Notes on and Solutions to Selected Problems In: Applied Linear Regression by Sanford Weisberg

John L. Weatherwax*

December 15, 2014

*wax@alum.mit.edu

To my family.

Introduction

This is a very nice book on linear regression. It was written to serve the needs of people who want to be able to apply linear regression techniques in a practical setting. Many of the results presented are not derived mathematically in great detail but I believe that is a conscious design decision of Weisberg. Instead of presenting derivations which many students find uninteresting and a distraction that clouds the application of linear regression Weisberg has chosen a different approach. He has chosen to present (and derive in an appendix) the equations for the scalar regression case and then to simply state some of the results in more complicated cases. He then provides a great number of very practical examples *using* linear regression (both successfully and unsuccessfully) and documents cases and situations the practicing statistical should be aware of before sending out results. This is much more in line with how a practicing statistician would use linear regression. It is valuable to know how to derive results but in todays age of easy to come by statistical software it is more important to know how to *use* these results. Two interesting applications (among many) presented by Weisberg are of using linear regression to predict the amount of snow to fall latter in the season given the amount of snow that has already fallen and predicting the length of time a visitor must wait at the Old Faithful geyser's to observe the next eruption.

The use of the `R` statistical language is felt through and the book provides a very nice summary set of notes that accompany the book by explicitly describing the `R` commands used to duplicate the results presented in the textbook. This is a great way to learn the `R` language and to get started solving practical problems very quickly. The code snippets for various exercises can be found at the following location:

`http://www.waxworksmath.com/Authors/N_Z/Weisberg/weisberg.html`

I found this book very enjoyable and would recommend it highly. I feel that it is a great way to develop a deeper understanding of linear regression. I believe that with this knowledge one will find a great number of applications and extensions in your own work. Included here are a few problems I had time to write up.

I've worked hard to make these notes as good as I can, but I have no illusions that they are perfect. If you feel that that there is a better way to accomplish or explain an exercise or derivation presented in these notes; or that one or more of the explanations is unclear, incomplete, or misleading, please tell me. If you find an error of any kind – technical, grammatical, typographical, whatever – please tell me that, too. I'll gladly add to the acknowledgments in later printings the name of the first person to bring each problem to my attention.

Acknowledgments

Special thanks to Jon Eskreis-Winkler for his corrections and suggestions.

Chapter 1 (Scatterplots)

See the R functions chap_prob_1.R – chap_prob_5.R for some very simple implementations of the problems from this chapter.

Chapter 2 (Simple Linear Regression)

Notes On The Text

Here we demonstrate how to derive a few of the expression presented in this chapter as well as how to compute some of the specific numbers presented in the textbook using R commands. This hopefully will save the student time and explain some of the mystery around how some of the calculations are performed.

Notes on properties of the least squares estimates

Note that $\hat{E}(Y|X=\bar{x})$ can be computed easily by using Equation 109 to replace $\hat{\beta}_0$ in the expression for $\hat{E}(Y|X=x)$. We find

$$\hat{E}(Y|X=\bar{x}) = \hat{\beta}_0 + \hat{\beta}_1\bar{x} = (\bar{y} - \hat{\beta}_1\bar{x}) + \hat{\beta}_1\bar{x} = \bar{y}\,,$$

for the expected fitted value for the mean $\bar{x}$ and as stated in the text

Notes on confidence intervals and tests

The book defines the expression $t(\alpha/2, d)$ to represent the value that cuts off $\alpha/2 \times 100\%$ of the probability from the *upper* tail of the t-distribution with d degrees of freedom. This can be computed in R by either of

$$\texttt{qt}(1-\alpha/2, \texttt{d})\,,$$

or

$$\texttt{qt}(\alpha/2, \texttt{d}, \texttt{lower.tail} = \texttt{FALSE})\,.$$

Using these facts we can numerically verify some of the results given in the book. For example, for Forbes' data since $n = 17$ when we attempt to compute a 90% confidence interval on the estimate of $\hat{\beta}_0$ we would have $\alpha = 0.1$ and $d = n - 2 = 15$ so the needed value in the expression of the confidence interval is

$$t(0.05, 15) = \texttt{qt}(1 - 0.1/2, 15) = \texttt{qt}(0.95, 15) = 1.7530\,.$$

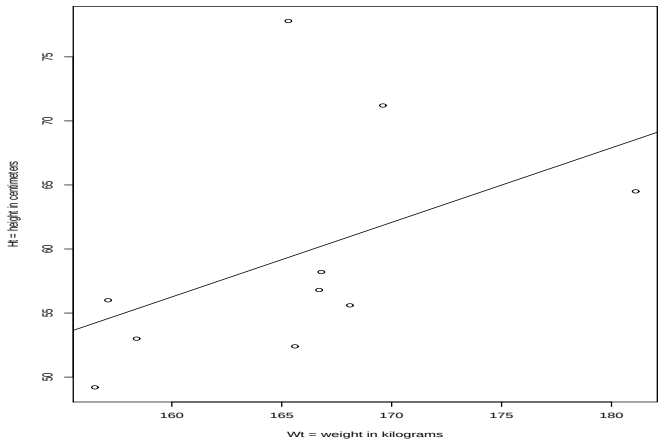

Figure 1: A scatterplot of Wt as a function of Ht for Problem 2.1.

When we compute the t-statistic for the intercept $\hat{\beta}_0$ we get the value $t = 2.137$. The p-value for this numerical quantity can be computed in `R` as

$$2 * \texttt{pt}(-2.137, 17 - 2) = 2 * 0.02473897 = 0.04947794\,.$$

which is close to the value of 0.05 as claimed in the book.

To compute a 95% confidence interval for the slope estimate of β_1 we have $\alpha = 0.05$ and $d = n - 2 = 15$ so the needed expression in the confidence interval is

$$t(0.025, 15) = \texttt{qt}(1 - 0.05/2, 15) = 2.131450 = \texttt{qt}(0.05/2, 15, \texttt{lower.tail} = \texttt{FALSE})\,.$$

In the example using the Ft. Collins data the book computes a t-statistic of $t = 1.553$ which will then have a p-value given by

$$2 * \texttt{pt}(-1.553, 93 - 2) = 0.1238942\,,$$

matching the value given in the book.

Problem Solutions

2.1 (height and weight data)

2.1.1: See the Figure 1 for a scatterplot of this data and a OLS line. A linear fit looks reasonable but there will certainly be some outliers. In addition, the data set size is very small $n = 10$ making decisions on the accuracy of our results that much harder.

2.1.2: These can be computed using the formulas given in this chapter. This is done in the `R` script `chap_2_prob_1.R`.

2.1.3: t-tests for the coefficients β_0 and β_1 are given by computing (for the null or zero hypothesis)

$$t_0 = \frac{\hat{\beta}_0}{\text{se}(\hat{\beta}_0)} \quad \text{and} \quad t_1 = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)} ,$$

when we compute these we get $t_0 = -0.572$ and $t_1 = 1.496$ respectively. These values agree with the results printed by the `R summary` command. The `R` commands to calculate the p-values associated with these to t-statistics are given by the following `R` expressions (with $n = 10$)

```
2 * pt(-abs(-0.572),n-2) # gives 0.58305

2 * pt(-abs(1.496),n-2) # gives 0.17310
```

These p-values agree with the results given by the `summary` command.

2.1.4: The `anova` table for this linear model looks like

```
Analysis of Variance Table

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

From which we see that the F statistic is given by $F = 2.237$ and equals $t_1^2 = (1.496)^2$ as it must.

See the `R` script `chap_2_prob_1.R` for the various parts of this problem.

2.2 (more with Forbes' data)

2.2.1: See the Figure 2 (left) for a scatterplot and the ordinary least squares fit of $Lpres$ as a function of $Temp$.

2.2.2: Using the `R` command `lm` we can use the `summary` command where we find a F-statistic given by 3157 which has a p-value less than $2.2\,, 10^{-16}$ which is very strong evidence against rejecting the null hypothesis.

2.2.3: See the Figure 2 (right) for the requested plot.

2.2.4: Using the Hooker data and the `R` function `lm` we find a mean function given by

$$E(Lpres|u1) = 724.9 - 215470\, u1\,.$$

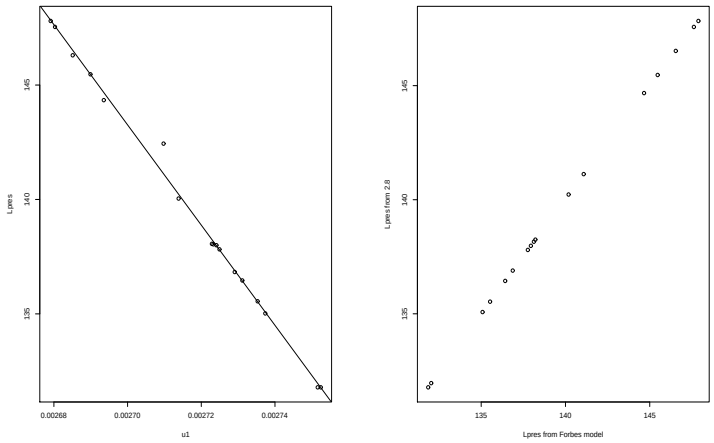

Figure 2: **Left:** A scatterplot and OLS fit of $LPres$ as a function of $u1$ requested in Problem 2.2. **Right:** A plot of $Lpres$ predicted from equation 2.8 as a function of $Lpres$ predicted from Forbes' model.

2.2.5: We will use the explicit formula provided in the text to program the standard error of prediction. This formula is given by

$$\text{sepred}(\tilde{y}_*|x_*) = \hat{\sigma}\left(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\text{SXX}}\right)^{1/2} . \tag{1}$$

When we do this and compute the z values we find the mean and standard deviation given by -0.00074 and 0.9549 respectively. These two results are rather close to the expected values of 0 and 1.

2.2.6: For this part of the problem we use the Hooker model to compute predictions on the Forebe's data set. The `R` code to do this is given by

```
u1_temp <- 1 / ( (5/9) * forbes$Temp + 255.37 )
forbes_hooker_predictions <- predict( m_hooker, newdata=data.frame(u1=u1_temp) )
sepred <- sigma_hat * sqrt( 1 + 1/n + ( u1_temp - mean(u1) )^2/ SXX )

z <- ( forbes$Lpres - forbes_hooker_predictions ) / sepred
```

where the values for `n`, `mean(u1)`, and `SXX` are based on the data used to compute the model in this case the *Hooker* data. The mean and standard deviation of the values in `z` are given by 0.19529 and 0.98453 respectively. The consequence of this is that the model we are using to predict the Forbes' data seems to be biased (has a non-zero mean).

See the `R` script `chap_2_prob_2.R` for the various parts of this problem.

2.3 (derivations from the mean)

2.3.1: The parameter α represents the value of y when x is equal to its mean value $\bar{x}$. This would be $\bar{y}$, see page 5.

2.3.2: The least squares estimates of α and $\hat{\beta}_1$ can be obtained by creating a new data set where the dependent variables X has had its mean subtracted. Then using Equation 109 to compute α we see that $\alpha = \bar{y}$. Using Equation 110 to compute $\hat{\beta}_1$ since SXX and SXY are the same whether or not the mean is subtracted from the values of X we have $\hat{\beta}_1$ given by Equation 110 as we were to show.

2.3.3: Since $\hat{\alpha}$ is the sample mean it has a variance given by

$$\mathrm{Var}(\hat{\alpha}) = \mathrm{Var}(\bar{y}) = \mathrm{Var}(\frac{1}{n}\sum y_i) = \frac{\sigma^2}{n}\,.$$

The variance of $\hat{\beta}_1$ is the same as given in Equation 111, and the covariance of $\hat{\alpha}$ and $\hat{\beta}_1$ is given by

$$\mathrm{Cov}(\hat{\alpha}, \hat{\beta}_1) = \mathrm{Cov}(\bar{y}, \hat{\beta}_1) = 0\,,$$

as given by Equation 112.

2.4 (heights of mothers and daughters)

2.4.1: Since all of the information requested is provided by the `R` command `summary` we will report that output here. After fitting a linear model on *Dheight* given *Mheight* we find the `R summary` looks like

```
Call:
lm(formula = Dheight ~ Mheight)

Residuals:
     Min       1Q   Median       3Q      Max
-7.39740 -1.52866  0.03600  1.49211  9.05250

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.91744    1.62247   18.44   <2e-16 ***
Mheight      0.54175    0.02596   20.87   <2e-16 ***
---

Residual standard error: 2.266 on 1373 degrees of freedom
Multiple R-Squared: 0.2408,     Adjusted R-squared: 0.2402
F-statistic: 435.5 on 1 and 1373 DF,  p-value: < 2.2e-16
```

From that output we see have the estimates of coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$, their standard errors, the value of the coefficient of determination $R^2 = 0.2408$ and the variance estimate 2.266. The hypothesis test that $E(Dheight|Mheight) = \beta_0$ versus the alternative that $E(Dheight|Mheight) = \beta_0 + \beta_1 Mheight$ is summarized by the F-statistic which in this case is given by 435.5 and has a p-value less than $2.2\,10^{-16}$ or strong evidence against the hypothesis that $E(Dheight|Mheight) = \beta_0$ in favor of the other hypothesis.

2.4.2: If y_i is written in mean differenced form given by

$$y_i = \alpha + \beta_1(x_i - \bar{x}) + e_i = E(y) + \beta_1(x_i - \bar{x}) + e_i\,,$$

then as the variable x_i is the mother's height if $\beta_1 = 1$ then any amount by which this given mothers height is greater than the average will directly translate numerically one-to-one into an amount by which the daughters height is greater than the average $E(y)$. This later expression is the *populations* average height. If β_1 was greater than one this would imply that taller mothers (one's whos heights were greater than the average) would product children whos heights were greater than the average also. Mothers who's heights were less that then average $x_i < \bar{x}$ would on average produce daughters who's height is less than the population average i.e. $y_i < E(y)$. Taken together this would seem to indicate that taller mothers will produce taller daughters while shorter mothers would produce shorter daughters. If $\beta_1 < 1$ then the opposite behavior would be observed in that taller mothers would produce children who while still taller than the average would not be proportionally taller. In the case when $\beta_1 < 1$ the daughters heights seem to be "regressing to the mean $E(y)$". It is observations like this that lead to the term "regression".

Recall that the standard error of β_1 is given by

$$se(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\text{SXX}}}\,, \tag{2}$$

with which the $(1-\alpha)\times 100\%$ confidence interval for β_1 is given by

$$\hat{\beta}_1 - t(\alpha/2, n-2)se(\hat{\beta}_1) \le \beta_1 \le \hat{\beta}_1 + t(\alpha/2, n-2)se(\hat{\beta}_1)\,.$$

which in this case becomes

$$0.4747836 < \beta_1 < 0.6087104\,,$$

both limits of which are considerably less than one.

2.4.3: We can easily do this with the `predict` command. We find the `R` prediction given by

```
          fit      lwr      upr
[1,] 64.58925 58.74045 70.43805
```

See the `R` script `chap_2_prob_4.R` for the various parts of this problem.

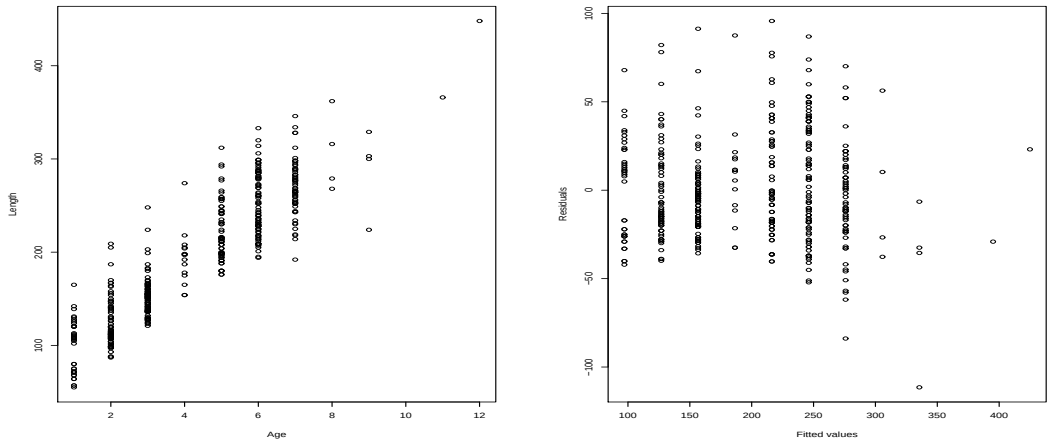

Figure 3: **Left:** The `wblake2` data set with a OLS line. **Right:** The resulting residuals for the `wblake2` data set. Note the appearance of "curvature" in this plot indicating a linear model may not be appropriate.

2.5 (smallmouth bass)

2.5.1: We can load the `wblake` data set and obtain a confidence interval around the mean length at the various ages with the following `R` command

```
predict(m,data.frame(Age=c(2,4,6)), interval="confidence", level=0.95)
       fit      lwr      upr
1 126.1749 122.1643 130.1856
2 186.8227 184.1217 189.5237
3 247.4705 243.8481 251.0929
```

2.5.2: Using the same call as above we find that the 95% confidence interval for the mean length at age 9 is given by.

