Angle*, the actual angle; *BallWt*, the weight of the ball in grams used on that particular test; *BallDia*, the diameter in inches of the ball used on that test; *Dist*, distance in feet of the flight of the ball.

6.23.1. Summarize any evidence that manipulating the fans can change the distance that a baseball travels. Be sure to explain how you reached your conclusions, and provide appropriate summary statistics that might be useful for a newspaper reporter (a report of this experiment is given in the Minneapolis *StarTribune* for July 27, 2003).

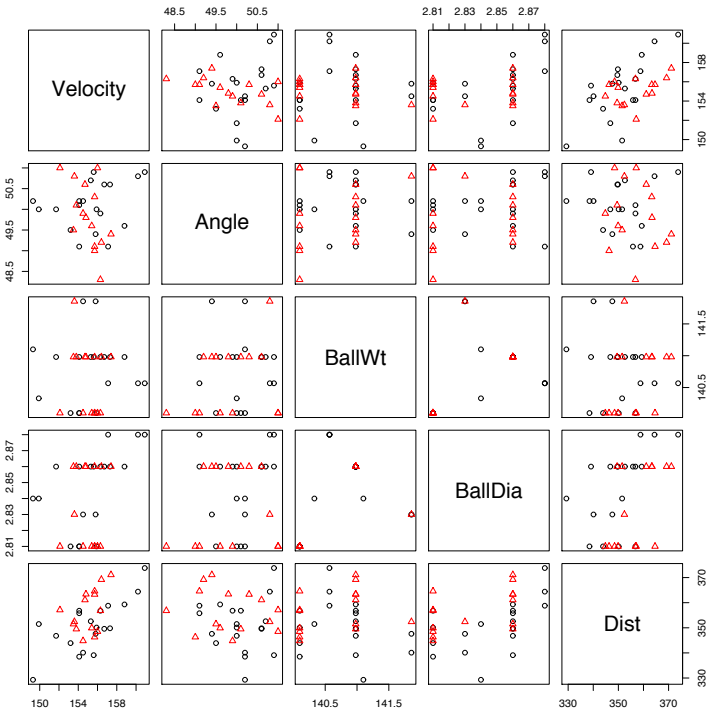
Solution: A reasonable place to start is a with a boxplot of the response *Dist* for each value of *Cond*:

```
> boxplot(Dist ~ Cond, data=domedata)
```



From this figure, it appears plausible that there is an advantage for tailwind hits over headwind hits. We next examine the scatterplot matrix of the

response and the continuous predictors, using *Cond* to color and mark the points.



The key features of this graph are: (1) Distance seems linearly related to velocity, and the red points for tailwind are generally above the black points for headwind; (2) effects of other variables, if any, are small; (3) the variables are linearly related among themselves.

This example is an ideal candidate for analysis via POD models. If you do not have access to POD software, you can start with a model like `Dist ~ Cond*(Velocity + Angle + BallWt + BallDia)`, and get to the same answer, after quite a bit of extra work. Here is the computation in R using POD models.

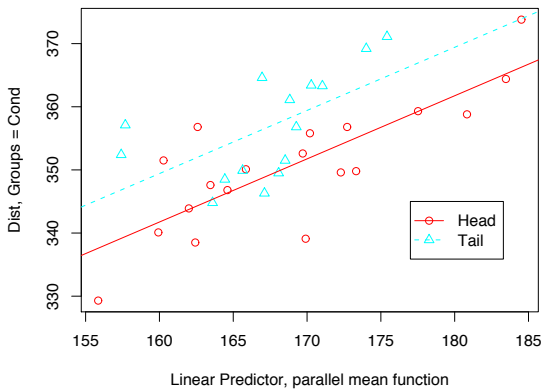
```
> m1 <- pod(Dist~Velocity+Angle+BallWt+BallDia,data=domedata, group=Cond,
+             control=nls.control(maxiter=50,tol=7e-4,minFactor=1/1024))
> anova(m1)
POD Analysis of Variance Table for Dist, grouped by Cond

1: Dist ~ Velocity + Angle + BallWt + BallDia
2: Dist ~ Velocity + Angle + BallWt + BallDia + factor(Cond)
3: Dist ~ eta0 + eta1 * Velocity + eta2 * Angle + eta3 * BallWt +
       eta4 * BallDia + CondTail * (th02 + th12 * (eta1 * Velocity +
       eta2 * Angle + eta3 * BallWt + eta4 * BallDia))
```

```
4: Dist ~ (Velocity + Angle + BallWt + BallDia) * factor(Cond)

             Res.Df  RSS Df Sum of Sq        F Pr(>F)
1: common        29 1747
2: parallel      28 1297  1       450     9.62 0.0049
3: pod           27 1297  1  -5.4e-04  1.2e-05 0.9973
4: pod + 2fi     24 1124  3       172     1.23 0.3220
```

Although not shown, the POD algorithm did not converge with the default settings for the algorithm. The `control` argument, which is passed to the nonlinear regression fitting method was used to increase the value of `tol` from 1×10^{-5} to 7×10^{-4}, and then convergence was attained. The POD analysis of variance suggests that nothing more complicated than the parallel regression model is required for these data, and

```
> m2 <- update(m1,mean.function="parallel")
> summary(m2)
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        181.744    335.696    0.54   0.5925
Velocity             1.728      0.543    3.18   0.0036
Angle               -1.601      1.799   -0.89   0.3811
BallWt              -3.986      2.670   -1.49   0.1466
BallDia            190.372     62.512    3.05   0.0050
factor(Cond)Tail     7.670      2.459    3.12   0.0042

Residual standard error: 6.8 on 28 degrees of freedom
Multiple R-Squared: 0.592,
F-statistic: 8.12 on 5 and 28 DF,  p-value: 7.81e-05
```

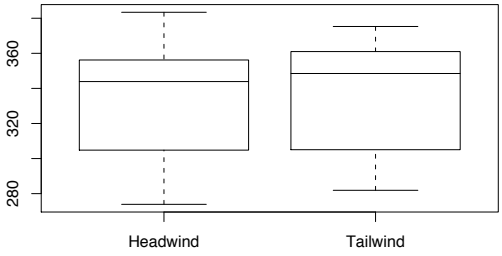
The tailwind effect, adjusted for the other variables, is about 7.7 feet, with a p-value of about 0.0042. While it is possible to refine the result, by deleting the unimportant predictors like *Angle* and *BallWt*, the advantage for tailwinds was clear in these data. A summary graph is shown below.



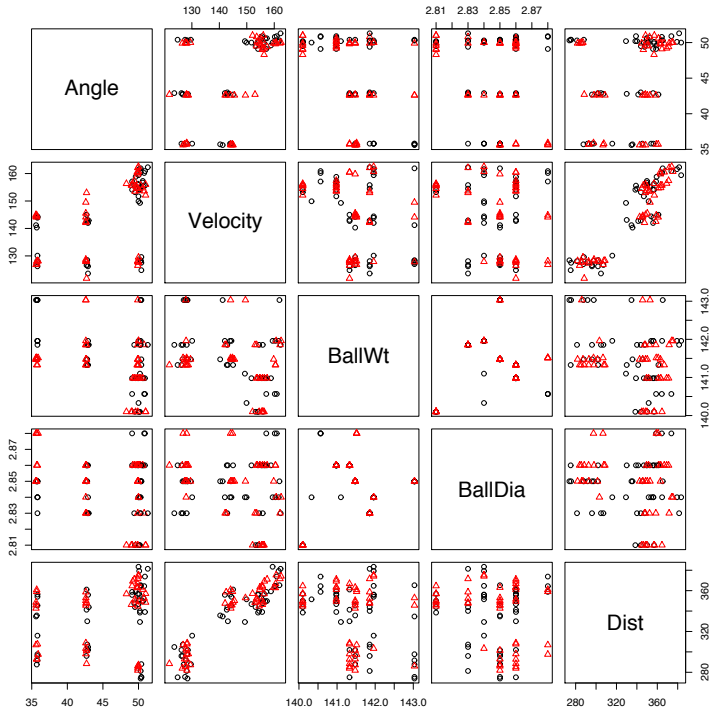
6.23.2. In light of the discussion in Section 6.5, one could argue that this experiment by itself cannot provide adequate information to decide if the fans can affect length of a fly ball. The treatment is *manipulating the fans*; each condition was set up only once, and then repeatedly observed. Resetting the fans after each shot is not practical because of the need to wait at least 20 minutes for the air flows to stabilize.

A second experiment was carried out in May, 2003, using a similar experimental protocol. As before, the fans were first set to provide a headwind, and then, after several trials, the fans were switched to a tailwind. Unlike the first experiment, however, the nominal *Angle* and *Velocity* were varied according to a 3×2 factorial design; actual angles and velocities are again measured. The data file `domedata1.txt` contains the results from both the first experiment and the second experiment, with an additional column called *Date* indicating which sample is which. Analyze these data, and write a brief report of your findings.

Solution: We can duplicate the analysis from the last sub-problem, if we ignore the *Date* effect. The boxplot is



which shows no marginal effect of tailwind. The scatterplot matrix



reinforces this finding of no clear tailwind effect. We again fit the POD models,

```
> m1 <- pod(Dist~Velocity+Angle+BallWt+BallDia,data=domedata1, group=Cond)
> anova(m1)
POD Analysis of Variance Table for Dist, grouped by Cond

1: Dist ~ Velocity + Angle + BallWt + BallDia
2: Dist ~ Velocity + Angle + BallWt + BallDia + factor(Cond)
3: Dist ~ eta0 + eta1 * Velocity + eta2 * Angle + eta3 * BallWt +
       eta4 * BallDia + CondTailwind * (th02 + th12 * (eta1 * Velocity +
       eta2 * Angle + eta3 * BallWt + eta4 * BallDia))
4: Dist ~ (Velocity + Angle + BallWt + BallDia) * factor(Cond)

             Res.Df  RSS Df Sum of Sq    F Pr(>F)
1: common        91 7833
2: parallel      90 7758  1        75 0.87   0.35
3: pod           89 7755  1         3 0.03   0.85
4: pod + 2fi     86 7415  3       340 1.32   0.27
```

This time, the common regression model is adequate, as all the p-values are large, and so there is no tailwind effect at all. Adding *Date* to the problem as a blocking effect does not change the outcome.

We don't really have enough data to interpret the results of this experiment. The data on the first day led to different conclusions than did the data on day two. We don't know if this is normal day-to-day variation, meaning that the effect will be real sometimes, or if the first day was, somehow, abnormal. Only more replications, meaning more days, can answer this question. ∎

7

Transformations

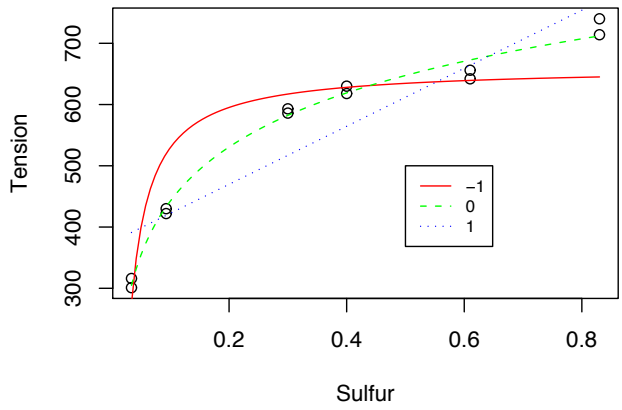
If you and your students use R for computing, there are several functions, now in the **car** package, that correspond exactly the computations in following chapters. These functions are discussed in the *R Primer for Applied Linear Regression*, which you can get by entering the command `alrWeb("primer")` when **alr3** is loaded into R. Even more information on these functions is contained in the book *An R Companion to Applied Regression*; see `http://tinyurl.com/carbook`, which is devoted to the **car** package.

Problems

7.1 The data in the file `baeskel.txt` were collected in a study of the effect of dissolved sulfur on the surface tension of liquid copper (Baes and Kellogg, 1953). The predictor *Sulfur* is the weight percent sulfur, and the response is *Tension*, the decrease in surface tension in dynes per cm. Two replicate observations were taken at each value of *Sulfur*. These data were previously discussed by Sclove (1972).

7.1.1. Draw the plot of *Tension* versus *Sulfur* to verify that a transformation is required to achieve a straight-line mean function.

Solution:



■

7.1.2. Set $\lambda = -1$, and fit the mean function

$$\mathrm{E}(Tension|Sulfur) = \beta_0 + \beta_1 Sulfur^{\lambda}$$

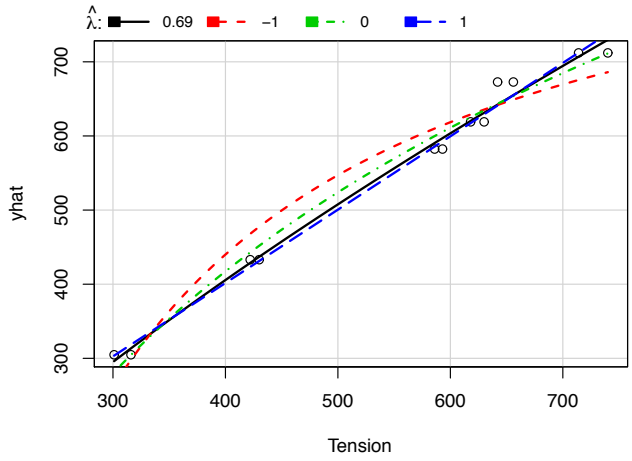
using OLS; that is, fit the OLS regression with *Tension* as the response and $1/Sulfur$ as the predictor. Let *new* be a vector of 100 equally spaced values between the minimum value of *Sulfur* and its maximum value. Compute the fitted values from the regression you just fit, given by $Fit.new = \beta_0 + \beta_1 new^{\lambda}$. Then, add to the graph you drew in Problem 7.1.1 the line joining the points $(new, Fit.new)$. Repeat for $\lambda = 0, 1$. Which of these three choices of λ gives fitted values that match the data most closely?

Solution: From the above figure, only the log transformation closely matches the data. ■

7.1.3. Replace *Sulfur* by its logarithm, and consider transforming the response *Tension*. To do this, draw the inverse response plot with the fitted values from the regression of *Tension* on log(*Sulphur*) on the vertical axis and *Tension* on the horizontal axis. Repeat the methodology of Problem 7.1.2 to decide if further transformation of the response will be helpful.

Solution: As pointed out in the text, with a single predictor the inverse response plot is equivalent to a plot of the response on the horizontal axis and the predictor on the vertical axis. The plot can be drawn most easily with the `invResPlot` function

```
> invResPlot(lm(Tension ~ log(Sulfur), baeskel))
      lambda       RSS
1  0.6860853  2202.113
2 -1.0000000 10594.340
3  0.0000000  3658.171
4  1.0000000  2509.564
```



Untransformed, $\lambda = 1$, matches well, almost as well as the optimal valie of about 2/3, suggesting no further need to transform. This could be verified by performing a lack of fit test from the regression of *Tension* on log(*Sulfur*),

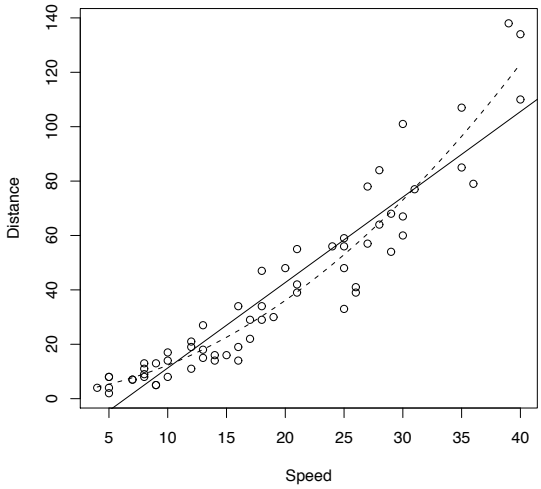
```
> m1 <- lm(Tension~log(Sulfur)+factor(Sulfur))
> anova(m1)
Analysis of Variance Table

Response: Tension
               Df Sum Sq Mean Sq F value  Pr(>F)
log(Sulfur)     1 241678  241678 2141.90 6.8e-09
factor(Sulfur)  4   1859     465    4.12   0.061
Residuals       6    677     113
```

The lack-of-fit test has p-value of 0.06. ∎

7.2 The (hypothetical) data in the file `stopping.txt` give stopping times for $n = 62$ trials of various automobiles traveling at *Speed* miles per hour and the resulting stopping *Distance* in feet (Ezekiel and Fox, 1959).

7.2.1. Draw the scatterplot of *Distance* versus *Speed*. Add the simple regression mean function to your plot. What problems are apparent? Compute a test for lack of fit, and summarize results.

Solution:



The solid line is for simple regression, and the dashed line is a quadratic fit. A lack of fit test can be done using a pure error analysis, since there are replications, or by comparing the quadratic mean function to the simple linear regression mean function.

```
> m1<-lm(Distance~Speed,stopping)
> m2 <- lm(Distance~Speed+I(Speed^2), data=stopping)
> pureErrorAnova(m1)
Analysis of Variance Table

Response: Distance
            Df Sum Sq Mean Sq F value Pr(>F)
Speed        1  59639   59639  625.95 <2e-16
Lack.of.Fit 26   5071     195    2.05  0.025
Residuals   34   3239      95
> anova(m2)
Analysis of Variance Table

Response: Distance
           Df Sum Sq Mean Sq F value  Pr(>F)
Speed       1  59639   59639   605.2 < 2e-16
I(Speed^2)  1   2496    2496    25.3 4.8e-06
Residuals  59   5814      99
```

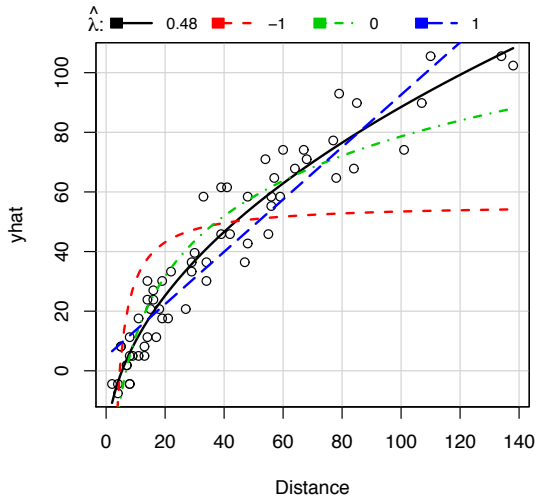
Both methods indicate that the simple regression mean function is not adequate.∎

7.2.2. Find an appropriate transformation for *Distance* that can linearize this regression.

Solution: Using the inverse response plot method:

```
> invResPlot(m1)  # suggests square root of Distance
```

```
      lambda       RSS
1  0.4849737  4463.944
2 -1.0000000 33149.061
3  0.0000000  7890.434
4  1.0000000  7293.835
```

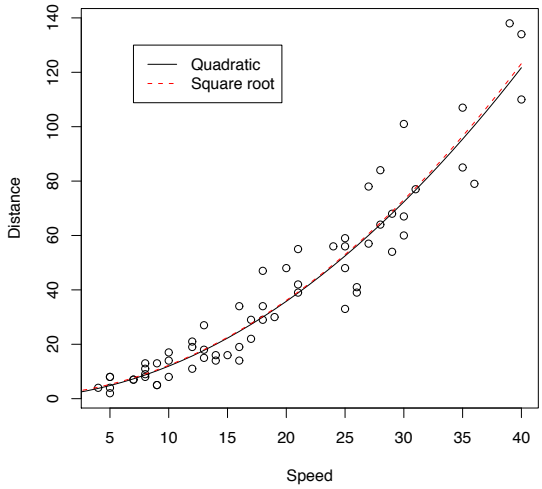


The optimal transformation is at about $\hat{\lambda} = .49$

This suggests using the square root scale for *Distance*.∎

7.2.3. Hald (1960) has suggested on the basis of a theoretical argument that the mean function $\mathrm{E}(Distance|Speed) = \beta_0 + \beta_1 Speed + \beta_2 Speed^2$, with $\mathrm{Var}(Distance|Speed) = \sigma^2 Speed^2$ is appropriate for data of this type. Compare the fit of this model to the model found in Problem 7.2.2. For *Speed* in the range 0 to 40 mph, draw the curves that give the predicted *Distance* from each model, and qualitatively compare them.

Solution:

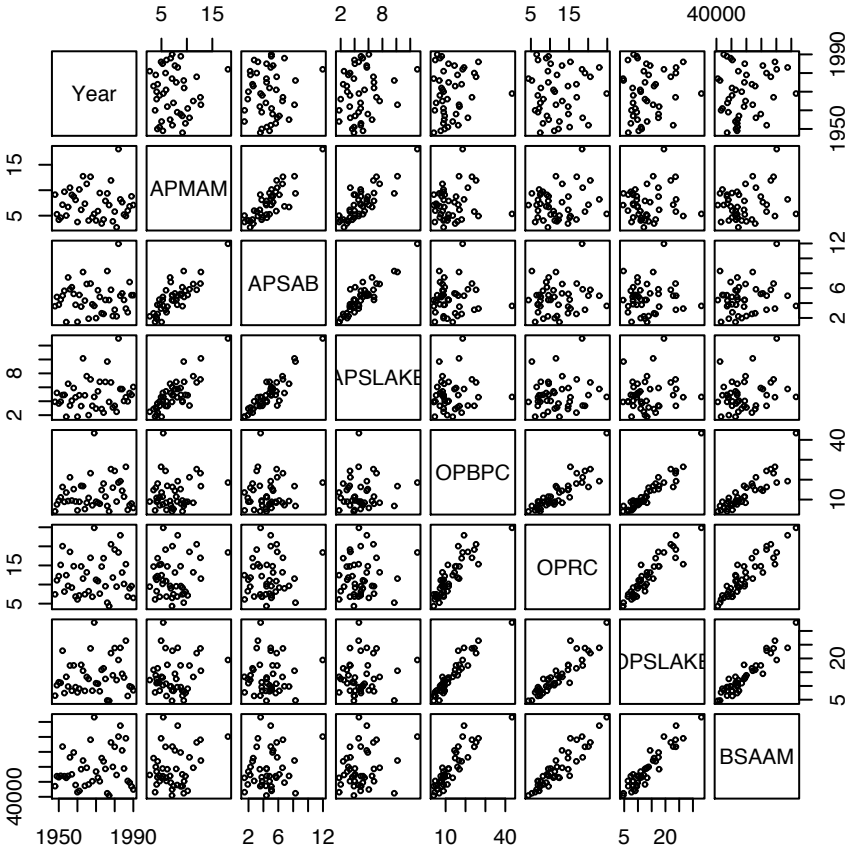


The plot of fitted values from the weighted quadratic model and the squares of the fitted values of the unweighted analysis in square root scales are virtually identical. ∎

7.3 This problem uses the data discussed in Problem 1.5. A major source of water in Southern California is the Owens Valley. This water supply is in turn replenished by spring runoff from the Sierra Nevada mountains. If runoff could be predicted, engineers, planners and policy makers could do their jobs more efficiently. The data in the file `water.txt` contains 43 years of precipitation measurements taken at six sites in the mountains, in inches of water, and stream runoff volume at a site near Bishop, California. The three sites with name starting with "O" are fairly close to each other, and the three sites starting with "A" are also fairly close to each other.

7.3.1. Load the data file, and construct the scatterplot matrix of the six snowfall variables, which are the predictors in this problem. Using the methodology for automatic choice of transformations outlined in Section 7.2.2, find transformations to make the predictors as close to linearly related as possible. Obtain a test of the hypothesis that all $\lambda_j = 0$ against a general alternative, and summarize your results. Do the transformations you found appear to achieve linearity? How do you know?

Solution:



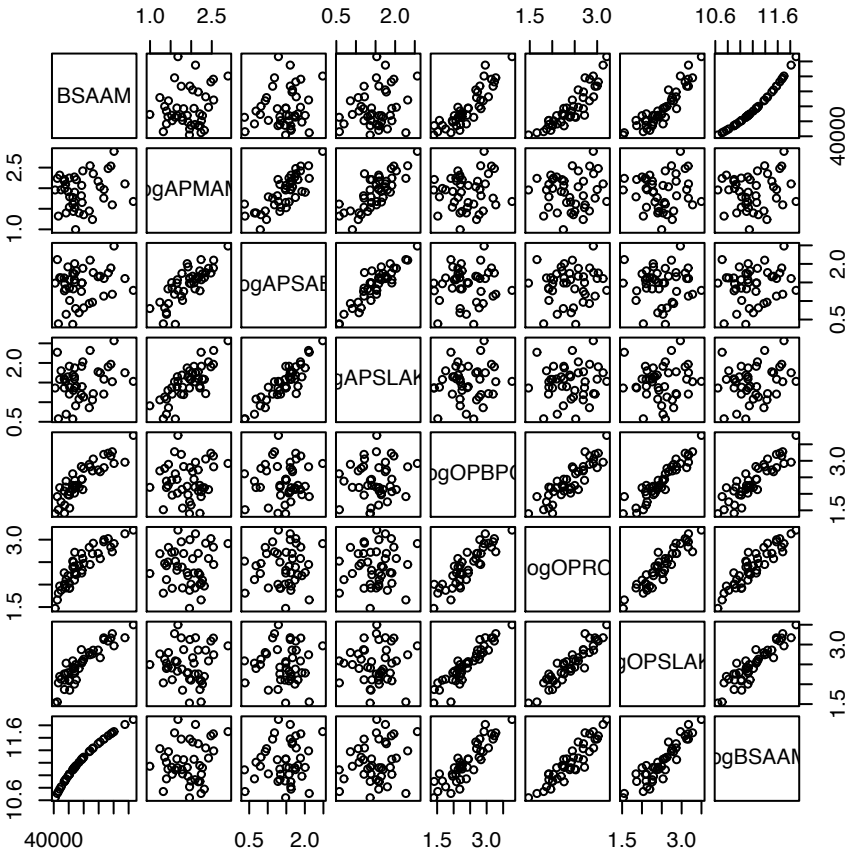
The key messages from the scatterplot matrix are: (1) the "O" measurements are very highly correlated, but the "A" measurements are less highly correlated; (2) there is no obvious dependence on time; (3) evidence of curvature in the marginal response plots, the last row of the scatterplot matrix, is weak.

Code for the automatic choice of a transformation is available in at least two sources: in the program Arc described by Cook and Weisberg (1999), and in the `alr3` package for R and S-Plus included with this book. Both give essentially identical output:

```
> summary(ans <- powerTransform( as.matrix(water[ , 2:7]) ~ 1))
bcPower Transformations to Multinormality

        Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
APMAM      0.0982   0.2861          -0.4625           0.6589
APSAB      0.3450   0.2032          -0.0533           0.7432
```

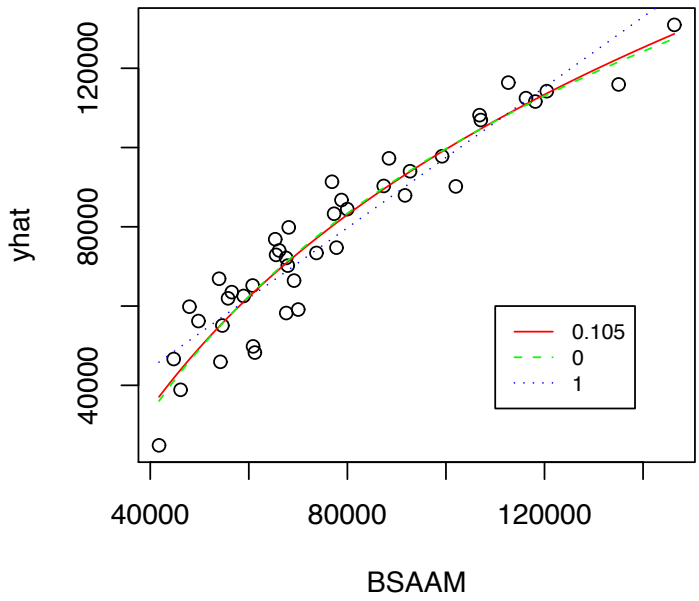
```
APSLAKE     0.0818    0.2185          -0.3466           0.5101
OPBPC       0.0982    0.1577          -0.2109           0.4073
OPRC        0.2536    0.2445          -0.2255           0.7328
OPSLAKE     0.2534    0.1763          -0.0921           0.5988

Likelihood ratio tests about transformation parameters
                                        LRT df         pval
LR test, lambda = (0 0 0 0 0 0)  5.452999  6 4.871556e-01
LR test, lambda = (1 1 1 1 1 1) 61.203125  6 2.562905e-11
```

The indication is to transform all the predictors to log scale, since the p-value for the LR test is about .49.



7.3.2. Given log transformations of the predictors, show that a log transformation of the response is reasonable.

Solution: Either the Box-Cox method, or the inverse response plot method, will indicate that the log transformation matches the data. Here is the inverse response plot produced using the function `inverse.response.plot` in R:



The lines shown on the plot are for $\hat{\lambda} = .10$, the nonlinear LS estimate, and for $\lambda = 0, 1$. The standard error of the estimate is about .30, so zero, logarithms, is about 1/3 of a standard error from the estimate. On the plot, the lines for logs and for $\hat{\lambda} = .10$ are virtually identical. ∎

7.3.3. Consider the multiple linear regression model with mean function given by

$$\begin{aligned} \mathrm{E}(\log(y)|\mathbf{x}) \quad = \quad & \beta_0 + \beta_1 \log(APMAM) + \beta_2 \log(APSAB) \\ & + \beta_3 \log(APSLAKE) + \beta_4 \log(OPBPC) \\ & + \beta_5 \log(OPRC) + \beta_6 \log(OPSLAKE) \end{aligned}$$

with constant variance function. Estimate the regression coefficients using OLS. You will find that two of the estimates are negative; which are they? Does a negative coefficient make any sense? Why are the coefficients negative?