```
          fit      lwr      upr
[1,] 338.4422 331.4231 345.4612
```

The age value of 9 is outside of the range of the data we have samples for and therefore will may have extrapolation errors.

2.5.3: To show that a linear model is *not* appropriate we will consider a scatterplot of the residuals $y_i - \hat{y}_i$ considered as a function of the fitted values $\hat{y}_i$. Under an appropriate model this scatterplot should be a null-plot. In Figure 3 (left) we show the raw data provided in the `wblake2` data. In Figure 3 (right) we show a plot of the residuals versus the fitted values. As stated in the text the observed curvature may indicate that the assumed mean function

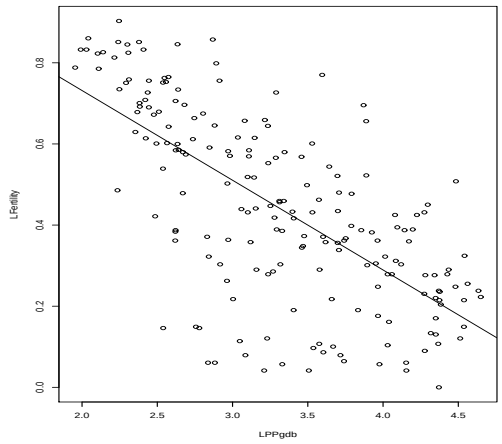

Figure 4: The `UN` data set with a OLS line. The x-axis is the base-10 logarithm of the per person gross domestic product $PPgdp$. The y-axis is the base-10 logarithm of the birth rate per 1000 females in the year 2000.

is inappropriate. This curvature might indicate that a *quadratic* term should be added to improve the regression results.

See the `R` script `chap_2_prob_5.R` for the various parts of this problem.

2.6 (the united nations data)

2.6.1: We use the `anova` command in `R` to find the following analysis of variance table

```
Analysis of Variance Table

Response: LFertility
           Df Sum Sq Mean Sq F value    Pr(>F)
LPPgdb      1 4.8004  4.8004  162.15 < 2.2e-16 ***
Residuals 191 5.6546  0.0296
```

This gives strong evidence against the fact that this data is generated with uniform variance.

2.6.2: See Figure 4 for the plot of the data and the OLS fitted line for the `UN` data set.

2.6.3: Our alternative (AH) and null hypothesis (NH) for the requested test on the value of β_1 would be given by

$$\begin{aligned} \text{NH}: \quad \beta_1 &= 0 \\ \text{AH}: \quad \beta_1 &< 0\,. \end{aligned}$$

Under the null hypothesis the t-statistic

$$t = \frac{\hat{\beta}_1 - 0}{\text{se}(\hat{\beta}_0)} = \frac{-0.22116}{0.01737} = -12.73230\,.$$

The t-statistic above should be distributed as a t-distribution with $n-2$ degrees of freedom. For the `UN` data set we have $n = 193$ samples so the t-distribution will have 191 degrees of freedom. To compute the *one-sided* p-value defined as

$$p = \Pr\{T < t|\text{NH}\} = \Pr\{T < -12.73230|\text{NH}\}\ .$$

For this t-statistic we can use the following `R` command

$$\texttt{pt}(-12.73230, 191) = 1.378382\,10^{-27}\,,$$

providing a very strong argument against the null hypothesis.

We cannot use the F-statistic to test the NH against the one sided hypothesis that $\beta_1 < 0$ since the F-statistic is testing the significance of the model with a *non-zero value* for β_1 against the NH that β_1 is in fact zero. Under the null hypothesis the F-statistic defined by

$$F = \frac{(\text{SYY} - \text{RSS})/1}{\hat{\sigma}^2}\,,$$

will be drawn from an $F(1, n-2)$ distribution and tests the NH that $E(Y|X = x) = \beta_0$ against the AH that $E(Y|X = x) = \beta_0 + \beta_1 x$. When we compute this value using the above formula we get a value of $F = 162.1460$. Thus we want to compute the p-value for this statistic

$$\begin{aligned} p &= \Pr\{f \geq F = 162.1460|\text{NH}\} = 1 - \Pr\{f < F = 162.1460|\text{NH}\} \\ &= 1 - \texttt{pf}(F, 1, n-2) = 1 - \texttt{pf}(162.1460, 1, 191) = 0.0\,. \end{aligned}$$

Note also that $t^2 = 162.1114 = F$ as it should.

2.6.4: Using the `R summary` command we find the coefficient of determinism $R^2 = 0.4591$ explains that almost 46% of the variance in the response $\log_{10}(Fertility)$ is explained by this regression.

2.6.6: We can use the `R` command `predict` to compute prediction intervals. Computing a 95% prediction interval of $\log_{10}(Fertility)$ gives the interval $(0.17, 0.51, 0.85)$. Using these values as exponents we find a 95% condidence interval and point prediction for $Fertility$ given by $(1.47, 3.23, 7.094)$.

2.6.7: Looking for the suggested local we find Niger has the largest fertility and Hong Kong has the smallest fertility.

2.7 (regression through the origin)

2.7.1: The least square estimator for β_1 is obtained by finding the value of $\hat{\beta}_1$ such that $\text{RSS}(\beta_1)$ is minimized. Taking the derivative of the given expression for $\text{RSS}(\hat{\beta}_1)$ with respect

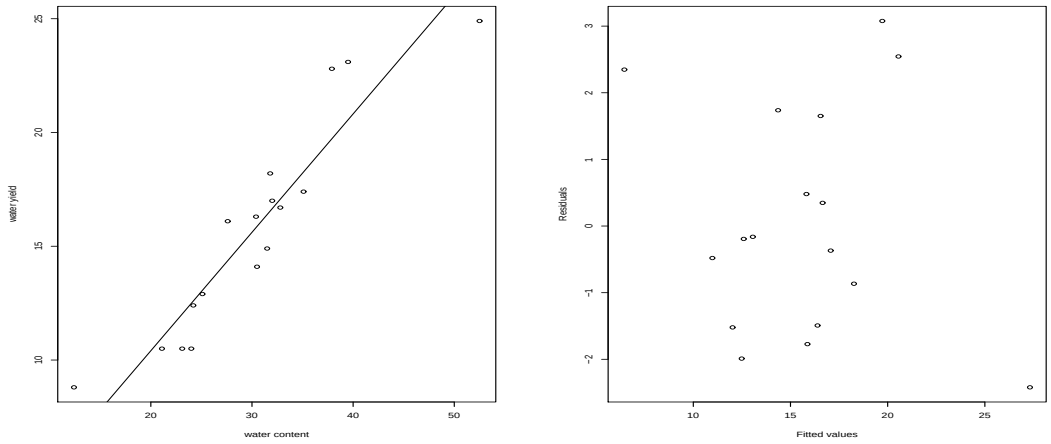

Figure 5: **Left:** The `snake` data set with a OLS line. The x-axis is the water content of snow on April 1st and the y-axis is the water yield from April to July (in inches). **Right:** A plot of the residuals $\hat{e}$ as a function of the fitted values $\hat{y}$ for the `snake` data set.

to $\hat{\beta}_1$ and setting the resulting expression equal to zero we find

$$\frac{d}{d\hat{\beta}_1}\text{RSS}(\hat{\beta}_1) = 2\sum(y_i - \hat{\beta}_1 x_i)(-x_i) = 0\,,$$

or

$$-\sum y_i x_i + \hat{\beta}_1 \sum x_i^2 = 0\,.$$

Solving this expression for $\hat{\beta}_1$ we find

$$\hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2}\,. \tag{3}$$

To study the bias introduced by this estimator of β_1 we compute

$$E(\hat{\beta}_1) \;=\; \frac{\sum x_i E(y_i)}{\sum x_i^2} = \beta_1 \frac{\sum x_i^2}{\sum x_i^2} = \beta_1\,,$$

showing that this estimator is unbiased. To study the variance of this estimator we compute

$$\begin{aligned}\text{Var}(\hat{\beta}_1) &= \frac{1}{\left(\sum x_i^2\right)^2}\sum_i \text{Var}(x_i y_i) = \frac{1}{\left(\sum x_i^2\right)^2}\sum_i x_i^2 \text{Var}(y_i) \\ &= \frac{\sigma^2}{\left(\sum x_i^2\right)^2}\sum_i x_i^2 = \frac{\sigma^2}{\sum_i x_i^2}\,,\end{aligned} \tag{4}$$

the requested expression. An estimate of $\hat{\sigma}$ is given by the usual

$$\hat{\sigma}^2 = \frac{\text{RSS}}{n-1}\,,$$

which has $n-1$ degrees of freedom.

2.7.3: We can use the `R` command `lm` to fit a regression model without an intercept by using the command `m <- lm( Y ~ X - 1 )`. A plot of the data and the OLS fit is shown in Figure 5 (left). The `summary` command then gives estimates of the coefficient $\hat{\beta}_1$ and $\hat{\sigma}^2$. We find that $\hat{\beta}_1 = 0.52039$ and $\hat{\sigma}^2 = 1.7$. The 95% confidence interval for β_1 is given by

$$\hat{\beta}_1 - t(\alpha/2, n-1)se(\hat{\beta}_1) \leq \beta_1 \leq \hat{\beta}_1 + t(\alpha/2, n-1)se(\hat{\beta}_1)\,.$$

Note the $n-1$ argument to the quantile of the t distribution that comes because we have only one free parameter (here the slope value β_1). In this case evaluating these we find

$$0.492451 < \beta_1 < 00.548337\,.$$

The `summary` command also produces the t-statistic for this value of $\hat{\beta}_1$ and is given by 39.48. The p-value for such a t-statistic is smaller than machine precision giving very strong evidence against rejecting the null hypothesis (that $\beta_1 = 0$).

2.7.4: When we plot the residuals $\hat{e}$ as a function of the fitted values $\hat{y}$, we obtain the plot given in Figure 5 (right). This looks reasonably like a null-plot. We also compute that $\sum \hat{e}_i^2 = 0.918 \neq 0$ as expected since we don't include a intercept in this regression.

2.8 (scale invariance)

2.8.1: Define "primed" variables as the ones that are "scaled" relative to the original variables. That is $x' = cx$. Then from the definitions given in the text we see that

$$\begin{aligned}
\bar{x}' &= c\,\bar{x} \\
\mathrm{SXX}' &= c^2\,\mathrm{SXX} \\
\mathrm{SD}_{x'} &= c\,\mathrm{SD}_x \\
\mathrm{SXY}' &= c\,\mathrm{SXY} \\
s_{x'y} &= c\,s_{xy}\,.
\end{aligned}$$

Thus the standard OLS estimate of $\hat{\beta}_0'$ and $\hat{\beta}_1'$ become

$$\begin{aligned}
\hat{\beta}_1' &= \frac{\mathrm{SXY}'}{\mathrm{SXX}'} = \frac{c\,\mathrm{SXY}}{c^2\,\mathrm{SXX}} = \frac{1}{c}\frac{\mathrm{SXY}}{\mathrm{SXX}} = \frac{1}{c}\hat{\beta}_1 \\
\hat{\beta}_0' &= \bar{y}' - \hat{\beta}_1'\bar{x}' = \bar{y} - \frac{1}{c}\hat{\beta}_1(c\bar{x}) = \hat{\beta}_0\,.
\end{aligned}$$

To determine how $\hat{\sigma}$ transforms recall that it is equal to $\frac{\mathrm{RSS}}{n-2}$ so we need to determine how RSS transforms. This in tern is given by $\sum_i \hat{e}_i^2$ and so we need to determine how $\hat{e}_i$ transforms. We find

$$\begin{aligned}
\hat{e}_i' &= y_i - \hat{y}_i' \\
&= y_i - (\hat{\beta}_0' + \hat{\beta}_1' x_i') \\
&= y_i - (\hat{\beta}_0 + \left(\frac{1}{c}\hat{\beta}_1\right) cx_i) \\
&= y_i - \hat{y}_i = \hat{e}_i\,.
\end{aligned}$$

Thus $\text{RSS}' = \text{RSS}$, and $\hat{\sigma}$, is unchanged by this transformation. Finally, recalling that $R^2 = 1 - \frac{\text{RSS}}{\text{SYY}}$ we have that $R'^2 = R^2$.

The t-test for statistical significance of the estimates of the individual components β_0 and β_1 with the others *in* the model is based on computing t-statistics. For example for β_0 we would compute

$$t = \frac{\hat{\beta}_i}{\text{se}(\hat{\beta}_i)} \quad \text{for} \quad i = 0, 1\,,$$

and comparing this value to the quantiles of the t-distribution with $n-2$ degrees of freedom. To answer how these transform when $x' = cx$ we need to determine how the standard error of $\hat{\beta}_i$ transforms under this mapping. We find

$$\begin{aligned}
\text{se}(\hat{\beta}'_0) &= \hat{\sigma}' \left(\frac{1}{n} + \frac{\bar{x'}^2}{\text{SXX}'} \right)^{1/2} = \hat{\sigma} \left(\frac{1}{n} + \frac{c^2 \bar{x}^2}{c^2 \text{SXX}} \right)^{1/2} = \text{se}(\hat{\beta}_0) \\
\text{se}(\hat{\beta}'_1) &= \frac{\hat{\sigma}'}{\sqrt{\text{SXX}'}} = \frac{\hat{\sigma}}{c\sqrt{\text{SXX}}} = \frac{1}{c}\text{se}(\hat{\beta}_1)\,.
\end{aligned}$$

Thus the t-statistic transform as

$$\begin{aligned}
t_{\beta'_0} &= \frac{\hat{\beta}'_0}{\text{se}(\hat{\beta}'_0)} = \frac{\hat{\beta}_0}{\text{se}(\hat{\beta}_0)} = t_{\beta_0} \\
t_{\beta'_1} &= \frac{\hat{\beta}'_1}{\text{se}(\hat{\beta}'_1)} = \frac{(1/c)\hat{\beta}_1}{(1/c)\text{se}(\hat{\beta}_1)} = t_{\beta_1}
\end{aligned}$$

that is they don't change. This is to be expected since by just scaling the variable differently should not affect the significance of their values.

2.8.2: In this case we see that

$$\begin{aligned}
\bar{y}' &= d\,\bar{y} \\
\text{SXX}' &= \text{SXX} \\
\text{SYY}' &= d^2\,\text{SYY} \\
\text{SD}_{y'} &= d\,\text{SD}_y \\
\text{SXY}' &= d\,\text{SXY}\,.
\end{aligned}$$

Thus the standard OLS estimate of $\hat{\beta}'_0$ and $\hat{\beta}'_1$ become

$$\begin{aligned}
\hat{\beta}'_1 &= \frac{\text{SXY}'}{\text{SXX}'} = \frac{d\,\text{SXY}}{\text{SXX}} = d\hat{\beta}_1 \\
\hat{\beta}'_0 &= \bar{y}' - \hat{\beta}'_1 \bar{x}' = d\bar{y} - d\hat{\beta}_1 \bar{x} = d\hat{\beta}_0\,.
\end{aligned}$$

To determine how $\hat{\sigma}$ transforms recall that it is equal to $\frac{\text{RSS}}{n-2}$ so we need to determine how RSS transforms. This in tern is given by $\sum_i \hat{e}_i^2$ and so we need to determine how $\hat{e}_i$ transforms. We find

$$\begin{aligned}
\hat{e}'_i &= y'_i - \hat{y}'_i \\
&= dy_i - (\hat{\beta}'_0 + \hat{\beta}'_1 x'_i) \\
&= dy_i - d(\hat{\beta}_0 + \hat{\beta}_1 x_i) \\
&= dy_i - d\hat{y}_i = d\hat{e}_i\,.
\end{aligned}$$

Thus $\mathrm{RSS}' = d^2\mathrm{RSS}$, and $\hat{\sigma}' = d\hat{\sigma}$. Finally, recalling that $R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{SYY}}$ we have that

$$R'^2 = 1 - \frac{\mathrm{RSS}'}{\mathrm{SYY}'} = 1 - \frac{d^2\mathrm{RSS}}{d^2\mathrm{SYY}} = R^2 .$$

To answer how the t-statics transform when $y' = dy$ we need to determine how the standard error of $\hat{\beta}_i$ transforms under this mapping. We find

$$\begin{aligned} \mathrm{se}(\hat{\beta}_0') &= \hat{\sigma}' \left(\frac{1}{n} + \frac{\bar{x'}^2}{\mathrm{SXX}'} \right)^{1/2} = d\hat{\sigma} \left(\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SXX}} \right)^{1/2} = d\,\mathrm{se}(\hat{\beta}_0) \\ \mathrm{se}(\hat{\beta}_1') &= \frac{\hat{\sigma}'}{\sqrt{\mathrm{SXX}'}} = \frac{d\hat{\sigma}}{\sqrt{\mathrm{SXX}}} = d\,\mathrm{se}(\hat{\beta}_1) . \end{aligned}$$

Thus the t-statistic transform as

$$\begin{aligned} t_{\beta_0'} &= \frac{\hat{\beta}_0'}{\mathrm{se}(\hat{\beta}_0')} = \frac{d\,\hat{\beta}_0}{d\,\mathrm{se}(\hat{\beta}_0)} = t_{\beta_0} \\ t_{\beta_1'} &= \frac{\hat{\beta}_1'}{\mathrm{se}(\hat{\beta}_1')} = \frac{d\,\hat{\beta}_1}{d\,\mathrm{se}(\hat{\beta}_1)} = t_{\beta_1} \end{aligned}$$

that is they don't change. Again this is to be expected.

2.9 (verifying an expression for RSS)

We want to show that

$$\mathrm{RSS} = \mathrm{SYY} - \frac{\mathrm{SXY}^2}{\mathrm{SXX}} = \mathrm{SYY} - \hat{\beta}_1^2\mathrm{SXX} . \tag{5}$$

To do this consider the definition of RSS

$$\begin{aligned} \mathrm{RSS} &= \sum \hat{e}_i^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \sum (y_i^2 - 2y_i(\hat{\beta}_0 + \hat{\beta}_1 x_i) + (\hat{\beta}_0 + \hat{\beta}_1 x_i)^2) \\ &= \sum y_i^2 - 2\hat{\beta}_0 \sum y_i - 2\hat{\beta}_1 \sum x_i y_i + \sum (\hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 x_i + \hat{\beta}_1^2 x_i^2) \\ &= \sum y_i - 2\hat{\beta}_0 n\bar{y} - 2\hat{\beta}_1 \sum x_i y_i + n\hat{\beta}_0^2 + 2\hat{\beta}_0\hat{\beta}_1 n\bar{x} + \hat{\beta}_1^2 \sum x_i^2 . \end{aligned}$$

Now using Equations 104, 105, and 106 we have RSS given by

$$\begin{aligned} \mathrm{RSS} &= \mathrm{SYY} + n\bar{y}^2 - 2\hat{\beta}_0 n\bar{y} - 2\hat{\beta}_1(\mathrm{SXY} + n\bar{x}\bar{y}) \\ &+ n\hat{\beta}_0 2\hat{\beta}_0\hat{\beta}_1 n\bar{x} + \hat{\beta}_1^2(\mathrm{SXX} + n\bar{x}^2) . \end{aligned}$$

When we use Equation 109 the above becomes

$$\begin{aligned} \mathrm{RSS} &= \mathrm{SYY} + n\bar{y}^2 - 2n\bar{y}(\bar{y} - \hat{\beta}_1\bar{x}) - 2\hat{\beta}_1(\mathrm{SXY} + n\bar{x}\bar{y}) \\ &+ n(\bar{y} - \hat{\beta}_1\bar{x})^2 + 2\hat{\beta}_1(\bar{y} - \hat{\beta}_1\bar{x})n\bar{x} + \hat{\beta}_1^2(\mathrm{SXX} + n\bar{x}^2) \\ &= \mathrm{SYY} - n\bar{y}^2 + 2n\bar{x}\bar{y}\hat{\beta}_1 - 2\hat{\beta}_1\mathrm{SXY} - 2\hat{\beta}_1 n\bar{x}\bar{y} \\ &+ n\bar{y}^2 - 2n\bar{y}\bar{x}\hat{\beta}_1 + n\hat{\beta}_1^2\bar{x}^2 + 2\hat{\beta}_1 n\bar{x}\bar{y} - 2\hat{\beta}_1\bar{x}^2 n + \hat{\beta}_1^2\mathrm{SXX} + \hat{\beta}_1^2 n\bar{x}^2 \\ &= \mathrm{SYY} + 2\hat{\beta}_1\mathrm{SXY} + \mathrm{SXX}\hat{\beta}_1^2 . \end{aligned}$$

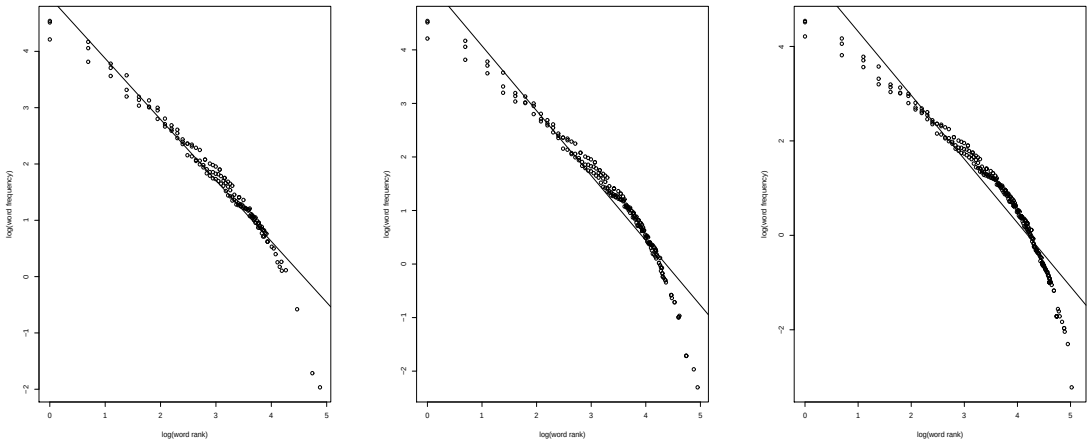

Figure 6: **Left:** A scatterplot (and ordinary least squares best fit line) of $\log(f_i)$ (log word frequency) as a function a function of $\log(i)$ (log word rank), when selecting the top 50 words according to Hamilton. **Center:** The same but now selecting the top 75 words according to the writings of Hamilton. **Right:** The same but now selecting the top 100 words. Note that as we take more words the linear fit of $\log(f)$ to $\log(i)$ becomes increasingly worse.

Using Equation 110 twice we can write the expression for RSS as

$$\begin{aligned} \text{RSS} &= \text{SYY} - 2\frac{\text{SXY}^2}{\text{SXX}} + \frac{\text{SXY}^2}{\text{SXX}} = \text{SXY} - \frac{\text{SXY}^2}{\text{SXX}} \\ &= \text{SXY} - \left(\frac{\text{SXY}^2}{\text{SXX}}\right)\text{SXX} \qquad (6) \\ &= \text{SXY} - \hat{\beta}_1 \text{SXX}\,, \qquad (7) \end{aligned}$$

the desired expressions.

2.10 (Zipf's law)

2.10.1: See the plot in Figure 6 (left) where we have used the words such that $HamiltonRank \leq 50$ and which results in the following `R` summary for the linear model

```
Call:
lm(formula = txt_freq ~ txt_ranks)

Residuals:
     Min       1Q   Median       3Q      Max
-1.63962 -0.03072  0.04052  0.10207  0.26603

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

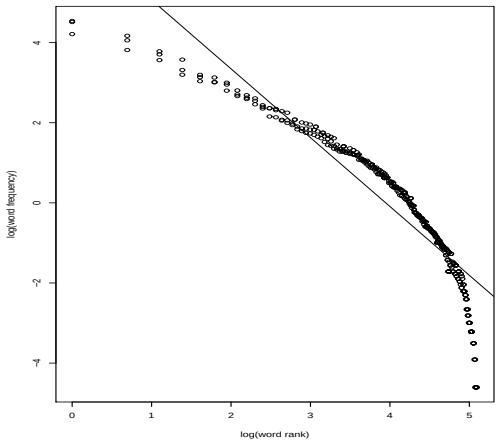

Figure 7: A scatterplot (and least squares best fit line) of $\log(f_i)$ (log word frequency) as a function a function of $\log(i)$ (log word rank), when selecting *all* words in the corpus `MWwords`.

```
(Intercept)  4.95697    0.06659   74.44   <2e-16 ***
txt_ranks    1.08290    0.02110   51.32   <2e-16 ***
---

Residual standard error: 0.2407 on 148 degrees of freedom
Multiple R-Squared: 0.9468,     Adjusted R-squared: 0.9464
F-statistic:  2634 on 1 and 148 DF,  p-value: < 2.2e-16
```

It certainly appears from this summary that the value $b = 1$ is a reasonable number.

2.10.2: To test the two sided hypothesis that $b = 1$, we will consider the t-statistic

$$t = \frac{\hat{b} - 1}{\text{se}(\hat{b})} = \frac{1.08289839 - 1}{0.02110104} = 0.07655\,,$$

which is to be compared to the quantiles of a t-distribution with $n-2 = 150-2 = 148$ degrees of freedom. We find the probability of interest given by `2 * pt(-0.07655,148) = 0.93`. Which in words states that the probability of obtaining this result purely *by chance* is around 94%. Thus this data gives almost no evidence supporting the rejection of the null hypothesis that $b = 1$. This is somewhat like saying that we should accept the null hypothesis.

2.10.3: See the plots in Figure 6 (center) and (right) where we can graphically see that when we take more and more words the linear fit performs more and more poorly. If we take this experiment to its logical extreme by including *all* terms we get the plot shown in Figure 7. There we can see that the linear fit is particularly poor.

See the `R` script `chap_2_prob_10.R` for the various parts of this problem.

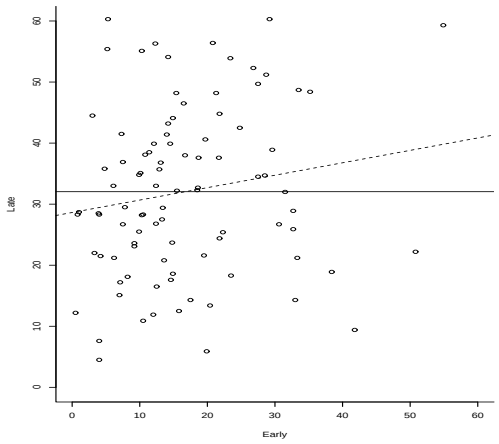

Figure 8: A scatterplot (and mean value and ordinary least squares best fit line) of the amount of snowfall observed late in the season as a function of the amount of snowfall observed early in the season. This plot duplicates Figure 1.6 in the book.

2.11 (snow fall at Ft. Collins)

In Figure 8 we plot the Ft. Collins snow fall data set. Given this scatterplot looks so much like a null plot we desire to study the statistical significance of the least squared computed slope coefficient β_1. We can test for the significance of any of the terms β_0 or β_1 by using t-statistics. When we run the `R` command `summary` on the linear model we get a t-statistic for this parameter of 1.553 which equates to a p-value of 0.124. Thus in greater than 12% of the time we could get an estimated value of this slope this large or larger *by chance* when the true slope is in fact zero. With this value we find the t-value squared given by $1.553^2 = 2.4118$ which exactly equals the F-statistic of 2.411 presented by the `summary` command.

See the `R` script `chap_2_prob_11.R` for the various parts of this problem.

2.12 (old faithful)

Note to solve this problem, we need to recall the definition (and differences between) two terms: **A Confidence Interval** and **A Prediction Interval**.

- A **Confidence Interval**, estimates *population* parameters and is used to report ranges one can be certain these parameters will fall.
- A **Prediction Interval**, estimates the *future* value of a dependent variable based on a single instance of the independent variables. This interval incorporates two sources of errors: the natural spread present in our independent variable and errors in our estimates of model parameter.

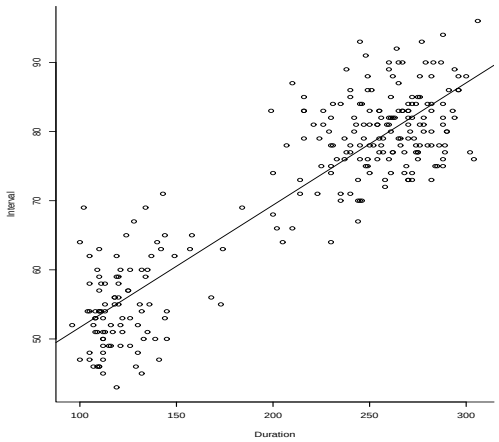

Figure 9: A scatterplot and least squares best fit line of the amount of time (in minutes) until the *next* eruption at "old faithful" in Yellowstone national park, as a function of the duration (in seconds) of the *current* eruption.

2.12.1: See Figure 9 where we plot a the Old Faithful data and the OLS linear fit, computed using the `R` command `lm`. That function also provides the following approximate pointwise prediction equation relating *duration* to *interval*

$$E(interval|duration) = 33.9878 + 0.1769\, duration\,.$$

2.12.2: This part of the problem is asking to compute a 95% *confidence* interval for the true mean of the dependent variable *interval* given the independent variable *duration*. Note that the part **2.12.3** aims at computing the corresponding *prediction* interval. Computing a confidence interval can be done easily with the `R` command `predict` and the `interval` option `confidence`, as follows

```
predict( m, newdata=data.frame(Duration=250), interval="confidence", level=0.95 )
          fit      lwr      upr
[1,] 78.20354 77.36915 79.03794
```

2.12.3: This can be done easily with the `R` command `predict` and the `interval` option `prediction`. We find

```
> predict( m, newdata=data.frame(Duration=250), interval="prediction", level=0.95 )
          fit      lwr      upr
[1,] 78.20354 66.35401 90.05307
```

This states that the next eruption will "most likely" happen sometime *after* 78 minutes have pasted and sometime *before* 90 minutes have passed.

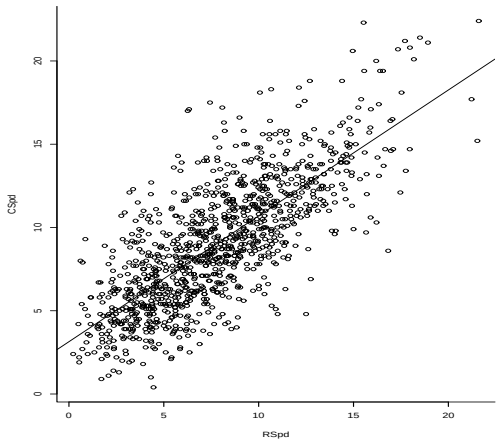

Figure 10: A scatterplot and an ordinary least squares best fit line of the response *CSpd* versus the predictor *RSpd*.

2.12.4: We are told to assume that the conditional distribution of $interval|duration = 250$ will be normal with a mean given by 77.36 (extracted from the `confidence` call in Part **2.12.2** above) and a standard error given by a variance given by

$$\text{sefit}(\hat{y}|x) = \hat{\sigma}\left(\frac{1}{n} + \frac{x - \bar{x}}{\text{SXX}}\right)^{1/2} . \tag{8}$$

Using this we recall that the confidence region for the mean function is the set of values of y such that

$$(\hat{\beta}_0 + \hat{\beta}_1 x) - [2F(\alpha, 2, n-2)]^{1/2}\,\text{sefit}(\hat{y}|x) \le y \le (\hat{\beta}_0 + \hat{\beta}_1 x) + [2F(\alpha, 2, n-2)]^{1/2}\,\text{sefit}(\hat{y}|x)\,,$$

where $F(\alpha, n_1, n_2)$ is defined to represent the value that cuts off $\alpha \times 100\%$ of the probabilty from the *upper* tail of the F-distribution having n_1 and n_2 degrees of freedom. This can be computed with the `R` command

$$\texttt{qf}(1-\alpha, n_1, n_2)\,.$$

For this problem we find $\alpha = 0.1$ and $n = 270$ so $F(\alpha, 2, n-2)$ is given by `qf(1-0.1,2,268)=2.32`. Thus we can compute the needed values to compute the various parts of the confidence region to compute the requested region.

See the `R` script `chap_2_prob_12.R` for the various parts of this problem.

2.13 (windmills)

2.13.1: In Figure 10 we see a plot of the response *CSpd* as a function of *RSpd* and an ordinary least squares linear fit to this data. A linear fit appears to perform reasonably well.

2.13.2: We can fit *CSpd* as a function of *RSpd* using the `R` function `lm`. Doing so and then using the `R` function `summary` gives the following summary of this linear fit

```
Call:
lm(formula = CSpd ~ RSpd)

Residuals:
    Min      1Q  Median      3Q     Max
-7.7877 -1.5864 -0.1994  1.4403  9.1738

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.14123    0.16958   18.52   <2e-16 ***
RSpd         0.75573    0.01963   38.50   <2e-16 ***
---

Residual standard error: 2.466 on 1114 degrees of freedom
Multiple R-Squared: 0.5709,     Adjusted R-squared: 0.5705
F-statistic:  1482 on 1 and 1114 DF,  p-value: < 2.2e-16
```

2.13.3: A 95% *prediction* interval is given by using the `R` command `predict` with the prediction option. We find

```
> predict( m, newdata=data.frame(RSpd=7.4285), interval="prediction", level=0.95 )
          fit      lwr      upr
[1,] 8.755197 3.914023 13.59637
```

which gives the interval $(3.91, 13.59)$.

2.13.4: Each prediction y_{*i} is given by

$$y_{*i} = \hat{\beta}_0 + \hat{\beta}_1 x_{*i} \, .$$

Thus the average of the m values of y_{*i} denoted by $\bar{y}_*$ is given by

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

which shows that the average of m predictions is equal to the prediction taken at the average value of $\bar{x}_*$ as claimed. To derive the standard error of the *prediction* of $\bar{y}_*$ we recognize that the variance of each individual predicted value y_{*i} has two terms: one due to the noise e_{*i} and one due to the errors in the fitted coefficients $\hat{\beta}_i$. Thus the variance in $\bar{y}_*$ is given by

$$\begin{aligned}
\mathrm{Var}[\bar{y}_*] &= \frac{1}{m^2}\sum_{i=1}^{m} \mathrm{Var}[y_{*i}] \\
&= \frac{1}{m^2}\sum_{i=1}^{m} \sigma^2 + \mathrm{Var}[\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_*] \\
&= \frac{1}{m}\sigma^2 + \mathrm{Var}[\hat{\beta}_0] + \mathrm{Var}[\hat{\beta}_1 \bar{x}_*] + 2\mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1 \bar{x}_*] \\
&= \frac{1}{m}\sigma^2 + \mathrm{Var}[\hat{\beta}_0] + \bar{x}_*^2 \mathrm{Var}[\hat{\beta}_1] + 2\bar{x}_* \mathrm{Cov}[\hat{\beta}_0, \hat{\beta}_1] \, .
\end{aligned}$$

Now using Equations 114, 111, and 115 from the appendix in the above expression we find

$$\begin{aligned}
\mathrm{Var}[\bar{y}_*] &= \frac{\sigma^2}{m} + \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\mathrm{SXX}}\right) + \frac{\sigma^2}{\mathrm{SXX}}\bar{x}_*^2 - \frac{2\overline{x_*}\sigma^2\bar{x}}{\mathrm{SXX}} \\
&= \frac{\sigma^2}{m} + \sigma^2\left(\frac{1}{n} + \frac{1}{\mathrm{SXX}}(\bar{x}_*^2 - 2\bar{x}\bar{x}_* + \bar{x}^2)\right) \\
&= \frac{\sigma^2}{m} + \sigma^2\left(\frac{1}{n} + \frac{(\bar{x}_* - \bar{x})^2}{\mathrm{SXX}}\right), \qquad (9)
\end{aligned}$$

which when we take the square root of the above expression gives the desired result.

2.13.5: Since we have the standard error and the predicted value with $\bar{y}_* = \hat{\beta}_0 + \hat{\beta}_1\bar{x}_*$, we could compute a 95% confidence interval like previously.

See the `R` script `chap_2_prob_13.R` for the various parts of this problem.

Chapter 3 (Multiple Regression)

Notes On The Text

The Ordinary Least Squares Estimate

In this subsection of these notes we derive some of the results stated in the book but provided without proof. This hopefully will remove some of the mystery of where these results come from and provide further understanding. We begin by deriving the equations for the least squares coefficient estimates $\hat{\beta}^*$ and $\hat{\beta}_0$ in terms of the mean reduced data matrices $\mathcal{X}$ and $\mathcal{Y}$. We begin by recalling the definition of the data matrix X

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, \tag{10}$$

and a similar definition for the vector Y. Then since the normal equations require directly computing

$$\hat{\beta} = (X'X)^{-1}(X'Y)\,, \tag{11}$$

or equivalently solving the system

$$(X'X)\hat{\beta} = X'Y\,, \tag{12}$$

for $\hat{\beta}$ we can begin our search for expressions relating $\hat{\beta}_0$ and $\hat{\beta}^*$ by first *partitioning* $\hat{\beta}$ as

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}^* \end{bmatrix}.$$

Using this as motivation, we can conformally partition X into two pieces. We group the left-most column of all ones in a vector called $\mathbf{i}$ and the other variables will be held in the matrix V such that X now looks like

$$X = \begin{bmatrix} \mathbf{i} & V \end{bmatrix}.$$

From this decomposition we have that the matrix V is given by $V = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \cdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$.