Solution:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   9.4667     0.1235   76.63   <2e-16
logAPMAM     -0.0203     0.0660   -0.31   0.7597
logAPSAB     -0.1030     0.0894   -1.15   0.2567
logAPSLAKE    0.2206     0.0896    2.46   0.0187
logOPBPC      0.1113     0.0817    1.36   0.1813
logOPRC       0.3616     0.1093    3.31   0.0021
logOPSLAKE    0.1861     0.1314    1.42   0.1652

Residual standard error: 0.102 on 36 degrees of freedom
Multiple R-Squared: 0.91,       Adjusted R-squared: 0.895
F-statistic: 60.5 on 6 and 36 DF,  p-value: <2e-16
```

The negative coefficients are for two of the (nonsignificant) "A" terms. The negative signs are due to correlations with other terms already included in the mean function. ∎

7.3.4. In the OLS fit, the regression coefficient estimates for the three predictors beginning with "O" are approximately equal. Are there conditions under which one might expect these coefficients to be equal? What are they? Test the hypothesis that they are equal against the alternative that they are not all equal.

Solution: Fit two models, one with six terms plus the intercept, the other replacing the logarithms of the "O" terms by their sum. The anova comparing these models is:

```
Model 1: logBSAAM ~ logAPMAM + logAPSAB + logAPSLAKE + water$sum
Model 2: logBSAAM ~ logAPMAM + logAPSAB + logAPSLAKE + logOPBPC + logOPRC +
    logOPSLAKE
  Res.Df      RSS Df Sum of Sq       F  Pr(>F)
1     38 0.405364
2     36 0.372435  2  0.032929 1.59149 0.21762
```

The sum is as good as the individuals. Suppose that all three "O" measurements were depth of snowfall in the same mountain valley. The total snow, which is proportional to the amount of runoff at Bishop, the response, is the depth times the surface area. If all three are in the same valley, they correspond to the same surface area. Thus the average of the three might give a better estimate of average depth in the whole valley, and so the average or sum could do as well as the three measurements. The average of the logarithms corresponds to the log of the geometric means of the depths.∎

7.3.5. Write one or two paragraphs that summarize the use of the snowfall variables to predict runoff. The summary should discuss the important predictors, give useful graphical summaries, and give an estimate of variability. Be creative.

7.4 The data in the file `salarygov.txt` give the maximum monthly salary for 495 non-unionized job classes in a midwestern governmental unit in 1986. The variables are described in Table 7.3.

7.4.1. The data as given has as its unit of analysis the *job class*. In a study of the dependence of maximum salary on skill, one might prefer to have as

Table 7.3 The governmental salary data.

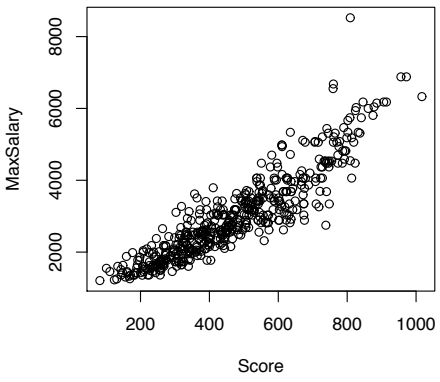
Variable	Description
MaxSalary	Maximum salary in dollars for employees in this job class, the response
NE	Total number of employees currently employed in this job class
NW	Number of women employees in the job class
Score	Score for job class based on difficulty, skill level, training requirements and level of responsibility as determined by a consultant to the governmental unit. This value for these data is in the range between 82 to 1017.
JobClass	Name of the job class; a few names were illegible or partly illegible

unit of analysis the *employee*, not the job class. Discuss how this preference would change the analysis.

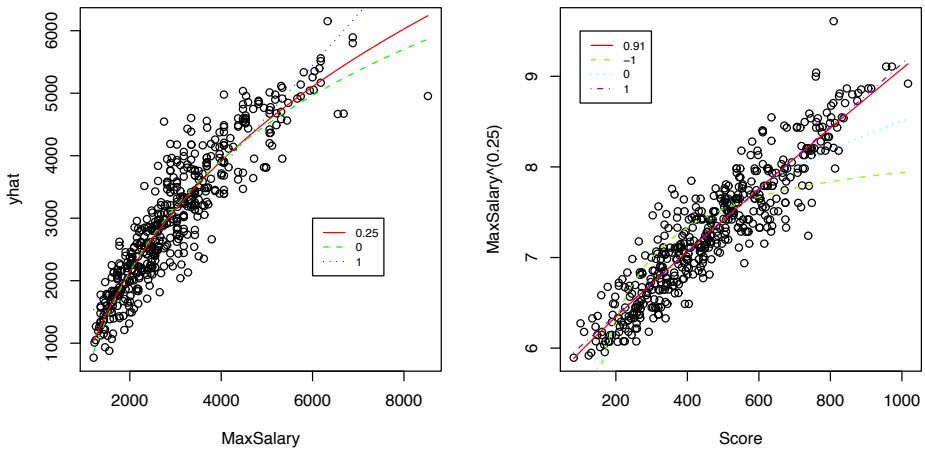
Solution: If the unit of analysis were the employee, then the data file should have $\sum NE$ cases, with the ith job class repeated NE_i times. The same results are obtained by using weighted least squares, with weights given by the *NE*. ∎

7.4.2. Examine the scatterplot of *MaxSalary* versus *Score*. Find transformation(s) that would make the mean function for the resulting scatter plot approximately linear. Does the transformation you choose also appear to achieve constant variance?

Solution:



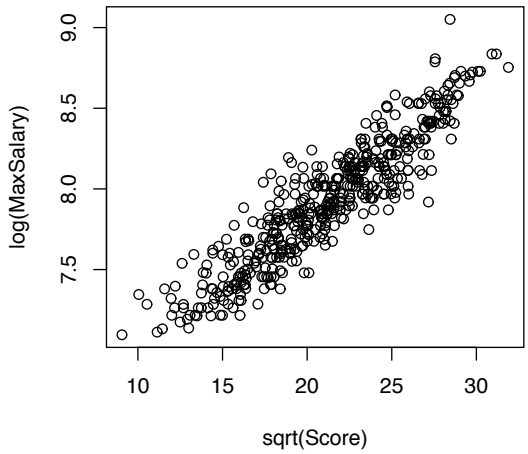
Variability increases from left to right. The mean function for this graph might be a polynomial, like a quadratic, so a power transformation of *Score* with powers in $(-1, 1)$ will not be helpful. We can start by transforming *Salary*. Using an response plot,



The left plot is the inverse response plot, while the right plot is of $Salary^{1/4}$ versus *Score*. Linearity seems to be achieved here, as the difference between untransformed horizontal axis and transforming the horizontal axis to the optimal 0.9 power is clearly unimportant.

Alternatively, one could transform the two variables simultaneously toward bivariate normality:

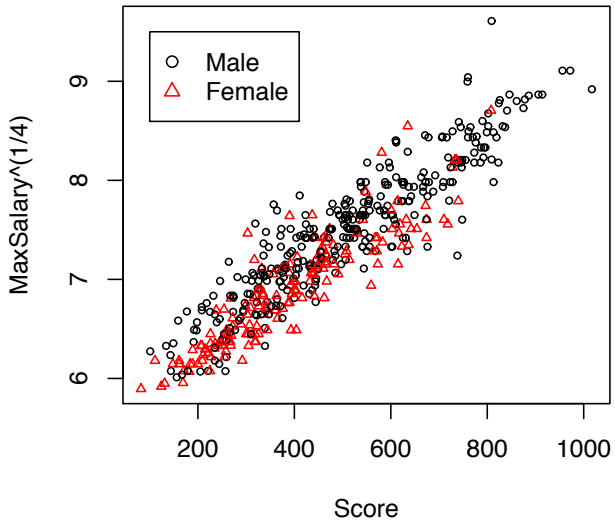
```
> summary(tran1 <- powerTransform(cbind(Score, MaxSalary) ~ 1, salarygov))
bcPower Transformations to Multinormality

           Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Score         0.5974   0.0691           0.4619           0.7329
MaxSalary    -0.0973   0.0770          -0.2483           0.0537

Likelihood ratio tests about transformation parameters
                                  LRT df       pval
LR test, lambda = (0 0)    125.090145  2 0.00000000
LR test, lambda = (1 1)    211.070400  2 0.00000000
LR test, lambda = (0.5 0)    7.615758  2 0.02219521

> plot(sqrt(Score),log(MaxSalary))
```



This suggests transforming *Score* to square root scale, and *Salary* to log-scale. This will also achieve linearity. Either approach is useful. Both approaches also seem to overcome the problem of nonconstant variance as well. ∎

7.4.3. According to Minnesota statutes, and probably laws in other states as well, a job class is considered to be female dominated if 70% of the employees or more in the job class are female. These data were collected to examine whether female-dominated positions are compensated at a lower level, adjusting for *Score*, than are other positions. Create a factor with two levels that divides the job classes into female dominated or not, fit appropriate models, and summarize your results. Be mindful of the need to transform variables, and the possibility of weighting.

Solution: We begin by drawing the scatterplot with the points colored and marked according to group:



We see that the female-dominated job classes are generally lower-score classes. What we don't see in the graph is that the higher score classes tend to be very small, some with just one position represented.

We use WLS to compare the four models of no sex effect (model 4), parallel lines (model 2), common intercept (model 3), and the general model (model 1). The results are:

```
        Res.Df   RSS  Df Sum of Sq      F Pr(>F)
Model 4    493 307.2
Model 2    492 244.3   1      62.9 126.45 <2e-16
Model 1    491 244.2   1       0.1   0.27    0.6

        Res.Df   RSS  Df Sum of Sq     F  Pr(>F)
Model 4    493 307.2
Model 3    492 251.4   1      55.8 112.2 < 2e-16
Model 1    491 244.2   1       7.2  14.5 0.00016
```

Model 1 is a clear improvement over model 3, but models 2 and 1 are not different; model 4 is not acceptable. We proceed with the parallel mean functions model.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.74e+00   3.17e-02   181.1   <2e-16
Score        3.53e-03   6.88e-05    51.3   <2e-16
fclass      -2.29e-01   2.03e-02   -11.3   <2e-16
```

```
Residual standard error: 0.705 on 492 degrees of freedom
Multiple R-Squared: 0.891
F-statistic: 2.02e+03 on 2 and 492 DF,  p-value: <2e-16
```

According to this the intercept for female-dominated classes is .229 lower than the intercept for male-dominated classes. For example, for a score of 500, the 95% prediction intervals for male and female-dominated classes are:

```
                    fit    lwr    upr
Male dominated   7.5072 6.1224 8.8920
Female dominated 7.2785 5.8935 8.6635
```

Raising the end-points to the fourth power translates to the original dollar scale:

```
Transformed to dollar scale:
                     fit    lwr    upr
Male dominated    3176.3 1405.1 6251.7
Female dominated  2806.5 1206.4 5633.6
```

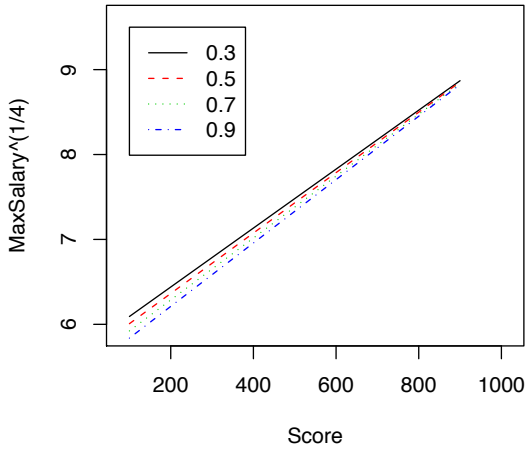
For a job class with score 500, the lower end-point of the interval is about \$200 less, and the upper end-point is about \$600 less. ∎

7.4.4. An alternative to using a factor for female dominated jobs is to use a term NW/NE, the fraction of women in the job class. Repeat the last problem, but encoding the information about sex using this variable in place of the factor.

Solution: We can fit the same four models replacing a factor for female-dominated class by a new variable NW/NE. Again using WLS, with NE as weights,

```
Analysis of Variance Table

Model 1: MaxSalary^(1/4) ~ Score
Model 2: MaxSalary^(1/4) ~ Score + ffrac
Model 3: MaxSalary^(1/4) ~ Score + ffrac + Score:ffrac
  Res.Df   RSS  Df Sum of Sq      F Pr(>F)
1    493 307.2
2    492 253.0   1      54.2 106.36 <2e-16
3    491 250.2   1       2.8   5.41  0.020
> anova(n4,n3,n1)
Analysis of Variance Table

Model 1: MaxSalary^(1/4) ~ Score
Model 2: MaxSalary^(1/4) ~ Score + Score:ffrac
Model 3: MaxSalary^(1/4) ~ Score + ffrac + Score:ffrac
  Res.Df RSS  Df Sum of Sq    F  Pr(>F)
1    493 307
2    492 270   1        37 72.6 < 2e-16
3    491 250   1        20 39.2 8.4e-10
```

The p-value for comparing the general model to the parallel model suggests some evidence that the general model is to be preferred. Here is a plot of the predicted *MaxSalary* for *ffrac* $\in (.3, .5, .7, .9)$.



For job classes with a large score, there is little or no difference in expected salary for different values of *NW/NE*. For job classes with low scores, the larger the fraction of women, the lower the expected salary. This analysis differs slightly from that of the last subproblem, but the general conclusions are similar. Using different transformations, or ignoring weights, could lead to different conclusions. ∎

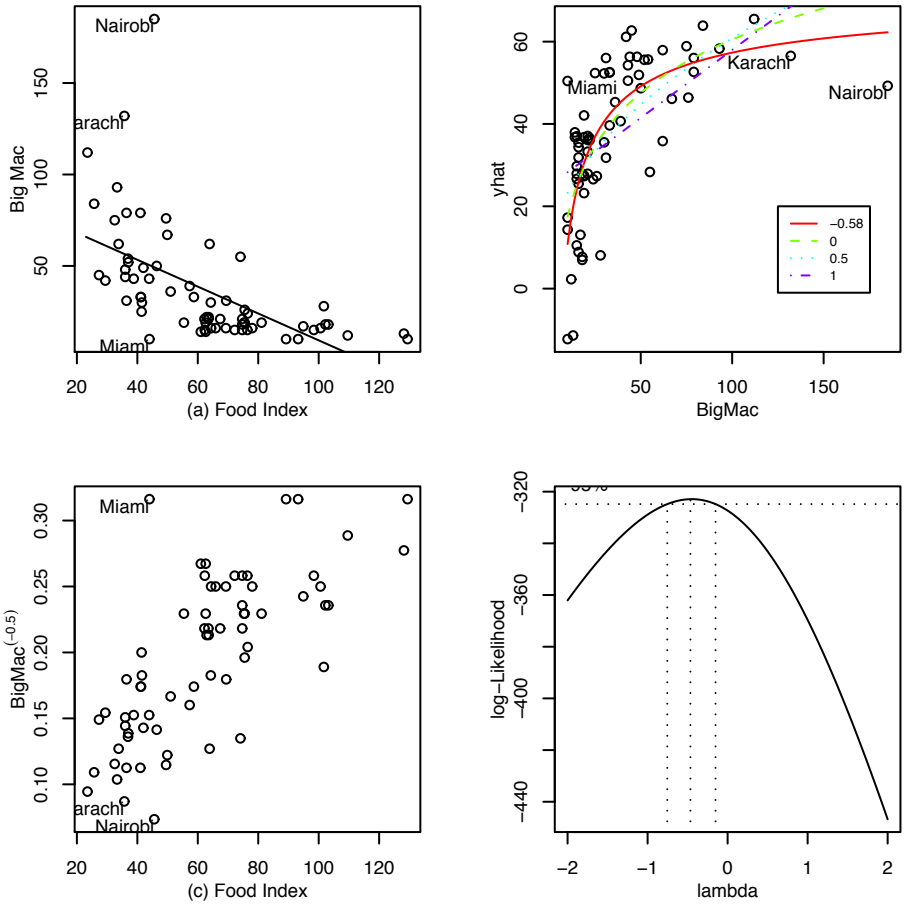
7.5 World cities The Union Bank of Switzerland publishes a report entitled *Prices and Earnings Around the Globe* on their internet web site, www.ubs.com. The data in the file `BigMac2003.txt` and described in Table 7.4 are taken from their 2003 version for 69 world cities.

7.5.1. Draw the scatterplot with *BigMac* on the vertical axis and *FoodIndex* on the horizontal axis. Provide a qualitative description of this graph. Use an inverse response plot and the Box-Cox method to find a transformation of *BigMac* so that the resulting scatterplot has a linear mean function. Two of the cities, with very large values for *BigMac*, are very influential for selecting a transformation. You should do this exercise with all the cities, and with those two cities removed.

Solution:

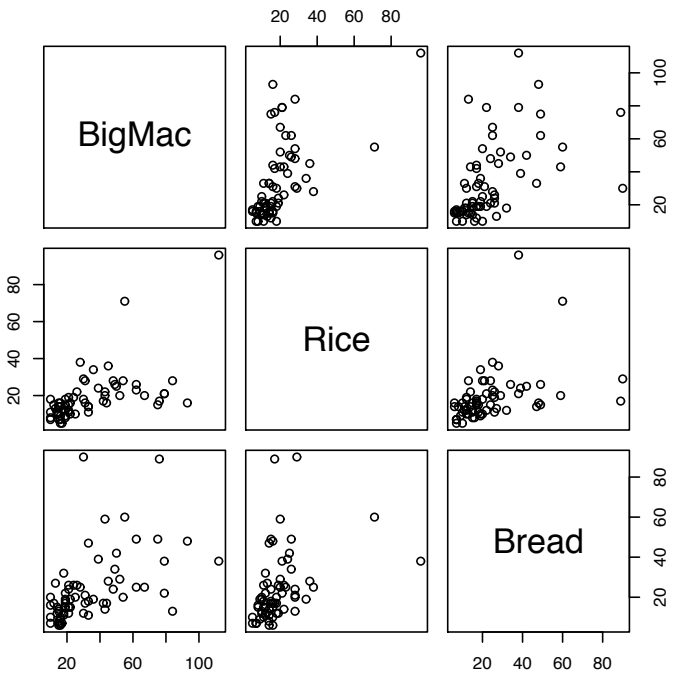
Table 7.4 Global price comparison data. Most of the data are from the Union Bank of Switzerland publication *Prices and Earnings Around the Globe*, 2003 edition, from www.ubs.com.

Variable	Description
BigMac	Minutes of labor to buy a BigMac hamburger based on a typical wage averaged over thirteen occupations
Bread	Minutes of labor to buy one kg bread
Rice	Minutes of labor to buy 1 kg of rice
Bus	Lowest cost of 10 km public transit
FoodIndex	Food price index, Zurich=100
TeachGI	Primary teacher's gross annual salary, thousands of US dollars
TeachNI	Primary teacher's net annual salary, thousands of US dollars
TaxRate	$100 \times (TeachGI - TeachNI)/TeackGI$. In some places, this is negative, suggesting a government subsidy rather than tax
TH	Teacher's hours per week of work
Apt	Monthly rent in US dollars of a typical three room apartment
City	City name



Plot (a) shows the scatterplot, and it indicates that the real cost of a Big Mac, which is the amount of work required to buy one, declines with overall food prices; the Big Mac is cheapest, for the local people, in the wealthiest countries. The inverse response plot in (b) is used to select a transformation; four choices are shown, and the most extreme, with power of about $-.5$, appears to match the best, although the improvement over the logarithmic transformation is small. This choice is influenced by Nairobi and Karachi, and without these points a log transformation is consistent with the plots. Plot (c) shows that in the transformed scale linearity is achieved, and (d) shows that the Box-Cox procedure essentially agrees with the inverse response plot. In summary, either a log transform or the inverse square root scale seem to be appropriate. ∎

7.5.2. Draw the scatterplot matrix of the three variables (*BigMac*, *Rice*, *Bread*), and use the multivariate Box-Cox procedure to decide on normalizing transformations. Test the null hypothesis that $\boldsymbol{\lambda} = (1, 1, 1)'$ against a general alternative. Does deleting Karachi and Nairobi change your conclusions?
Solution:



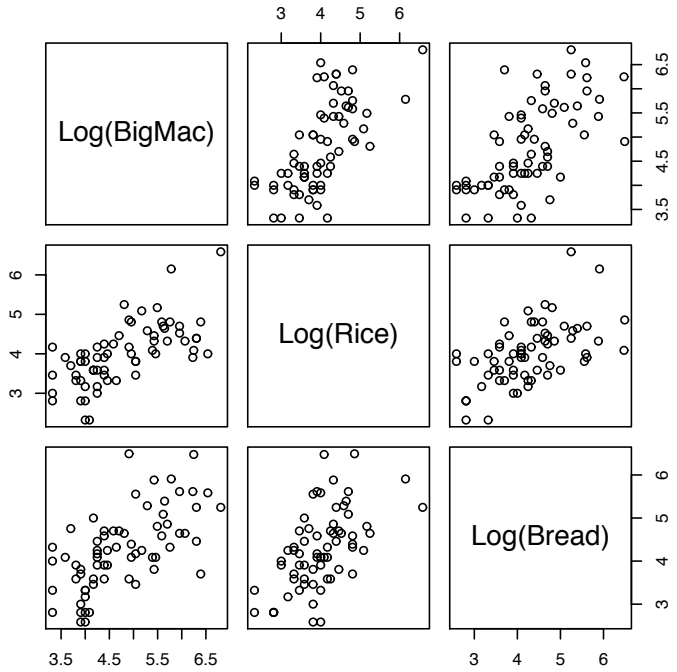
The scatterplot matrix indicates the need to transform because the points are clustered with curvature obvious. The results of the multivariate Box-Cox procedure are,

```
> summary(pows <- powerTransform(cbind(BigMac, Rice, Bread) ~ 1, BigMac2003))
bcPower Transformations to Multinormality

       Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
BigMac   -0.3035   0.1503          -0.5980          -0.0089
Rice     -0.2406   0.1345          -0.5043           0.0230
Bread    -0.1566   0.1466          -0.4439           0.1307

Likelihood ratio tests about transformation parameters
                                    LRT df       pval
LR test, lambda = (0 0 0)      7.683155  3 0.05303454
LR test, lambda = (1 1 1)    204.555613  3 0.00000000
LR test, lambda = (-0.5 0 0)   6.605247  3 0.08560296
> summary(pow1s<-powerTransform(cbind(BigMac, Rice, Bread) ~ 1, BigMac2003,
+    subset=-c(26, 46)))
bcPower Transformations to Multinormality

       Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
BigMac   -0.2886   0.1742          -0.6301           0.0529
Rice     -0.2465   0.1413          -0.5235           0.0305
Bread    -0.1968   0.1507          -0.4922           0.0986

Likelihood ratio tests about transformation parameters
                                   LRT df       pval
LR test, lambda = (0 0 0)   7.083917  3 0.06927061
LR test, lambda = (1 1 1) 181.891304  3 0.00000000
```

In these data in R, the city names are given as the row labels, so we can refit the power transformation diagnostics without these cities using:

```
> summary(pows <- powerTransform(cbind(BigMac, Rice, Bread) ~ 1,
        data=BigMac2003[-c("Karachi", "Nairobi"), ]))
```

The resulting transformations, not shown here, are not very different from the transformations using all the data, and logs of all three seem to be appropriate. The scatterplot matrix for the transformed variables is



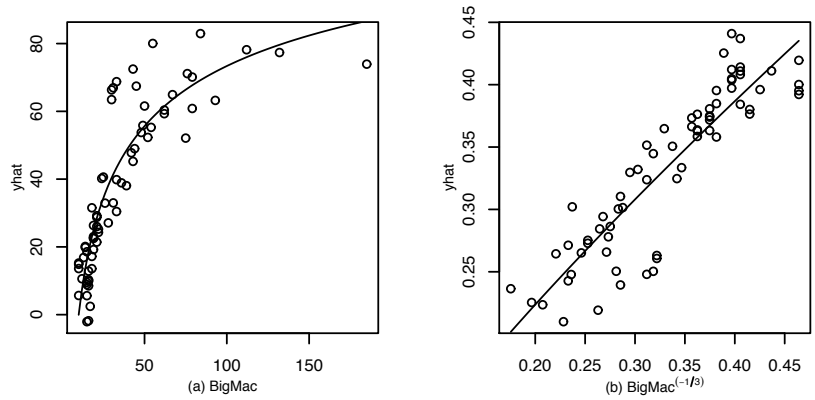
■

7.5.3. Set up the regression using the four terms, log(*Bread*), log(*Bus*), log(*TeachGI*), and $Apt^{0.33}$, and with response *BigMac*. To get $Apt^{0.33}$ using R, you need to use the *AsIs* function `I()`,

```
> m3 <- lm(BigMac ~ log(Bread) + log(Bus) + log(TeachGI) +
>           I(Apt^0.33), BigMac)
```

Draw the inverse response plot of $\hat{y}$ versus *BigMac*. Estimate the best power transformation. Check on the adequacy of your estimate by refitting the regression model with the transformed response and drawing the inverse response plot again. If transformation was successful, this second inverse response plot should have a linear mean function.

Solution:

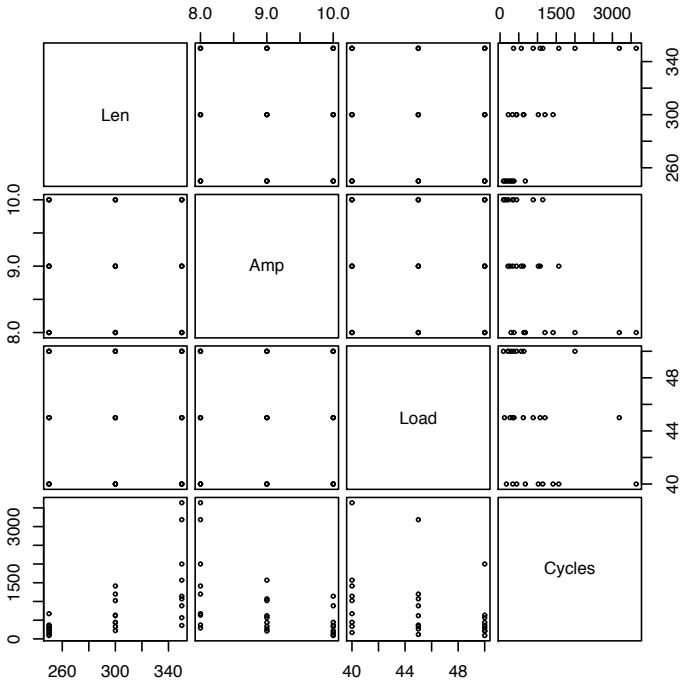


Panel (a) is the inverse response plot before transformation. The matching curve corresponds to $\hat{\lambda} \approx -1/3$. Panel (b) is the inverse response plot after transforming *BigMac*. No further transformation seems necessary. ∎

7.6 The data in the file `wool.txt` were introduced in Section 6.3. For this problem, we will start with *Cycles*, rather than its logarithm, as the response.

7.6.1. Draw the scatterplot matrix for these data and summarize the information in this plot.

Solution:



The regular pattern of the points is typical for a designed experiment like this one. Transformations of predictors are not appropriate for these data, as the untransformed predictors are already linear predictors. The mean functions for the plots including *Cycles* are curved, so transforming *Cycles* might help. ∎

7.6.2. View all three predictors as factors with three levels, and *without transforming* Cycles, fit the second-order mean function with terms for all main effects and all two-factor interactions. Summarize results.

Solution:

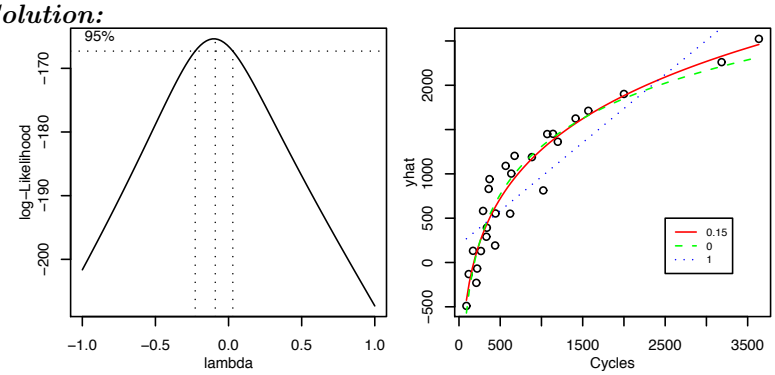
	Df	Sum Sq	Mean Sq	F value	Pr(>F)
AmpF	2	5624248.96	2812124.48	231.95	0.0000
LoadF	2	1753096.96	876548.48	72.30	0.0000
LenF	2	8182252.52	4091126.26	337.44	0.0000
AmpF:LoadF	4	283609.04	70902.26	5.85	0.0168
AmpF:LenF	4	3555537.48	888884.37	73.32	0.0000
LoadF:LenF	4	732881.48	183220.37	15.11	0.0008
Residuals	8	96991.85	12123.98		

All main effects and all interactions are significant. ∎

7.6.3. Fit the first-order mean function consisting only of the main effects. From Problem 7.6.2, this mean function is not adequate for these data based

on using *Cycles* as the response. Use both the inverse response plot and the Box-Cox method to select a transformation for *Cycles* based on the first-order mean function.