Using this factorization of X we see that the product $X'X$ required by the left-hand-side of Equation 12 is given by

$$X'X = \begin{bmatrix} \mathbf{i}' \\ V' \end{bmatrix} \begin{bmatrix} \mathbf{i} & V \end{bmatrix} = \begin{bmatrix} n & \mathbf{i}'V \\ V'\mathbf{i} & V'V \end{bmatrix}. \tag{13}$$

Since $\mathbf{i}$ is a vector of all ones the product of $V'\mathbf{i}$ in the above expression can be explicitly computed. We find

$$V'\mathbf{i} = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} = \begin{bmatrix} \sum x_{i1} \\ \sum x_{i2} \\ \vdots \\ \sum x_{ip} \end{bmatrix} = n \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \equiv n\bar{\mathbf{x}} \,.$$

Where $\bar{x}$ is a vector where each component is the mean of the corresponding predictor. From this result we see that the block product in $X'X$ becomes

$$X'X = \begin{bmatrix} n & n\bar{x}' \\ n\bar{x} & V'V \end{bmatrix} . \tag{14}$$

We next compute the product of $X'Y$ needed in the right-hand-side of Equation 12. Note that this product is given by

$$X'Y = \begin{bmatrix} \mathbf{i}' \\ V' \end{bmatrix} Y = \begin{bmatrix} \mathbf{i}'Y \\ V'Y \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ V'Y \end{bmatrix} .$$

With both of these results our full system required in Equation 12 is

$$\begin{bmatrix} n & n\bar{x}' \\ n\bar{x} & V'V \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}^* \end{bmatrix} = \begin{bmatrix} n\bar{y} \\ V'Y \end{bmatrix} . \tag{15}$$

To produce formulas for $\hat{\beta}_0$ and $\hat{\beta}^*$ we will perform the first step in the Gaussian elimination procedure on this coefficient matrix above and produce an equivalent system for these two variables. We begin by multiplying on the left by the block matrix $\begin{bmatrix} 1/n & 0 \\ 0 & I \end{bmatrix}$ to get

$$\begin{bmatrix} 1/n & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} n & n\bar{x}' \\ n\bar{x} & V'V \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}^* \end{bmatrix} = \begin{bmatrix} 1/n & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} n\bar{y} \\ V'Y \end{bmatrix} ,$$

or

$$\begin{bmatrix} 1 & \bar{x}' \\ n\bar{x} & V'V \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}^* \end{bmatrix} = \begin{bmatrix} \bar{y} \\ V'Y \end{bmatrix} . \tag{16}$$

Note that the first equation can now be solved for $\hat{\beta}_0$ to give

$$\hat{\beta}_0 = \bar{y} - \bar{x}'\hat{\beta}^* \,, \tag{17}$$

which since $\bar{x}'\hat{\beta}^*$ is a scalar product we can transpose it to show that $\bar{x}'\hat{\beta}^* = \hat{\beta}^{*'}\bar{x}$ so the above is equivalent to

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}^{*'}\bar{x} \,, \tag{18}$$

which is an equation we desired to prove. Putting this expression for $\hat{\beta}_0$ given by Equation 17 back into the second equation in 16 gives

$$n\bar{x}(\bar{y} - \bar{x}'\hat{\beta}^*) + V'V\hat{\beta}^* = V'Y \,,$$

or remembering that all y expressions are scalars and so commute with the vector $\bar{x}$ we have

$$(V'V - n\bar{x}\bar{x}')\hat{\beta}^* = V'Y - n\bar{y}\bar{x} \,. \tag{19}$$

Lets introduce the books notation of the mean centered matrices $\mathcal{X}$ and $\mathcal{Y}$. Now $\mathcal{X}$ is defined by and can be written as

$$\begin{aligned}\mathcal{X} &\equiv \begin{bmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{bmatrix} \qquad (20) \\ &= \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{1n} & x_{n2} & \cdots & x_{np} \end{bmatrix} - \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & \vdots & & \vdots \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix} \\ &= V - \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix} = V - \mathbf{i}\bar{x}' \,. \qquad (21)\end{aligned}$$

In the same way $\mathcal{Y}$ is defined by and can be seen to be equivalent to

$$\mathcal{Y} \equiv \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix} = Y - \bar{y}\mathbf{i} \,. \qquad (22)$$

We solve Equations 21 for V and 22 for Y in terms of $\mathcal{X}$ and $\mathcal{Y}$ respectively and put the resulting expressions into Equation 19 to get

$$((\mathcal{X} + \mathbf{i}\bar{x}')'(\mathcal{X} + \mathbf{i}\bar{x}') - n\bar{x}\bar{x}')\hat{\beta}^* = (\mathcal{X} + \mathbf{i}\bar{x}')'(\mathcal{Y} + \bar{y}\mathbf{i}) - n\bar{y}\bar{x} \,.$$

On expanding the products we find

$$(\mathcal{X}'\mathcal{X} + \mathcal{X}'\mathbf{i}\bar{x}' + \bar{x}\mathbf{i}'\mathcal{X} + \bar{x}\mathbf{i}'\mathbf{i}\bar{x}' - n\bar{x}\bar{x}')\hat{\beta}^* = \mathcal{X}'\mathcal{Y} + \bar{y}\mathcal{X}'\mathbf{i} + \bar{x}\mathbf{i}'\mathcal{Y} + \bar{y}\bar{x}\mathbf{i}'\mathbf{i} - n\bar{y}\bar{x} \,. \qquad (23)$$

This result can be simplified greatly by observing that many products of $\mathcal{X}$ and $\mathcal{Y}$ with the vector $\mathbf{i}$ are zero. For example, we have $\mathbf{i}'\mathcal{X} = 0$ similarly we have $\mathcal{X}'\mathbf{i} = 0$ (the transpose of the previous result) and $\mathbf{i}'\mathcal{Y} = 0$. To prove these simple observations consider the meaning of the product $\mathbf{i}'\mathcal{X}$. We find

$$\begin{aligned}\mathbf{i}'\mathcal{X} &= \begin{bmatrix} 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^n (x_{i1}-\bar{x}_1) & \sum_{i=1}^n (x_{i2}-\bar{x}_2) & \cdots & \sum_{i=1}^n (x_{ip}-\bar{x}_p) \end{bmatrix} = 0' \,.\end{aligned}$$

A final simplification we can apply to Equation 23 is to note that $\mathbf{i}'\mathbf{i} = n$. Using these we finally arrive at

$$(\mathcal{X}'\mathcal{X})\hat{\beta}^* = \mathcal{X}'\mathcal{Y} \,, \qquad (24)$$

which is the equation for $\hat{\beta}^*$ we desired to prove.

It can be useful to derive (and we will use these results later) the componentwise relationships that are implied by Equation 24. To do this we will first note that the ij-th component of the demeaned data matrix defined in Equation 21 is given by $\mathcal{X}_{ij} = x_{ij} - \bar{x}_j$, for $1 \leq i \leq n$ and $1 \leq j \leq p$. We can use this representation to write the product on the left-hand-side of Equation 24 as

$$(\mathcal{X}'\mathcal{X})_{jk} = \sum_{i=1}^{n} (\mathcal{X}')_{ji}\mathcal{X}_{ik} = \sum_{i=1}^{n} \mathcal{X}_{ij}\mathcal{X}_{ik} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)\,. \tag{25}$$

With this expression we can compute a componentwise expression for $(\mathcal{X}'\mathcal{X})\hat{\beta}^*$. We find

$$\begin{aligned} ((\mathcal{X}'\mathcal{X})\hat{\beta}^*)_i &= \sum_{k=1}^{p} (\mathcal{X}'\mathcal{X})_{ik}\hat{\beta}^*_k \\ &= \sum_{k=1}^{p}\sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)\hat{\beta}^*_k \qquad (26) \\ &= \sum_{j=1}^{n} (x_{ji} - \bar{x}_i) \sum_{k=1}^{p} (x_{jk} - \bar{x}_k)\hat{\beta}^*_k\,. \qquad (27) \end{aligned}$$

Where we have written our desired summation expressions in two different forms. Next we evaluate the componentwise representation for the right-hand-side of Equation 24. We have

$$(\mathcal{X}'\mathcal{Y})_i = \sum_{j=1}^{n} (\mathcal{X}')_{ij}\mathcal{Y}_j = \sum_{j=1}^{n} \mathcal{X}_{ji}\mathcal{Y}_j = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(y_j - \bar{y})\,. \tag{28}$$

These two expressions will be used in later derivations.

Properties of the Estimates

We can derive the variance of our estimated regression coefficients $\hat{\beta}$ by considering one of its equivalent expressions. Using Equation 11 and remembering that the data matrix X is not random we have

$$\begin{aligned} \mathrm{Var}(\hat{\beta}) &= (X'X)^{-1}X'\mathrm{Var}(Y)X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}\,. \end{aligned}$$

Since know that $\hat{\beta}^*$ has exactly the equivalent formula (given by Equation 24) or $\hat{\beta}^* = (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\mathcal{Y}$ it will have an analogous variance calculation given by

$$\mathrm{Var}(\hat{\beta}^*) = \sigma^2(\mathcal{X}'\mathcal{X})^{-1}\,.$$

We next derive several equivalent expressions for the residual sum of squares (RSS). We begin directly using the definition of RSS and the least squares solution for $\hat{\beta}$ given by Equation 11. In matrix notation the RSS in term of the vector of residuals $\hat{\mathbf{e}}$ is given by

$$\mathrm{RSS} = \hat{\mathbf{e}}'\hat{\mathbf{e}}$$

$$
\begin{aligned}
&= (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) \\
&= (\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y})'(\mathbf{Y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} + \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} \qquad (29) \\
&= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}\,, \qquad (30)
\end{aligned}
$$

which is the second equation near in the books equation 3.15 in the list of equivalent expressions for RSS and we have used the fact that $X'Y = (X'X)\hat{\beta}$ in the second to last equation to derive the last equation. Now replacing X and Y in this last expression with $\mathcal{X}$ and $\mathcal{Y}$ obtained from Equations 21 and 22 we find

$$
\begin{aligned}
\mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{Y} &= (\mathcal{Y} + \bar{y}\mathbf{i})'(\mathcal{Y} + \bar{y}\mathbf{i}) - \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}^{*\prime} \end{bmatrix} \begin{bmatrix} \mathbf{i} & \mathcal{X} + \mathbf{i}\bar{x}' \end{bmatrix}' (\mathcal{Y} + \bar{y}\mathbf{i}) \\
&= \mathcal{Y}'\mathcal{Y} + \bar{y}\mathcal{Y}'\mathbf{i} + \bar{y}\mathbf{i}'\mathcal{Y} + \bar{y}^2\mathbf{i}'\mathbf{i} - \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}^{*\prime} \end{bmatrix} \begin{bmatrix} \mathbf{i}' \\ \mathcal{X}' + \bar{x}\mathbf{i}' \end{bmatrix} (\mathcal{Y} + \bar{y}\mathbf{i}) \\
&= \mathcal{Y}'\mathcal{Y} + n\bar{y}^2 - \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}^{*\prime} \end{bmatrix} \begin{bmatrix} \mathbf{i}'\mathcal{Y} + \bar{y}\mathbf{i}'\mathbf{i} \\ \mathcal{X}'\mathcal{Y} + \bar{y}\mathcal{X}'\mathbf{i} + \bar{x}\mathbf{i}'\mathcal{Y} + \bar{y}\bar{x}\mathbf{i}'\mathbf{i} \end{bmatrix} \\
&= \mathcal{Y}'\mathcal{Y} + n\bar{y}^2 - \begin{bmatrix} \hat{\beta}_0 & \hat{\beta}^{*\prime} \end{bmatrix} \begin{bmatrix} n\bar{y} \\ \mathcal{X}'\mathcal{Y} + n\bar{y}\bar{x} \end{bmatrix} \\
&= \mathcal{Y}'\mathcal{Y} + n\bar{y}^2 - n\bar{y}\hat{\beta}_0 - \hat{\beta}^{*\prime}\mathcal{X}'\mathcal{Y} - n\bar{y}\hat{\beta}^{*\prime}\bar{x} \\
&= \mathcal{Y}'\mathcal{Y} + n\bar{y}^2 - n\bar{y}^2 + n\bar{y}\hat{\beta}^{*\prime}\bar{x} - \hat{\beta}^{*\prime}\mathcal{X}'\mathcal{Y} - n\bar{y}\hat{\beta}^{*\prime}\bar{x} \\
&= \mathcal{Y}'\mathcal{Y} - \hat{\beta}^{*\prime}\mathcal{X}'\mathcal{Y} \qquad (31) \\
&= \mathcal{Y}'\mathcal{Y} - \hat{\beta}^{*\prime}(\mathcal{X}'\mathcal{X})\hat{\beta}^{*}\,, \qquad (32)
\end{aligned}
$$

which is the expression presented in equation 3.15 in the book. If we consider $\mathcal{Y}'\mathcal{Y}$ presented above we get

$$
\begin{aligned}
\mathcal{Y}'\mathcal{Y} &= (Y' - \bar{y}\mathbf{i}')(Y - \bar{y}\mathbf{i}) \\
&= Y'Y - \bar{y}Y'\mathbf{i} - \bar{y}\mathbf{i}'Y + \bar{y}^2\mathbf{i}'\mathbf{i} \\
&= Y'Y - \bar{y}n\bar{y} - \bar{y}n\bar{y} + \bar{y}^2 n \\
&= Y'Y - n\bar{y}^2\,.
\end{aligned}
$$

Thus solving for $Y'Y$ gives

$$
Y'Y = \mathcal{Y}'\mathcal{Y} + n\bar{y}^2\,.
$$

When we put this into Equation 30 we get

$$
\mathrm{RSS} = \mathcal{Y}'\mathcal{Y} - \hat{\beta}'(X'X)\hat{\beta} + n\bar{y}^2\,, \qquad (33)
$$

which is the last equivalence for RSS we need to derive.

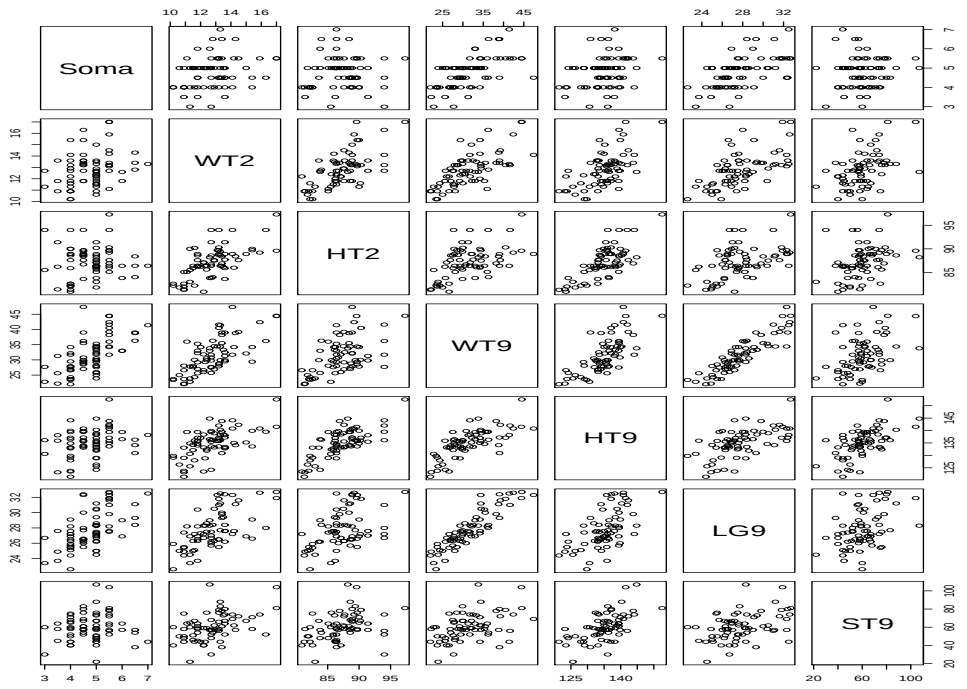

Figure 11: A scatterplot of Wt as a function of Ht for Problem 2.1.

Problem Solutions

3.1 (the Berkeley Guidance Study)

3.1.1: In Figure 11 we present a scatter plot matrix of the variables $WT2$, $HT2$, $WT9$, $HT9$, $LG9$, $ST9$, $Soma$. There we see very strong correlation between the $WT2$ and $HT2$ variables, between the $WT9$ and $HT9$ variables, and in fact all variables of the same age. There is a weaker correlation but still some between the variables of one year and later years.

3.1.2: The methods from `R` that are needed to construct an added variable plot is described nicely in the `RSprimer` that comes with this textbook. For notation lets assume that X is a variable representing the term in the regression you are considering adding. The basic idea in added-variable plots is to create a regression that contains all variables *but* the term you are interested in adding i.e. X. From this model extract the residuals that correspond to the resulting regression. Next fit a linear model where the dependent variable is X and all other terms are the predictors and extract this models residual. Finally, plot these two sequence of residuals on one graph. For this example we do this in Figure 12. The plot in the lower-right corresponds to the added-variable plot. From the looks of this plot it appears that adding the variable $LG9$ will not significantly improve predictability of the value of $Soma$.

3.1.3: To fit this multiple regression problem we will use the `lm` function provided by the `R` statistical language and display the resulting output from the `summary` command. When we do this we find

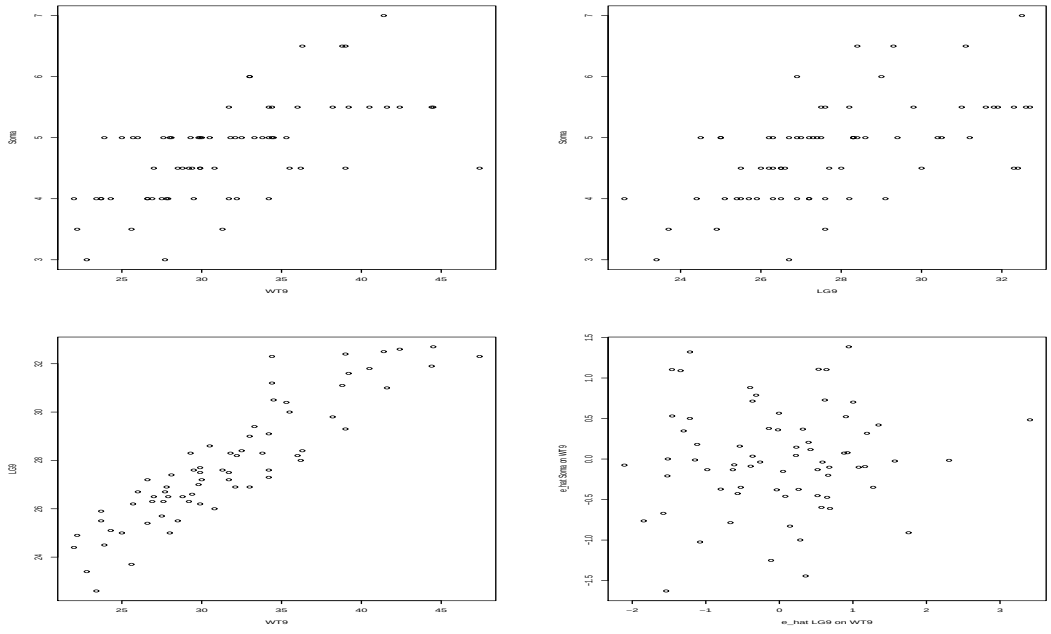

Figure 12: **Upper Left:** A scatter plot of *Soma* vs. $WT9$. **Upper Right:** A scatter plot of *Soma* vs. $LG9$. **Lower Left:** A scatter plot of $WT9$ vs. $LG9$. Note the very strong correlation between these two variables, implying that knowledge of one is tantamount to knowing the other. **Lower Right:** The added-variable plot which closely resembles a null-plot.

```
> summary(m)

Call:
lm(formula = Soma ~ HT2 + WT2 + HT9 + WT9 + ST9)

Residuals:
     Min       1Q   Median       3Q      Max
-2.03132 -0.34062  0.01917  0.43939  0.97266

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,     Adjusted R-squared: 0.4837
F-statistic: 13.93 on 5 and 64 DF,  p-value: 3.309e-09
```

From the above `summary` command we observe that $\hat{\sigma} = 0.5791$ and $R^2 = 0.5211$. The overall F-test has a statistic given by 13.93 and a p-value of $3.3\,10^{-9}$. The t-statistics for the null hypothesis that $\beta_i = 0$ given all other $\beta_j \neq 0$ is also presented in this table. From the above table we can conclude that the variable $HT9$ (given that all the others are in the model) can be probably be dropped. In addition, since the variable $WT9$ as it has the largest t-value it is probably one of the most important factors at predicting $Soma$.

We can obtain the F-statistic and obtain the overall analysis of variance table by constructing a model with only a constant term and then calling the `anova` command with two arguments: the model with only the constant term and the model with all terms. When we do this we find

```
> anova(m0,m)
Analysis of Variance Table

Model 1: Soma ~ 1
Model 2: Soma ~ HT2 + WT2 + HT9 + WT9 + ST9
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     69 44.818
2     64 21.462  5    23.356 13.929 3.309e-09 ***
```

Note that this table gives the same F-statistic and p-value as we found from the `summary` command as it should.

3.1.4/5: We can obtain the sequential analysis of variance table by calling the `anova` command on the initial model since the left to right order found in 3.25 was how this model was constructed. We find

```
> anova(mLR)
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
```

If we fit the terms in the opposite order (that is from right-to-left) we find the sequential anova table given by

```
> anova(mRL)
```

```
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
```

Some conclusions from these two tables is that the variable $WT9$ (when inserted in any order) explains a lot of the variance of $Soma$. It explains "less" when included in the regression after other variables like $HT2$, $WT2$, $HT9$. In addition, once $WT9$ has been included in the regression the additional variable contribute significantly less to the residual sum of squares.

See the `R` script `chap_3_prob_1.R` for the various parts of this problem.

3.2 (added-variable plots)

3.2.1: For this problem we will add the variable $\log(PPgdp)$ *after* the variable $Purban$, which is backwards to the example presented in the book where $Purban$ was added after $\log(PPgdp)$. We begin by computing a linear model of $\log(Fertility)$ using both $\log(PPgdp)$ and $Purban$ as predictors and find the estimated coefficient $\hat{\beta}_1$ of $\log(PPgdp)$ given by -0.125475.

To compute the added-variable plot for $\log(PPgdp)$ after $Purban$ we compute the *residuals* of the following two models (in `R` notation)

- `logFertility ~ Purban`
- `logPPgdp ~ Purban`

In an added-variable plot we are assuming that we have begun our analysis with a subset of variables (here only one $Purban$) and are considering adding another one (here $\log(PPgdp)$). The residuals in the linear regression of $\log(Fertility)$ on $Purban$ represents numerically what we *don't* know about $\log(Fertility)$ when we are told the value of $Purban$. In the same way the residuals of $\log(PPgdp)$ on $Purban$ represents what information is in $\log(PPgdp)$ that is *not* already contained in $Purban$. Thus these residuals represent the addition information we can use to predict values of $\log(Fertility)$. When we fit a linear model on the two residuals above we find that the slope coefficient given by -0.1255 (numerically equivalent to the above number).

See the `R` script `chap_3_prob_2.R` for the various parts of this problem.

3.3 (a specific mean function)

3.3.1: To make an added-variable plot we would compute the residuals from the two regressions `Y ~ X1`, and `X2 ~ X1`. Since X_2 is deterministically given by X_1 this second regression will have *no* error (or a zero residual). Thus when we plot `res( Y ~ X1 )` vs. `res( X2 ~ X1 )` all of the points will have $x = 0$ and estimating a slope (the added-variable value) will be difficult/impossible.

3.3.2: As in the previous problem when we make an added-variable plot we would compute the residuals from the two regressions `Y ~ X1`, and `X2 ~ X1`. In this case since Y is deterministically given by $3X_1$ this first regression will have *no* error and all residuals will be zero. Thus when we plot `res( Y ~ X1 )` vs. `res( X2 ~ X1 )` all of the points will have $y = 0$ giving an the estimate of the regression coefficient in front of X_2 of $\hat{\beta}_2 \approx 0$.

3.3.3: If X_2 is *independent* of X_1 then the regression `X2 ~ X1` will not be able to explain *any* of the variation in X_2 about its mean and the residuals of this regression will directly represent the variability in X_2. In addition, by the independence of X_1 and X_2, the residual of the regression of Y on X_1 will have all of the variation of Y itself with respect to the variable X_2. Thus in this independent case the scatter plot of Y versus X_2 will have the same shape as the added-variable plot for X_2 after X_1.

3.3.4: It depends. In an added-variable plot we plot the residuals of the regression of Y onto $X_1, X_2, \ldots X_n$ against the residuals of the regression of X_{n+1} onto these same variables. If the first regressions on $X_1, X_2, \ldots, X_n$ has *no* variance reduction while the variable X_{n+1} *does*, one could have the vertical variation of Y vs. X_{n+1} smaller than that in the residual of Y vs. $X_1, X_2, \ldots X_n$. Thus to answer this question depends on how much the variables $X_1, X_2, \ldots X_n$ reduce the variance in Y relative to how much the variable X_{n+1} reduces the variance of Y. The variables $X_1, X_2, \ldots X_n$ can reduce the variance of Y more, an equal amount, or less than how much X_{n+1} does.

3.4 (dependent variables with zero correlation)

3.4.1: We can evaluate the slopes we would obtain in performing each of these simple regressions using Equation 108. In addition, the constant term is given by Equation 107. We find that (defining some expressions as we do this)

$$\begin{aligned}
\hat{\beta}_1(Y \sim X_1) &\equiv \hat{\beta}_{YX_1} = \frac{\sum(x_{i1} - \bar{x}_1)(y_i - \bar{y})}{\sum(x_{i1} - \bar{x}_1)^2} \\
\hat{\beta}_1(Y \sim X_2) &\equiv \hat{\beta}_{YX_2} = \frac{\sum(x_{i2} - \bar{x}_2)(y_i - \bar{y})}{\sum(x_{i2} - \bar{x}_2)^2} \\
\hat{\beta}_1(X_2 \sim X_1) &\equiv \hat{\beta}_{X_2X_1} = \frac{\sum(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum(x_{i1} - \bar{x}_1)^2} = 0\,.
\end{aligned}$$

Where the last expression follows because X_1 and X_2 are uncorrelated.

3.4.2: The residuals for the regressions of Y on X_1 and for X_2 on X_1 are given by

$$\begin{aligned}
\text{res}_i(Y \sim X_1) &= y_i - (\bar{y} - \hat{\beta}_{YX_1}\bar{x}_1 + \hat{\beta}_{YX_1}x_{i1}) \\
&= y_i - \bar{y} - \hat{\beta}_{YX_1}(x_{i1} - \bar{x}_1) \\
\text{res}_i(X_2 \sim X_1) &= x_{i2} - (\bar{x}_2 - \hat{\beta}_{X_2X_1}\bar{x}_1 + \hat{\beta}_{X_2X_1}x_{i1}) \\
&= x_{i2} - \bar{x}_2 - \hat{\beta}_{X_2X_1}(x_{i1} - \bar{x}_1) \\
&= x_{i2} - \bar{x}_2 \,.
\end{aligned}$$

3.4.3: Note that since we included a constant term in these regressions the mean of both the residuals above is zero. So the coefficient we want to evaluate is given by

$$\hat{\beta}_1 \left(\text{res}(Y \sim X_1) \sim \text{res}(X_2 \sim X_1)\right) = \frac{\sum_i \text{res}_i(X_2 \sim X_1)\text{res}_i(Y \sim X_1)}{\sum_i \text{res}_i(X_2 \sim X_1)^2} \,. \tag{34}$$

The denominator is easy to evaluate and equals SX_2X_2. Next the numerator becomes (again using the fact that X_1 and X_2 are uncorrelated)

$$\begin{aligned}
\sum_i \text{res}_i(X_2 \sim X_1)\text{res}_i(Y \sim X_1) &= \sum_i (x_{i2} - \bar{x}_2)(y_i - \bar{y} - \hat{\beta}_{YX_1}(x_{i1} - \bar{x}_1)) \\
&= \sum_i (x_{i2} - \bar{x}_2)(y_i - \bar{y}) - \hat{\beta}_{YX_1} \sum_i (x_{i2} - \bar{x}_2)(x_{i1} - \bar{x}_1) \\
&= \text{SX}_2\text{Y} \,.
\end{aligned}$$

Thus the ratio required by Equation 34 is given by

$$\frac{\text{SX}_2\text{Y}}{\text{SX}_2\text{X}_2} \,,$$

which exactly equals $\hat{\beta}_1(Y \sim X_2)$ as we were to show. The intercept of the added variable plot is given by an expression like Equation 107 but where the variables X and Y are redefined to represent the residuals of interest. Since both of the residuals $\text{res}(Y \sim X_1)$ and $\text{res}(X_2 \sim X_1)$ have zero mean this coefficient as a combination of these two values must also be zero.

3.5 (predicting $BSAAM$ from $OPBPC$, $OPRC$, and $OPSLAKE$)

3.5.1: See Figure 13 for the requested scatterplot matrix. From that plot we see that the response $BSAAM$ seems to be positively correlated with *every* variable. In addition, each of the predictors $OPBPC$, $OPRC$, and $OPSLAKE$ seems to be positively correlated with each other. Thus we anticipate the added-variable plots for these variables will show that the addition of alternative variables (after the first one) will not provide much predictability. In summary, it looks like the correlations among all variables will be large (near one) and positive. Computing a correlation matrix with the `R` command `round(cor(water[,c(8,5,6,7)]),4)` we obtain

```
       BSAAM  OPBPC   OPRC OPSLAKE
```

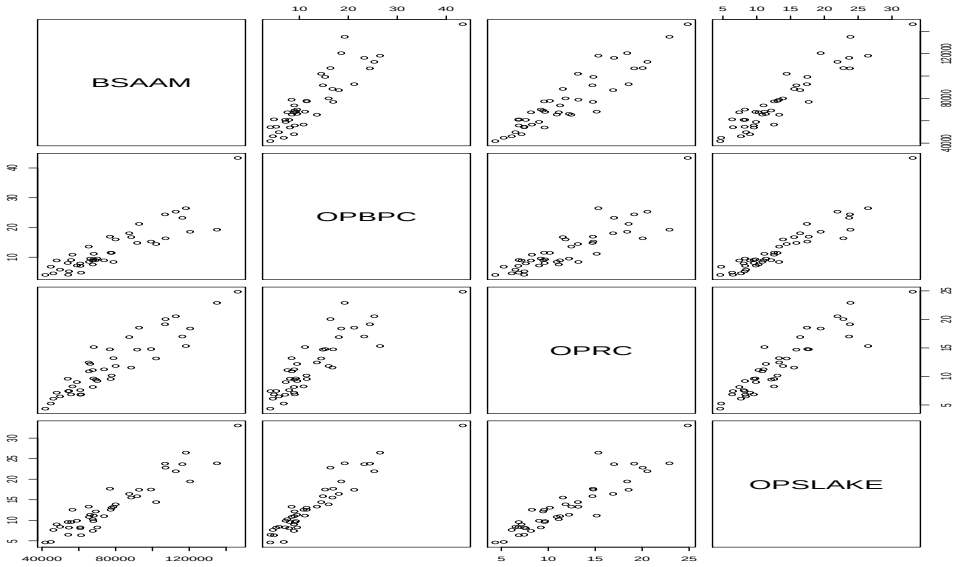

Figure 13: A scatter plot matrix that displays the scatter dependence of the variables $BSAAM$, $OPBPC$, $OPRC$, and $OPSLAKE$. See the corresponding text for discussion.

```
BSAAM   1.0000 0.8857 0.9196  0.9384
OPBPC   0.8857 1.0000 0.8647  0.9433
OPRC    0.9196 0.8647 1.0000  0.9191
OPSLAKE 0.9384 0.9433 0.9191  1.0000
```

which verifies the statements make above. Computing a "summary" of a linear regression of $BSAAM$ on $OPBPC$, $OPRC$, and $OPSLAKE$ we find

```
Call:
lm(formula = BSAAM ~ OPBPC + OPRC + OPSLAKE)

Residuals:
     Min       1Q   Median       3Q      Max
-15964.1  -6491.8   -404.4   4741.9  19921.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 22991.85    3545.32   6.485  1.1e-07 ***
OPBPC          40.61     502.40   0.081  0.93599
OPRC         1867.46     647.04   2.886  0.00633 **
OPSLAKE      2353.96     771.71   3.050  0.00410 **
---

Residual standard error: 8304 on 39 degrees of freedom
Multiple R-Squared: 0.9017,     Adjusted R-squared: 0.8941
F-statistic: 119.2 on 3 and 39 DF,  p-value: < 2.2e-16
```

The t-values shown in this output indicate that the intercept is very significant and thus we can conclude that the mean of $BSAAM$ is very likely (almost certainly) non-zero. The t-values for the two terms $OPSLAKE$ and $OPRC$ are also relatively large and probably provide some explanatory power. Notice that these are also the two variable with the highest correlation with the response $BSAAM$ and the magnitude of their t-values is in the same order relationship as their correlations.

3.5.2: The F-statistic above represents the result of an overall test that $BSAAM$ is independent of the three other terms. The p-value for this statistic is very small, indicating that there is very little chance that this observed variance reduction is due to chance alone. We can conclude that $BSAAM$ is *not* independent of the other three variables.

3.5.3: For this part of the problem we want to compute three different analysis of variance tables for three different orders possible of the variables $OPBPC$, $OPRC$, and $OPSLAKE$. The language `R` performs sequential analysis of variance with the `anova` command where the order of the terms removed is determined by the order in which they are specified in the `lm` command. See the `R` script `chap_3_prob_5.R` for more details. Running this we find (printed all on one page for clarity)

```
> m1 <- lm( BSAAM ~ OPBPC + OPRC + OPSLAKE )
> anova(m1)
Analysis of Variance Table

Response: BSAAM
          Df     Sum Sq    Mean Sq  F value    Pr(>F)
OPBPC      1 2.1458e+10 2.1458e+10 311.1610 < 2.2e-16 ***
OPRC       1 2.5616e+09 2.5616e+09  37.1458 3.825e-07 ***
OPSLAKE    1 6.4165e+08 6.4165e+08   9.3045  0.004097 **
Residuals 39 2.6895e+09 6.8962e+07
---
> m2 <- lm( BSAAM ~ OPBPC + OPSLAKE + OPRC )
> anova(m2)
Analysis of Variance Table

Response: BSAAM
          Df     Sum Sq    Mean Sq  F value    Pr(>F)
OPBPC      1 2.1458e+10 2.1458e+10 311.1610 < 2.2e-16 ***
OPSLAKE    1 2.6288e+09 2.6288e+09  38.1203 2.967e-07 ***
OPRC       1 5.7444e+08 5.7444e+08   8.3299  0.006326 **
Residuals 39 2.6895e+09 6.8962e+07
---
> m3 <- lm( BSAAM ~ OPSLAKE + OPRC + OPBPC )
> anova(m3)
Analysis of Variance Table

Response: BSAAM
          Df     Sum Sq    Mean Sq  F value    Pr(>F)
OPSLAKE    1 2.4087e+10 2.4087e+10 349.2806 < 2.2e-16 ***
OPRC       1 5.7405e+08 5.7405e+08   8.3242  0.006342 **
OPBPC      1 4.5057e+05 4.5057e+05   0.0065  0.935990
Residuals 39 2.6895e+09 6.8962e+07
---
```

Note that the t-statistic *squared* for each coefficient taken individually will equal the F-statistics shown in the sequential ANOVA table above which has that same variable added *last*. To show this, we can compute the t-statistic squared. We find for each of the three terms this is given by

```
(Intercept)       OPBPC        OPRC     OPSLAKE
42.05700100  0.00653367  8.32990239  9.30448873
```

When we compare this to the corresponding row in the sequential anova tables above we see that they are equal.

3.5.4: We can test for the significance of the two terms $OPRC$ and $OPBPC$ taken together against the model where the are absent using the `R` command `anova`. The call and results for this are given by

```
> mLarger  = lm( BSAAM ~ OPSLAKE + OPRC + OPBPC ) # all three inputs
> mSmaller = lm( BSAAM ~ OPSLAKE )
> anova(mSmaller,mLarger)
Analysis of Variance Table

Model 1: BSAAM ~ OPSLAKE
Model 2: BSAAM ~ OPSLAKE + OPRC + OPBPC
  Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
1     41 3264010454
2     39 2689509185  2  574501270 4.1654 0.02293 *
```

From the output presented here we see that the F-value for the model with the two additional terms is 4.165 compared to the model with just the term $OPSLAKE$ has a 0.02 probability of occurring by chance when in fact these two terms should not be included. Since this probability is relatively small we can reasonable conclude that these two terms are in fact significant and do indeed reduce the variance of the residuals over what might be observed by chance.

See the `R` script `chap_3_prob_5.R` for the various parts of this problem.