Solution:



The first plot is the profile log-likelihood for the Box-Cox method, and the second is the inverse response plot, with the fitted line from setting $\lambda = 0$. Log transformations are suggested. ∎

7.6.4. In the transformed scale, refit the second-order model, and show that none of the interactions are required in this scale. For this problem, the transformation leads to a much simpler model than is required for the response in the original scale. This is an example of *removable nonadditivity*.

Solution:

```
> m4 <- lm(log(Cycles) ~ AmpF+LoadF+LenF, data=wool)
> m5 <- lm(log(Cycles) ~ (AmpF+LoadF+LenF)^2, data=wool)
> anova(m4,m5)
Analysis of Variance Table

Model 1: log(Cycles) ~ AmpF + LoadF + LenF
Model 2: log(Cycles) ~ (AmpF + LoadF + LenF)^2
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     20 0.717
2      8 0.166 12     0.552 2.22   0.13
```

∎

7.7 Justify transforming *Miles* in the Fuel data.

Solution: The range for *Miles* is from 1534 to 300,767, and according to the log rule, transformation of *Miles* to log scale is justified as a starting point because the range is about two orders of magnitude. We can see if further transformation is desirable using the multivariate Box-Cox method:

```
> summary(b1 <- powerTransform(cbind(Tax, Dlic, Income, log2(Miles)) ~ 1,
+         data=fuel2001))
bcPower Transformations to Multinormality

       Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
```

Table 7.5 Description of variables in the data file `UN3.txt`. The data were collected from http://unstats.un.org/unsd/demographic, and refer to values collected between 2000 and 2003.

Variable	Description
Locality	Country/locality name
ModernC	Percent of unmarried women using a modern method of contraception
Change	Annual population growth rate, percent
PPgdp	Per capita gross national product, US dollars
Frate	Percent of females over age 15 economically active
Pop	Total 2001 population, 1000s
Fertility	Expected number of live births per female, 2000
Purban	Percent of population that is urban, 2001

```
Tax       1.8493   0.4803           0.9079           2.7907
Dlic      2.2669   1.3671          -0.4127           4.9464
Income   -0.5104   0.8432          -2.1631           1.1423
          6.4715   1.4063           3.7151           9.2280

Likelihood ratio tests about transformation parameters
                                        LRT df         pval
LR test, lambda = (0 0 0 0)    47.18163  4 1.397693e-09
LR test, lambda = (1 1 1 1)    25.41973  4 4.141995e-05
LR test, lambda = (1 1 1 6.47)  7.68158  4 1.039638e-01
```

The suggested transformation parameter for log(*Miles*) is well outside the usual range of -2 to 2, and so we would conclude that no further transformation is needed.

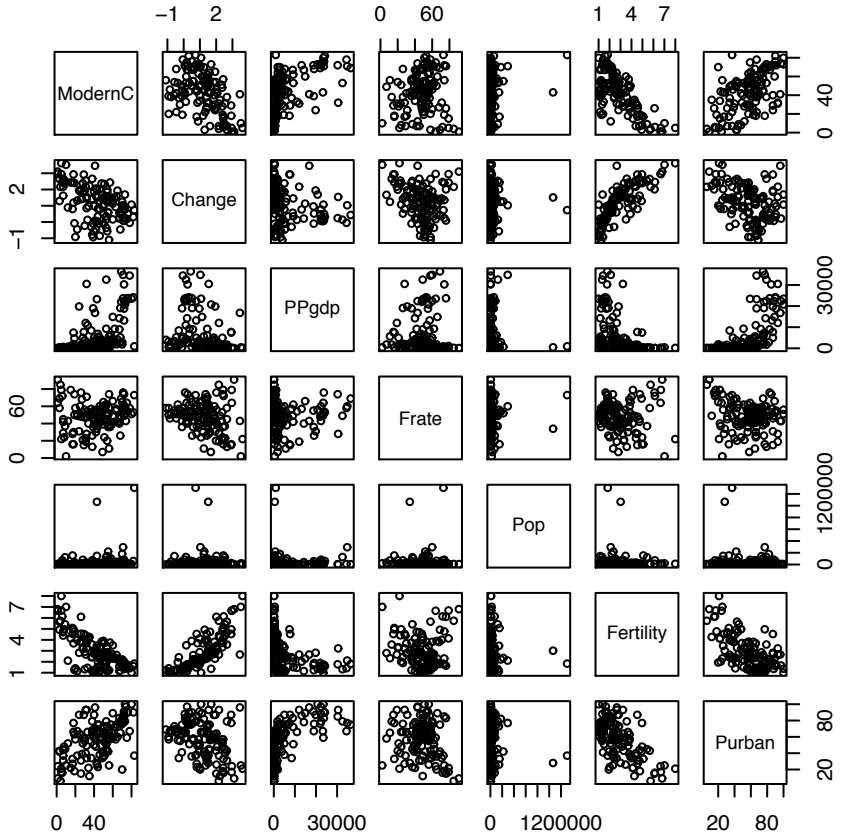
If you start with the Box-Cox method before replacing *Miles* with log(*Miles*), a square root transformation is suggested as better than the logarithmic. However, changes in scale for the predictors are less important than changes in scale for the response, and there is probably little difference between using these two transformations. The logarithmic is preferred because it is easier to interpret. ∎

7.8 The file `UN3.txt` contains data described in Table 7.5. There are data for $n = 125$ localities, mostly UN member countries, for which values are observed for all the variables recorded.

Consider the regression problem with *ModernC* as the response variable, and the other variables in the file as defining terms.

7.8.1. Select appropriate transformations of the predictors to be used as terms. (Hint: Since *Change* is negative for some localities, the Box-Cox family of transformations cannot be used without either adding a constant or using the Yeo-Johnson family of transformations.)

Solution: Start by drawing the scatterplot matrix,

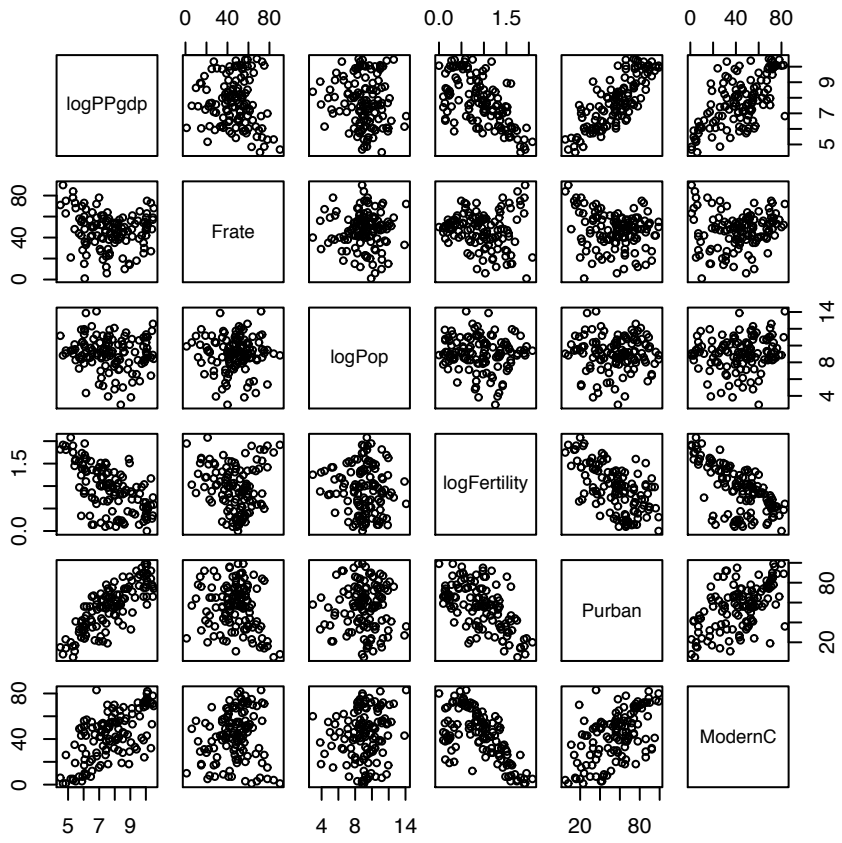


Change is sometimes negative, but it also has a fairly narrow range, so transforming it is unlikely to help. *Pop*, on the other hand, is highly variable, and almost certain to need to be transformed. Since we won't consider transforming *Change* we can use the multivariate Box-Cox method with the Box-Cox family of transformations rather than the more complex Yeo-Johnson transformations:

```
> summary(bc <- powerTransform(cbind(PPgdp, Frate, Pop, Fertility, Purban) ~ 1,
+    data=UN3))
bcPower Transformations to Multinormality

           Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
PPgdp        -0.0731   0.0462          -0.1636           0.0173
Frate         1.0659   0.1550           0.7621           1.3696
Pop           0.0391   0.0312          -0.0221           0.1002
Fertility     0.0868   0.1428          -0.1930           0.3666
```

```
Purban        0.8432   0.1384           0.5720           1.1145

Likelihood ratio tests about transformation parameters
                                          LRT df       pval
LR test, lambda = (0 0 0 0 0)  142.766852  5 0.0000000
LR test, lambda = (1 1 1 1 1) 1170.351668  5 0.0000000
LR test, lambda = (0 1 0 0 1)    5.269362  5 0.3838989
```

Transforming only *PPgdp*, *Pop* and *Fertility* using logarithms has an LRT with p-value of about 0.05, so the rounding to this convenient values is somewhat worse than using the values shown in the table under "Est. Power" for the transformation parameters. Nevertheless, this set of transformations is a distinct improvement over no transformation at all, and we will use these as an initial set of transformed terms. Notice in the scatterplot matrix below, the mean function of each plot appears reasonably linear.

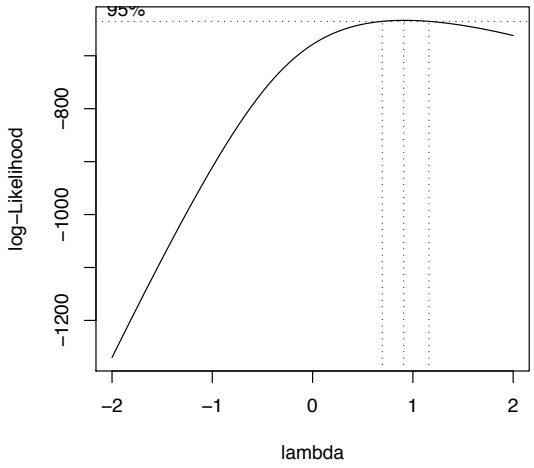


7.8.2. Given the transformed predictors as terms, select a transformation for the response.

Solution: Fitting the tentative model

```
> m1 <- lm(ModernC~logb(PPgdp,2)+Frate+logb(Pop,2)+logb(Fertility,2)
+              +Purban,data=un0)
```

both the Box-Cox method and the inverse response plot will suggest no further transformation; here is the plot for the Box-Cox method:



7.8.3. Fit the regression using the transformations you have obtained and summarize your results.

Solution: The fitted regression is

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -19.3558    15.2267   -1.27  0.20615
logb(PPgdp, 2)       4.4903     1.0442    4.30  3.5e-05
Frate                0.1678     0.0827    2.03  0.04468
logb(Pop, 2)         1.2221     0.4670    2.62  0.01002
logb(Fertility, 2)  -9.3489     2.4422   -3.83  0.00021
Purban               0.0157     0.1034    0.15  0.87970

Residual standard error: 14.5 on 119 degrees of freedom
Multiple R-Squared: 0.559
F-statistic: 30.1 on 5 and 119 DF,  p-value: <2e-16
```

All the coefficient estimates are clearly non-zero, apart from the coefficient for *Purban*; this term could probably be dropped form the mean function without

any loss. Model checking and residual plots are needed to verify that this is a reasonable fitted mean function, but that is the topic of Chapter 8. ∎

8

Regression Diagnostics: Residuals

Problems

8.1 Working with the Hat matrix

8.1.1. Prove the results given by (8.8) and (8.9).

Solution: This problem uses the matrix algebra result that for any matrices A, B and C, $\text{tr}(ABC) = \text{tr}(BCA)$, where "tr" means trace of a matrix, or the sum of its diagonal elements. Then:

$$\sum h_{ii} = \text{tr}(\mathbf{H}) = \text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \text{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) = \text{tr}(\mathbf{I}_{p'}) = p'$$

As in the text, $\mathbf{HX} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{XI} = \mathbf{X}$, so if $\mathbf{X}_j$ is any column of $\mathbf{X}$, $\mathbf{HX}_j = \mathbf{X}_j$. Let $\mathbf{1}$ be the column of ones, which is included in $\mathbf{X}$ because the mean function has an intercept, and so $\mathbf{H1} = \mathbf{1}$, which is in scalar form the same as (8.9). ∎

8.1.2. Prove that $1/n \le h_{ii} \le 1/r$, where h_{ii} is a diagonal entry in $\mathbf{H}$, and r is the number of rows in $\mathbf{X}$ that are exactly the same as $\mathbf{x}_i$.

Solution: That $1/n \le h_{ii}$ follows directly from (8.11). To prove the upper bound, we use the properties that $\mathbf{H} = \mathbf{H}^2 = \mathbf{HH}'$, $h_{ij} = h_{ji}$, and, if $\mathbf{x}_i = \mathbf{x}_j$, then $h_{ij} = h_{ji} = h_{ii}$. We can write

$$h_{ii} = \sum_{i=1}^{n} h_{ij}h_{ji} = \sum_{i=1}^{n} h_{ij}^2 \ge r h_{ii}^2$$

which simplifies to $h_{ii} \le 1/r$. ∎

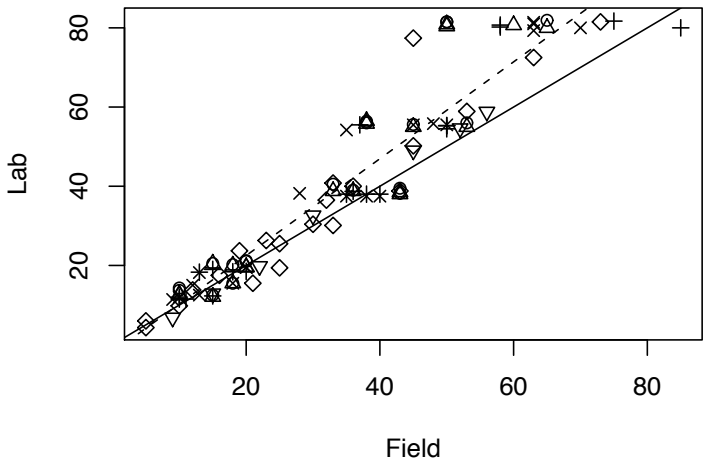
8.2 If the linear trend were removed from Figure 8.5f, what would the resulting graph look like?

Solution: It is be the same as Figure 8.5e. ∎

8.3 This example compares in-field ultrasonic measurements of the depths of defects in the Alaska oil pipeline to measurements of the same defects in a laboratory. The lab measurements were done in six different batches. The goal is to decide if the field measurement can be used to predict the more accurate lab measurement. Use the Lab measurement as the response variable and the Field measurement as the predictor variable. The data, in the file `pipeline.txt`, were given at www.itl.nist.gov/div898/handbook/pmd/section6/pmd621.htm. The three variables are called *Field*, the in-field measurement *Lab* the more accurate in-lab measurement, and *Batch*, the batch number.

8.3.1. Draw the scatterplot of *Lab* versus *Field*, and comment on the applicability of the simple linear regression model.

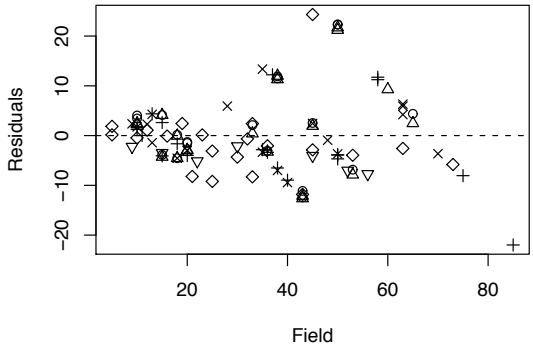
Solution:



Although not requested in the problem, a separate symbol has been used for each batch. A linear mean function seems plausible, but constant variance is unlikely. The solid line is the 45-degree line, and the dashed line is the OLS line. It appears that the field measurement underestimates depth for the deeper faults. ∎

8.3.2. Fit the simple regression model, and get the residual plot. Compute the score test for nonconstant variance and summarize your results.

Solution:



Here is the computer output for this problem using the `car` library in R/S-plus:

```
> m1 <- lm(Lab ~ Field, pipeline)
> ncvTest(m1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 29.586    Df = 1     p = 5.3499e-08
```

The score test for variance as a function of *Field* is $S = 29.59$ with 1 df, for a very small p-value. The conclusion is that variance increases with *Field*; deeper faults are less well measured. ∎

8.3.3. Fit the simple regression mean function again, but this time assume that $\mathrm{Var}(Lab|Field) = \sigma^2 \times Field$. Get the score test for the fit of this variance function. Also test for nonconstant variance as a function of batch; since the batches are arbitrarily numbered, be sure to treat *Batch* as a factor. (Hint: Both these tests are extensions of the methodology outlined in the text. The only change required is to be sure that the residuals defined by (8.13) are used when computing the statistic.)

Solution:

```
> m2 <- update(m1, weights=1/Field)
> ncvTest(m2)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 9.0315    Df = 1     p = 0.0026536
> ncvTest(m2, ~ factor(Batch))
Non-constant Variance Score Test
Variance formula: ~ factor(Batch)
Chisquare = 6.955    Df = 5     p = 0.22401
```

$S = 9.03$ with 1 df, and a tiny p-value, so this weighting scheme is not successful at modeling the nonconstant variance. The score statistics for *Batch*

is $S = 6.96$ with 5 df, for a p-value of about 0.23, so there is no evidence that the variability differs between batches. ∎

8.3.4. Repeat Problem 8.3.3, but with $\mathrm{Var}(Lab|Field) = \sigma^2 \times Field^2$.

Solution:

```
> m3 <- update(m1, weights=1/Field^2)
> ncvTest(m3)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.026989    Df = 1     p = 0.8695
```

The scores tests are $S = 0.027$ for *Field* with 1 df, and $S = 1.85$ with 5 df for *Batch*, both with large p-values. There is no evidence of an incorrectly specified variance function. ∎

8.4 Refer to Problem 7.2, page 111. Fit Hald's model, given in Problem 7.2.3, but with constant variance, $\mathrm{Var}(Distance|Speed) = \sigma^2$. Compute the score test for nonconstant variance for the alternatives that (a) variance depends on the mean; (b) variance depends on *Speed*; and (c) variance depends on *Speed* and $Speed^2$. Is adding $Speed^2$ helpful?

Solution:

```
> ncvTest(m1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 22.97    Df = 1     p = 1.6454e-06
> ncvTest(m1, ~ Speed)
Non-constant Variance Score Test
Variance formula: ~ Speed
Chisquare = 23.392    Df = 1     p = 1.3212e-06
> ncvTest(m1, ~ Speed + I(Speed^2))
Non-constant Variance Score Test
Variance formula: ~ Speed + I(Speed^2)
Chisquare = 23.466    Df = 2     p = 8.0262e-06
```

The change in the score test for adding $Speed^2$ is $23.466 - 23.392 = 0.047$ with 1 df, an insignificant improvement.∎

8.5 Consider the simple regression model, $\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x$ with $\mathrm{Var}(Y|X = x) = \sigma^2$.

8.5.1. Find a formula for the h_{ij} and for the leverages h_{ii}.

Solution: $\mathbf{H}$ is $n \times n$ even for simple regression. Using (3.17) to get $(\mathbf{X}'\mathbf{X})^{-1}$, a formula may be obtained for an individual h_{ij}. We find

$$\begin{aligned} h_{ij} &= \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \\ &= (1\ x_i)\begin{pmatrix} \frac{\sum x_i^2}{nSXX} & -\frac{\bar{x}}{SXX} \\ -\frac{\bar{x}}{SXX} & \frac{1}{SXX} \end{pmatrix}\begin{pmatrix} 1 \\ x_j \end{pmatrix} \\ &= \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{SXX} \end{aligned}$$

By setting j equal to i,

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX} \tag{8.25}$$

∎

8.5.2. In a 2D plot of the response versus the predictor in a simple regression problem, explain how high-leverage points can be identified.

Solution: Cases with large h_{ii} will have $(x_i - \bar{x})^2$ large, and will therefore correspond to observations at the extreme left or right of the scatterplot. ∎

8.5.3. Make up a predictor X so that the value of the leverage in simple regression for one of the cases is equal to one.

Solution: Let the predictor X consist of the value zero appearing $(n-1)$ times and the value 1 appearing exactly once. For this variable, $\bar{x} = 1/n$, and $SXX = 1 - 1/n$, and the leverage for the case with value one is $h = 1/n + (1 - 1/n)^2/(1 - 1/n) = 1$. ∎

8.6 Using the QR factorization defined in Appendix A.98, show that $\mathbf{H} = \mathbf{QQ}'$ Hence, if q_i is the ith row of $\mathbf{Q}$,

$$h_{ii} = q_i' q_i \qquad\qquad h_{ij} = q_i' q_j$$

Thus if the QR factorization of $\mathbf{X}$ is computed, the h_{ii} and the h_{ij} are easily obtained.

Solution: Since $\mathbf{X} = \mathbf{QR}$ and $(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{R}'\mathbf{Q}'\mathbf{QR})^{-1} = (\mathbf{R}'\mathbf{R})^{-1} = \mathbf{R}^{-1}(\mathbf{R}')^{-1}$, and so

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{QRR}^{-1}(\mathbf{R}')^{-1}\mathbf{R}'\mathbf{Q}' = \mathbf{QQ}'$$

∎

8.7 Let $\mathbf{U}$ be an $n \times 1$ vector with one as its ith element and zeroes elsewhere. Consider computing the regression of $\mathbf{U}$ on an $n \times p'$ full rank matrix $\mathbf{X}$. As usual, let $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ be the Hat matrix with elements h_{ij}.

8.7.1. Show that the elements of the vector of fitted values from the regression of $\mathbf{U}$ on $\mathbf{X}$ are the h_{1j}, $j = 1, 2, \ldots, n$.

Solution: The fitted values are $\mathbf{HU}$, which will pick out the i-th column of $\mathbf{H}$. ∎

8.7.2. Show that the vector of residuals have $1 - h_{11}$ as the first element, and the other elements are $-h_{1j}, j > 1$.

Solution: The residuals are $\mathbf{U} - \mathbf{HU}$, giving the values specified in the problem. ∎

8.8 Two $n \times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ are *orthogonal* if $\mathbf{AB} = \mathbf{BA} = \mathbf{0}$. Show that $\mathbf{I} - \mathbf{H}$ and $\mathbf{H}$ are orthogonal. Use this result to show that as long as the intercept is in the mean function, the slope of the regression of $\hat{\mathbf{e}}$ on $\hat{\mathbf{Y}}$ is zero. What is the slope of the regression of $\hat{\mathbf{e}}$ on $\mathbf{Y}$?

Solution:

$$\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{H} - \mathbf{H}^2 = \mathbf{H} - \mathbf{H} = \mathbf{0}$$

so $\mathbf{H}$ and $\mathbf{I}-\mathbf{H}$ are orthogonal. The numerator of the OLS slope in the simple regression of $\hat{\mathbf{e}}$ on $\hat{\mathbf{Y}}$ is $(\hat{\mathbf{e}}-\bar{\hat{e}}\mathbf{1})'(\hat{\mathbf{Y}}-\bar{\hat{\mathbf{Y}}}\mathbf{1})$, where $\mathbf{1}$ is a column of ones. As long as the intercept is in the mean function, $\bar{\hat{e}}=0$, and the numerator reduces to $\hat{\mathbf{e}}'\hat{\mathbf{Y}}=\mathbf{Y}(\mathbf{I}-\mathbf{H})\mathbf{H}\mathbf{Y}=\mathbf{0}$.

The slope of the regression of $\hat{\mathbf{e}}$ on $\mathbf{Y}$ is $\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}/(\mathbf{Y}-\bar{y}\mathbf{1})'(\mathbf{Y}-\bar{y}\mathbf{1})=RSS/SYY$. ∎

8.9 Suppose that $\mathbf{W}$ is a known positive diagonal matrix of positive weights, and we have a weighted least squares problem,

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{e}\quad \mathrm{Var}(\mathbf{e})=\hat{\sigma}^2\mathbf{W}^{-1} \tag{8.26}$$

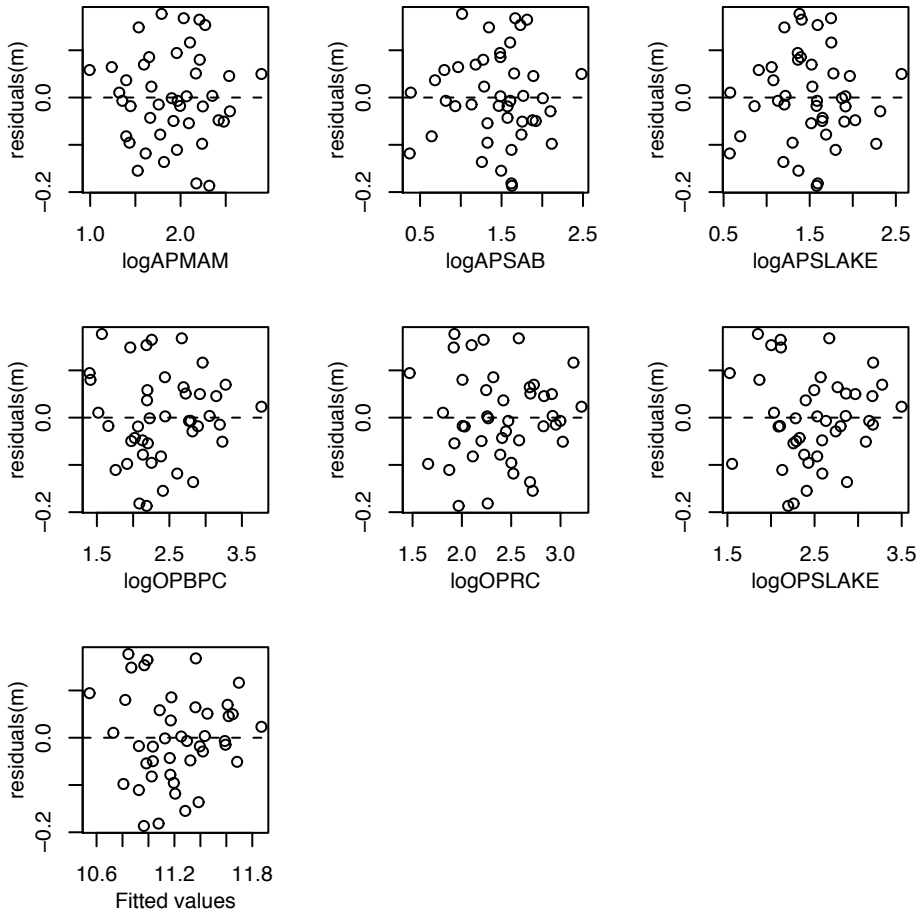
Using the transformations as in Section 5.1, show that the hat matrix is given by (8.12).

Solution: Writing $\mathbf{Z}=\mathbf{W}^{1/2}\mathbf{Y}$ and $\mathbf{M}=\mathbf{W}^{1/2}\mathbf{X}$, Section 5.1 shows that the regression of $\mathbf{Z}$ on $\mathbf{M}$ has a constant variance function and can be solved using OLS. For the unweighted problem, the hat matrix is $\mathbf{H}=\mathbf{M}(\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'$, which is the same as (8.12). ∎

8.10 Draw residuals plots for the mean function described in Problem 7.3.3 for the California water data, and comment on your results. Test for curvature as a function of fitted values. Also, get marginal model plots for this model.

Solution:

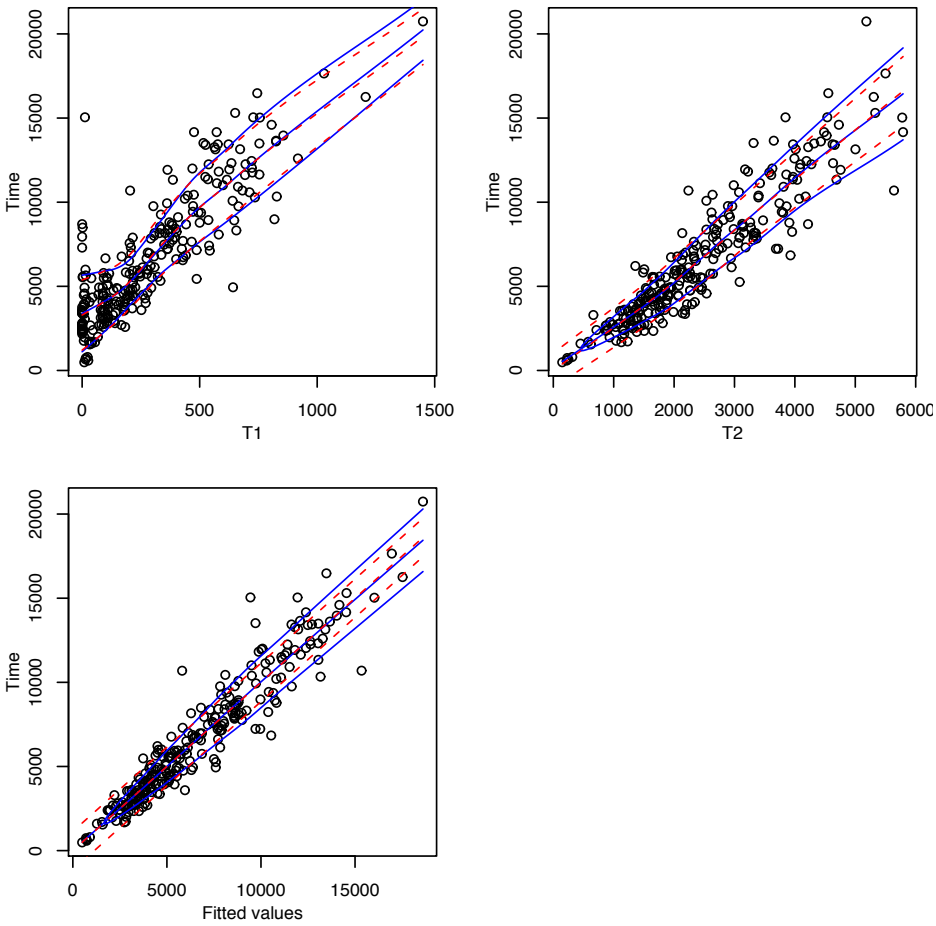
```
           Test stat Pr(>|t|)
logAPMAM     0.44994  0.65553
logAPSAB    -0.46471  0.64502
logAPSLAKE  -0.85245  0.39976
logOPBPC     1.38484  0.17487
logOPRC      0.83865  0.40735
logOPSLAKE   1.62951  0.11217
Tukey test   1.83863  0.06597
```



All the plots look like null plots, and none of the curvature tests are significant. ∎

8.11 Refer to the transactions data discussed in Section 4.6.1. Fit the mean function (4.16) with constant variance, and use marginal model plots to examine the fit. Be sure to consider both the mean function and the variance function. Comment on the results.

Solution: The marginal model plots are given below:



The model does a good job of reproducing the mean function in all three plots, but a very poor job with the standard deviation functions: the constant variance assumptions overstates variance with fitted values are small, and understates variance when fitted values are large. ∎

8.12 The number of crustacean zooplankton species present in a lake can be different, even for two nearby lakes. The data in the file `lakes.dat`, provided by S. Dodson and discussed in part in Dodson (1992), gives the number of known crustacean zooplankton species for 69 world lakes. Also included are a number of characteristics of each lake. There are some missing values, indicated with a "?" in the data file. The goal of the analysis to understand how the number of species present depends on the other measured variables that are characteristics of the lake. The variables are described in Table 8.5.

Table 8.5 Crustacean zooplankton species data, from Dodson (1992).

Variable	Description
Species	Number of zooplankton species
MaxDepth	Maximum lake depth, m
MeanDepth	Mean lake depth, m
Cond	Specific conductance, micro Siemans
Elev	Elevation, m
Lat	N latitude, degrees
Long	W longitude, degrees
Dist	distance to nearest lake, km
NLakes	number of lakes within 20 km
Photo	Rate of photosynthesis, mostly by the ^{14}C method
Area	Surface area of the lake, in hectares
Lake	Name of Lake

Decide on appropriate transformations of the data to be used in this problem. Then, fit appropriate linear regression models, and summarize your results.

Solution: The predictors should mostly be transformed, using logs of everything except *Photo*, *Dist*, *Long* and *Lat* (I added 2 to *Elev* because one lake had elevation -1). I transformed *Photo* to $Photo^{-.33}$ and *Dist* to $\text{Dist}^{.33}$. Transforming *Long* and *Lat* doesn't make much sense. The Box-Cox method does not suggest further transforming the response.

Only $Dist^{-.33}$, $\log(Area)$, $\log(NLakes)$, and $Photo^{.33}$ appear to be important. There is some non-constant variance; the score test has p-value of about 0.04. One might expect nonconstant variance because the response is a count. One approach at this point is to use Poisson regression, but that is not a topic of this book. Another alternative is to use a variance stabilizing transformation, probably the square root. The concern is that stabilizing variance may destroy linearity of the mean function. We fit in both the untransformed scale and in square root scale. Using marginal model plots, both seem to match the data equally well, but the square root scale also seems to have reasonably constant variance, since the p-value for the score test is about 0.67. The residual plots appear to be a little better in square root scale as well. The regression summary is