Chapter 4 (Drawing Conclusions)

Notes On The Text

sampling from a multivariate normal population

Here we will derive the distribution of y_i given x_i where the vector $(x_i, y_i)^T$ is distributed as a joint two-dimensional normal. This result that we derive is simply stated in the book and is verified in these notes below. We begin with the fact that the distribution of $(x_i, y_i)^T$ is jointly Gaussian. This means that the density of this pair of variables is given by

$$p(x_i, y_i) = \frac{1}{(2\pi)^{2/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2} \left(\begin{bmatrix} x_i \\ y_i \end{bmatrix} - \mu \right)' \Sigma^{-1} \left(\begin{bmatrix} x_i \\ y_i \end{bmatrix} - \mu \right) \right\}, \tag{35}$$

where μ is the mean vector which in components is given by $\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$ and Σ is the covariance matrix given by $\Sigma = \begin{bmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix}$. With these expressions we begin with the definition of conditional probability. Dropping the i subscripts, recall that the density we would like to evaluate $p(y|x)$ can be expressed as

$$p(y|x) = \frac{p(x,y)}{p(x)}. \tag{36}$$

Lets begin to evaluate this by simplifying the expression above for $p(x, y)$. To simplify the notation above lets define the exponent of the above expression to be $\mathcal{E}$. Thus $\mathcal{E}$ is given by

$$\mathcal{E} = -\frac{1}{2} \left(\begin{bmatrix} x \\ y \end{bmatrix} - \mu \right)' \Sigma^{-1} \left(\begin{bmatrix} x \\ y \end{bmatrix} - \mu \right).$$

Since this expression requires the inverse of Σ we compute it and find

$$\Sigma^{-1} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho_{xy}^2)} \begin{bmatrix} \sigma_y^2 & -\rho_{xy}\sigma_x\sigma_y \\ -\rho_{xy}\sigma_x\sigma_y & \sigma_x^2 \end{bmatrix}. \tag{37}$$

So $\mathcal{E}$ becomes (in terms of a new variable we introduce $\hat{\mathcal{E}}$)

$$\mathcal{E} = -\frac{1}{2} \left(\frac{1}{\sigma_x^2\sigma_y^2(1-\rho_{xy}^2)} \right) \hat{\mathcal{E}}.$$

Finally, we compute the value of $\hat{\mathcal{E}}$ as

$$\begin{aligned}
\hat{\mathcal{E}} &\equiv \begin{bmatrix} x-\mu_x, & y-\mu_y \end{bmatrix} \begin{bmatrix} \sigma_y^2 & -\rho_{xy}\sigma_x\sigma_y \\ -\rho_{xy}\sigma_x\sigma_y & \sigma_x^2 \end{bmatrix} \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix} \\
&= \begin{bmatrix} x-\mu_x, & y-\mu_y \end{bmatrix} \begin{bmatrix} \sigma_y^2(x-\mu_x) - \rho_{xy}\sigma_x\sigma_y(y-\mu_y) \\ -\rho_{xy}\sigma_x\sigma_y(x-\mu_x) + \sigma_x^2(y-\mu_y) \end{bmatrix} \\
&= \sigma_y^2(x-\mu_x)^2 - \rho_{xy}\sigma_x\sigma_y(x-\mu_x)(y-\mu_y) - \rho_{xy}\sigma_x\sigma_y(x-\mu_x)(y-\mu_y) + \sigma_x^2(y-\mu_y)^2 \\
&= \sigma_y^2(x-\mu_x)^2 - 2\rho_{xy}\sigma_x\sigma_y(x-\mu_x)(y-\mu_y) + \sigma_x^2(y-\mu_y)^2.
\end{aligned}$$

In Equation 36 since x is given and y is variable this observation will motivate the grouping of the terms and we will write this expression in a special way. We have $\mathcal{E}$ given by

$$\begin{aligned}\mathcal{E} &= -\frac{1}{2}\left(\frac{1}{\sigma_y^2(1-\rho_{xy}^2)}\right)\left((y-\mu_y)^2 - 2\rho_{xy}\frac{\sigma_y}{\sigma_x}(y-\mu_y)(x-\mu_x)\right) \\ &- \frac{1}{2}\left(\frac{1}{\sigma_x^2(1-\rho_{xy}^2)}\right)(x-\mu_x)^2 .\end{aligned}$$

We next complete the square of the y expression to obtain

$$\begin{aligned}\mathcal{E} &= -\frac{1}{2}\left(\frac{1}{\sigma_y^2(1-\rho_{xy}^2)}\right)\left((y-\mu_y)^2 - 2\rho_{xy}\frac{\sigma_y}{\sigma_x}(y-\mu_y)(x-\mu_x) + {\rho_{xy}}^2\frac{\sigma_y^2}{\sigma_x^2}(x-\mu_x)^2\right) \\ &+ \frac{1}{2}\left(\frac{1}{\sigma_y^2(1-\rho_{xy}^2)}\right)\left({\rho_{xy}}^2\frac{\sigma_y^2}{\sigma_x^2}(x-\mu_x)^2\right) - \frac{1}{2}\left(\frac{1}{\sigma_x^2(1-\rho_{xy}^2)}\right)(x-\mu_x)^2 \\ &= -\frac{1}{2}\left(\frac{1}{\sigma_y^2(1-\rho_{xy}^2)}\right)\left(y-\mu_y-\rho_{xy}\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)^2 + \frac{1}{2}\left(\frac{1}{\sigma_x^2(1-\rho_{xy}^2)}\right)({\rho_{xy}}^2-1)(x-\mu_x)^2 \\ &= -\frac{1}{2}\left(\frac{1}{\sigma_y^2(1-\rho_{xy}^2)}\right)\left(y-\mu_y-\rho_{xy}\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)^2 - \frac{1}{2}\left(\frac{1}{\sigma_x^2}\right)(x-\mu_x)^2 . \qquad (38)\end{aligned}$$

Now that we have evaluated $p(x,y)$ lets compute $p(x)$ the marginal density of x. Recalling its definition we have

$$\begin{aligned}p(x) &= \int p(x,y)dy = \frac{1}{2\pi|\Sigma|^{1/2}}e^{-\frac{1}{2}\left(\frac{1}{\sigma_x^2}\right)(x-\mu_x)^2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{1}{\sigma_y^2(1-\rho_{xy}^2)}\right)\left(y-\mu_y-\rho_{xy}\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)^2}dy \\ &= \frac{1}{2\pi|\Sigma|^{1/2}}e^{-\frac{1}{2}\left(\frac{1}{\sigma_x^2}\right)(x-\mu_x)^2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{1}{\sigma_y^2(1-\rho_{xy}^2)}\right)y^2}dy .\end{aligned}$$

To evaluate this last integral let $v = \frac{y}{\sigma_y\sqrt{1-\rho_{xy}^2}}$, then $dy = \sigma_y\sqrt{1-{\rho_{xy}}^2}dv$ and the above becomes

$$p(x) = \frac{\sigma_y\sqrt{1-{\rho_{xy}}^2}}{2\pi|\Sigma|^{1/2}}e^{-\frac{1}{2}\left(\frac{1}{\sigma_x^2}\right)(x-\mu_x)^2}\int_{-\infty}^{\infty} e^{-\frac{1}{2}v^2}dv .$$

This last integral is $\sqrt{2\pi}$ and from the definition Σ we have that $|\Sigma| = \sigma_x^2\sigma_y^2(1-\rho_{xy}^2)$ and the above simplifies to

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma_x}e^{-\frac{1}{2}\left(\frac{1}{\sigma_x^2}\right)(x-\mu_x)^2} ,$$

or another one-dimensional Gaussian. Combining this expression with the previously derived expression for $p(x,y)$ we find

$$\begin{aligned}p(y|x) &= \frac{p(x,y)}{p(x)} \\ &= \frac{\left(\frac{1}{2\pi}\right)\left(\frac{1}{\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}}\right)e^{-\frac{1}{2}\left(\frac{1}{\sigma_y^2(1-\rho_{xy}^2)}\right)\left(y-\mu_y-\rho_{xy}\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)^2}e^{-\frac{1}{2}\left(\frac{1}{\sigma_x^2}\right)(x-\mu_x)^2 .}}{\left(\frac{1}{\sqrt{2\pi}}\right)\left(\frac{1}{\sigma_x}\right)e^{-\frac{1}{2}\left(\frac{1}{\sigma_x^2}\right)(x-\mu_x)^2 .}} \\ &= \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\sigma_y\sqrt{1-\rho_{xy}^2}}\right)e^{-\frac{1}{2}\left(\frac{1}{\sigma_y^2(1-\rho_{xy}^2)}\right)\left(y-\mu_y-\rho_{xy}\frac{\sigma_y}{\sigma_x}(x-\mu_x)\right)^2} .\end{aligned}$$

This later expression is a normal with a mean of

$$\mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x - \mu_x)\,, \tag{39}$$

and a variance given by

$$\sigma_y^2(1 - \rho_{xy}^2)\,, \tag{40}$$

which is the result we desired to show.

The above result, expresses the density for y given x when x and y are jointly normal. In the case where the value y (again a scalar) and $\mathbf{x}$ (now a *vector*) are jointly normal the distribution of y given $\mathbf{x}$ can be derived in a similar way. The result of this derivation is stated in the book but without proof. In this section of the notes we will verify the given expressions. We begin with the quoted expression for $p(\mathbf{x}, y)$ and integrate with respect to y to obtain the conditional distribution $p(\mathbf{x})$. Then using this in Equation 36 we will produce an explicit expression for $p(y|\mathbf{x})$ the density of interest.

Since we are to assume that the covariance matrix, Σ, for the vector $\begin{bmatrix} \mathbf{x} \\ y \end{bmatrix}$ is given in block form as

$$\Sigma \equiv \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}' & \sigma_y^2 \end{bmatrix}.$$

Where Σ_{xx} is a $p \times p$ matrix and Σ_{xy} is a $p \times 1$ column vector. Using this we can express the joint density $p(\mathbf{x}, y)$ in block form as

$$p(\mathbf{x}, y) = \frac{1}{(2\pi)^{\frac{p+1}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2}\left(\mathbf{x}' - \mu_x{}', y - \mu_y\right) \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}' & \sigma_y^2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x} - \mu_x \\ y - \mu_y \end{bmatrix} \right\}. \tag{41}$$

Thus to further simplify this we need to derive an expression for Σ^{-1}. To compute this inverse we will multiply Σ on the left by a block matrix with some variable entries which we hope we can find suitable values for and thus derive the block inverse. As an example of this lets multiply Σ on the left by the block matrix $\begin{bmatrix} \Sigma_{xx}^{-1} & 0 \\ b' & d \end{bmatrix}$, where b is a $p \times 1$ dimensional vector and d is a scalar. Currently, the values of these two variables are unknown. When we multiply by this matrix we desire to find b and d such that

$$\begin{bmatrix} \Sigma_{xx}^{-1} & 0 \\ b' & d \end{bmatrix} \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}' & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & 1 \end{bmatrix}. \tag{42}$$

Equating the block multiplication result on the left to the components of the block matrix on the right gives

$$b'\Sigma_{xx} + d\Sigma_{xy}' = 0\,.$$

for the $(2, 1)$ component. This later equation can be solved for b by taking transposes and inverting Σ_{xx} as

$$b = -\Sigma_{xx}^{-1}\Sigma_{xy}d\,.$$

If we take $d = 1$ and b given by the solution above, the product on the left-hand-side given by Equation 42 does not becomes the identity but is given by

$$\begin{bmatrix} \Sigma_{xx}^{-1} & 0 \\ -\Sigma_{xy}'\Sigma_{xx}^{-1} & 1 \end{bmatrix} \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}' & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} I & \Sigma_{xx}^{-1}\Sigma_{xy} \\ 0 & \sigma_y^2 - \Sigma_{xy}'\Sigma_{xx}^{-1}\Sigma_{xy} \end{bmatrix}. \tag{43}$$

Note what we have just done is the "forward solve" step in Gaussian elimination. Taking the inverse of both sides of this later equation we find

$$\begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}' & \sigma_y^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{xx}^{-1} & 0 \\ -\Sigma_{xy}'\Sigma_{xx}^{-1} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} I & \Sigma_{xx}^{-1}\Sigma_{xy} \\ 0 & \sigma_y^2 - \Sigma_{xy}'\Sigma_{xx}^{-1}\Sigma_{xy} \end{bmatrix}^{-1} .$$

or

$$\begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}' & \sigma_y^2 \end{bmatrix}^{-1} = \begin{bmatrix} I & \Sigma_{xx}^{-1}\Sigma_{xy} \\ 0 & \sigma_y^2 - \Sigma_{xy}'\Sigma_{xx}^{-1}\Sigma_{xy} \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{xx}^{-1} & 0 \\ -\Sigma_{xy}'\Sigma_{xx}^{-1} & 1 \end{bmatrix} .$$

Thus it remains to find the inverse of the block matrix $\begin{bmatrix} I & \Sigma_{xx}^{-1}\Sigma_{xy} \\ 0 & \sigma_y^2 - \Sigma_{xy}'\Sigma_{xx}^{-1}\Sigma_{xy} \end{bmatrix}$. This inverse is the well known "backwards solve" in Gaussian elimination. Note that this inverse is given by

$$\begin{bmatrix} I & \Sigma_{xx}^{-1}\Sigma_{xy} \\ 0 & \frac{1}{\alpha} \end{bmatrix}^{-1} = \begin{bmatrix} I & -\alpha\Sigma_{xx}^{-1}\Sigma_{xy} \\ 0 & \alpha \end{bmatrix} ,$$

where we have made the definition of the scalar α such that $\frac{1}{\alpha} \equiv \sigma_y^2 - \Sigma_{xy}'\Sigma_{xx}^{-1}\Sigma_{xy}$. Using this result we have that

$$\begin{aligned} \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}' & \sigma_y^2 \end{bmatrix}^{-1} &= \begin{bmatrix} I & -\alpha\Sigma_{xx}^{-1}\Sigma_{xy} \\ 0 & \alpha \end{bmatrix} \begin{bmatrix} \Sigma_{xx}^{-1} & 0 \\ -\Sigma_{xy}'\Sigma_{xx}^{-1} & 1 \end{bmatrix} \\ &= \begin{bmatrix} \Sigma_{xx}^{-1} + \alpha\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{xy}'\Sigma_{xx}^{-1} & -\alpha\Sigma_{xx}^{-1}\Sigma_{xy} \\ -\alpha\Sigma_{xy}'\Sigma_{xx}^{-1} & \alpha \end{bmatrix} . \end{aligned} \tag{44}$$

Using this expression one of the required product in the exponential of $p(\mathbf{x}, y)$ is given by

$$\begin{aligned} \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}' & \sigma_y^2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x} - \mu_x \\ y - \mu_y \end{bmatrix} &= \begin{bmatrix} \Sigma_{xx}^{-1} + \alpha\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{xy}'\Sigma_{xx}^{-1} & -\alpha\Sigma_{xx}^{-1}\Sigma_{xy} \\ -\alpha\Sigma_{xy}'\Sigma_{xx}^{-1} & \alpha \end{bmatrix} \begin{bmatrix} \mathbf{x} - \mu_x \\ y - \mu_y \end{bmatrix} \\ &= \begin{bmatrix} (\Sigma_{xx}^{-1} + \alpha\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{xy}'\Sigma_{xx}^{-1})(\mathbf{x} - \mu_x) - \alpha\Sigma_{xx}^{-1}\Sigma_{xy}(y - \mu_y) \\ -\alpha\Sigma_{xy}'\Sigma_{xx}^{-1}(\mathbf{x} - \mu_x) + \alpha(y - \mu_y) \end{bmatrix} \\ &= \begin{bmatrix} d + \alpha\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{xy}'d - \alpha\Sigma_{xx}^{-1}\Sigma_{xy}(y - \mu_y) \\ -\alpha\Sigma_{xy}'d + \alpha(y - \mu_y) \end{bmatrix} . \end{aligned}$$

Where since the product $\Sigma_{xx}^{-1}(\mathbf{x} - \mu_x)$ appears a great number of times we defined it to be d, so $d \equiv \Sigma_{xx}^{-1}(\mathbf{x} - \mu_x)$. The computing the product needed to produce the quadratic term in the exponential of $p(\mathbf{x}, y)$ we get

$$\begin{aligned} (\mathbf{x}' - \mu_x', y - \mu_y) \begin{bmatrix} d + \alpha\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{xy}'d - \alpha\Sigma_{xx}^{-1}\Sigma_{xy}(y - \mu_y) \\ -\alpha\Sigma_{xy}'d + \alpha(y - \mu_y) \end{bmatrix} &= (\mathbf{x} - \mu_x)'d \\ &+ \alpha(\mathbf{x} - \mu_x)'\Sigma_{xx}^{-1}\Sigma_{xy}\Sigma_{xy}'d \\ &- \alpha(\mathbf{x} - \mu_x)'\Sigma_{xx}^{-1}\Sigma_{xy}(y - \mu_y) \\ &- \alpha(y - \mu_y)\Sigma_{xy}'d \\ &+ \alpha(y - \mu_y)^2 . \end{aligned}$$

Taking the transpose of either term we see that the third and fourth scalar products in the above expressions are equal. Combining these we get

$$(\mathbf{x} - \mu_x)'d + \alpha d'\Sigma_{xy}\Sigma_{xy}'d + \alpha(y - \mu_y)^2 - 2\alpha d'\,\Sigma_{xy}(y - \mu_y) .$$

Completing the square of the expression with respect to $y-\mu_y$ we have this expression given by

$$\alpha\left[(y-\mu_y)-d'\Sigma_{xy}\right]^2-\alpha(d'\Sigma_{xy})^2+\alpha d'\Sigma_{xy}\Sigma_{xy}'d+(\mathbf{x}-\mu_x)'d=\alpha\left[(y-\mu_y)-d'\Sigma_{xy}\right]^2+(\mathbf{x}-\mu_x)'d\,.$$

Thus using this and the definition of d we see that $p(\mathbf{x},y)$ is given by

$$p(\mathbf{x},y)=\frac{1}{(2\pi)^{\frac{p+1}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{x}})'\Sigma_{xx}^{-1}(\mathbf{x}-\mu_{\mathbf{x}})\right\}\exp\left\{-\frac{\alpha}{2}\left[y-\mu_y-d'\Sigma_{xy}\right]^2\right\} \quad (45)$$

From this expression we can derive $p(\mathbf{x})$ by integrating out y. This requires evaluating the following integral

$$\begin{aligned}\int\exp\left\{-\frac{\alpha}{2}\left[y-\mu_y-d'\Sigma_{xy}\right]^2\right\}dy &= \int\exp\left\{-\frac{\alpha}{2}y^2\right\}dy \\ &= \frac{1}{\sqrt{\alpha}}\int\exp\left\{-\frac{1}{2}y^2\right\}dy=\frac{\sqrt{2\pi}}{\sqrt{\alpha}}\,.\end{aligned}$$