```
> summary(m2)
Call:
lm(formula = sqrt(Species) ~ I(Dist^(-1/3)) + logb(Area, 2) +
    logb(NLakes, 2) + I(Photo^(1/3)), data = d)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.1234     0.3224    3.48  0.00117
```

```
I(Dist^(-1/3))    0.5797     0.2349    2.47  0.01774
logb(Area, 2)     0.0565     0.0112    5.02  9.9e-06
logb(NLakes, 2)   0.0872     0.0448    1.95  0.05824
I(Photo^(1/3))    0.0995     0.0257    3.87  0.00037

Residual standard error: 0.505 on 42 degrees of freedom
Multiple R-Squared: 0.72,       Adjusted R-squared: 0.693
F-statistic:   27 on 4 and 42 DF,  p-value: 4.05e-11
```

Bigger values of *Area*, *NLakes* and *Photo* lead to more species, while isolated lakes have fewer species. The variable *Photo* is missing for about one third of the lakes, so one might want to examine models that ignore *Photo*. The analysis given is reasonable if a missing at random assumption is tenable here; we don't really have enough information to decide if it is tenable or not. ∎

9

Outliers and Influence

Problems

9.1 In an unweighted regression problem with $n = 54$, $p' = 5$, the results included $\hat{\sigma} = 4.0$, and the following statistics for four of the cases:

$\hat{e}_i$	h_{ii}
1.000	0.9000
1.732	0.7500
9.000	0.2500
10.295	0.185

For each of these four cases, compute r_i, D_i, and t_i. Test each of the four cases to be an outlier. Make a qualitative statement about the influence of each case on the analysis.

Solution:

```
> ehat <- c(1.000, 1.732, 9, 10.295)
> lev <- c(.9, .75, .25, .185)
> sig <- 4
> r <- ehat/(sig*sqrt(1-lev))
> D <- (1/5) * r^2 * (lev/(1-lev))
> ti <- r * sqrt((54-5-1)/(54-5-r^2))
> data.frame(ehat,lev,r,D,ti)
    ehat   lev       r       D      ti
1  1.000 0.900 0.79057 1.12500 0.78750
```

```
2  1.732 0.750 0.86600 0.44997 0.86375
3  9.000 0.250 2.59808 0.45000 2.76923
4 10.295 0.185 2.85094 0.36899 3.08954
```

Case 1 is likely to be most influential because of the large value of D. Cases 4 and 3 are most likely to be outliers because of the large values of t_i.∎

9.2 In the fuel consumption data, consider fitting the mean function

$$\mathrm{E}(Fuel|X) = \beta_0 + \beta_1 Tax + \beta_2 Dlic + \beta_3 Income + \beta_4 \log(Miles)$$

For this regression, we find $\hat{\sigma} = 64.891$ with 46 df, and the diagnostic statistics for four states and the District of Columbia were:

	Fuel	$\hat{e}_i$	h_{ii}
Alaska	514.279	−163.145	0.256
New York	374.164	−137.599	0.162
Hawaii	426.349	−102.409	0.206
Wyoming	842.792	183.499	0.084
Dist. of Col.	317.492	−49.452	0.415

Compute D_i and t_i for each of these cases, and test for one outlier. Which is most influential?

Solution:

```
                    y     ehat      r      t     h     D
Alaska          514.279 -163.145 -2.915 -3.193 0.256 0.585
New_York        374.164 -137.599 -2.317 -2.438 0.162 0.208
Hawaii          426.349 -102.409 -1.771 -1.814 0.206 0.162
Wyoming         842.792  183.499  2.954  3.246 0.084 0.160
Dist._of_Col.   317.492  -49.452 -0.996 -0.996 0.415 0.141
```

The largest outlier test is 3.246, and the Bonferroni p-values are, for all five states,

```
> pmin(51*2*pt(-abs(out$Ti[subset]),46),1)
      Alaska     New_York       Hawaii      Wyoming Dist._of_Col.
     0.12958      0.95272      1.00000      0.11145      1.00000
```

None would be declared outliers. Alaska has the largest influence on the regression. ∎

9.3 The matrix $(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})$ can be written as $(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}) = \mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}'_i$, where $\mathbf{x}'_i$ is the ith row of $\mathbf{X}$. Use this definition to prove that (A.37) holds.

Solution: (A.37) asserts that

$$(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}}$$

Multiply on the right by $(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})$ and on the right by $\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}'_i$, and simplify. The LHS equals $\mathbf{I}$, and the RHS is

$$\begin{aligned}\left((\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}}{1-h_{ii}}\right)(\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}'_i) = \\ &= \mathbf{I} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}'_i}{1-h_{ii}}(1 - 1 + h_{ii} - h_{ii}) \\ &= \mathbf{I}\end{aligned}$$

∎

9.4 The quantity $y_i - \mathbf{x}'_i\hat{\boldsymbol{\beta}}_{(i)}$ is the residual for the ith case when $\boldsymbol{\beta}$ is estimated without the ith case. Use (A.37) to show that

$$y_i - \mathbf{x}'_i\hat{\boldsymbol{\beta}}_{(i)} = \frac{\hat{e}_i}{1-h_{ii}}$$

This quantity is called the *predicted residual*, or the *PRESS residual*.

Solution: Using (A.38),

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{(i)} &= \hat{\boldsymbol{\beta}} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\hat{e}_i}{1-h_{ii}} \\ y_i - \mathbf{x}'_i\hat{\boldsymbol{\beta}}_{(i)} &= y_i - \mathbf{x}'_i\hat{\boldsymbol{\beta}} + \frac{\mathbf{x}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\hat{e}_i}{1-h_{ii}} \\ &= \hat{e}_i + \frac{h_{ii}}{1-h_{ii}}\hat{e}_i \\ &= \frac{\hat{e}_i}{1-h_{ii}}\end{aligned}$$

∎

9.5 Use (A.37) to verify (9.8).

Solution: Using (A.37), $\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\hat{e}_i/(1-h_{ii})$. Substitute into (9.6) to get

$$\begin{aligned}D_i &= \frac{1}{p'\hat{\sigma}^2}\frac{h_{ii}}{(1-h_{ii})^2}\hat{e}_i^2 \\ &= \frac{1}{p'}\left(\frac{h_{ii}}{1-h_{ii}}\right)r_i^2\end{aligned}$$

where $r_i^2 = \hat{e}_i^2/\hat{\sigma}^2(1-h_{ii})$. ∎

9.6 Suppose that interest centered on $\boldsymbol{\beta}^*$ rather than $\boldsymbol{\beta}$, where $\boldsymbol{\beta}^*$ is the parameter vector excluding the intercept. Using (5.21) as a basis, define a distance measure D_i^* like Cook's D_i and show that (Cook, 1979)

$$D_i^* = \frac{r_i^2}{p}\left(\frac{h_{ii} - 1/n}{1 - h_{ii} + 1/n}\right)$$

where p is the number of terms in the mean function excluding the intercept.

Solution: Equation (5.21) is the confidence region for $\boldsymbol{\beta}^*$,

$$\frac{(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^*)'(\mathcal{X}'\mathcal{X})(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}^*)}{p\hat{\sigma}^2} \leq F(\alpha; p, n-p)$$

We use the left-side of this equation to define D_i^*,

$$D_i^* = \frac{(\hat{\boldsymbol{\beta}}_{(i)}^* - \hat{\boldsymbol{\beta}}^*)'(\mathcal{X}'\mathcal{X})(\hat{\boldsymbol{\beta}}_{(i)}^* - \hat{\boldsymbol{\beta}}^*)}{p\hat{\sigma}^2}$$

We need an updating formula like (A.37) that excludes the intercept. Using (8.11), it follows that

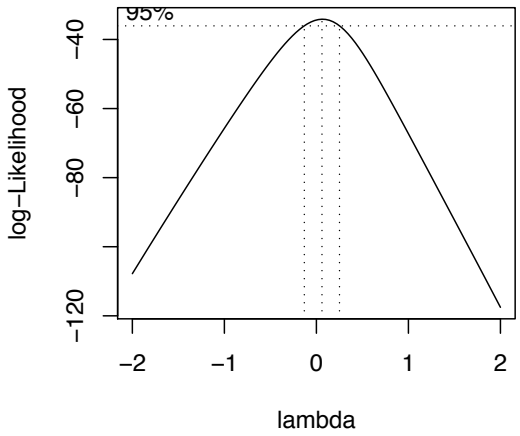
$$(\mathcal{X}_{(i)}'\mathcal{X}_{(i)})^{-1} = (\mathcal{X}'\mathcal{X})^{-1} + \frac{(\mathcal{X}'\mathcal{X})^{-1}\mathbf{x}_i^*(\mathbf{x}_i^*)'(\mathcal{X}'\mathcal{X})^{-1}}{1 - h_{ii} - 1/n}$$

where $(\mathbf{x}_i^*)'$ is the i-th row of $\mathcal{X}$. Using this, the result follows as in Problem 9.5. ∎

9.7 Refer to the lathe data in Problem 6.2.

9.7.1. Starting with the full second-order model, use the Box-Cox method to show that an appropriate scale for the response is the logarithmic scale.

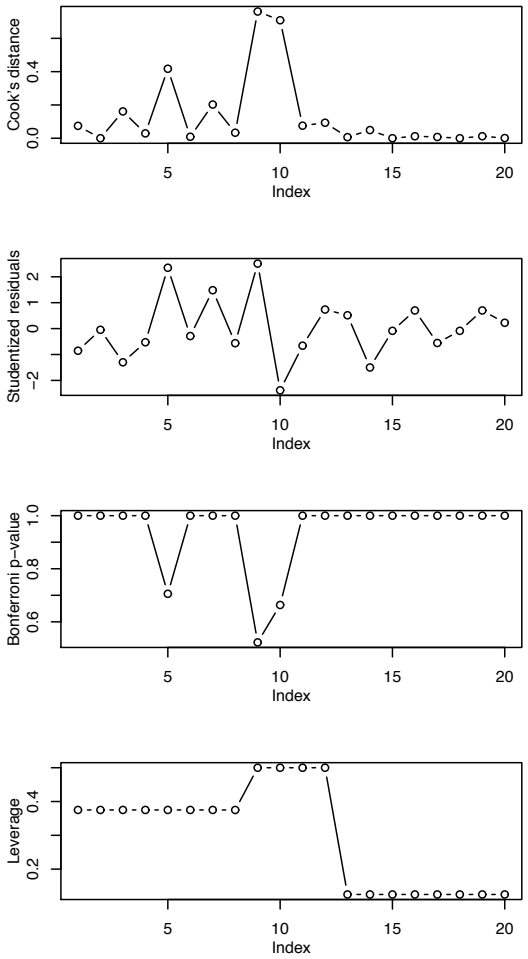
Solution:



This is the graph of the profile log-likelihood for the transformation parameter using the Box-Cox method for the second-order lathe model. The confidence interval for λ is very narrow and includes zero, suggesting a log transformation. ∎

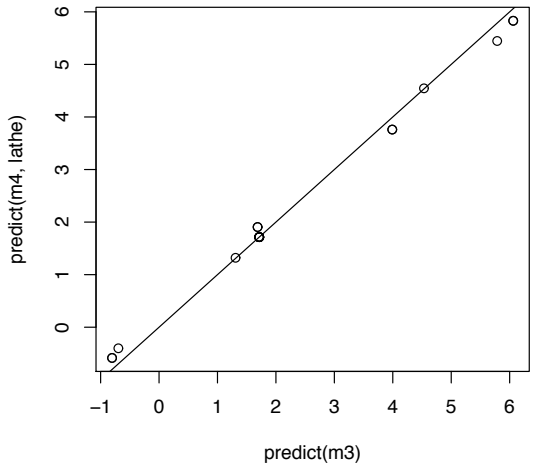
9.7.2. Find the two cases that are most influential in the fit of the quadratic mean function, and explain why they are influential. Delete these points from the data, refit the quadratic mean function, and compare to the fit with all the data.

Solution:



Cases 9–12, the unreplicated "star points," have very high leverage. Two of these, numbers 9 and 10, also had large residuals, one positive and one negative, and these two cases have the largest Cook's distances. One way to assess their impact is to delete them, and refit to the smaller data set. We can then compare the fitted values:

```
> m4 <- update(m3, subset=-c(9,10))
> plot(predict(m3),predict(m4,lathe))
> abline(0,1)
```



The change in fitted values, including the fitted values for the two deleted cases, is generally not very large, and so the effect of deletion is minor by this measure. One could also look at changes on coefficients, in tests, and so on.

■

9.8 Florida election 2000

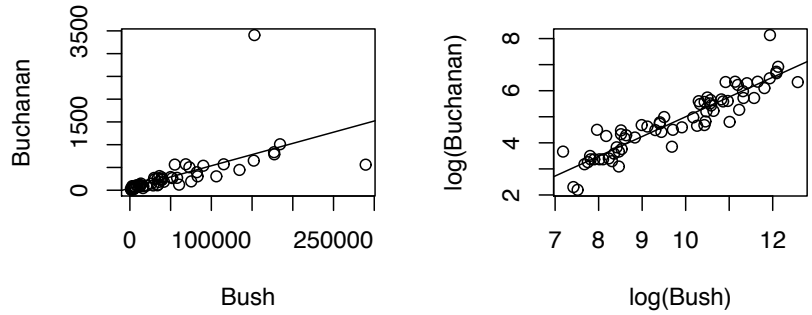
In the 2000 election for US president, the counting of votes in Florida was controversial. In Palm Beach county in south Florida, for example, voters used a so-called butterfly ballot. Some believe that the layout of the ballot caused some voters to cast votes for Buchanan when their intended choice was Gore.

The data in the file `florida.txt`[1] has four variables, *County*, the county name, and *Gore*, *Bush* and *Buchanan*, the number of votes for each of these three candidates. Draw the scatterplot of *Buchanan* versus *Bush*, and test the hypothesis that Palm Beach county is an outlier relative to the simple linear regression mean function for E(*Buchanan*|*Bush*). Identify another county with an unusual value of the Buchanan vote given its Bush vote, and test that county to be an outlier. State your conclusions from the test, and its relevance, if any, to the issue of the butterfly ballot.

Next, repeat the analysis, but first consider transforming the variables in the plot to better satisfy the assumptions of the simple linear regression model. Again test to see if Palm Beach County is an outlier, and summarize.

Solution: The scatterplots of the original data, and the data in log scale are:

[1]Source: http://abcnews.go.com/sections/politics/2000vote/general/FL_county.html.



The clearly separated point in the figure at the left is for Palm Beach County; the separated point at the right of this figure is for Dade County, which apparently had a very low vote for Buchanan. In the right figure, these differences are less clear. The outlier tests for these two counties (with the Bonferroni correction), are:

	Untransformed		Log Scale	
	Outlier test	p-val	Outlier test	p-val
Palm Beach	24.08	0.00	4.07	0.00
Dade	−3.28	0.06	−1.39	1.00

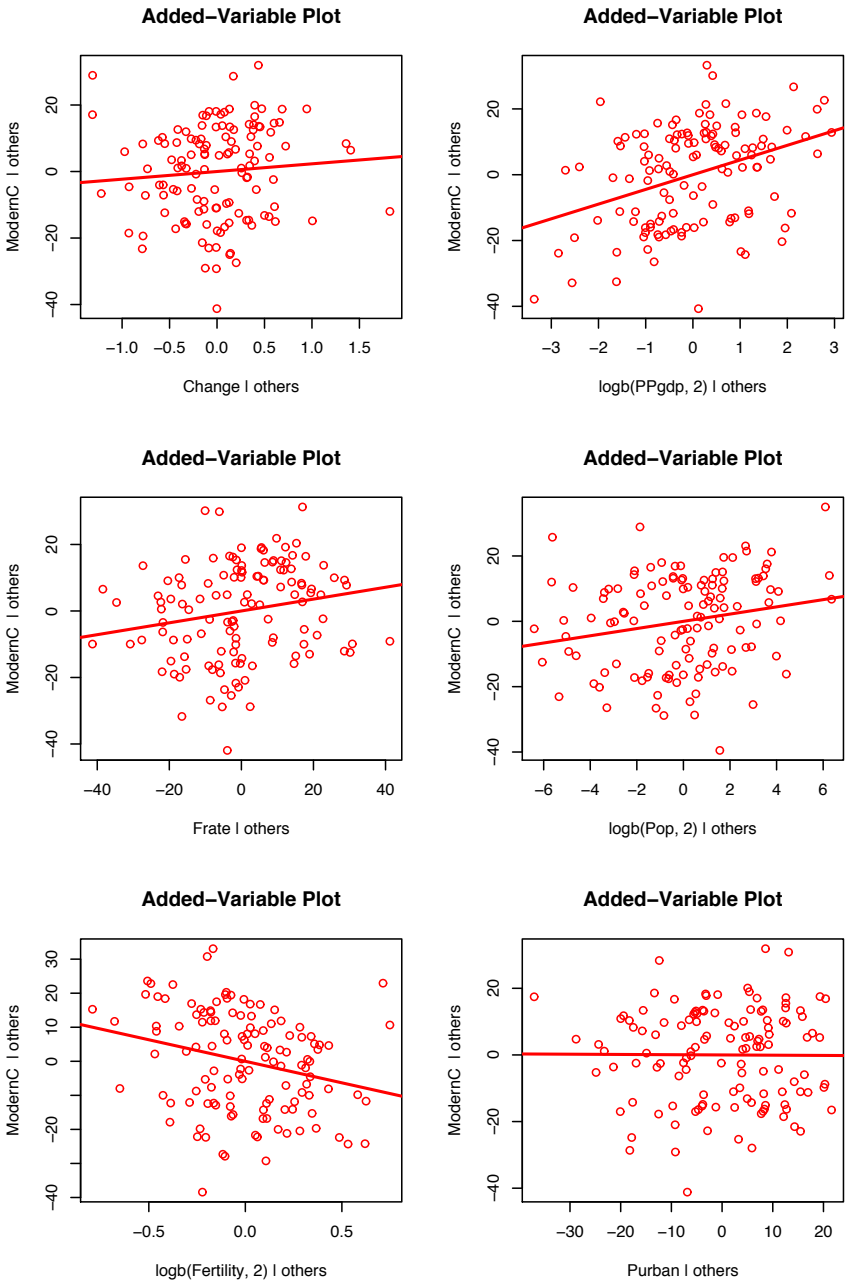
Palm Beach apparently had too many Buchanan votes, using the questionable untransformed values, or the more appropriate log scale. The value of Dade county does not appear to be an outlier in the log scale. ∎

9.9 Refer to the United Nations data described in Problem 7.8, and consider the regression with response *ModernC*, and predictors (log(*PPgdp*), *Change*, *Pop*, *Fertility*, *Frate*, *Purban*).

9.9.1. Examine added-variable plots for each of the terms in the regression model and summarize. Is it likely that any of the localities are influential for any of the terms? Which localities? Which terms?

Solution:

```
> avPlots(m1,id.n=2)
```



Separated cases at the right or left of an added-variable plot would indicate influence; no such cases appear in these plots. This is confirmed by an index plot of Cook's distance; none of the localities is overly influential. ∎

9.9.2. Are there any outliers in the data?
Solution:

```
> outlierTest(m1)

No Studentized residuals with Bonferonni p < 0.05
Largest |rstudent|:
        rstudent unadjusted p-value Bonferonni p
Poland -2.970918          0.0036041      0.45052
```

Poland had the largest positive Studentized residual, with corresponding p-value, after Bonferroni correction, of about 0.45. Although modern contracption use is high in Poland, it is apparently not an outlier. ∎

9.9.3. Complete analysis of the regression of *ModernC* on the terms in the mean function.

Solution: Both *Change* and *Purban* can be dropped from the mean function as unimportant. For the remaining variables

```
> summary(m3 <- update(m2, ~.-Change))
Call:
lm(formula = ModernC ~ logb(PPgdp, 2) + Frate + logb(Pop, 2) +
    logb(Fertility, 2), data = UN3)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         -19.469     15.146   -1.29  0.20114
logb(PPgdp, 2)        4.596      0.773    5.95  2.7e-08
Frate                 0.165      0.080    2.06  0.04155
logb(Pop, 2)          1.229      0.463    2.65  0.00902
logb(Fertility, 2)   -9.442      2.354   -4.01  0.00011

Residual standard error: 14.5 on 120 degrees of freedom
Multiple R-Squared: 0.558,      Adjusted R-squared: 0.544
F-statistic: 37.9 on 4 and 120 DF,  p-value: <2e-16
```

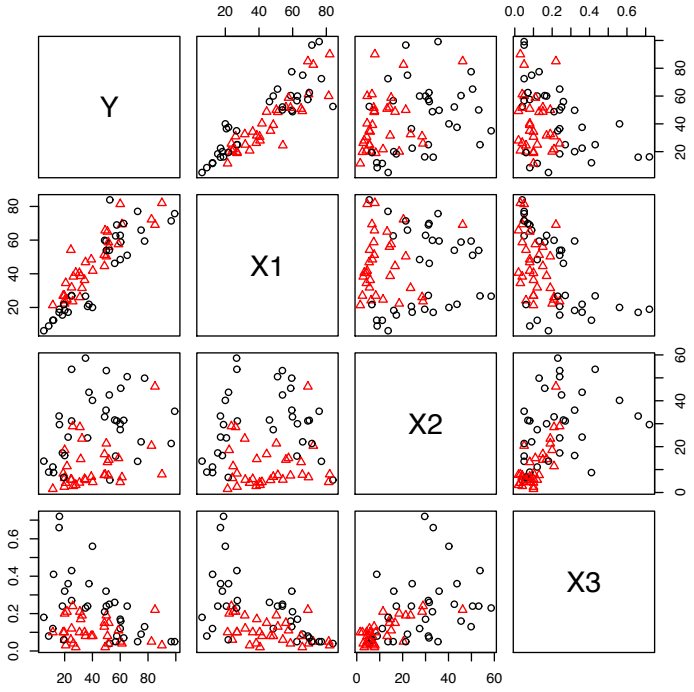
The residual error is about 14.5, so the fitted rate of modern contraceptive use is not estimated very precisely. Use increases with per person GDP, but a doubling of GDP is expected to increase contraception use by only 5%. The positive coefficient for *Frate* suggests a positive relation between female economic activity *ModernC*. More populous localities have higher use of *ModernC*, while, as expected, higher *Fertility* is associated with lower *ModernC*. Of course the data are observational, so we cannot infer causation here. ∎

9.10 The data in the data file `landrent.txt` were collected by Douglas Tiffany to study the variation in rent paid in 1977 for agricultural land planted to alfalfa. The variables are Y = average rent per acre planted to alfalfa X_1 = average rent paid for all tillable land X_2 = density of dairy cows (number per square mile) X_3 = proportion of farmland used as pasture $X_4 = 1$ if liming is required to grow alfalfa; 0, otherwise.

The unit of analysis is a county in Minnesota; the 67 counties with appreciable rented farmland are included. Alfalfa is a high protein crop that is suitable feed for dairy cows. It is thought that rent for land planted to alfalfa relative to rent for other agricultural purposes would be higher in areas with a high density of dairy cows and rents would be lower in counties where liming is required, since that would mean additional expense. Use all the techniques learned so far to explore these data with regard to understanding rent structure. Summarize your results.

Solution: As usual, we begin with a scatterplot matrix. We use X_4, which is a dummy variable, as a marking variable. This is done in R using the command

```
pairs(Y~X1+X2+X3,data=landrent,col=landrent$X4+1,pch=landrent$X4+1)
```



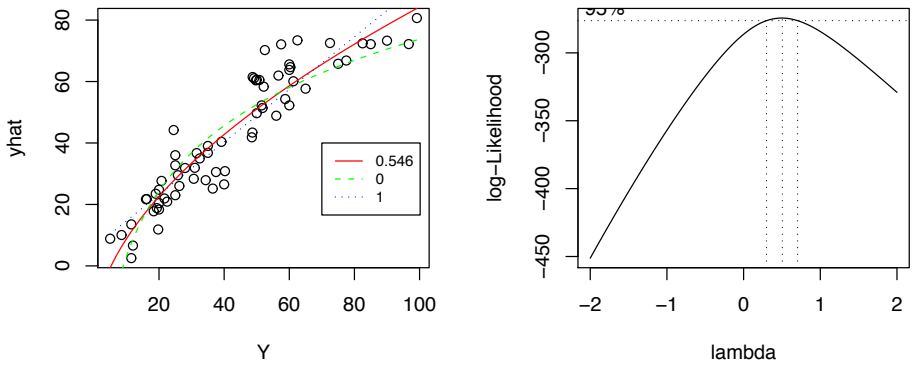
The mean functions in each of the plots of predictors versus other predictors, either conditioning on point color or ignoring it, seems to be somewhat curved, so transformations of the predictors seem likely to be useful. The results of the multivariate Box-Cox

method are:

```
> summary(b1 <- powerTransform(cbind(X1, X2, X3) ~ 1, data=landrent))
bcPower Transformations to Multinormality
```

```
   Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
X1    0.7903   0.2030           0.3924           1.1882
X2    0.2371   0.1218          -0.0016           0.4759
X3    0.0825   0.0991          -0.1118           0.2768

Likelihood ratio tests about transformation parameters
                                LRT df         pval
LR test, lambda = (0 0 0)  23.155504  3 3.747847e-05
LR test, lambda = (1 1 1) 102.374387  3 0.000000e+00
LR test, lambda = (1 0 0)   5.253666  3 1.541374e-01
```

which suggests replacing X_1 and X_2 by their logarithms. Ignoring for the moment the indicator X_4, we now turn to transforming Y. Given below are both the inverse response plot for the mean function `Y X1+logb(X2,2)+logb(X3,2)`, and the Box-Cox likelihood plot.



Both figures suggest using a transformation of Y close to the square root. The inverse response plot suggests that the improvement of the square root over untransformed is relatively small, and the decision not to transform may be reasonable. In this solution, however, we use the square root transformations for the response.

Next, we turn to adding the indicator variable. To decide how to do this, we can use POD models[2],

```
> anova(m2 <- pod(sqrt(Y)~X1+logb(X2,2)+logb(X3,2),data=rent,group=X4))
POD Analysis of Variance Table for sqrt(Y), grouped by X4

1: sqrt(Y) ~ X1 + logb(X2, 2) + logb(X3, 2)
2: sqrt(Y) ~ X1 + logb(X2, 2) + logb(X3, 2) + X4
```

[2]POD models are not required here, but if available they provide a convenient way of studying a smaller sequence of models than would be required without them.