Thus $p(\mathbf{x})$ is given by

$$p(\mathbf{x})=\frac{\sqrt{\alpha}}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{x}})'\Sigma_{xx}^{-1}(\mathbf{x}-\mu_{\mathbf{x}})\right\}\,.$$

To simplify this further consider the expression $\frac{\sqrt{\alpha}}{|\Sigma|^{\frac{1}{2}}}=\sqrt{\frac{\alpha}{|\Sigma|}}$. We will use Equation 43 to simplify this. Taking the determinant of both sides of that equation we find

$$|\Sigma_{xx}^{-1}|\begin{vmatrix}\Sigma_{xx} & \Sigma_{xy}\\ \Sigma_{xy}' & \sigma_y^2\end{vmatrix}=\sigma_y^2-\Sigma_{xy}'\Sigma_{xx}^{-1}\Sigma_{xy}=\frac{1}{\alpha}\,,$$

or solving for $\frac{\alpha}{|\Sigma|}$ we find $\frac{\alpha}{|\Sigma|}=\frac{1}{|\Sigma_{xx}|}$. Thus we finally obtain

$$p(\mathbf{x})=\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_{xx}|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_{\mathbf{x}})'\Sigma_{xx}^{-1}(\mathbf{x}-\mu_{\mathbf{x}})\right\}\,.$$

So we see that $p(\mathbf{x})$ is another multidimensional Gaussian this one with a mean of μ_x and a covariance matrix Σ_{xx} as the density for $\mathbf{x}$. Now that we have an explicit expression for $p(\mathbf{x})$ and using Equation 45 we can derive an *explicit* representation for $p(y|\mathbf{x})$. We find

$$\begin{aligned}p(y|\mathbf{x}) &= \frac{p(\mathbf{x},y)}{p(\mathbf{x})}\\ &= \frac{(2\pi)^{\frac{p}{2}}|\Sigma_{xx}|^{1/2}}{(2\pi)^{\frac{p+1}{2}}|\Sigma|^{1/2}}\exp\left\{-\frac{\alpha}{2}\left[y-\mu_y-d'\Sigma_{xy}\right]^2\right\}\\ &= \frac{\alpha}{(2\pi)^{\frac{1}{2}}}\exp\left\{-\frac{\alpha}{2}\left[y-\mu_y-d'\Sigma_{xy}\right]^2\right\}\\ &= \frac{1}{(2\pi)^{\frac{1}{2}}(\sigma_y^2-\Sigma_{xy}'\Sigma_{xx}^{-1}\Sigma_{xy})}\exp\left\{-\frac{1}{2}\frac{\left[y-\mu_y-(\mathbf{x}-\mu_{\mathbf{x}})'\Sigma_{xx}^{-1}\Sigma_{xy}\right]^2}{\sigma_y^2-\Sigma_{xy}'\Sigma_{xx}^{-1}\Sigma_{xy}}\right\}\,. \quad (46)\end{aligned}$$

The point of Equation 46 is that the distribution of $p(y|\mathbf{x})$ is a multivariate Gaussian with a mean given by

$$\mu_y+\Sigma_{xy}'\Sigma_{xx}^{-1}(\mathbf{x}-\mu_{\mathbf{x}})=\mu_y-\Sigma_{xy}'\Sigma_{xx}^{-1}\mu_{\mathbf{x}}+\Sigma_{xy}'\Sigma_{xx}^{-1}\mathbf{x}\,,$$

which is equivalent to what is given in the book, and a variance given by

$$\sigma_y^2 - \Sigma_{xy}' \Sigma_{xx}^{-1} \Sigma_{xy} \,,$$

this later result is somewhat different than what the book has and I believe there is typo in the book.

Problem Solutions

4.1 (the regression of $Soma$ on AVE, LIN, and $QUAD$)

Considering the three models given in the text plus the linear model of $Soma$ with predictors AVE, LIN, and $QUAD$, we can compare the various methods by computing some summary statistics such as the coefficient of determinism R^2 and the estimated variance of the residuals $\hat{\sigma}^2$. In the `R` script `chap_4_prob_1.R` we compute these statistics for each of these models. As expected, all models have identical summary statistics.

4.2 (the regression of y on x)

4.2.1: When we solve for x_i we find

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

4.2.2: The conditional distribution of x_i given y_i is given by

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

Thus equation in 4.2.1 (given by inverting the regression of y on x) will be equivalent to the mean of the conditional distribution $x|y$ if and only if

$$\rho_{xy} = \frac{1}{\rho_{xy}} \quad \text{or} \quad \rho_{xy}^2 = 1 \quad \text{or} \quad \rho_{xy} = \pm 1 \,.$$

This means that for this equivalence to hold x and y must be perfectly correlated, equivalently they are multiples of each other $x_i = c y_i$ for all i.

4.3 (linear regression on the transaction data)

4.3.1: The model M4 includes terms that are linearly dependent. The code that performs linear regression determines this and cannot determine coefficients for A and D. This is

because T1 and T2 are listed first, if they had been listed last the R code would have given NA values for the coefficients of these variables.

4.3.2/3: Since the coefficient β_i in a linear regression represent the amount of increase the dependent variable experiences when the variable X_i increases by one unit (all other variables held constant) we expect that the coefficients satisfy some relationships. Some simple ones are given below

$$\begin{aligned}
\beta_{11} &= \frac{1}{2}\beta_{32} + \beta_{42} \\
\beta_{21} &= \frac{1}{2}\beta_{32} - \beta_{42} \\
\beta_{32} &= \beta_{23} \\
\beta_{11} &= \beta_{43}\,.
\end{aligned}$$

From this we see that the values of some coefficients are the same while others are different between models. The coefficient of T_2 is different in M1 and M3 because in M1 it represents the change in Y holding T_1 constant, while in M3 it represents this change holding D *constant*. Since holding D constant actually places a restriction on how T_2 can change we don't expect the change in Y to be the same in both cases and hence the coefficients must be different.

4.5 (bootstrap confidence intervals of the fuel data)

We can use the provide R function `boot.case` to compute bootstrap samples from our data set and their corresponding regression coefficients. This function by default, computes a "linear model" on each bootstrapped sample determining the coefficients of the linear model. These coefficients make up the elements of the matrix that corresponds to the output of the `boot.case` function call. We can then compute an estimate of the 95% confidence interval of the coefficients by using the R function `quantile`. When we run the R script `chap_4_prob_5.R` we compute

```
      (Intercept)         Tax      Dlic       Income  logMiles
2.5%    -160.3563 -10.1112307 0.1031144 -0.009551768 -4.608278
97.5%    770.7148   0.5430165 0.7963736 -0.002627101 32.236726
```

From this we see that we can be relatively confident about the signs of only two variables: *Dlic* and *Income*.

4.6 (bootstrap confidence intervals of the windmill data)

For this problem we want to use the bootstrap to estimate the long-term average wind speed at the candidate site. Procedurally to do this we will draw B bootstrap samples from the provided `wm1` dataset, compute the linear regression coefficients corresponding to this

bootstrap sample and then predict the value of the variable $CSpd$ when $RSpd = 7.4285$ using the standard expression

$$\hat{\beta}_0 + \hat{\beta}_1 RSpd\,.$$

The discussion at the end of Problem 2.13.4 shows that the prediction of the average of m predictors is equal to the average the m predictions *and* the standard error of this estimate gets closer and closer to the standard error of the mean of the m predictors (a single predictor) as $m \to \infty$. Since the `R` function `boot.case` does not calculate the desired predictions at a fixed point we implement this prediction as a function. The `R` code for this function is given by

```
wwx_prediction_function <- function(m) {
    predict( m,
            newdata=data.frame(RSpd=7.4285),
            interval="prediction", level=0.95 )[1]
}
```

With this function defined we can now call the `boot.case` as

```
pred.boot <- boot.case( m, B=999, f=wwx_prediction_function ),
```

to get 999 bootstrapped predicted mean values in the vector `pred.boot`. The bootstrapped estimate of the long-term average wind speed is then given by taking the mean of this vector while the 95% confidence interval can be obtained by a call to the `quantiles` function. We find

```
> mean( pred.boot )
[1] 8.754117
> quantile( pred.boot, c(0.025,0.975) )
    2.5%    97.5%
8.613919 8.894611
```

See the `R` script `chap_4_prob_6.R` for the various parts of this problem.

4.7 (the effect of dropping terms in a linear regression)

If the *true* mean function is given by

$$E(Y|X_1 = x_1, X_2 = x_2) = 3 + 4x_1 + 2x_2\,. \tag{47}$$

Then the mean function observed just over the variable X_1 can be computed from the conditional expectation theorem

$$E(Y|X_1 = x_1) = E\left(E(Y|X_1 = x_1, X_2 = x_2)|X_1 = x_1\right)\,.$$

This may seem notationally confusing but it basically means that to derive the reduced expectation function we take the expectation of Equation 47 with respect to X_2. For general coefficients β_i this expectation is given by

$$E(\beta_0 + \beta_1 x_1 + \beta_2 X_2 | X_1 = x_1) = \beta_0 + \beta_1 x_1 + \beta_2 E(X_2 | X_1 = x_1)\,.$$

In this expression we see that the *correlation* between X_1 and X_2 expressed by the term $E(X_2|X_1 = x_1)$ is how the missing X_2 term affects our regression. Thus excluding variables cause more difficulties when they are highly correlated with the ones included in the regression. For the specific mean function given here, the explicit expression for $E(Y|X_1 = x_1)$ is

$$E(Y|X_1 = x_1) = 3 + 4x_1 + 3E(X_2|X_1 = x_1)\,.$$

If we have $E(X_2|X_1 = x_1) = \gamma_0 + \gamma_1 x_1$, so the value of X_2 depends linearly on the value of X_1 then this becomes

$$E(Y|X_1 = x_1) = (3 + \gamma_0) + (4 + \gamma_1)x_1\,.$$

The coefficient of x_1 will be negative if $\gamma_1 < -4$.

4.8 (the meaning of the sign of the regression coefficients)

If the true mean function is the one with predictors Sex and $Years$ then when we fit to the reduced mean function we will have a biased estimate. Namely the Sex only predictor should look like

$$E(Salary|Sex) = \beta_0 + \beta_1 Sex + \beta_2 E(Age|Sex)\,.$$

If the predictors Age and Sex are related such that $E(Age|Sex) = \gamma_0 + \gamma_1 Sex$, which would state that on average the age of a worker is γ_0, and the value of γ_1 represents is the correction to this average age experience by women since the average age of women is $\gamma_0 + \gamma_1$. Thus if $\gamma_1 < 0$ women are on average younger than men, while if $\gamma_1 > 0$ then they are on average older than men. In this case the estimated mean function would become

$$\begin{aligned} E(Salary|Sex) &= \beta_0 + \beta_1 Sex + \beta_2(\gamma_0 + \gamma_1 Sex) \\ &= \beta_0 + \gamma_0\beta_2 + (\beta_1 + \gamma_1\beta_2)Sex\,. \end{aligned}$$

So if $\beta_1 + \gamma_1\beta_2 < 0$, then the reduced model (with only Sex as the predictor) would show a negative coefficient for the variable Sex. This could happen if we expect β_2 to be very large (relative to β_1) due to a strong dependence of Age on salary (older worker earning more) and γ_1 were negative stating that on average women are younger than men. In that case one might have

$$\beta_1 + \gamma_1\beta_2 < 0\,,$$

and the reduced mean function might have a negative coefficient. If this were true it would mean that women appear to earn a lesser salary than men not because they are women but because on average they are *younger* than their male counterparts. A very different conclusion.

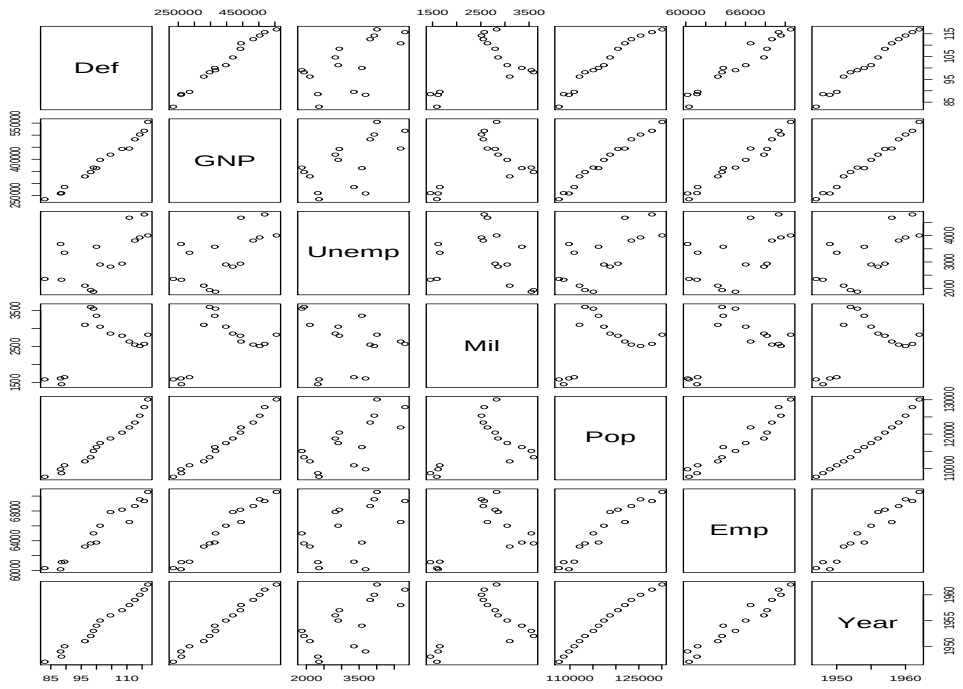

Figure 14: A scatterplot of the variables in the `longley` dataset of Problem 4.1. Note that several of the variables are very highly correlated.

4.9 (missing at random for the sleep data)

4.9.1: Missing at random would be reasonable if there was *not* a systematic reason why certain species did not have data.

4.9.2: Missing at random would not be reasonable if there *was* a systematics reason why certain species were not represented. For example, often very small or very large species have not be studied due to the difficult in doing so. If in fact it is these "extreme" species that are excluded the assumption of missing at random may not hold.

4.9.3: If these 62 species were a sampled in some specific way, this sampling would probably bias the sample of species away from the population mean and in this case the "missing" samples would be from the species not sampled in the original set of 62.

4.10 (the `longley` data)

4.10.1: See Figure 14 for a plot of the requested scatterplot.

4.10.2/3: Fitting a linear model using the default variables provided without modification in the `long` dataset we can extract the standard errors from the `summary` command. In addition, we implement the suggested simulation and compute the standard deviations of

the estimated coefficients. Displaying the standard errors first and the standard deviations of the bootstrap samples second we find

```
> print( s$coefficients[,2], digits=4 )
(Intercept)         GNP       Unemp         Mil         Pop         Emp
  8.187e+01   6.090e-05   8.190e-04   7.800e-04   6.447e-04   7.461e-04

> print( apply(b0,2,std), digits=4 )
(Intercept)         GNP       Unemp         Mil         Pop         Emp
  4.059e+01   2.989e-05   3.261e-04   3.077e-04   3.372e-04   2.919e-04
```

One would expect that if the regression is a robust one and not sensitive to rounding errors the standard deviations of the bootstrap samples would be smaller than the standard errors in the coefficients estimates. When we compare the two tables we see that the bootstrap samples are on the same order of magnitude as the standard error. This leads one to question the stability of the initial results.

See the `R` script `chap_4_prob_10.R` for implementations of various parts of this problem.

Chapter 5 (Weights, Lack of Fit, and More)

Notes On The Text

Testing for Lack of Fit, Variance Unknown

From the numbers given in the text we can calculate the lack-of-fit (LOF) sum-of-squares $\mathrm{SS}_{\mathrm{lof}}$, and degrees of freedom $\mathrm{df}_{\mathrm{lof}}$ from the sum-of-squares pure error (PE) as

$$\begin{aligned} \mathrm{SS}_{\mathrm{lof}} &= \mathrm{RSS} - \mathrm{SS}_{\mathrm{pe}} = 4.21266 - 2.3585 = 1.8581 \\ \mathrm{dof}_{\mathrm{lof}} &= n - p' - \mathrm{df}_{\mathrm{pe}} = 10 - 2 - 6 = 2\,, \end{aligned}$$

The F-test takes the ratio of the mean-square-error for the lack-of-fit or

$$\frac{\mathrm{SS}_{\mathrm{lof}}}{\mathrm{dof}_{\mathrm{lof}}} = \frac{1.8581}{2} = 0.92905\,,$$

to the mean-square-error of pure error or

$$\frac{\mathrm{SS}_{\mathrm{pe}}}{\mathrm{dof}_{\mathrm{pe}}} = \frac{2.3585}{6} = 0.3930\,.$$

This ratio is distributed as an F-distribution

$$\frac{\mathrm{SS}_{\mathrm{lof}}/\mathrm{dof}_{\mathrm{lof}}}{\mathrm{SS}_{\mathrm{pe}}/\mathrm{dof}_{\mathrm{pe}}} \sim F(\mathrm{dof}_{\mathrm{lof}}, \mathrm{dof}_{\mathrm{pe}})\,.$$

This ratio using the above numbers is given by 2.363, to be compared with the value of the F-distribution with 2 and 6 degrees of freedom that contains $1 - 0.05 = 0.95$ percent of the density. Using the `R` command `qf( 1-0.05, 2, 6 )` we find that the 5% threshold value is 5.143253, which heuristically means that the lack of fit sum-of-squares is small enought and that the given model should *not* be rejected because it "does not fit the data".

Problem Solutions

5.1 (Galton's sweet peas)

5.1.1: We plot a scatterplot of *Progeny* vs. *Parent* in Figure 15.