```
3: sqrt(Y) ~ eta0 + eta1 * X1 + eta2 * logb(X2, 2) + eta3 * logb(X3,
3:     2) + X41 * (th02 + th12 * (eta1 * X1 + eta2 * logb(X2, 2) +
3:     eta3 * logb(X3, 2)))
4: sqrt(Y) ~ X1 + logb(X2, 2) + logb(X3, 2) + X4 + X1:X4 + logb(X2,
4:     2):X4 + logb(X3, 2):X4
             Res.Df   RSS Df Sum of Sq    F Pr(>F)
1: common        63 25.21
2: parallel      62 24.33  1      0.88 2.30   0.13
3: pod           61 24.29  1      0.04 0.10   0.75
4: pod + 2fi     59 22.62  2      1.67 2.17   0.12
```

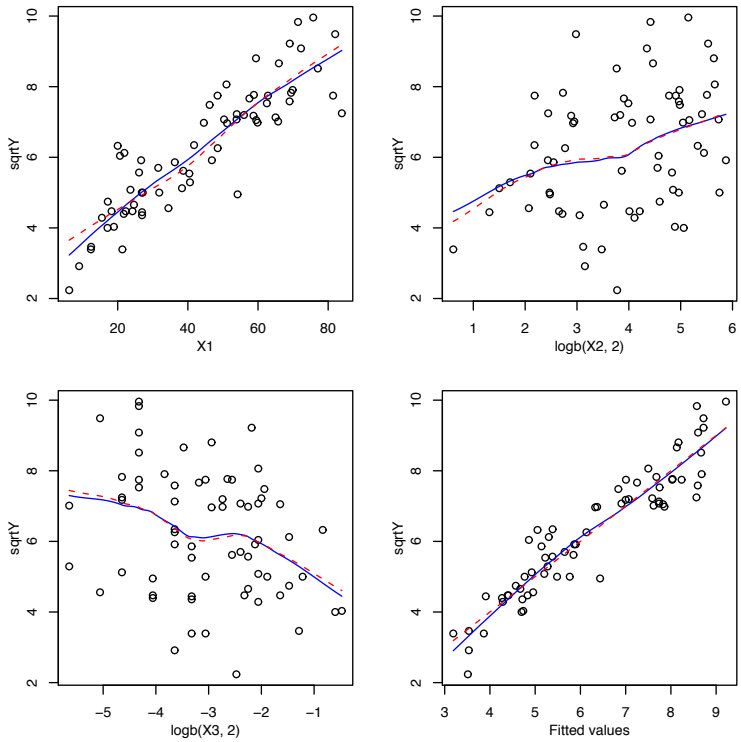
All the p-values are large, suggesting no effect due to X_4, so it can be ignored in the fitting because the "common" model is as good as any of the others.

The summary of the fitted regression is

```
> summary(m3 <- lm(sqrt(Y) ~ X1 + logb(X2,2) + logb(X3,2), rent))

Call:
lm(formula = sqrt(Y) ~ X1 + logb(X2, 2) + logb(X3, 2), data = rent)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.08831    0.55899    1.95    0.056
X1           0.07030    0.00539   13.05   <2e-16
logb(X2, 2)  0.46485    0.09558    4.86    8e-06
logb(X3, 2) -0.09527    0.12036   -0.79    0.432

Residual standard error: 0.633 on 63 degrees of freedom
Multiple R-Squared: 0.879
F-statistic:  153 on 3 and 63 DF,  p-value: <2e-16
```

The variable $\log(X_3)$ could also be dropped from the mean function.

We turn to model checking, which would suggest looking for influential observations, outliers, and lack of fit of the mean function. We show only the marginal model plots, which show no problems. The fitted model matches the data very closely.



In summary, rent paid increases with X_1 = average rent paid in the county and X_2 = density of dairy cows. Neither liming nor amount of pasture in the county are of any importance. ∎

9.11 The data in the file `cloud.txt` summarize the results of the first Florida Area Cumulus Experiment, or FACE-1, designed to study the effectiveness of cloud seeding to increase rainfall in a target area (Woodley, Simpson, Biondini, and Berkeley, 1977). A fixed target area of approximately 3000 square miles was established to the north and east of Coral Gables, Florida. During the summer of 1975, each day was judged on its suitability for seeding. The decision to use a particular day in the experiment was based primarily on a suitability criterion S depending on a mathematical model for rainfall. Days with $S > 1.5$ were chosen as experimental days; there were 24 days chosen in 1975. On each day, the decision to seed was made by flipping a coin; as it turned out, 12 days were seeded, 12 unseeded. On seeded days, silver iodide was injected into the clouds from small aircraft. The predictors and the response are defined in Table 9.3.

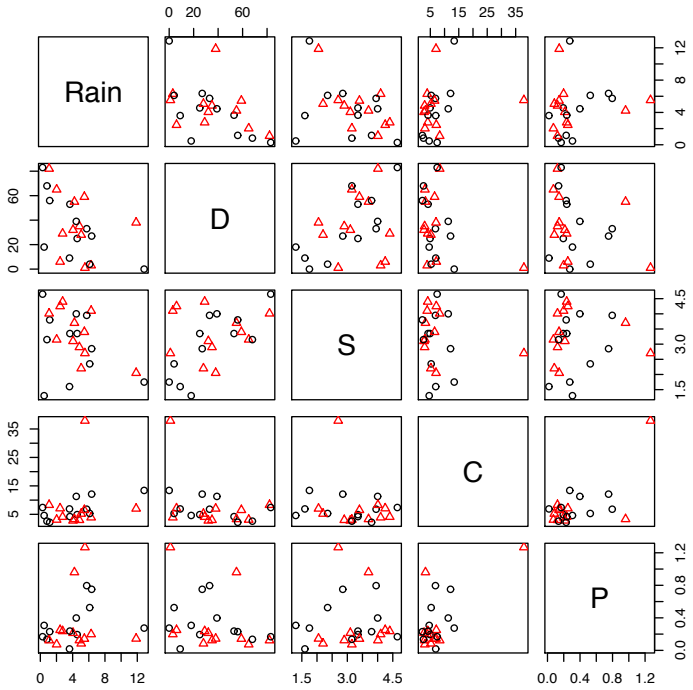
The goal of the analysis is to decide if there is evidence that cloud seeding is effective in increasing rainfall. Begin your analysis by drawing appropriate graphs. Obtain appropriate transformations of predictors. Fit appropriate

Table 9.3 The Florida Area Cumulus experiment on cloud seeding.

Variable	Description
A	Action, 1 = seed, 0 = do not seed
D	Days after the first day of the experiment (June 16, 1975=0)
S	Suitability for seeding
C	Percent cloud cover in the experimental area, measured using radar in Coral Gables, Florida
P	Prewetness, amount of rainfall in the hour preceding seeding in 10^7 cubic meters
E	Echo motion category, either l or 2, a measure of the type of cloud
Rain	Rainfall following the action of seeding or not seeding in 10^7 cubic meters

mean functions, and summarize your results. (Hint: Be sure to check for influential observations and outliers.)

Solution: (An alternative solution is given by Cook and Weisberg, (1982), *Residuals and Influence in Regression*, London: Chapman and Hall, available for download from `www.stat.umn.edu/rir`). The only variables that could be transformed are S, C, P and the response variable *Rain*. Transformation of D, the day number, is unlikely to be helpful. Since P and *Rain* are both measures of rainfall, we will require that if they are transformed, we use the same transformation for each.



In the scatterplot matrix, seeded days are shown with red triangles, unseeded with black circles. We see: (1) rainfall generally declines over the summer, from the plot of *Rain* versus D; (2) the suitability for seeding S generally increases over the summer, so better days for seeding appear to occur when expected rainfall is lower; (3) one seeded day early in the summer had an extremely large value of C. The experimenters recognized this as a "disturbed" day. This case is likely to be influential in fitting and in selecting transformations.

We turn to selecting transformations, keeping in mind that *Rain* and P should be similarly transformed. We don't have any special software for this, although such software could be written. Unlike earlier work, we will transform both the predictors and the response simultaneously, so we are transforming for multivariate normality, to put *Rain* and P on an equal footing.

```
> summary(b1 <- powerTransform(cbind(Rain, S, C, P) ~ 1, data=cloud))
bcPower Transformations to Multinormality

     Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Rain    0.5529   0.1968          0.1672           0.9387
S       1.5518   0.6547          0.2686           2.8350
C      -0.4410   0.2910         -1.0113           0.1293
P       0.1144   0.1707         -0.2201           0.4489
```

```
Likelihood ratio tests about transformation parameters
                                    LRT df         pval
LR test, lambda = (0 0 0 0)   17.258538  4 1.721668e-03
LR test, lambda = (1 1 1 1)   55.801387  4 2.206946e-11
LR test, lambda = (0.5 1 0 0)  4.078017  4 3.955503e-01
> summary(b1 <- powerTransform(cbind(Rain, S, C, P) ~ 1, data=cloud, subset=-2))
bcPower Transformations to Multinormality

     Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Rain    0.5720   0.2020           0.1760           0.9680
S       1.7428   0.6785           0.4130           3.0727
C      -0.0957   0.4286          -0.9357           0.7442
P       0.1363   0.1800          -0.2165           0.4891

Likelihood ratio tests about transformation parameters
                                    LRT df         pval
LR test, lambda = (0 0 0 0)   15.684617  4 3.472917e-03
LR test, lambda = (1 1 1 1)   28.591327  4 9.463253e-06
LR test, lambda = (0.5 1 0 0)  2.072199  4 7.224815e-01
```

We have used the multivariate Box-Cox method twice, the second time excluding the disturbed day, which was case number two. This case is influential for the choice of transformation for C only, changing the point estimate from about $-.4$ to about $-.1$.

We will replace C by its logarithm, and leave S untransformed. We could probably use any power between $\lambda = 0$ for logarithms to $\lambda = 0.5$ for *Rain* and C. We will use cube-root powers for these variables, because these variables are *volumes*, or cubic variables, so the cube-root transformation makes some dimensional sense.

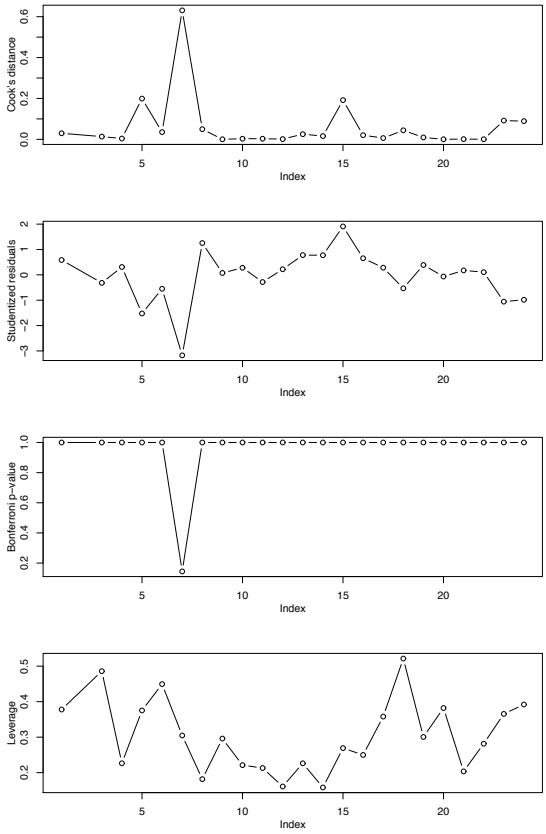
We will begin with response $Rain^{1/3}$ and predictors $(S, D, \log(C), P^{1/3})$. Since the primary interest is in the action variable A, a reasonable modelling approach is to examine pod models, although this is not the only approach possible.

```
> anova(p1 <- pod(m1,group=A))
POD Analysis of Variance Table for Rain^(1/3), grouped by A

1: Rain^(1/3) ~ S + D + logb(C, 2) + I(P^(1/3)) + E
2: Rain^(1/3) ~ S + D + logb(C, 2) + I(P^(1/3)) + E + A
3: Rain^(1/3) ~ eta0 + eta1 * S + eta2 * D + eta3 * logb(C, 2) +
3:     eta4 * I(P^(1/3)) + eta5 * E + A1 * (th02 + th12 * (eta1 *
3:     S + eta2 * D + eta3 * logb(C, 2) + eta4 * I(P^(1/3)) + eta5 *
3:     E))
4: Rain^(1/3) ~ S + D + logb(C, 2) + I(P^(1/3)) + E + A + S:A +
4:     D:A + logb(C, 2):A + I(P^(1/3)):A + E:A
             Res.Df   RSS Df Sum of Sq     F Pr(>F)
1: common        17 2.269
2: parallel      16 1.712  1     0.557 11.65 0.0058
```

```
3: pod          15 1.453  1     0.259  5.40 0.0402
4: pod + 2fi    11 0.526  4     0.927  4.84 0.0169
```

suggesting a clear seeding effect, since the p-value for comparing the common to the parallel models is so large, but also the possibility of the need for a more complex pod or general models. At this point, we look for influential observations with a goal of possibly simplifying the result.

```
m2 <- update(m1, ~.+A, subset=-2)
```

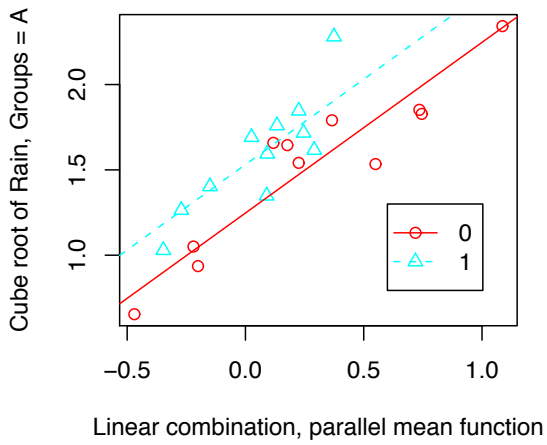


The residual analysis is done for the parallel regression model because easy to use software for residual analysis for the pod model is not (yet) available. We see that case number 7 is quite influential, with D_7 larger than 0.8. This was an unseeded day with typical values for the predictors ($h_{7,7}$ is not large), but with very low observed rainfall (the residual is large and negative). If we delete this one day, we get:

```
> anova(p2 <- update(p1,subset=-c(2,7)))
POD Analysis of Variance Table for Rain^(1/3), grouped by A

1: Rain^(1/3) ~ S + D + logb(C, 2) + I(P^(1/3)) + E
```

```
2: Rain^(1/3) ~ S + D + logb(C, 2) + I(P^(1/3)) + E + A
3: Rain^(1/3) ~ eta0 + eta1 * S + eta2 * D + eta3 * logb(C, 2) +
3:     eta4 * I(P^(1/3)) + eta5 * E + A1 * (th02 + th12 * (eta1 *
3:     S + eta2 * D + eta3 * logb(C, 2) + eta4 * I(P^(1/3)) + eta5 *
3:     E))
4: Rain^(1/3) ~ S + D + logb(C, 2) + I(P^(1/3)) + E + A + S:A +
4:     D:A + logb(C, 2):A + I(P^(1/3)):A + E:A
             Res.Df   RSS Df Sum of Sq     F Pr(>F)
1: common        16 1.006
2: parallel      15 0.634  1     0.373 13.36 0.0044
3: pod           14 0.590  1     0.044  1.56 0.2394
4: pod + 2fi     10 0.279  4     0.311  2.79 0.0858
```

suggesting that the parallel model is now adequate. Here is the pod summary graph:



which presents a convincing picture that, after deleting two observations, nearly all the points for seeded days had higher rainfall than similar days, according to the linear combination used, for unseeded days. The regression summary is:

```
> summary(p3)

Call:
lm(formula = Rain^(1/3) ~ S + D + logb(C, 2) + I(P^(1/3)) + E +
    A, data = cloud, subset = -c(2, 7))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.24719    0.35548    3.51   0.0032
S           -0.25545    0.06271   -4.07   0.0010
```

Table 9.4 The drug cost data.

Variable	Description
COST	Ave. cost to plan for one prescription for one day, dollars.
RXPM	Average number of prescriptions per member per year
GS	Percent generic substitution used by the plan
RI	Restrictiveness index (0=none, 100=total)
COPAY	Average member co-payment for prescriptions
AGE	Average member age
F	percent female members
MM	Member months, a measure of the size of the plan
ID	An identifier for the name of the plan

```
D            -0.00486    0.00216   -2.25   0.0399
logb(C, 2)    0.09578    0.06635    1.44   0.1694
I(P^(1/3))    1.13431    0.31270    3.63   0.0025
E             0.21902    0.11883    1.84   0.0852
A             0.28374    0.09552    2.97   0.0095

Residual standard error: 0.206 on 15 degrees of freedom
Multiple R-Squared: 0.816
F-statistic: 11.1 on 6 and 15 DF,  p-value: 8.84e-05
```

Further checking, like looking at residuals again, or looking at marginal model plots, can be useful here. ∎

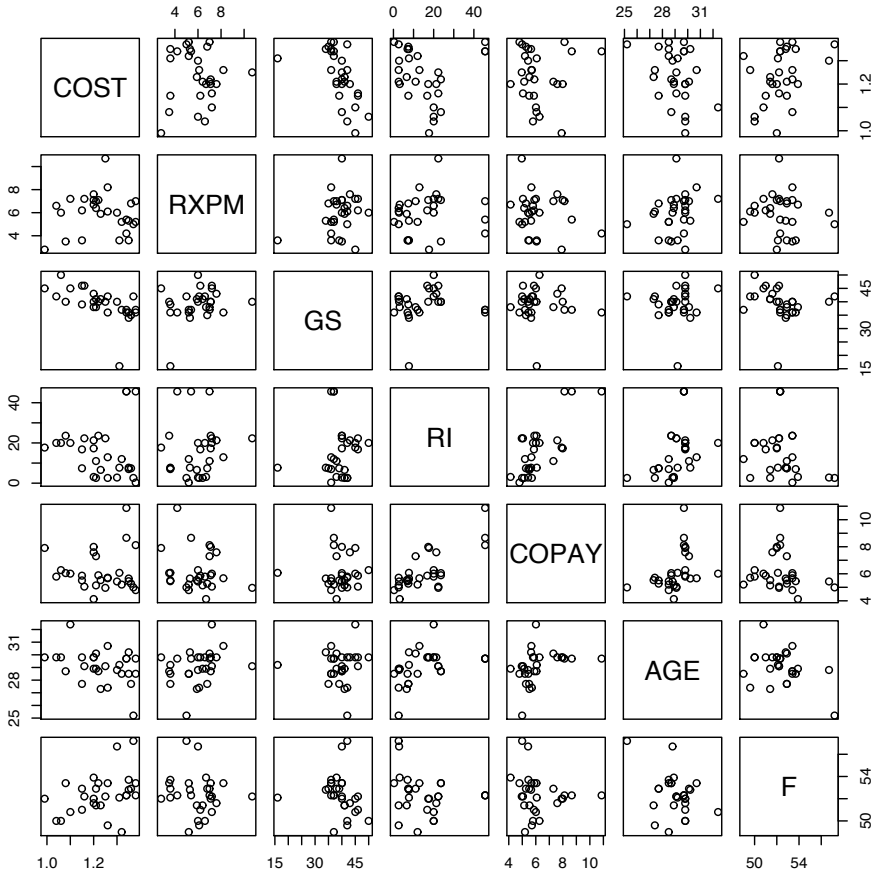
9.12 Health plans use many tools to try to control the cost of prescription medicines. For older drugs, *generic substitutes* that are equivalent to name-brand drugs are sometimes available at a lower cost. Another tool that may lower costs is restricting the drugs that physicians may prescribe. For example, if three similar drugs are available for treating the same symptoms, a health plan may require physicians to prescribe only one of them. Since the usage of the chosen drug will be higher, the health plan may be able to negotiate a lower price for that drug.

The data in the file `drugcost.txt`, provided by Mark Siracuse, can be used to explore the effectiveness of these two strategies in controlling drug costs. The response variable is *COST*, the average cost of drugs per prescription per day, and predictors include *GS*, the extent to which the plan uses generic substitution, a number between zero, no substitution, and 100, always use a generic substitute if available, and *RI*, a measure of the restrictiveness of the plan, from zero, no restrictions on the physician, to 100, the maximum possible restrictiveness. Other variables that might impact cost were also collected, and are described in Table 9.4. The data are from the mid 1990s, and are for 29 plans throughout the United States with pharmacies administered by a national insurance company.

Provide a complete analysis if these data, paying particular regard to possible outliers and influential cases. Summarize your results with regard to the

importance of *GS* and *RI*. In particular, can we infer that more use of *GS* and *RI* will reduce drug costs?

Solution: Nearly all the variables have a very restricted range, so transformations are likely to be of little value. Second, the unit of analysis is not clear. The unit might be the *medical plan* or it might be the *patient in a medical plan*. The latter case would suggest that weighting is required in fitting models, while the former does not require weighting. You can use either approach, but the issue should be discussed. Using the plan as the unit of analysis is appropriate for a policy maker interested in understanding how plans cope with prescription costs. Using the member as the unit of analysis might be appropriate for a consumer or someone studying how the health *community* pays for drugs delivered to individuals.



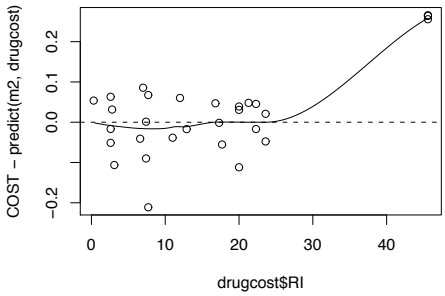
Three of the plans (MN1, MN2 and MN3) have very high values of RI, and also very high costs. One plan, DE, is much lower on GS than all the other

plans. At the first stage, I removed these four plans. In this scaling, there is no need for transformations (I have NOT used MM as a predictor, although it appears to be irrelevant anyway).

The unweighted analysis is particularly straightforward. The scatterplot matrix of the remaining data, not shown here, suggests all variables are either approximately linearly related or unrelated, so a linear regression model without further transformation will work well. The fitted model, after removing predictors that seem to be irrelevant is:

```
Call:
lm(formula = COST ~ RXPM + GS + RI + AGE, data = drugcost,
     subset = -c(mn.plans,de))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.39398    0.34365    6.97  9.2e-07
RXPM         0.01983    0.00804    2.47   0.0229
GS          -0.01382    0.00373   -3.70   0.0014
RI          -0.00450    0.00213   -2.12   0.0471
AGE         -0.02357    0.01096   -2.15   0.0439

Residual standard error: 0.0619 on 20 degrees of freedom
Multiple R-Squared: 0.727
F-statistic: 13.3 on 4 and 20 DF,  p-value: 1.88e-05
> confint(m2)
                   2.5 %        97.5 %
(Intercept)  1.677131944  3.1108341533
RXPM         0.003049227  0.0366051519
GS          -0.021612770 -0.0060328316
RI          -0.008932843 -0.0000630616
AGE         -0.046443207 -0.0007052995
```

Based on the confidence intervals, increasing GS by 10% will lower prescription per day cost by around \$0.06 to \$0.22, and increasing the restricted formulary by 10% will decrease costs up to \$0.09. Here we have summarized the effect of a change of a more clinically meaningful 10% rather than a change of 1%. We return to the effects of the Minnesota clinics.



The Minnesota clinics are the three points (two are over-printed) in the upper right corner of the plot. The fitted model does not seem to work for these clinics. If they are included, the regression summary is:

```
> summary(m3 <- update(m2, subset=NULL))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.71737    0.39615    6.86  4.3e-07
RXPM         0.01854    0.00993    1.87  0.07431
GS          -0.01190    0.00275   -4.32  0.00023
RI           0.00167    0.00135    1.23  0.22950
AGE         -0.03967    0.01373   -2.89  0.00807

Residual standard error: 0.0823 on 24 degrees of freedom
Multiple R-Squared: 0.521,
F-statistic: 6.53 on 4 and 24 DF,  p-value: 0.00105
```

The fit with these cases included is much worse ($\hat{\sigma}$ increases from about 0.06 to about 0.08), and RI no longer has coefficient estimate that is clearly different from zero. We are led to conclude that the evidence in favor of using RI to decrease drug costs is very weak, but increasing GS appears to be very useful. ∎

10

Variable Selection

Problems

10.1 Generate data as described for the two simulated data sets in Section 10.1, and compare the results you get to the results given in the text.

Solution: Here are the R commands that will reproduce the results given in the text.

```
> set.seed(1013185)
> case1 <- data.frame(x1=rnorm(100),x2=rnorm(100),
+                     x3=rnorm(100),x4=rnorm(100))
> e <- rnorm(100)
> case1$y <- 1 + case1$x1 + case1$x2 + e
> m1 <- lm(y~x1+x2+x3+x4,data=case1)
>
> X <- as.matrix(case1[,-5]) # change from data.frame to a matrix, drop y
> Var2 <- matrix(c(1,   0, .95,   0,
+                  0,   1,   0,-.95,
+                .95,   0,   1,   0,
+                  0,-.95,   0,   1), ncol=4)
> s1 <- chol(Var2)  # cholesky factor of Var2
> X <- X %*% s1
> dimnames(X)[[2]] <- paste("x",1:4,sep="")
> case2 <- data.frame(X)
> case2$y <- 1 + case2$x1 + case2$x2 + e
> m2 <- lm(y~x1+x2+x3+x4,data=case2)
```

The `set.seed` command initializes the random number generator to be sure the same numbers are used as in the book. For case 2, we have reused the same random numbers to make the results for the two cases correlated. *Var2* is the matrix (10.2). In the next line we found the Cholesky decomposition of *Var2*, so $Var2 = s1's1$, and so $Xs1$ is like a sample from N(0, *Var2*). The next line makes sure the columns of X have the right names. Then y is recomputed, again using the same errors as for case 1, and the model is fit. ∎

Table 10.12 Mantel's data for Problem 10.2.

	Y	X1	X2	X3
1	5.00	1.00	1004.00	6.00
2	6.00	200.00	806.00	7.30
3	8.00	−50.00	1058.00	11.00
4	9.00	909.00	100.00	13.00
5	11.00	506.00	505.00	13.10

10.2 Using the Şdata Ť in Table 10.12 with a response Y and three predictors X_1, X_2 and X_3 from Mantel (1970) in the file `mantel.txt`, apply the BE and FS algorithms, using C_p as a criterion function. Also, find AIC and C_p for all possible models, and compare results. What is $X_{\mathcal{A}}$?

Solution: Using the `step` method in R/S-Plus, here is the result for forward selection:

```
> m0 <- lm(Y ~ 1, data=mantel)
> step(m0,scope=~X1+X2+X3,direction="forward")
Start:  AIC= 9.59
 Y ~ 1

       Df Sum of Sq   RSS   AIC
+ X3    1     20.69  2.11 -0.31
+ X1    1      8.61 14.19  9.22
+ X2    1      8.51 14.29  9.25
<none>              22.80  9.59

Step:  AIC= -0.31
 Y ~ X3

       Df Sum of Sq    RSS    AIC
<none>               2.112 -0.309
+ X2    1     0.066  2.046  1.532
+ X1    1     0.065  2.048  1.536

Call:
lm(formula = Y ~ X3, data = mantel)
```

```
Coefficients:
(Intercept)           X3
      0.798        0.695
```

This method uses AIC to select models, but since all the terms have a single df, the ordering of models with AIC and C_p is identical. Starting with the mean function with no predictors, at the first step we consider adding the one term that makes AIC as small as possible, which is X_3. At the second step, we consider adding another term after X_3 if it further reduces AIC; in this problem adding either X_1 or X_2 actually increases AIC, so we would select $X_{\mathcal{A}} = \{X_3\}$.

Using backward elimination,

```
> m1 <- lm(Y~X1+X2+X3, data=mantel)
> step(m1,scope=~1, direction="backward")
Start:  AIC= -314.77
 Y ~ X1 + X2 + X3

       Df Sum of Sq     RSS    AIC
<none>              4.6e-28 -314.8
- X3    1   1.7e-27 2.1e-27 -309.2
- X1    1       2.0     2.0    1.5
- X2    1       2.0     2.0    1.5

Call:
lm(formula = Y ~ X1 + X2 + X3, data = mantel)

Coefficients:
(Intercept)           X1           X2           X3
  -1.00e+03     1.00e+00     1.00e+00     4.11e-15
```

It appears that the backward elimination algorithm selects to remove *none* of the terms, as AIC is lowest for the mean function will all terms. *However, the residual sum of squares for both the full mean function, and the mean function without X_3, are zero, within rounding error.* Consequently, the difference in AIC between the full mean function and the mean function without X_3 is due to rounding error only. Consequently, X_3 can be deleted, and still give an exact fit. Using backward elimination, therefore, $X_{\mathcal{A}} = \{X_1, X_2\}$.

These two computational algorithms give different answers. We would certainly prefer the choice $X_{\mathcal{A}} = \{X_1, X_2\}$ from backward elimination because it gives an exact fit. ∎

10.3 Use BE with the highway accident data and compare with the results in Table 10.7.

Solution: Using AIC as the selection criterion,

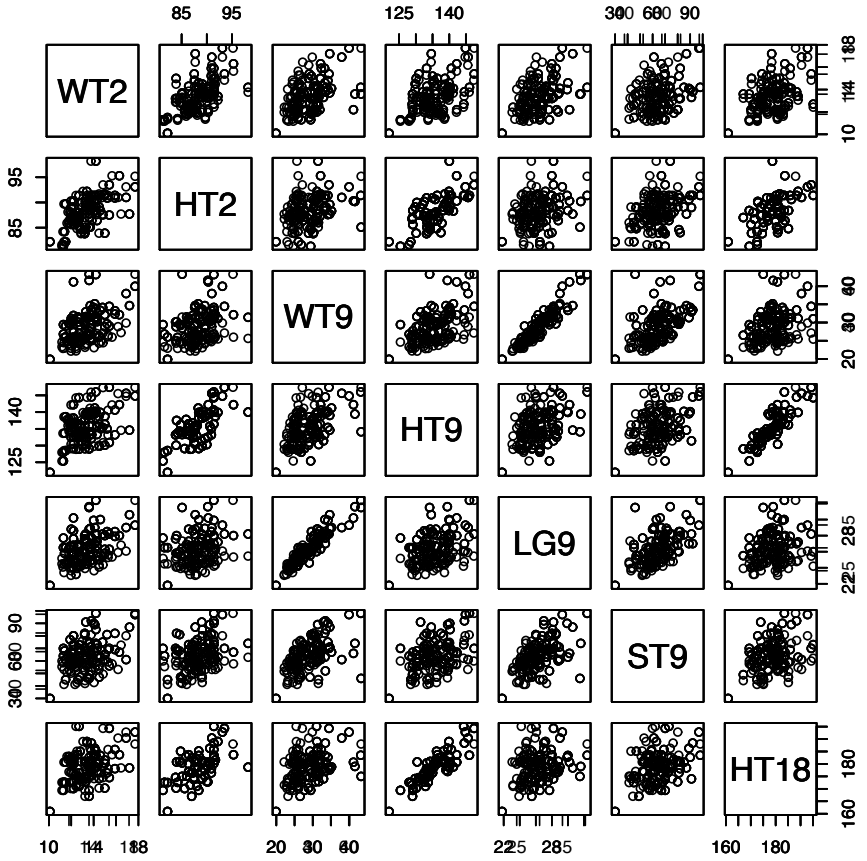
```
> ansf2 <- step(m1,scope=list(lower=~logLen,
+          upper=~logLen+logADT+logTrks+logSigs1+Slim+Shld+
+                 Lane+Acpt+Itg+Lwid+Hwy),
```

```
+                 direction="backward", data=a)
Start:  AIC= -65.61
 logRate ~ logLen + logADT + logTrks + logSigs1 + Slim + Shld +
    Lane + Acpt + Itg + Lwid + Hwy

           Df Sum of Sq   RSS   AIC
- Shld      1    0.0011   3.5 -67.6
- Itg       1    0.0031   3.5 -67.6
- Lane      1    0.0054   3.5 -67.6
- Lwid      1    0.0134   3.6 -67.5
- Acpt      1       0.1   3.6 -66.8
- logTrks   1       0.1   3.6 -66.6
<none>                    3.5 -65.6
- Hwy       3       0.6   4.2 -65.3
- logADT    1       0.3   3.8 -64.7
- Slim      1       0.4   3.9 -63.7
- logSigs1  1       0.9   4.5 -58.6

Step:  AIC= -67.6
 logRate ~ logLen + logADT + logTrks + logSigs1 + Slim + Lane +
    Acpt + Itg + Lwid + Hwy

           Df Sum of Sq   RSS   AIC
- Itg       1    0.0028   3.5 -69.6
- Lane      1    0.0057   3.5 -69.5
- Lwid      1    0.0149   3.6 -69.4
- Acpt      1       0.1   3.6 -68.5
- logTrks   1       0.1   3.7 -68.3
<none>                    3.5 -67.6
- Hwy       3       0.7   4.2 -66.7
- logADT    1       0.3   3.8 -66.6
- Slim      1       0.7   4.2 -62.5
- logSigs1  1       1.0   4.5 -59.9

Step:  AIC= -69.57
 logRate ~ logLen + logADT + logTrks + logSigs1 + Slim + Lane +
    Acpt + Lwid + Hwy

           Df Sum of Sq   RSS   AIC
- Lane      1    0.0052   3.5 -71.5
- Lwid      1    0.0140   3.6 -71.4
- Acpt      1       0.1   3.6 -70.5
- logTrks   1       0.1   3.7 -70.3
<none>                    3.5 -69.6
- logADT    1       0.3   3.9 -68.2
- Hwy       3       1.2   4.7 -64.5
- Slim      1       0.8   4.3 -63.7
- logSigs1  1       1.0   4.5 -61.9
```

```
Step:  AIC= -71.51
 logRate ~ logLen + logADT + logTrks + logSigs1 + Slim + Acpt +
    Lwid + Hwy

           Df Sum of Sq   RSS   AIC
- Lwid      1     0.016   3.6 -73.3
- Acpt      1       0.1   3.6 -72.5
- logTrks   1       0.1   3.7 -72.3
<none>                    3.5 -71.5
- logADT    1       0.4   3.9 -69.5
- Slim      1       0.8   4.3 -65.6
- Hwy       3       1.3   4.8 -65.5
- logSigs1  1       1.0   4.6 -63.7

Step:  AIC= -73.33
 logRate ~ logLen + logADT + logTrks + logSigs1 + Slim + Acpt +
    Hwy

           Df Sum of Sq   RSS   AIC
- Acpt      1       0.1   3.7 -74.2
- logTrks   1       0.1   3.7 -73.9
<none>                    3.6 -73.3
- logADT    1       0.4   3.9 -71.5
- Slim      1       0.8   4.4 -67.5
- Hwy       3       1.3   4.8 -67.4
- logSigs1  1       1.0   4.6 -65.4

Step:  AIC= -74.21
 logRate ~ logLen + logADT + logTrks + logSigs1 + Slim + Hwy

           Df Sum of Sq   RSS   AIC
- logTrks   1       0.1   3.8 -74.7
<none>                    3.7 -74.2
- logADT    1       0.3   4.0 -73.0
- Hwy       3       1.5   5.2 -66.7
- logSigs1  1       1.2   4.8 -65.5
- Slim      1       1.2   4.9 -65.1

Step:  AIC= -74.71
 logRate ~ logLen + logADT + logSigs1 + Slim + Hwy

           Df Sum of Sq   RSS   AIC
<none>                    3.8 -74.7
- logADT    1       0.3   4.1 -73.9
- Hwy       3       1.7   5.5 -66.4
- Slim      1       1.2   5.0 -66.4
- logSigs1  1       1.6   5.4 -63.3
```

10.4 For the boys in the Berkeley Guidance Study in Problem 3.1, find a model for *HT18* as a function of the other variables for ages two and nine. Perform a complete analysis, including selection of transformations and diagnostic analysis, and summarize your results.

Solution: We begin as usual with a scatterplot matrix of the relevant variables.



There is a separated point in most of the frames of the scatterplot matrix. Identification of points in R and S-Plus can't be done from a scatterplot matrix, so we drew the plot of *HT9* versus *WT9* to discover that case #60 is the unusual child, who was among the tallest children at age nine, but much heavier than any other child. We temporarily remove this child from the data.

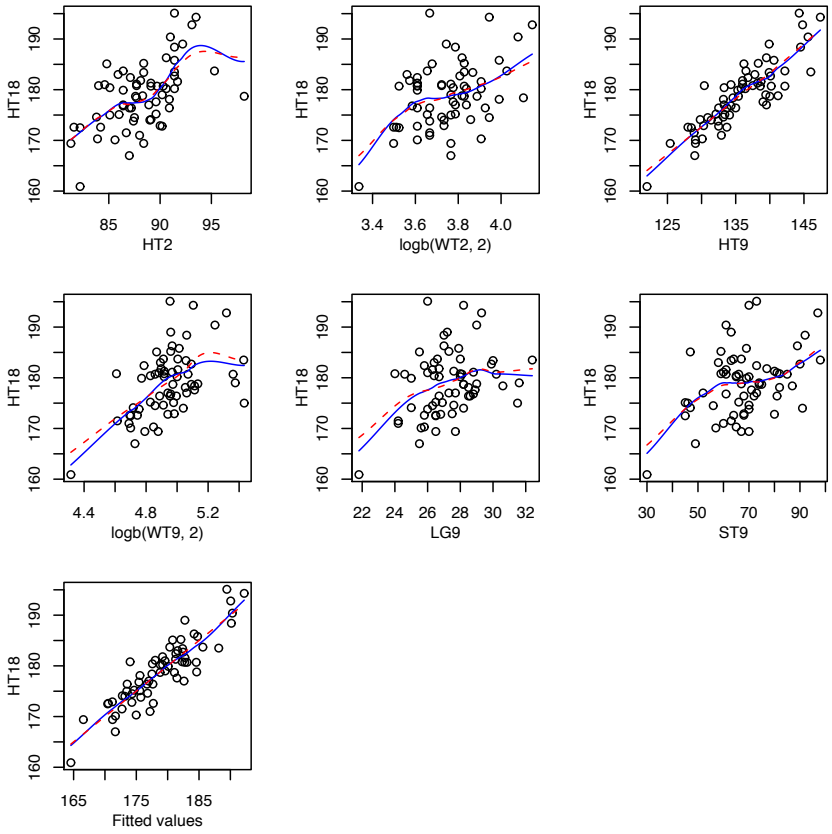
Next, we consider transformations of the predictors. Because the ranges of the predictors are so narrow, and the visual linearity of the frames in the scatterplot matrix, we would not expect that much improvement is possibly

via transformation. However, the multivariate Box-Cox method indicates that transformations can be desirable:

```
> summary(b1 <- powerTransform(cbind(HT2, WT2, HT9, WT9, LG9, ST9) ~ 1,
+         data=BGSboys, subset=-60))
bcPower Transformations to Multinormality

    Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
HT2   -2.3128   2.2847          -6.7909           2.1652
WT2   -1.3903   0.8588          -3.0735           0.2928
HT9   -1.8501   2.2731          -6.3053           2.6051
WT9   -1.0664   0.4123          -1.8745          -0.2582
LG9   -1.3051   1.0438          -3.3509           0.7407
ST9    0.8712   0.4453          -0.0016           1.7440

Likelihood ratio tests about transformation parameters
                                        LRT df         pval
LR test, lambda = (0 0 0 0 0 0)  13.873050  6 3.108687e-02
LR test, lambda = (1 1 1 1 1 1)  30.025456  6 3.887284e-05
LR test, lambda = (1 0 1 -1 0 1)  8.252498  6 2.201822e-01
```

The test for all logarithmic transformations has a p-value of about 0.03, while that for no transformations at all is very small. Examining the Wald tests for no transformations (powers equal to one), only the weight variables clearly require transformation. Shown are three additional likelihood ratio tests. The first transforms everything to logs except for *ST9*; the second transforms only the weight variables, and the third is intermediate between the two. There is little to choose between these three sets of transformations, so we will use the simplest, transforming only the weight variables to log scale.

We next turn to transforming the response. Using either the Box-Cox method or the inverse response plots, there is no evidence that any transformation will be helpful. The marginal model plots below suggest that this fitted mean function matches the data very well.



Finally, we can turn to subset selection.

```
> step(m1,lower=~1,data=BGSboys,subset=-60)
Start:  AIC= 150.83
 HT18 ~ HT2 + logb(WT2, 2) + HT9 + logb(WT9, 2) + LG9 + ST9

                Df Sum of Sq  RSS  AIC
- logb(WT9, 2)  1      0.24  534  149
- ST9           1        11  545  150
<none>                      534  151
- LG9           1        20  553  151
- logb(WT2, 2)  1        23  557  152
- HT2           1        28  562  152
- HT9           1       879 1413  212

Step:  AIC= 148.86
 HT18 ~ HT2 + logb(WT2, 2) + HT9 + LG9 + ST9

                Df Sum of Sq  RSS  AIC
- ST9           1        11  545  148
```

```
<none>                       534  149
- logb(WT2, 2)  1        23  557  150
- HT2           1        28  562  150
- LG9           1        89  623  157
- HT9           1      1190 1724  223

Step:  AIC= 148.24
 HT18 ~ HT2 + logb(WT2, 2) + HT9 + LG9

               Df Sum of Sq  RSS  AIC
<none>                       545  148
- HT2           1        24  569  149
- logb(WT2, 2)  1        24  569  149
- LG9           1        78  623  155
- HT9           1      1232 1778  223

Call:
lm(formula = HT18 ~ HT2 + logb(WT2, 2) + HT9 + LG9, data = BGSboys,
    subset = -60)

Coefficients:
 (Intercept)           HT2  logb(WT2, 2)           HT9           LG9
      32.922        -0.283         5.036         1.267        -0.729
```

We are left with four terms beyond the intercept for the candidate for the active predictors. This result is a bit surprising, as one might expect the age nine variables to be more relevant than the age two variables for age eighteen height.

Finally, we can examine the impact of case #60 by using the fitted value to predict height for that child:

```
> BGSboys$HT18[60]-predict(m2,data.frame(BGSboys[60,]))
[1] 3.3896
```

The error is about 3.4 cm, slightly more than one standard deviation, so the fitted model matches the unusual case fairly well. ∎

10.5

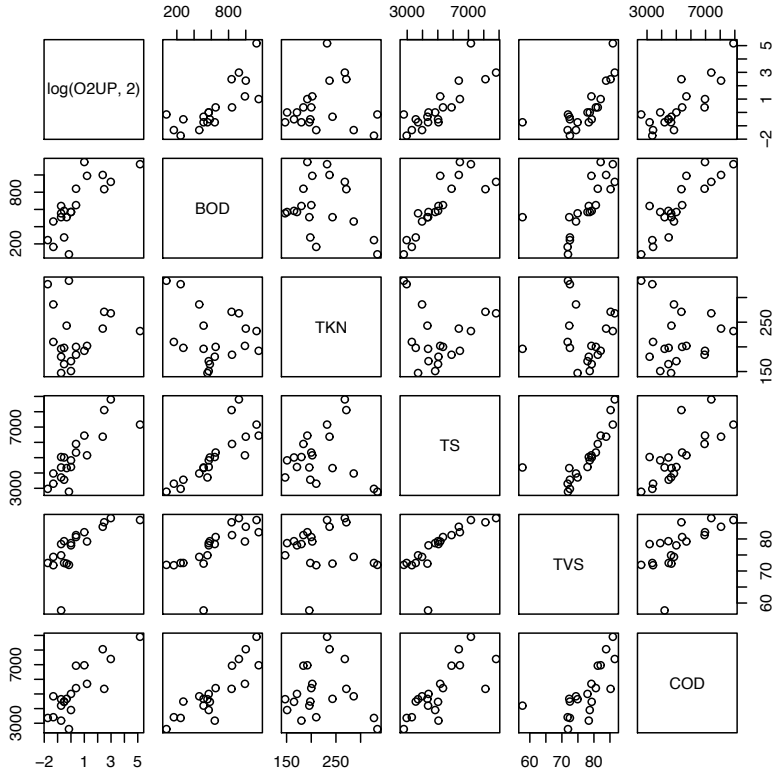
An experiment was conducted to study *O2UP*, oxygen uptake in milligrams of oxygen per minute, given five chemical measurements shown in Talbe 10.13 (Moore, 1975). The data were collected on samples of dairy wastes kept in suspension in water in a laboratory for 220 days. All observations were on the same sample over time. We desire an equation relating log(*O2UP*) to the other variables. The goal is to find variables that should be further studied with the eventual goal of developing a prediction equation; day cannot be used as a predictor. The data are given in the file `dwaste.txt`.

Complete the analysis of these data, including a complete diagnostic analysis. What diagnostic indicates the need for transforming *O2UP* to a logarithmic scale?

Table 10.13 Oxygen update experiment.

Variable	Description
Day	Day number
BOD	Biological oxygen demand
TKN	Total Kjeldahl nitrogen
TS	Total solids
TVS	Total volatile solids
COD	Chemical oxygen demand
O2UP	Oxygen uptake

Solution: As usual, we begin with a scatterplot matrix.



We have replaced *O2UP* by its logarithm based solely on the range of this variable. There are several separated points in the graph, which we would like to identify. R does not permit identifying points in a scatterplot matrix,

a facility that is greatly missed. The case with the very low value of TVS is case 17; deleting this case from the data, we get

```
> summary(b1 <- powerTransform(logb(O2UP,2)~BOD+TKN+TS+TVS+COD, data=dwaste,
  subset=-17))
box.cox Transformations to Multinormality

    Est.Power Std.Err. Wald(Power=0) Wald(Power=1)
BOD    0.6749   0.2469        2.7332       -1.3166
TKN   -0.5903   1.0466       -0.5640       -1.5195
TS     0.0668   0.4764        0.1403       -1.9589
TVS    2.3332   3.7079        0.6293        0.3596
COD    0.2722   0.5866        0.4640       -1.2408
                               LRT df  p.value
LR test, all lambda equal 0 11.123  5 0.049004
LR test, all lambda equal 1 10.880  5 0.053814
```

The p-values for both all logarithms and all untransformed are very close to 0.05. We interpret this to mean that there is very little information about the choice of transformation. We tentatively decide to continue without any further transformation. We can then justify the log-transform to the response using either the Box-Cox method or using an inverse response plot.

We next turn to residuals and influence. Examining residuals plots, and in particular using Tukey's test for nonadditivity, suggest that the mean function with predictors untransformed and the log of $O2UP$ as the response appears to be inadequate. An index plot of the influence statistics suggests that case #1 is highly influential for estimating coefficients; when case #1 is deleted, the resulting fit appears to be adequate. Using backward elimination, we are led to using only TS as the single active predictor. As a check, the plot of the fitted values from the mean function with all predictors versus the fitted values from the regression of with TS as the only term in the mean function is a straight line with relatively little scatter. We are led to include that TS might well be the only active term in the mean function.

We should now consider the deleted cases, seventeen and one. Case seventeen would have little impact on the mean function with TS as the only active term, since it was not unusual on TS. Case one is a little different because the data were ordered in time, and this day might well represent a different process that stabilized after a few hours. ∎

10.6 Prove the results (10.4)-(10.5). To avoid tedious algebra, start with an added-variable plot for X_j after all the other terms in the mean function. The estimated slope $\hat{\beta}_j$ is the OLS estimated slope in the added variable plot. Find the standard error of this estimate and show that it agrees with the given equations.

Solution: Let $\hat{e}(j)$ be the residuals from the regression of X_j on the other terms, and $\hat{e}(y)$ be the residuals from the regression of Y on all the terms except X_j. The added variable plot is of $\hat{e}(y)$ versus $\hat{e}(j)$, and the estimated

Table 10.14 Galpagos Island data.

Variable	Description
Island	Island name
NS	Number of species
ES	Number of endemic species (occurs only on that island)
Area	Surface area of island, hectares
Anear	Area of closest island, hectares
Dist	Distance to closest island, km
DistSC	Distance from Santa Cruz Island, km
Elevation	Elevation in m, missing values given as zero
EM	1 if elevation is observed, 0 if missing

slope is, from (2.5),

$$\hat{\beta}_j = \frac{\sum \hat{e}(j)_i \hat{e}(y)_i}{\sum \hat{e}(j)_i^2}$$

Correction of the sums of squares and cross-products for the mean is unnecessary because the averages of the two sets of residuals are both zero.

From (2.11), the standard error of $\hat{\beta}_j$ is $\sigma^2 / \sum \hat{e}(j)^2$. Now, $\sum \hat{e}(j)^2$ is just the residual sum of squares for the regression of X_j on the other terms, and so the result follows from (3.21), upon rearranging terms. ∎

10.7 Galápagos Islands

The Galápagos Islands off the coast of Ecuador provide an excellent laboratory for studying the factors that influence the development and survival of different life species. Johnson and Raven (1973) have presented data in the file `galapagos.txt` giving the number of species and related variables for 29 different islands. Counts are given for both the total number of species and the number of species that occur only on that one island (the endemic species).

Use these data to find factors that influence diversity, as measured by some function of the number of species and the number of endemic species, and summarize your results. One complicating factor is that elevation is not recorded for six very small islands, so some provision must be made for this. Four possibilities are: (1) find the elevations; (2) delete these six islands from the data; (3) ignore elevation as a predictor of diversity, or (4) substitute a plausible value for the missing data. Examination of large-scale maps suggests that none of these elevations exceed 200 m.

Solution: We substituted 40m for all the missing elevations. Starting with a scatterplot matrix, we concluded that all the variables should be replaced by their logarithms, including the response, which we took to be *NS*, the number of species. The regression of log(*NS*) on the log-predictors matches the data well, according to residual plots, marginal model plots and examination of Cook's distances and outlier statistics. We then used stepwise deletion of terms to get

```
> summary(m4)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9791     0.1827   16.31  3.6e-15
log(Area)     0.4105     0.0406   10.12  1.7e-10
log(Dist)    -0.1287     0.0873   -1.47     0.15

(Dispersion parameter for gaussian family taken to be 0.5503)

    Null deviance: 70.774  on 28  degrees of freedom
Residual deviance: 14.308  on 26  degrees of freedom
```

Bigger island have more species; separated islands may have fewer species. Analysis could also be done for the number of endemic species, perhaps using *NS* as a predictor.

We would be remiss to point out that the response might well be treated as a Poisson random variable, and study the number of species as a function of the predictors using Poisson regression, but that topic is not covered in this book. The Poisson model appears to work well, and suggests that *all* of the predictors might be important.∎

10.8 Suppose that (10.1) holds with $\boldsymbol{\beta}_{\mathcal{I}} = \mathbf{0}$, but we fit a subset model using the terms $X_{\mathcal{C}} \neq X_{\mathcal{A}}$; that is, $X_{\mathcal{C}}$ does not include all the relevant terms. Give general conditions under which the mean function $\mathrm{E}(Y|X_{\mathcal{C}})$ is a linear mean function. (Hint: See Appendix A.2.4.)

Solution:

$$\begin{aligned}\mathrm{E}(Y|X_{\mathcal{C}}) &= \mathrm{E}\left[\mathrm{E}(Y|X)|X_{\mathcal{C}}\right] \\ &= \boldsymbol{\beta}'_{\mathcal{A}}\mathrm{E}(\mathbf{x}_{\mathcal{A}}|X_{\mathcal{C}} = \mathbf{x}_{\mathcal{C}}) + \boldsymbol{\beta}'_{\mathcal{I}}\mathrm{E}(\mathbf{x}_{\mathcal{I}}|X_{\mathcal{C}} = \mathbf{x}_{\mathcal{C}}) \\ &= \boldsymbol{\beta}'_{\mathcal{A}}\mathrm{E}(\mathbf{x}_{\mathcal{A}}|X_{\mathcal{C}} = \mathbf{x}_{\mathcal{C}})\end{aligned}$$

We will get a linear regression mean function if the regression of each of the active predictors on $X_{\mathcal{C}}$ has a linear mean function. This is guaranteed if the Xs are multivariate normal, or at least approximately so, and this further justifies transforming predictors toward normality. If this is not done, then the linearity of the full model cannot guarantee the linearity of any subset model other than the one with active predictors only. ∎

10.9 For the highway accident data, fit the regression model with active predictors given by the subset with the smallest value of *PRESS* in Table 10.7. The coefficient estimate of *Slim* is *negative*, meaning that segments with higher speed limits lower *lower* accident rates. Explain this finding.

Solution: Since these data are not experimental, but rather observational, we cannot infer causation from these data, and so the negative coefficient estimate does not necessarily imply that raising speed limits causes fewer accidents. We might in fact want to infer that high accident rates *cause* lower speed limits, because changing a speed limit sign is an inexpensive response to high accident rates.∎

10.10 Reëxpress C_p as a function of the F statistic used for testing the null hypothesis (10.6) versus the alternative (10.1). Discuss.

Solution: We get

$$C_p = (k' - p)(F_p - 1) + p$$

where k' is the number of terms in the full mean function, p is the number of terms in the candidate for the active subset, and F_p is the F-statistic for comparing these two mean functions. For a fixed value of p, C_p orders mean functions in the same way as F_p. C_p will be smaller than p only if $F_p < 1$; under the null hypothesis that the smaller mean function is appropriate, the expected value of F_p is close to one, so subsets with values of C_p substantially smaller than p are not clearly superior to subsets with $C_p \approx p$.∎

10.11 In the windmill data discussed in Section 10.4, data were collected at the candidate site for about a year, for about 1200 observations. One issue is whether or not the collection period could be shortened to six months, about 600 observations, or three months, about 300 observations, and still give a reliable estimate of the long-term average wind speed.

Design and carry out a simulation experiment using the data described in Section 10.4.2 to characterize the increase in error due to shortening the collection period. For the purpose of the simulation, consider site #1 to be the "candidate" site, and site #2 to be the reference site, and consider only the use of *Spd2* to predict *Spd1*. (Hint: The sampling scheme used in Section 10.4.2 may not be appropriate for time periods shorter than a year because of seasonal variation. Rather than picking 600 observations at random to make up a simulated six-month period, a better idea might be to pick a starting observation at random, and then pick 600 consecutive observations to comprise the simulated six months.)

Solution: The file `wm5.txt` is *not* a part of the regular downloads of data files, but must be obtained separately from www.stat.umn.edu/alr/data/wm5.txt. Here is R code that will read the file, and create two functions for the simulation.

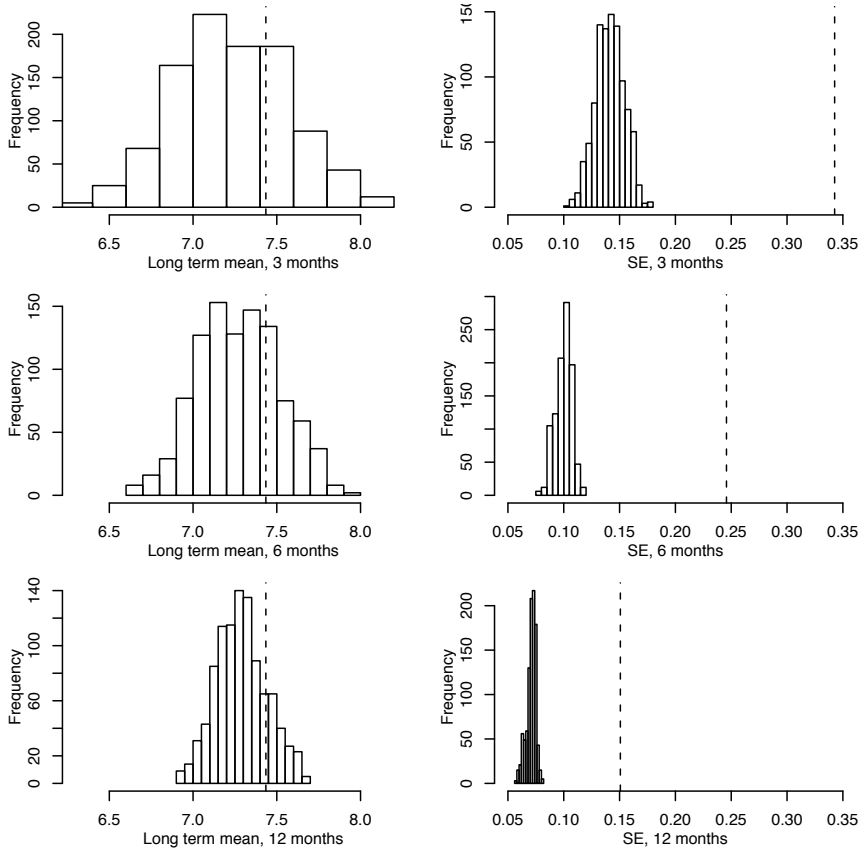
```
> wm5 <- read.table(url("http://www.stat.umn.edu/alr/data/wm5.txt"),
>       header=TRUE)
> # function to do the simulation
> # The sample size is 62039.  A simulation will consist of M
> # consecutive observations, where M is 300 (3 months),
> # 600 (6 months) or 1200 (one year)
> # (a) generate a random integer between 1 on 62039-M and then return M
> # consecutive integers starting with that number
> random.ints <- function(N=62039,M=600){
+      ans<-floor(runif(1)*(N-M))+1 ; ans:(ans+M-1)}
> m1 <- lm(Spd1 ~ Spd2, data=wm5)
> spd1.mean <- mean(wm5$Spd1)
> spd2.mean <- mean(wm5$Spd2)
> do.sim <- function(Nsim=1000,M=600){
```

```
+   ans <- NULL
+   for (j in 1:Nsim){
+    m <- update(m1,subset=random.ints(M=M))
+    ans <- rbind(ans,
+            unlist(predict(m,data.frame(Spd1=spd2.mean),se.fit=TRUE)))
+    }
+   ans}
```

If you are connected to the Internet, the `read.table` command will read the data file from the web. The function `random.ints` chooses a starting value at random, and then returns M consecutive integers. The function `do.sim` does the simulation by repeatedly updating the model `m1` to have only the cases specified by M, and then saves the output from the `predict` command. The `unlist` command turns the output from `predict` into a vector. The output from the function is a matrix with $Nsim$ rows. The function is run three times:

```
> ans300 <- do.sim(M=300)
> ans600 <- do.sim(M=600)
> ans1200 <- do.sim(M=1200)
```

and the results are summarized in the histograms given below, similar to the histograms in Figure 10.1.



The top row corresponds to samples of 300 consecutive observations, corresponding to about three months, the second row to six months and the third row to about a year. The left column is the histogram for the estimated long-term mean wind speed. All three methods seem to underestimate the true value, which is the actual mean for *Spd1* in the data, as indicated by the dashed line. The three-month data appears considerably more variable than the twelve month data, as might we expected; from three month data, and error of as much as 1 meter per second is possible.

The second column is also interesting, giving a histogram of the standard errors of the long-term wind speed as well as the dashed line for the standard deviation of the estimated means in the left histogram. *Unlike Figure 10.1, the standard error from the formula is substantially underestimating the error in the estimate.* If you repeat the simulation, but select cases at *random* rather than consecutive cases with a random starting point, then a result similar to Figure 10.1 is obtained. The formula assumes (1) no seasonal variation, (2) that the coefficients of the straight line are the same from year to year, and (3) that the correlation between adjacent observations is zero. Random sampling

assumes all three of these conditions hold, while the consecutive sampling does not. We conclude that one or more of these assumptions does not hold, accounting for the discrepancy between the simulated standard deviation and formula standard error. This strongly suggests that, if this methodology is to be used, short time periods like three months may produce discrepant answers, with an overly optimistic estimate of the error in the long term estimate of the mean. ∎

11

Nonlinear regression

Problems

11.1 Suppose we have a response Y, a predictor X, and a factor G with g levels. A generalization of the concurrent regression mean function given by Model 3 of Section 6.2.2, is, for $j = 1, \ldots, g$,

$$\mathrm{E}(Y|X = x, G = j) = \beta_0 + \beta_{1j}(x - \gamma) \tag{11.20}$$

for some point of concurrence γ.

11.1.1. Explain why (11.20) is a nonlinear mean function. Describe in words what this mean function specifies.

Solution: The mean function is nonlinear because γ multiplies β_{1j}. It describes a straight-line mean function for each level of G. Each group has its own slope β_{1j}, but all lines are concurrent at $x = \gamma$. ∎

11.1.2. Fit (11.20) to the sleep data discussed in Section 6.2.2, so the mean function of interest is

$$\mathrm{E}(TS|\log(BodyWt) = x, D = j) = \beta_0 + \beta_{1j}(x - \gamma)$$

(Hint: To get starting values, fit the concurrent regression model with $\gamma = 0$. The estimate of γ will be very highly variable, as is often the case with centering parameters like γ in this mean function.)

Solution:

```
> sleep1$logBodyWt <- log(sleep1$BodyWt)
> # The next line removes rows with missing values, and selects
```

```
> # only columns 3, 10 and 12 of sleep1
> sleep <- sleep1[!is.na(sleep1$TS),c(3,10,12)]
> sleep$fD <- factor(sleep$D, ordered=FALSE)
> attach(sleep)
> m0 <- lm(TS ~ logBodyWt:fD) # concurrent regressions
> summary(m0)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     11.626      0.546   21.30   <2e-16
logBodyWt:fD1   -0.289      0.279   -1.04    0.305
logBodyWt:fD2   -0.593      0.700   -0.85    0.401
logBodyWt:fD3   -0.932      0.352   -2.65    0.011
logBodyWt:fD4   -0.641      0.302   -2.12    0.038
logBodyWt:fD5   -1.659      0.332   -4.99    7e-06

Residual standard error: 3.69 on 52 degrees of freedom
Multiple R-Squared: 0.413
F-statistic: 7.33 on 5 and 52 DF,  p-value: 2.90e-05

> m1 <- nls(TS~ b0 + b11*((D==1)*(logBodyWt - gamma))
+                  + b12*((D==2)*(logBodyWt - gamma))
+                  + b13*((D==3)*(logBodyWt - gamma))
+                  + b14*((D==4)*(logBodyWt - gamma))
+                  + b15*((D==5)*(logBodyWt - gamma)),
+    data=sleep,
+    start=list(b0=11,b11=-.3,b12=-.6,b13=-.9,b14=-.6,
+               b15=-1.6, gamma=0))
> summary(m1)

Formula: TS ~ b0 + b11 * ((D == 1) * (logBodyWt - gamma)) + b12 * ((D ==
    2) * (logBodyWt - gamma)) + b13 * ((D == 3) * (logBodyWt -
    gamma)) + b14 * ((D == 4) * (logBodyWt - gamma)) + b15 *
    ((D == 5) * (logBodyWt - gamma))

Parameters:
      Estimate Std. Error t value Pr(>|t|)
b0      49.372    192.655    0.26  0.79877
b11     -0.590      0.258   -2.29  0.02610
b12     -0.630      0.167   -3.76  0.00044
b13     -0.650      0.192   -3.38  0.00138
b14     -0.652      0.191   -3.41  0.00128
b15     -0.705      0.388   -1.82  0.07492
gamma  -60.129    305.077   -0.20  0.84454

Residual standard error: 3.37 on 51 degrees of freedom
```

The estimate of γ has such as large variance that there is no reason to include γ in the mean function. ∎

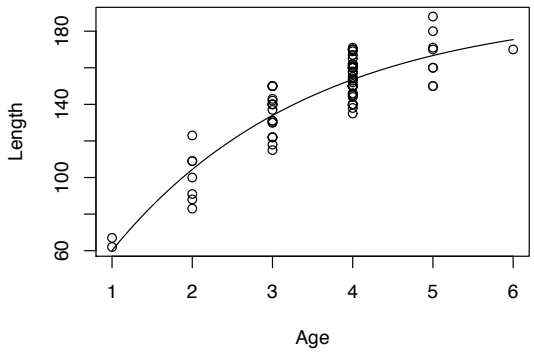
11.2 In fisheries studies, the most commonly used mean function for expected length of a fish at a given age is the *von Bertalanffy* function, von Bertalanffy (1938), Haddon (2001), given by

$$\mathrm{E}(Length|Age = t) = L_\infty(1 - \exp(-K(t - t_0)) \tag{11.21}$$

The parameter L_∞ is the expected value of *Length* for extremely large ages, and so it is the asymptotic or upper limit to growth, and K is a growth rate parameter that determines how quickly the upper limit to growth is reached. When $Age = t_0$, the expected length of the fish is zero, which allows fish to have non-zero length at birth if $t_0 < 0$.

11.2.1. The data in the file `lakemary.txt` gives the *Age* in years and *Length* in mm for a sample of 78 bluegill fish from Lake Mary, Minnesota, in 1981 (courtesy of Richard Frie). *Age* is determined by counting the number of rings on a scale of the fish. This is a cross-sectional data set, meaning that all the fish were measured once. Draw a scatterplot of the data.

Solution:



∎

11.2.2. Use nonlinear regression to fit the von Bertalanffy function to these data. To get starting values, first guess at L_∞ from the scatterplot to be a value larger than any of the observed values in the data. Next, divide both sides of (11.21) by the initial estimate of L_∞, and rearrange terms to get just $\exp(-K(t - t_0))$ on the right of the equation. Take logarithms, to let a linear mean function, and then use OLS for the linear mean function to get the remaining starting values. Draw the fitted mean function on your scatterplot.

Solution:

```
> LI <- 250
> z <- log(1-Length/LI)
> m0 <- lm(z ~ Age)
> K <- -coef(m0)[2]
> t0 <- coef(m0)[1]/coef(m0)[2]
```

```
> m1 <- nls(Length~LI*(1-exp(-K*(Age-t0))), data=m,
+           start=list(LI=LI, K=K, t0=t0))
Formula: Length ~ LI * (1 - exp(-K * (Age - t0)))
Parameters:
   Estimate Std. Error t value Pr(>|t|)
LI 192.8101    13.0800   14.74  < 2e-16
K    0.4063     0.0885    4.59  1.7e-05
t0   0.0809     0.2402    0.34     0.74
Residual standard error: 11 on 75 degrees of freedom
Correlation of Parameter Estimates:
       LI     K
K  -0.971
t0 -0.779 0.895
vals <- seq(1,6,length=100)
lines(vals,predict(m1,data.frame(Age=vals)))
```

The estimate of L_∞ is highly variable. The estimate of t_0 is essentially zero, and could probably be removed from the mean function. K appears to be fairly well estimated. ∎

11.2.3. Obtain a 95% confidence interval for L_∞ using the large-sample approximation, and using the bootstrap.

Solution: The 95% confidence intervals based on the large-sample approximation for all three parameters are

```
> confint(m1)
Waiting for profiling to be done...
          2.5%       97.5%
LI 174.3871573 233.3509747
K    0.2399484   0.5801005
t0  -0.5437455   0.4471603
```

For the bootstrap,

```
> boot <- bootCase(m1, B=999)
> out <- rbind(apply(boot, 2, mean),
+   apply(boot, 2, function(x) quantile(x, c(.025, .975))))
> colnames(out) <- names(coef(m1))
> rownames(out)[1] <- "Boot Mean"
> t(out)
     Boot Mean        2.5%       97.5%
LI 193.7126826 171.5047177 229.8803513
K    0.4292832   0.2551286   0.6859408
t0   0.1212150  -0.3707765   0.8049379
```

For L_∞, the asymptotic interval is too small, and the upper limit to growth could reasonably be much larger than 218 mm. Similarly, the growth parameter K could reasonably exceed the upper limit from the asymptotic method of 0.58. ∎

11.3 The data in the file `walleye.txt` give the *length* in mm and the *age* in years of a sample of over 3000 male walleye, a popular game fish, captured in Butternut Lake in Northern Wisconsin (LeBeau, 2004). The fish are also classified according to the time *period* in which they were captured, with *period* = 1 for pre-1990, *period* = 2 for 1990–1996, and *period* = 3 for 1997–2000. Management practices on the lake were different in each of the periods, so it is of interest to compare the length at age for the three time periods.

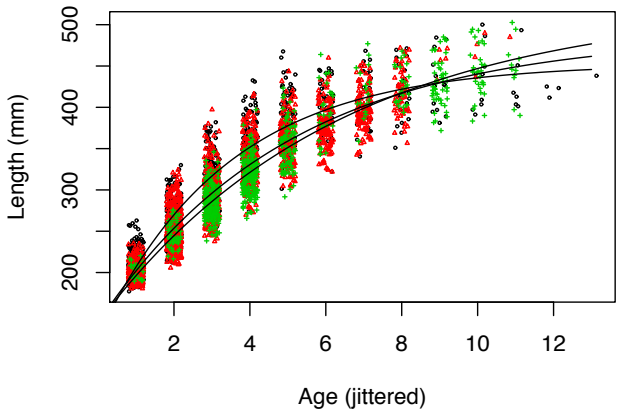
Using the von Bertalanffy length at age function (11.21), compare the three time periods. If different, are all the parameters different, or just some of them? Which ones? Summarize your results.

Solution: This requires specifying a sequence of models corresponding to the choices of mean function to be prepared. We considered five such mean functions, although many more are possible:

```
> d <- walleye
> attach(d)
> # Get the starting values
> LI <- max(length)+1
> z <- log(1-length/LI)
> m0 <- lm(z ~ age)
> K <- -coef(m0)[2]
> t0 <- coef(m0)[1]/coef(m0)[2]
> # Fit the models
> # c1:  no period effect
> c1 <- nls(length~LI*(1-exp(-K*(age-t0))),
+           start=list(LI=LI, K=K, t0=t0))
> # c2: All periods are different
> c2 <- nls(length~(period==1)*LI1*(1-exp(-K1*(age-t01))) +
+                  (period==2)*LI2*(1-exp(-K2*(age-t02))) +
+                  (period==3)*LI3*(1-exp(-K3*(age-t03))),
+                  start=list(LI1=LI,LI2=LI,LI3=LI,
+                             K1=K,K2=K,K3=K,
+                             t01=t0,t02=t0,t03=t0))
> # c3: Common LI
> c3 <- nls(length~(period==1)*LI*(1-exp(-K1*(age-t01))) +
+                  (period==2)*LI*(1-exp(-K2*(age-t02))) +
+                  (period==3)*LI*(1-exp(-K3*(age-t03))),
+                  start=list(LI=LI,
+                             K1=K,K2=K,K3=K,
+                             t01=t0,t02=t0,t03=t0))
> # c4:  Common K
> c4 <- nls(length~(period==1)*LI1*(1-exp(-K*(age-t01))) +
+                  (period==2)*LI2*(1-exp(-K*(age-t02))) +
+                  (period==3)*LI3*(1-exp(-K*(age-t03))),
+                  start=list(LI1=LI,LI2=LI,LI3=LI,
+                             K=K,
+                             t01=t0,t02=t0,t03=t0))
> # c5:  Common t0
> c5 <- nls(length~(period==1)*LI1*(1-exp(-K1*(age-t0))) +
```

```
+                   (period==2)*LI2*(1-exp(-K2*(age-t0))) +
+                   (period==3)*LI3*(1-exp(-K3*(age-t0))),
+                   start=list(LI1=LI,LI2=LI,LI3=LI,
+                              K1=K,K2=K,K3=K,
+                              t0=t0))
> # Compare models
> anova(c1,c3,c2)
Analysis of Variance Table

Model 1: length ~ LI * (1 - exp(-K * (age - t0)))
Model 2: length ~ (period == 1) * LI * (1 - exp(-K1 * (age - t01))) +
                  (period == 2) * LI * (1 - exp(-K2 * (age - t02))) +
                  (period == 3) * LI * (1 - exp(-K3 * (age - t03)))
Model 3: length ~ (period == 1) * LI1 * (1 - exp(-K1 * (age - t01))) +
                  (period == 2) * LI2 * (1 - exp(-K2 * (age - t02))) +
                  (period == 3) * LI3 * (1 - exp(-K3 * (age - t03)))
  Res.Df Res.Sum Sq   Df  Sum Sq F value  Pr(>F)
1   3195    2211448
2   3191    1994577    4  216871    86.7 < 2e-16
3   3189    1963513    2   31064    25.2 1.3e-11
> anova(c1,c4,c2)
Analysis of Variance Table

Model 1: length ~ LI * (1 - exp(-K * (age - t0)))
Model 2: length ~ (period == 1) * LI1 * (1 - exp(-K * (age - t01))) +
                  (period == 2) * LI2 * (1 - exp(-K * (age - t02))) +
                  (period == 3) * LI3 * (1 - exp(-K * (age - t03)))
Model 3: length ~ (period == 1) * LI1 * (1 - exp(-K1 * (age - t01))) +
                  (period == 2) * LI2 * (1 - exp(-K2 * (age - t02))) +
                  (period == 3) * LI3 * (1 - exp(-K3 * (age - t03)))
  Res.Df Res.Sum Sq   Df  Sum Sq F value Pr(>F)
1   3195    2211448
2   3191    2014863    4  196585    77.8 <2e-16
3   3189    1963513    2   51350    41.7 <2e-16
> anova(c1,c5,c2)
Analysis of Variance Table

Model 1: length ~ LI * (1 - exp(-K * (age - t0)))
Model 2: length ~ (period == 1) * LI1 * (1 - exp(-K1 * (age - t0))) +
                  (period == 2) * LI2 * (1 - exp(-K2 * (age - t0))) +
                  (period == 3) * LI3 * (1 - exp(-K3 * (age - t0)))
Model 3: length ~ (period == 1) * LI1 * (1 - exp(-K1 * (age - t01))) +
                  (period == 2) * LI2 * (1 - exp(-K2 * (age - t02))) +
                  (period == 3) * LI3 * (1 - exp(-K3 * (age - t03)))
  Res.Df Res.Sum Sq   Df  Sum Sq F value  Pr(>F)
1   3195    2211448
2   3191    1989989    4  221458    88.8 < 2e-16
3   3189    1963513    2   26476    21.5 5.3e-10
> detach(d)
```

The model `c1` ignores the period effect. `c5` has separate parameters for each period, and is the most general. Models `c2`–`c4` are intermediate, setting either the asymptote, rate or start parameters equal. In each case, we use the method suggested in previous problems to get starting values. The five models can be compared using analysis of variance. The most general model seems appropriate, so all three parameters differ in each period. Sample sizes here are very large, so the tests are very powerful and may be detecting relatively unimportant differences.



■

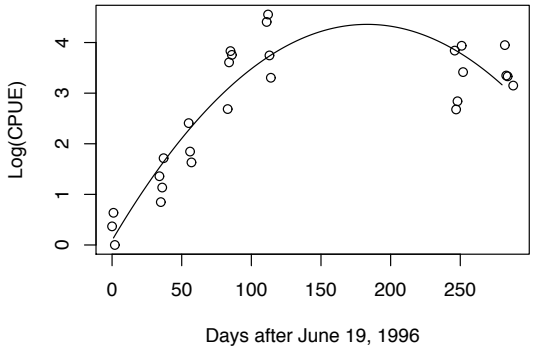
11.4 A Quadratic Polynomial as a Nonlinear Model The data in the file `swan96.txt` were collected by the Minnesota Department of Natural Resources to study the abundance of black crappies, a species of fish, on Swan Lake, Minnesota in 1996. The response variable is *LCPUE*, the logarithm of the catch of 200 mm or longer black crappies per unit of fishing effort. It is believed that *LCPUE* is proportional to abundance. The single predictor is *Day*, the day on which the sample was taken, measured as the number of days after June 19, 1996. Some of the measurements were taken the following spring on the same population of fish before the young of the year are born in late June. No samples are taken during the winter months when the lake surface was frozen.

11.4.1. For these data fit the quadratic polynomial

$$\mathrm{E}(LCPUE|Day = x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

assuming $\mathrm{Var}(LCPUE|Day = x) = \sigma^2$. Draw a scatterplot of *LCPUE* versus *Day*, and add the fitted curve to this plot.

Solution:



■

11.4.2. Using the deltaMethod described in Section 6.1.2, obtain the estimate and variance for the value of *Day* that maximizes E(*LCPUE*|*Day*).

Solution:

```
> n1 <- nls(LCPUE ~ b0 + b1*Day + b2*Day*Day, data=d,
+       start=list(b0=.09,b1=.05,b2=-.00013))
> summary(n1)
Formula: LCPUE ~ b0 + b1 * Day + b2 * Day * Day
Parameters:
      Estimate   Std. Error t value Pr(>|t|)
b0  9.2115e-02   2.7102e-01    0.34     0.74
b1  4.6605e-02   5.2002e-03    8.96    4e-09
b2 -1.2726e-04   1.6782e-05   -7.58    8e-08
Residual standard error: 0.572 on 24 degrees of freedom
> deltaMethod(n1, "-b1/(2*b2)")
           Estimate       SE
-b1/(2*b2) 183.1104 5.961452
```

■

11.4.3. Another parameterization of the quadratic polynomial is

$$\mathrm{E}(Y|X) = \theta_1 - 2\theta_2\theta_3 x + \theta_3 x^2$$

where the θs can be related to the βs by

$$\theta_1 = \beta_0, \quad \theta_2 = -\beta_1/2\beta_2, \quad \theta_3 = \beta_2$$

In this parameterization, θ_1 is the intercept, θ_2 is the value of the predictor that gives the maximum value of the response, and θ_3 is a measure of curvature. This is a nonlinear model because the mean function is a nonlinear function of the parameters. Its advantage is that at least two of the parameters, the intercept θ_1 and the value of x that maximizes the response θ_2, are

directly interpretable. Use nonlinear least squares to fit this mean function. Compare your results to the first two parts of this problem.

Solution:

```
> n2 <- nls(LCPUE  ~  th1-2*th2*th3*Day  + th3*Day^2, data=swan96,
+        start=list(th1=.09, th2=183, th3=-.00013))
> summary(n2)

Formula: LCPUE ~ th1 - 2 * th2 * th3 * Day + th3 * Day^2

Parameters:
      Estimate Std. Error t value Pr(>|t|)
th1  9.212e-02  2.710e-01   0.340    0.737
th2  1.831e+02  5.961e+00  30.716  < 2e-16 ***
th3 -1.273e-04  1.678e-05  -7.583 8.03e-08 ***
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

Residual standard error: 0.5716 on 24 degrees of freedom

Number of iterations to convergence: 2
Achieved convergence tolerance: 2.576e-07
```

The nonlinear least squares fit gives the same estimate and standard error as does the deltaMethod. ∎

11.5 Nonlinear regression can be used to select transformations for a linear regression mean function. As an example, consider the highway accident data, described in Table 7.1, with response $\log(Rate)$ and two predictors $X_1 = Len$ and $X_2 = ADT$. Fit the nonlinear mean function

$$\mathrm{E}(\log(Rate)|X_1 = x_1, X_2 = x_2, X_3 = x_3) = \beta_0 + \beta_1\psi_S(X_1, \lambda_1) + \beta_2\psi_S(X_2, \lambda_2)$$

where the scaled power transformations $\psi_S(X_j, \lambda_j)$ are defined at (7.3). Compare the results you get to results obtained using the transformation methodology in Chapter 7.

Solution:

```
> psi.s <- function(x,lambda) powtran(x,lambda,modified=FALSE)
> # starting values
> bstart <- coef(lm(logb(Rate,2) ~ psi.s(Len,1) + psi.s(ADT,1),
            data=highway))
> m2 <- nls(logb(Rate,2)~b0 + b1*psi.s(Len,lam1) + b2*psi.s(ADT,lam2),
            data=highway,start=list(b0=bstart[1],b1=bstart[2],
                                    b2=bstart[3],lam1=1,lam2=1))
> summary(m2)

Formula: logb(Rate, 2) ~ b0 + b1 * psi.s(Len, lam1) + b2 * psi.s(ADT,
    lam2)
```

```
Parameters:
     Estimate Std. Error t value Pr(>|t|)
b0      5.093      1.702    2.99   0.0051
b1     -1.672      1.970   -0.85   0.4018
b2     -0.566      0.692   -0.82   0.4193
lam1   -0.352      0.544   -0.65   0.5218
lam2   -0.693      0.879   -0.79   0.4363

Residual standard error: 0.546 on 34 degrees of freedom
```

The function `psi.s` matches the definition of ψ_S in the text, and it uses the `powtran` command in `alr3`. We get starting values by fitting via OLS assuming that $\lambda_1 = \lambda_2 = 1$. The nonlinear mean function is then specified using the starting values just obtained. The methods in Chapter 7 either transform one variable at a time for linearity in the regression of the response on the predictor, or else use the multivariate Box-Cox method to transform for multivariate normality. This method simultaneously transforms two predictors for linearity, and so is different from the other methods. The suggested transformations are $\lambda_1 \approx -1/3$ and $\lambda_2 \approx -2/3$, but both are within one standard error of zero for a log-transformation. ∎

11.6 POD models Partial one-dimensional mean functions for problems with both factors and continuous predictors were discussed in Section 6.4. For the Australian athletes data discussed in that section, the mean function (6.26),

$$\begin{aligned}\mathrm{E}(LBM|Sex, Ht, Wt, RCC) = \\ \beta_0 + \beta_1 Sex + \beta_2 Ht + \beta_3 Wt + \beta_4 RCC \\ + \eta_0 Sex + \eta_1 Sex \times (\beta_2 Ht + \beta_3 Wt + \beta_4 RCC)\end{aligned}$$

was suggested. This mean function is nonlinear because η_1 multiplies each of the βs. Problem 6.21 provides a simple algorithm for finding estimates using only standard linear regression software. This method, however, will not produce the large-sample estimated covariance matrix that is available using nonlinear least squares.

11.6.1. Describe a reasonable method for finding starting values for fitting (6.26) using nonlinear least squares.

Solution: If we assume that $\eta_1 = 0$, then the mean function becomes the parallel within-group regression mean function. Fit this model via OLS to get estimates of the βs. Assuming the estimates of the βs are known, write $L = \beta_2 Ht + \beta_3 Wt + \beta_4 RCC$, and the mean function becomes

$$\mathrm{E}(LBM|Sex, Ht, Wt, RCC) \quad = \quad \beta_0 + L + \eta_0 Sex + \eta_1 Sex \times L$$

and so starting values for β_0, η_0 and η_1 can be obtained from the OLS regression of LBM on L, Sex and their interaction. ∎

11.6.2. For the cloud seeding data, Problem 9.11, fit the partial one-dimensional model using the action variable A as the grouping variable, and summarize your results.

Solution: In R and S-Plus, this model can be fit using the `pod` command in the `alr3` package. It finds the starting values as specified in the last subproblem, and then fits the pod model using nonlinear least squares:

```
> summary(p1 <- pod(LBM~Ht+Wt+RCC, data=ais, group=Sex))

Formula: LBM ~ eta0 + eta1 * Ht + eta2 * Wt + eta3 * RCC +
    Sex1 * (th02 + th12 * (eta1 * Ht + eta2 * Wt + eta3 * RCC))

Parameters:
     Estimate Std. Error t value Pr(>|t|)
eta0 -14.6565     6.4645   -2.27  0.02447
eta1   0.1463     0.0342    4.27  3.0e-05
eta2   0.7093     0.0242   29.36  < 2e-16
eta3   0.7248     0.5854    1.24  0.21717
th02  12.8472     3.7634    3.41  0.00078
th12  -0.2587     0.0345   -7.51  2.1e-12

Residual standard error: 2.46 on 196 degrees of freedom
```

12

Logistic Regression

Problems

12.1 Downer data

For unknown reasons, dairy cows sometimes become recumbent—they lay down. Called *downers*, these cows may have a serious illness that may lead to death of the cow. These data are from a study of blood samples of over 400 downer cows studied at the Ruakura New Zealand Animal Health Laboratory during 1983-84. A variety of blood tests were performed, and for many of the animals the outcome (survived, died, or animal was killed) was determined. The goal is to see if survival can be predicted from the blood measurements. The variables in the data file `downer.txt` are described in Table 12.7. These data were collected from veterinary records, and not all variables were recorded for all cows.

12.1.1. Consider first predicting *Outcome* from *Myopathy*. Find the fraction of surviving cows of $Myopathy = 0$ and for $Myopathy = 1$.

Solution: The frequency table is:

```
                     Myopathy
                       No  Yes
Survive:           No 78   89
                   Yes 49    6
Survival fraction:  0.39 0.06
```

12.1.2. Fit the logistic regression with response *Outcome*, and the single predictor *Myopathy*. Obtain a 95% confidence interval for coefficient

Table 12.7 The recumbent cow data, from Clark, Henderson, Hoggard, Ellison and Young (1987).

Variable	n	Description
AST	429	Serum asparate amino transferase (U/l at 30C)
Calving	431	0 if measured before calving, 1 if after
CK	413	Serum creatine phosphokinase (U/l at 30C)
Daysrec	432	Days recumbent when measurements were done
Inflamat	136	Is inflammation present? 0=no, 1=yes
Myopathy	222	Is muscle disorder present? 1=yes, 0=no
PCV	175	Packed cell volume (Haemactocrit), percent
Urea	266	Serum urea (mmol/l)
Outcome	435	1 if survived, 0 if died or killed

for *Myopathy*, and compute the estimated decrease in odds of survival when $Myopathy = 1$. Obtained the estimated probability of survival when $Myopathy = 0$ and when $Myopathy = 1$, and compare with the observed survival fractions in Problem 12.1.1.

Solution: Using R,

```
> m1 <- glm(Outcome~Myopathy, data=d, family=binomial())
> summary(m1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.465      0.182   -2.55    0.011
Myopathy      -2.232      0.459   -4.86  1.2e-06
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 248.57  on 221  degrees of freedom
Residual deviance: 214.14  on 220  degrees of freedom

> exp(coef(m1))
(Intercept)    Myopathy
    0.62821     0.10731
```

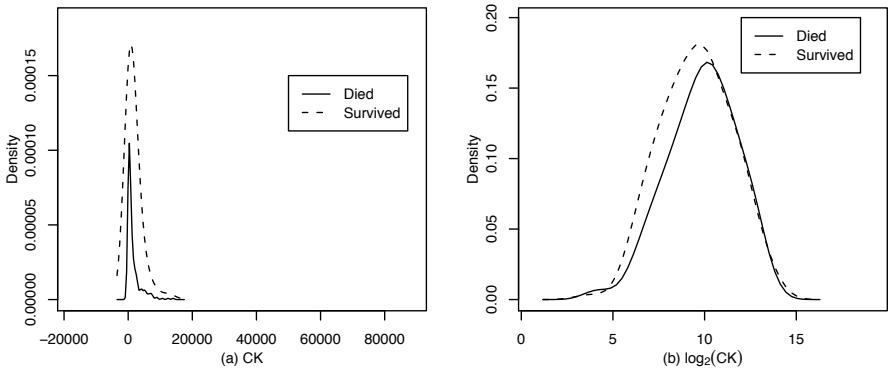
The survival odds are multiplied by 0.107 when *Myopathy* is present.

```
> predict(m1,data.frame(Myopathy=c(0,1)),type="response")
[1] 0.385827 0.063158
```

The estimated survival probabilities match the observed survival rates for the two conditions. ∎

12.1.3. Next, consider the regression problem with only *CK* as a predictor (*CK* is observed more often than is *Myopathy*, so this regression will be based on more cases than were used in the first two parts of this problem). Draw separate density estimates of *CK*, for $Outcome = 0$ and for $Outcome = 1$. Also, draw separate density estimates for $\log(CK)$ for the two groups. Comment on the graphs.

Solution:



Almost all the density for CK in both groups is concentrated in a small region, so the extreme values are virtually ignored. The graph for $\log(CK)$ seems more reasonable. The density of $Outcome = 1$ is slightly shifted to the left of the density estimate for $Outcome = 0$, suggesting that survivors tend to have lower values of $\log(CK)$. ∎

12.1.4. Fit the logistic regression mean function with $\log(CK)$ as the only term beyond the intercept. Summarize results.

Solution:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.00065    0.58089   6.887 5.69e-12 ***
log(CK, 2)  -0.42402    0.05497  -7.714 1.22e-14 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 550.49  on 412  degrees of freedom
Residual deviance: 475.18  on 411  degrees of freedom
AIC: 479.18
```

When CK *doubles*, so the base-two logarithm of CK increases by one unit, the survival odds are multiplied by $\exp(-.42402) = 0.65$. ∎

12.1.5. Fit the logistic mean function with terms for $\log(CK)$, *Myopathy* and a *Myopathy* $\times \log(CK)$ interaction. Interpret each of the coefficient estimates. Obtain a sequential deviance table for fitting the terms in the order given above, and summarize results. (Missing data can cause a problem here: if your computer program requires that you fit three separate mean functions to get the analysis of deviance, then you must be sure that each fit is based on the same set of observations, those for which CK and *Myopathy* are both observed.)

Solution:

```
Coefficients:
```

```
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.02810    1.16564   0.024   0.9808
log(CK, 2)          -0.04705    0.11001  -0.428   0.6689
Myopathy             5.31297    3.84652   1.381   0.1672
log(CK, 2):Myopathy -0.56346    0.30841  -1.827   0.0677 .
```

According to this mean function, if *CK* doubles, the odds of survival are multiplied by $\exp(-0.047) = .95$ if *Myopathy* is not present, and by $\exp(-.047 - .563) = .543$ if *Myopathy* is present. Interestingly, none of the Wald tests (labelled z above) are significant.

```
> anova(m3,test="Chisq")
Analysis of Deviance Table
                    Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                  217    246.271
log(CK, 2)           1   21.376       216    224.895 3.776e-06
Myopathy             1   12.615       215    212.280 3.826e-04
log(CK, 2):Myopathy  1    3.420       214    208.859     0.064
```

The main effect of $\log(CK)$, without adjustment, is significantly different from zero. *Myopathy* adjusted for $\log(CK)$ is also significantly different from zero. The interaction term, adjusting for both main effects, has significance level of 0.064. ∎

12.2 Starting with (12.6), prove (12.7).

Solution:

$$\begin{aligned}
\theta(\mathbf{x}_i) &= \frac{1}{1+\exp(-\boldsymbol{\beta}'\mathbf{x}_i)} \\
1-\theta(\mathbf{x}_i) &= \frac{\exp(-\boldsymbol{\beta}'\mathbf{x}_i)}{1+\exp(-\boldsymbol{\beta}'\mathbf{x}_i)} \\
\frac{\theta(\mathbf{x}_i)}{1-\theta(\mathbf{x}_i)} &= \exp(\boldsymbol{\beta}'\mathbf{x}_i) \\
\log\left(\frac{\theta(\mathbf{x}_i)}{1-\theta(\mathbf{x}_i)}\right) &= \boldsymbol{\beta}'\mathbf{x}_i
\end{aligned}$$

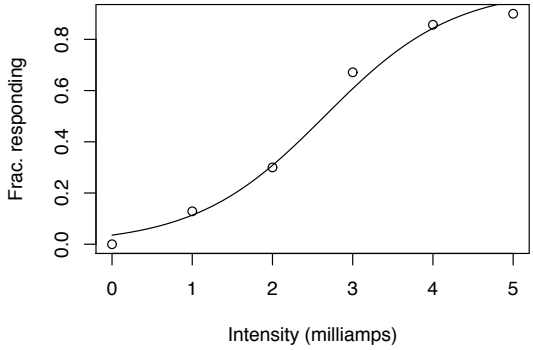
∎

12.3 Electric shocks A study carried out by R. Norell was designed to learn about the effect of small electrical currents on farm animals, with the eventual goal of understanding the effects of high-voltage power lines near farms. A total of $m = 70$ trials were carried out at each of six intensities, $0, 1, 2, 3, 4$ and 5 milliamps (shocks on the order of 15 milliamps are painful for many humans, Dalziel, 1941). The data are given in the file `shocks.txt` with columns *Intensity*, number of trials m, which is always equal to 70, and Y, the number of trials out of m for which the response, mouth movement, was observed.

Draw a plot of the fraction responding versus *Intensity*. Then, fit the logistic regression with predictor *Intensity*, and add the fitted curve to your

plot. Test the hypothesis that the probability of response is independent of *Intensity*, and summarize your conclusions. Provide a brief interpretation of the coefficient for *Intensity*. (Hint: The response in the logistic regression is the number of successes in m trials. Unless the number of trials is one for every case, computer programs will require that you specify the number of trials in some way. Some programs will have an argument with a name like "trials" or "weights" for this purpose. Others, like R and JMP, require that you specify a bivariate response consisting of the number of successes Y and the number of failures $m - Y$.)

Solution:




```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.3010     0.3238  -10.20   <2e-16 ***
Intensity     1.2459     0.1119   11.13   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

          Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                          5    250.487
Intensity  1  241.134         4      9.353 2.226e-54
```

The coefficient for *Intensity* is clearly non-zero (the Wald test and the likelihood-ratio tests both have p-values of essentially zero). A one milliamp increase in *Intensity* multiplies the odds of response by $\exp(1.2459) \approx 3.5$. The baseline odds of response when *Intensity* $= 0$ is $\exp(-3.3010) \approx 0.037$, so the response is rarely observed in the absence of electrical current. ∎

12.4 Donner party In the winter of 1846-47, about ninety wagon train emigrants in the Donner party were unable to cross the Sierra Nevada Mountains of California before winter, and almost half of them starved to death. The data in file `donner.txt` from Johnson (1996) include some information

about each of the members of the party. The variables include *Age*, the age of the person, *Sex*, whether male or female, *Status*, whether the person was a member of a family group, a hired worker for one of the family groups, or a single individual who did not appear to be a hired worker or a member of any of the larger family groups, and *Outcome*, coded one if the person survived and zero if the person died.

12.4.1. How many men and women were in the Donner Party? What was the survival rate for each sex? Obtain a test that the survival rates were the same against the alternative that they were different. What do you conclude?

Solution:

```
> attach(donner)
> print(counts <-table(Outcome,Sex))
       Sex
Outcome Female Male
      0 10     32
      1 25     24
> print(totals <- apply(counts,2,sum))
Female   Male
    35     56
> print(freqs <- counts[2,]/totals)
 Female    Male
0.71429 0.42857
> chisq.test(counts,correct=FALSE) # uncorrected Pearson's Chi-Square
        Pearson's Chi-squared test
data:  counts
X-squared = 7.0748, df = 1, p-value = 0.007817
```

There were 56 males and 35 females. The survival rate for females was about 71% and about 43% for males. We test for equality of rates using Pearson's X^2; the uncorrected test (not corrected for continuity) has p-value of about 0.008, so we reject the hypothesis that the survival rate was the same for the two sexes. ∎

12.4.2. Fit the logistic regression model with response *Outcome* and predictor *Age*, and provide an interpretation for the fitted coefficient for *Age*.

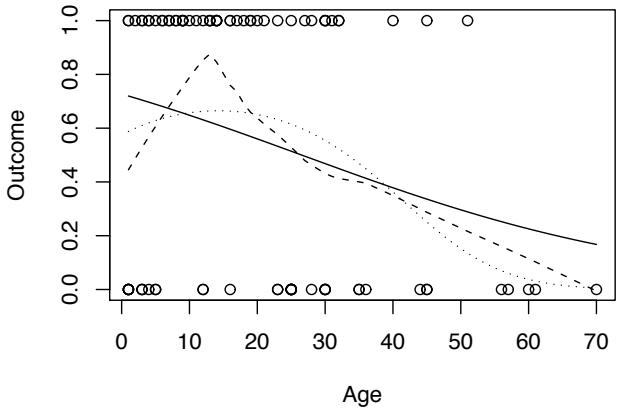
Solution:

```
> summary(m1 <- glm(Outcome ~ Age, data=donner, family=binomial()))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9792     0.3746    2.61    0.009
Age          -0.0369     0.0149   -2.47    0.013

    Null deviance: 120.86  on 87  degrees of freedom
Residual deviance: 114.02  on 86  degrees of freedom
```

The coefficient for *Age* is negative, suggesting that survival probability decreased with age. In particular, aging by one year multiplied the odds of survival by $\exp(-.0369) = .964$. ∎

12.4.3. Draw the graph of *Outcome* versus *Age*, and add both a smooth and a fitted logistic curve to the graph. The logistic regression curve apparently does not match the data: Explain what the differences are, and how this failure might be relevant to understanding who survived this tragedy. Fit again, but this time add a quadratic term in *Age*. Does the fitted curve now match the smooth more accurately?

Solution:



The solid line is for the logistic model with *Age* as the only term. The dashed line is for a lowess fit, and the dotted line is for logistic regression with both *Age* and Age^2 as terms. Survival probability seems to have been low for both the young and the old, and higher for those in the middle. The change in deviance between the linear and quadratic kernel mean functions is about 3.8 on 1 df, for a p-value of about 0.05. ∎

12.4.4. Fit the logistic regression model with terms for an intercept, *Age*, Age^2, *Sex*, and a factor for *Status*. Provide an interpretation for the parameter estimates for *Sex* and for each of the parameter estimates for *Status*. Obtain tests based on the deviance for adding each of the terms to a mean function that already includes the other terms, and summarize the results of each of the tests via a p-value and a one-sentence summary of the results.

Solution:

```
glm(formula = Outcome ~ Age + I(Age^2) + Sex + Status,
    family = binomial(),data = donner)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)   1.99e-01   6.17e-01    0.32    0.748
Age           1.67e-01   7.11e-02    2.36    0.018
I(Age^2)     -3.89e-03   1.53e-03   -2.55    0.011
SexMale      -6.64e-01   5.59e-01   -1.19    0.235
StatusHired  -1.63e+00   7.48e-01   -2.17    0.030
StatusSingle -1.85e+01   1.76e+03   -0.01    0.992

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 120.855  on 87  degrees of freedom
Residual deviance:  92.363  on 82  degrees of freedom
> drop1(m3,test="Chisq")
Single term deletions

Model:
Outcome ~ Age + I(Age^2) + Sex + Status
         Df Deviance   AIC   LRT Pr(Chi)
<none>          92.4 104.4
Age       1     99.3 109.3   6.9  0.0085
I(Age^2)  1    103.0 113.0  10.6  0.0011
Sex       1     93.8 103.8   1.4  0.2309
Status    2    103.9 111.9  11.6  0.0031
```

The survival rates for the two sexes do not appear to be different ($p = .23$). The risk factor of "Hired" (compared with "Family") is estimated to be $\exp(-1.63) = 0.19$, so hired people were about 1/5 as likely to survive. The risk for "Single" is $\exp(-18.5) \approx 0$; in fact, none of the Single people survived. ∎

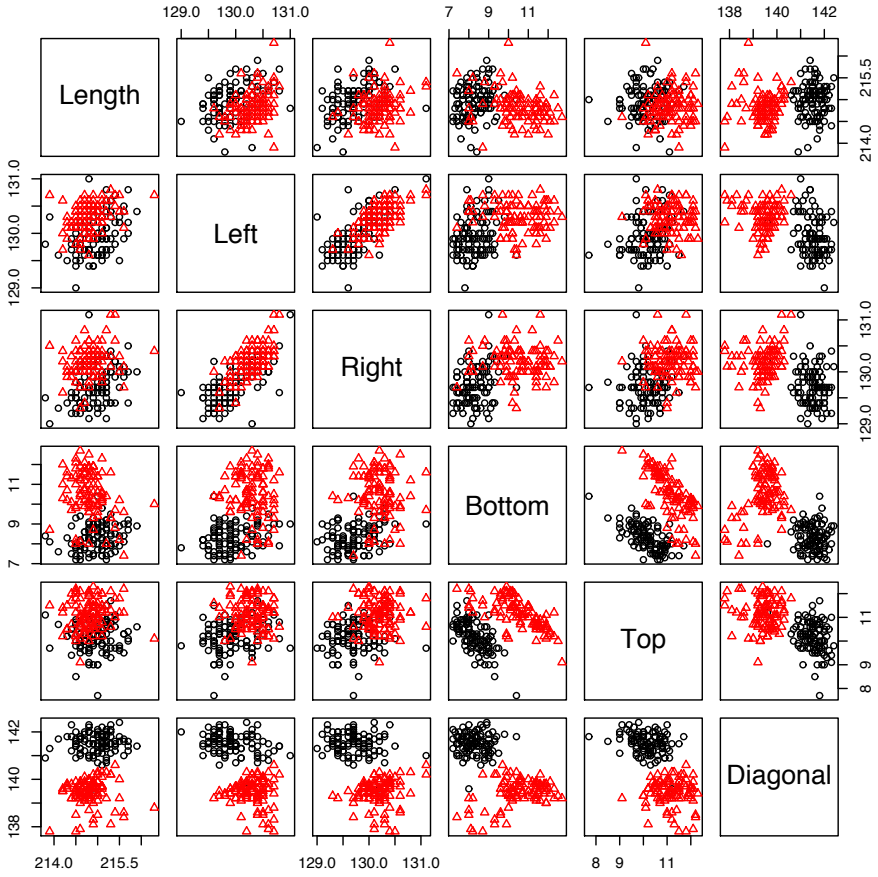
12.4.5. Assuming that the logistic regression model provides an adequate summary of the data, give a one-paragraph written summary on the survival of members of the Donner Party.

Solution: The hired and single people were the most likely to perish in the Donner party. These groups were almost all male (22 of 23 were male). Among family members, the survival rate for males and females as similar (56% versus 70%). Both the young and the old were less likely to survive. ∎

12.5 Counterfeit banknotes The data in the file `banknote.txt` contains information on one hundred counterfeit Swiss banknotes with $Y = 0$ and one hundred genuine banknotes with $Y = 1$. Also included are six physical measurements of the notes, including the *Length*, *Diagonal* and the *Left* and *Right* edges of the note, all in mm, and the distance from the image to the *Top* edge and *Bottom* edge of the paper, all in mm (Flury and Riedwyl, 1988). The goal of the analysis is estimate the probability or odds that a banknote is counterfeit given the values of the six measurements.

12.5.1. Draw a scatterplot matrix of six predictors, marking the points different colors for the two groups (genuine or counterfeit). Summarize the information in the scatterplot matrix.

Solution: This plot is much more informative in color, so you should redraw it on your computer screen.



There is little overlap between the red and black points; for example, in the plot of *Top* versus *Diagonal*, it appears that the two clouds of points are completely disjoint apart from one black point among the red ones. If these were completely disjoint, they there could be a *separating hyperplane*, meaning that we could classify points with perfect accuracy with just one linear combination of the terms. We expect that we will be able to separate genuine and counterfeit bills will almost perfect accuracy. ∎

12.5.2. Use logistic regression to study the conditional distribution of y given the predictors.

Solution: When we fit the logistic mean function with six predictors, we get the following confusing output:

```
> summary(m1 <- glm(Y ~ Length+Left+Right+Bottom+Top+Diagonal,
```

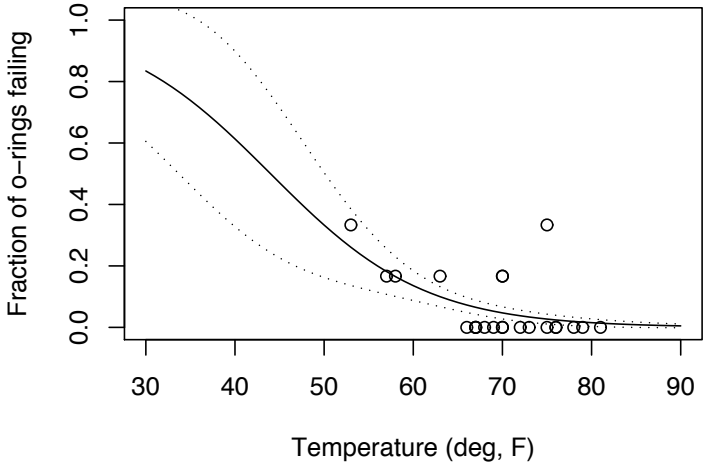
```
+           data=banknote,family=binomial()))
Deviance Residuals:
      Min         1Q     Median         3Q        Max
-8.22e-05  -2.11e-08   0.00e+00   2.11e-08   7.24e-05

Coefficients:
             Estimate Std. Error  z value Pr(>|z|)
(Intercept) -2.01e+03   3.06e+07 -6.6e-05        1
Length       3.08e+01   1.67e+05  1.9e-04        1
Left        -1.08e+01   3.25e+05 -3.3e-05        1
Right        1.06e+01   2.54e+05  4.2e-05        1
Bottom       5.88e+01   4.35e+04  1.3e-03        1
Top          4.98e+01   3.73e+04  1.3e-03        1
Diagonal    -4.04e+01   3.66e+04 -1.1e-03        1

    Null deviance: 2.7726e+02  on 199  degrees of freedom
Residual deviance: 2.0593e-08  on 193  degrees of freedom

Warning messages:
1: Algorithm did not converge in: glm.fit(x = X, y = Y,
   weights = weights, start = start, etastart = etastart,
2: fitted probabilities numerically 0 or 1 occurred in:
   glm.fit(x = X, y = Y, weights = weights, start = start...
```

The important points here are (1) *all* the (deviance) residuals are smaller than 0.0001 in absolute value; (2) the value of G^2 is zero to seven digits; (3) the program has warned of an exact fit, with fitted probabilities either zero or one. All these indicate a separating hyperplane. This will cause many programs to fail or give confusing results. To get a satisfactory fitted mean function, we used forward stepwise fitting and found that *Diagonal* and *Bottom* alone can separate genuine and counterfeit bills without error. ∎

12.6 Challenger The file `challeng.txt` from Dalal, Fowlkes, and Hoadley (1989) contains data on O-rings on 23 U. S. space shuttle missions prior to the Challenger disaster of January 20, 1986. For each of the previous missions, the temperature at take-off and the pressure of a pre-launch test were recorded, along with the number of O-rings that failed out of six.

Use these data to try to understand the probability of failure as a function of temperature, and of temperature and pressure. Use your fitted model to estimate the probability of failure of an O-ring when the temperature was 31°F, the launch temperature on January 20, 1986.

Solution:



This graph summarizes the information that the engineers should have used in deciding to launch the *Challenger*. The solid line is the fitted probabilities from the logistic regression of failures on *Temp*. The dotted lines are plus or minus one standard deviation from the fitted probabilities. Although variability in the estimates is much larger at the left of the graph, it is clear that based on the data observed there was substantial risk of failure at low temperatures. ∎

12.7 Titanic Refer to the Titanic data, described in Section 12.2.4.

12.7.1. Fit a logistic regression model with terms for factors *Sex*, *Age* and *Class*. On the basis of examination of the data in Table 12.5, explain why you expect that this mean function will be inadequate to explain these data.

Solution:

```
> summary(m1 <- glm(cbind(Surv,N-Surv)~Class+Age+Sex, data=dt,
      family=binomial()))

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.186      0.159    7.48  7.4e-14
ClassFirst     0.858      0.157    5.45  5.0e-08
ClassSecond   -0.160      0.174   -0.92     0.36
ClassThird    -0.920      0.149   -6.19  5.9e-10
AgeChild       1.062      0.244    4.35  1.4e-05
SexMale       -2.420      0.140  -17.24  < 2e-16

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 671.96  on 13  degrees of freedom
Residual deviance: 112.57  on  8  degrees of freedom
```

From Table 12.5, nearly all females survived, except in third class, where female survival was much lower. This implies a *Class* × *Sex* interaction. Other interactions might exist as well. ∎

12.7.2. Fit a logistic regression model that includes all the terms of the last part, plus all the two-factor interactions. Use appropriate testing procedures to decide if any of the two-factor interactions can be eliminated. Assuming that the mean function you have obtained matches the data well, summarize the results you have obtained by interpreting the parameters to describe different survival rates for various factor combinations. (Hint: How does the survival of the crew differ from the passengers? First class from third class? Males from females? Children versus adults? Did children in first class survive more often than children in third class?)

Solution:

```
> m2 <- update(m1,~(Class+Age+Sex)^2)
> drop1(m2,scope=~Class:Age+Class:Sex+Age:Sex, test="Chisq")
Single term deletions

Model:
cbind(Surv, N - Surv) ~ Class + Age + Sex + Class:Age + Class:Sex +
    Age:Sex
          Df Deviance   AIC   LRT Pr(Chi)
<none>        2.5e-10  70.6
Class:Age  2     37.3 103.9  37.3 8.1e-09
Class:Sex  3     65.0 129.6  65.0 5.0e-14
Age:Sex    1      1.7  70.3   1.7    0.19
> m3 <- update(m2, ~.-Age:Sex)
> drop1(m3,test="Chisq")
Single term deletions

Model:
cbind(Surv, N - Surv) ~ Class + Age + Sex + Class:Age + Class:Sex
          Df Deviance   AIC   LRT Pr(Chi)
<none>             1.7  70.3
Class:Age  2     45.9 110.5  44.2 2.5e-10
Class:Sex  3     76.9 139.5  75.2 3.3e-16
```

The *Age* × *Sex* interaction can apparently be dropped, but the other two interactions are required. Although not covered in the text, this is a model of conditional independence: given *Class*, *Age* and *Sex* are independent, meaning that within a fixed class survival does not depend on age or sex. Survival rates were highest for first class, lowest for third class. Overall, men were much less likely to survive than women. ∎

12.8 BWCAW blowdown The data file `blowAPB.txt` contains the data for Rich's blowdown data, as introduced at the beginning of this chapter, but for the two species $SPP = A$ for aspen, and $SPP = PB$ for paper birch.

12.8.1. Fit the same mean function used for balsam fir to each of these species. Is the interaction between S and *logD* required for these species?

Solution:

```
> m1 <- glm(y~logD+S+logD:S,data=trees,subset=SPP=="A",
+           family=binomial())
> m2 <- update(m1,subset=SPP=="PB")
> summary(m1)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.493      1.943   -2.83   0.0047
logD           1.402      0.623    2.25   0.0245
S              6.927      4.447    1.56   0.1193
logD:S        -0.730      1.430   -0.51   0.6097
> anova(m1,test="Chisq")
Analysis of Deviance Table
Terms added sequentially (first to last)
        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      435        531
logD     1       21       434        510   4.0e-06
S        1       85       433        425   2.4e-20
logD:S   1  2.7e-01       432        424         1
> summary(m2)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.990      1.719   -1.16     0.25
logD          -0.485      0.691   -0.70     0.48
S             -3.674      3.425   -1.07     0.28
logD:S         2.905      1.340    2.17     0.03
> anova(m2,test="Chisq")
Analysis of Deviance Table
Terms added sequentially (first to last)
        Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                      496        470
logD     1       20       495        450   7.7e-06
S        1       47       494        403   7.7e-12
logD:S   1        5       493        399   2.8e-02
```

For aspen, the interaction seems unnecessary, but it might be useful for paper birch. ∎

12.8.2. Ignoring the variable S, fit compare the two species, using the mean functions outlined in Section 6.2.2.

Solution:

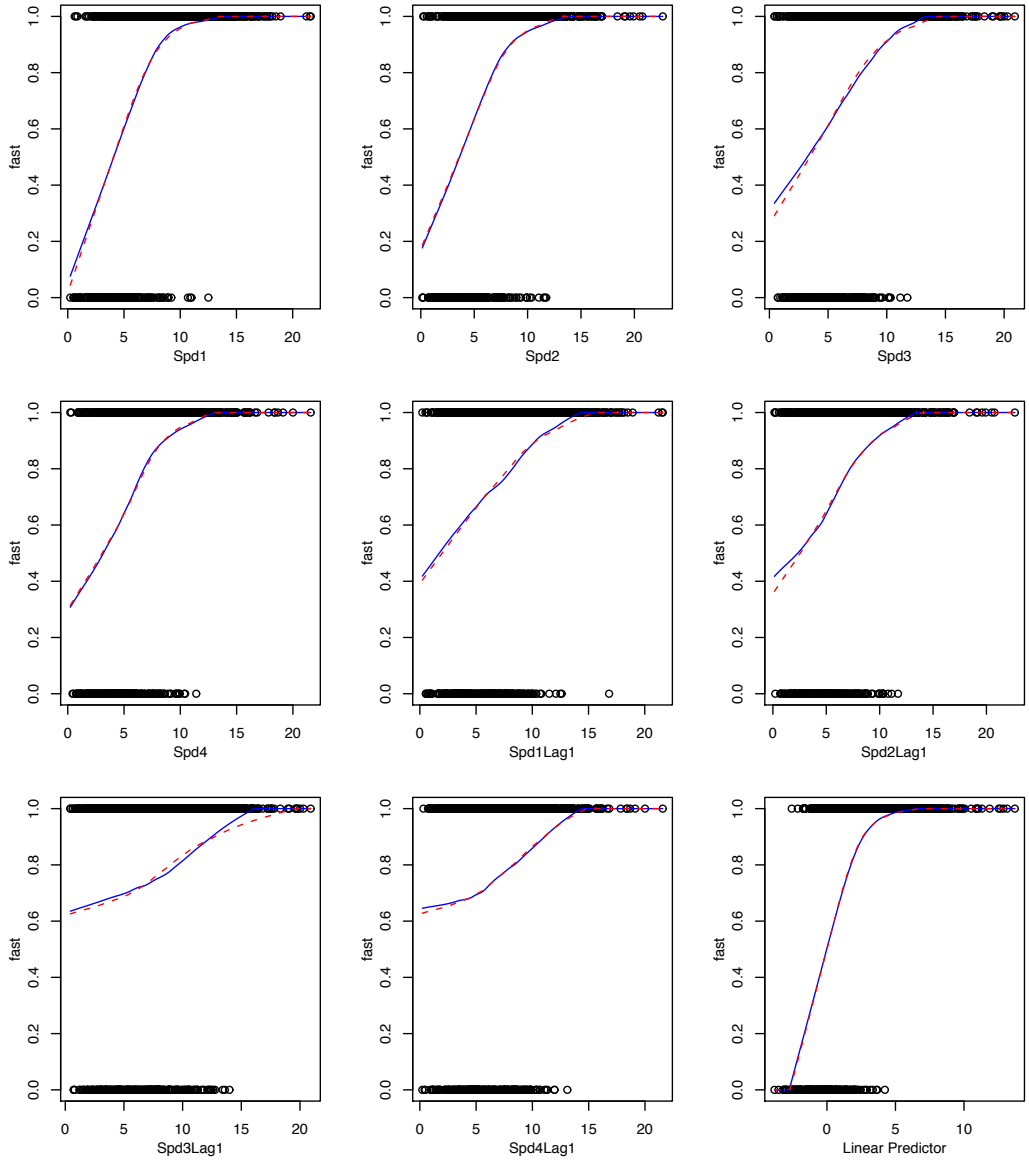
```
Analysis of Deviance Table
Model 1: y ~ logD
```

```
Model 2: y ~ SPP + logD
Model 3: y ~ SPP + logD + SPP:logD
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       931      181.5
2       930      157.8   1     23.7     3e-32
3       929      157.5   1      0.3       0.2
> anova(n4,n3,n1, test="Chisq")
Analysis of Deviance Table
Model 1: y ~ logD
Model 2: y ~ logD + SPP:logD
Model 3: y ~ SPP + logD + SPP:logD
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       931      181.5
2       930      157.6   1     23.9   1.5e-32
3       929      157.5   1      0.1       0.4
```

Either the mean function of parallel regressions (in logit scale), or the mean function of common intercept but different slopes, fit these data equally well. The common regression is not acceptable, and the most general mean function is not needed. ∎

12.9 Windmill data For the windmill data in the data file `wm4.txt`, use the four-site data to estimate the probability that the wind speed at the candidate site exceeds six meters per second, and summarize your results.

Solution: Here is one possible solution. We considered only four possible mean functions, (1) *Spd1* only; (2) all four speed variables; (3) speed plus bin information; and (4) speed, bins, and lagged speeds. Because of the very large samples, all the variables seem to be useful. The marginal model plots shown below provide excellent agreement between the data and the model



With a data set this large, there are many alternatives available in the analysis, including cross validation for variable selection. One could also estimate error rates, or otherwise summarize these data.