5.1.2: Often the motivation for using weighted-least-squares (WLS) is based on experimental setups where we are given m_i repeated measurements at a particular value of x_i. Then under the assumption that each measurement is an independent draw from the error distribution associated with our model we would expect the variance of an *average* response Y at x_i to be given by

$$\mathrm{Var}(\overline{Y}|X = x_i) = \frac{\sigma^2}{m_i}\,. \tag{48}$$

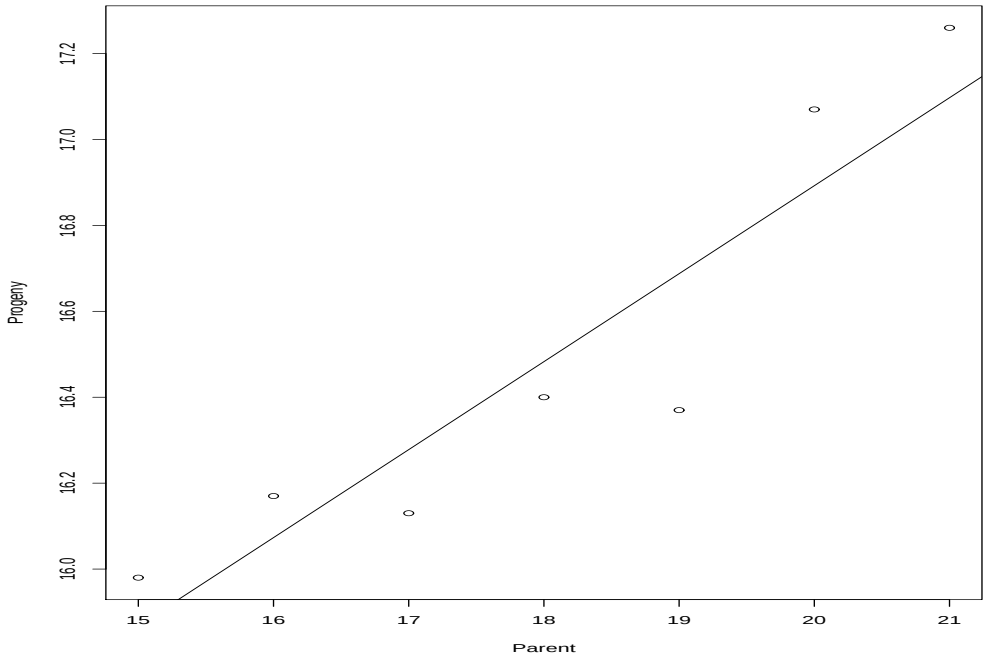

Figure 15: A scatterplot of the *Progeny* and *Parent* variables in the `galtonpeas` dataset of Problem 5.1 with a weighted least square fit line.

Dropping the bar notation and setting this equal to our definition of the regression weights w_i presented in this chapter during the introduction to weighted-least-squares of

$$\mathrm{Var}(Y|X = x_i) = \frac{\sigma^2}{w_i}\,, \tag{49}$$

we would have that $w_i = m_i$. For this problem we don't explicitly have values for the number of measurements m_i, but since we *know* the values for $\mathrm{Var}(Y|X = x_i)$ for each x_i value (we are given the standard deviation SD_i which we can square to get the variance) we can assign these to $\frac{\sigma^2}{w_i}$ as

$$\mathrm{Var}(Y|X = x_i) = \mathrm{SD}_i^2 = \frac{\sigma^2}{w_i}\,.$$

If we assume $\sigma^2 = 1$ as in the books example on the strong interaction force we see that w_i is given by

$$w_i = \frac{1}{\mathrm{SD}_i{}^2}\,. \tag{50}$$

We use the `weights` option to the `R` command `lm` to derive a least squares line.

5.1.3: To test the hypothesis that $\beta_1 = 1$ vs. the alternative that $\beta_1 < 1$ we can compute the t-statistic for the hypothesis that $\beta_1 = 1$. For the linear regression computed in this problem we find

$$t = \frac{\hat{\beta}_1 - 1}{\mathrm{se}(\hat{\beta}_1)} = -20.84137\,.$$

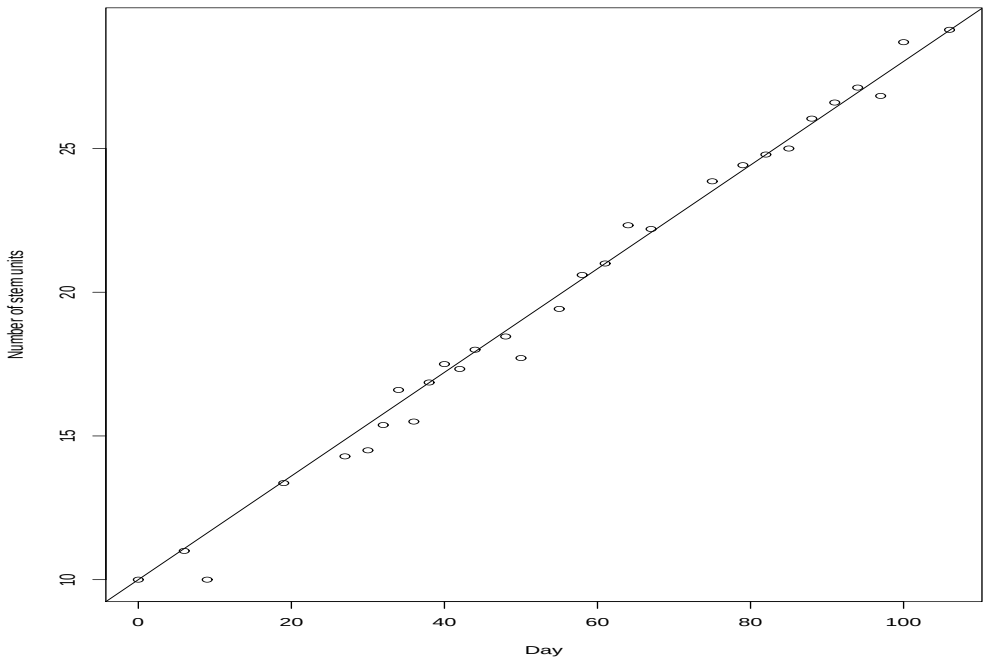

Figure 16: A scatterplot of the *ybar* and *Day* variables in the `shortshoots` dataset of Problem 5.2 with a weighted least square fit line.

This needs to be compared to the appropriate quantile of the t distribution with $n - p' = 5$ degrees of freedom. The p-value is the probability that we observe a t-statistics this negative or more negative by chance, when the null hypothesis is true and is given by the `R` call `pt( -20.84, 5 ) = 2.354991e-06`. Thus there is only a very small chance that this t-statistic happens "by chance" when $\beta_1 = 1$ and is therefore evidence *against* the hypothesis that this is the correct value. In fact, the estimated value from this data give $\hat{\beta}_1 = 0.20480$ with a standard-error of $\text{se}(\hat{\beta}_1) = 0.0381$ indicating heuristically that the value $\beta_1 = 1$ is not very likely.

5.2 (apple shoots)

5.2.1: See Figure 16 for a scatter plot of *ybar* vs. *Day* and the weighted (using the weights specified as in Equation 50) least-squares fit line. With or without weights a linear fit would appear to be a reasonable model for this data.

5.2.2: We fit both a weighted `wm` and an unweighted ûwm least squares models to this data. Edited results from the `summary` command are presented below.

```
> summary(wm)
Call:
lm(formula = ybar ~ Day, weights = 1/(SD^2))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.000e+01  3.805e-04 26283.8   <2e-16 ***
Day         1.803e-01  1.672e-03   107.8   <2e-16 ***
---

Residual standard error: 0.79 on 28 degrees of freedom
Multiple R-Squared: 0.9976,     Adjusted R-squared: 0.9975
F-statistic: 1.163e+04 on 1 and 28 DF,  p-value: < 2.2e-16

> summary(uwm)
Call:
lm(formula = ybar ~ Day)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.475879   0.200067   47.36   <2e-16 ***
Day         0.187238   0.003202   58.48   <2e-16 ***
---

Residual standard error: 0.5125 on 28 degrees of freedom
Multiple R-Squared: 0.9919,     Adjusted R-squared: 0.9916
F-statistic:  3420 on 1 and 28 DF,  p-value: < 2.2e-16
```

Some important points about these results. First the coefficient estimates for $\hat{\beta}$ are very similar between the two models as are the standard errors. There is a significant difference in the F-statistic however and the weighted least square model seems to be more certain than the unweighted model.

5.3 (nonparametric lack of fit)

5.3.1: This parametric bootstrap procedure is coded in the `R` file `param_bootstrap.R`.

5.3.2: The value of G obtained initially from the parametric and loess fit (using the data provided) has a value given by 15.32286. Whether this value is large or not can be determined using the parametric bootstrap samples. In the `R` script `chap_5_prob_3.R` we call the `R` function `param_bootstrap.R` with 10000 bootstrap samples and display the p-value that is generated. When we do this we find a p-value of 3×10^{-4}. This means that if the null hypothesis was true (that the linear parametric model is appropriate for this data) a value of G this large or larger would happen by chance only 0.03% of the time. The fact that this p-value is so small is an indication that the null-hypothesis should be rejected and maybe a

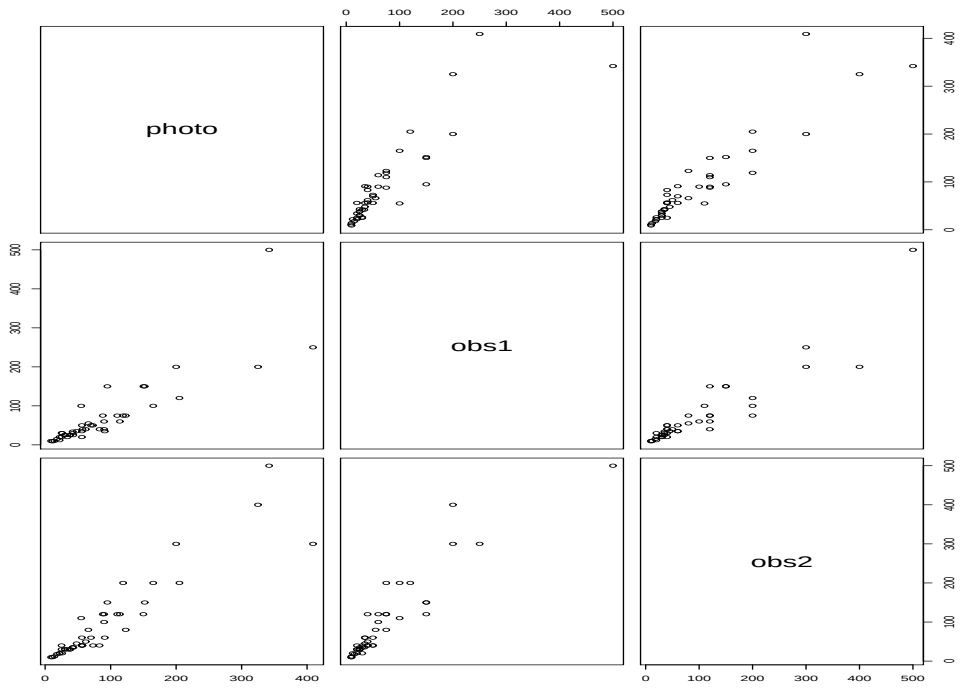

Figure 17: A scatterplot matrix of the variables *photo*, *obs*1, and *obs*2 in the `snowgeese` dataset of Problem 5.5.

regression more general than linear should be applied on this data.

5.5 (snow geese)

5.5.1: See the Figure 17 for an plot of the scatterplot matrix of the requested variables. This data looks like it could be nicely fit with a linear model but it appears that the variance in the response Y should *increase* with the value of the predictor (either *obs*1 or *obs*2). This would indicate the use that the appropriate technique to use is *weighted* least squares.

5.5.2: If we run the function from Problem 5.4 we obtain a p-value of 1.0, which means that we cannot reject the null hypothesis that the linear fit is appropriate for this data.

5.5.3: When we repeat the above exercise for $\sqrt{photo}$ and $\sqrt{obs1}$ we again get p-value near one indicating that it is not possible to reject the null hypothesis.

5.5.5: We use the `R` command `lm` to fit linear models of *photo* on *both* *obs*1 and *obs*2 and on each variable independently. When we do this a simple way to compare the three regressions is to look at the coefficient of determinism R^2 and the residual variance estimate $\hat{\sigma}^2$. For these three model (with the average and difference model) we find

$$\begin{aligned} obs1 &: R^2 = 0.7502 \quad \hat{\sigma}^2 = 1971.865 \\ obs2 &: R^2 = 0.8547 \quad \hat{\sigma}^2 = 1147.269 \end{aligned}$$

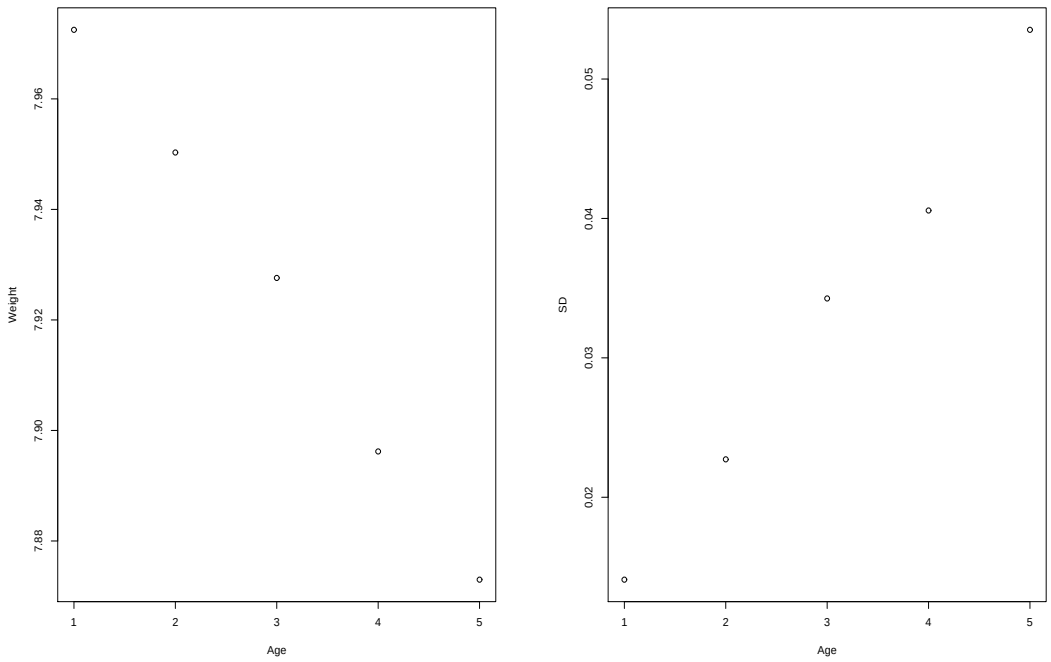

Figure 18: **Left:** A scatterplot of the variables $Weight$ vs. Age in the `jevons` coin dataset of Problem 5.6. As Age increases the average $Weight$ steadily decreases. **Right:** A scatterplot of the variables SD vs. Age for the `jevons` coin dataset of Problem 5.6. As Age increases so does the observed standard deviation of the weights of the coins at each age.

$$\begin{aligned} obs1 + obs2 &: \quad R^2 = 0.8558 \quad \hat{\sigma}^2 = 1165.103 \\ average + diff &: \quad R^2 = 0.8558 \quad \hat{\sigma}^2 = 1165.103\,. \end{aligned}$$

From these we see that in combining both prediction $obs1$ and $obs2$ to produce a prediction we obtain a larger R^2 and a $\hat{\sigma}^2$ that is close to the smallest seen. In addition, it looks like the second observer is better at counting geese than the first.

5.6 (Jevons' gold coins)

Warning: I'm not entirely sure that this result is correct, since the model lack-of-fit test when the variance is known seems to indicate that a straight line is not a good fit which seems counter intuitive for this data. Please email me if you see an error in this analysis.

5.6.1: See the Figure 18 for the two requested scatter plots.

5.6.2: Since the number of coin samples observed at each x_i (age) is relativly large we can assume that we *know* the standard deviation of the individual weights measurements at each age value. Since the data we are given represents the *average* of several individual coin

weights, we know that the variance of this average weight $\overline{Y}$ is given by

$$\text{Var}(\overline{Y}|X = x_i) = \frac{SD_i^2}{n_i}\,, \tag{51}$$

where n_i is the number of measurements taken with feature x_i (the same age) and SD_i is the sample standard deviation of these coin weights. Since weights, w_i, used in a weighted linear regression are defined as Equation 49 ($\text{Var}(Y|X = x_i) = \sigma^2/w_i$) we need to relate this expression to that given in Equation 51 above. To do this we could write

$$\text{Var}(\overline{Y}|X = x_i) = \frac{1}{\frac{n_i}{SD_i^2}}\,,$$

and assume that $\sigma^2 = 1$. By assuming that we *know* the value of σ^2 in this way we can use the results in the text about testing for lack of fit when the variance is known and the respective χ^2-tests. When we use the `lm` command with the `weights` option we get the following slightly edited `summary` output

```
Call:
lm(formula = Weight ~ Age, weights = 1/(n/SD^2))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.002701   0.005591 1431.35 7.52e-10 ***
Age         -0.026099   0.001292  -20.20 0.000265 ***
---

Residual standard error: 1.738e-05 on 3 degrees of freedom
Multiple R-Squared: 0.9927,     Adjusted R-squared: 0.9903
F-statistic: 408.2 on 1 and 3 DF,  p-value: 0.0002650

Analysis of Variance Table

Response: Weight
          Df     Sum Sq    Mean Sq F value    Pr(>F)
Age        1 1.2334e-07 1.2334e-07  408.23 0.0002650 ***
Residuals  3 9.0600e-10 3.0200e-10
```

What is a bit troubling is that the residuals sum square field displayed above is so much smaller than the hypothesized value of one. If we assume that there is nothing incorrect with these results we can still proceed as in the example in the text aimed at computing a lack-of-fit test based on the χ^2 statistic assuming the variance is *known*.

5.6.3/4: Since we have several measurements at the same value of x_i and we are assuming that we *know* the value of σ we can compute

$$X^2 = \frac{RSS}{\sigma^2} = \frac{9.0600\,10^{-10}}{1} = 9.0600\,10^{-10}\,.$$

This is to be compared with the χ^2 with $n-p' = 5-2 = 3$ degrees-of-freedom. We find using `pchisq( 9.0600e-10, 3 )` that the p-value for this test is $9.063702\,10^{-10}$, which indicates that the linear fit is not a good one? This seems to be in error.

5.6.5: The variance of a given weight will be given by two terms the first is the variance of the measurement about the mean and the variance of the mean due to a finite number of samples that went into its estimation. This means that to predict the weight W we would have

$$\mathrm{Var}(W|X=x) = SD_x^2 + \frac{SD_x^2}{n_x}\,,$$

and the expectation of W is given from the linear fit. For example when $x = 5$ (the oldest coin) to compute the probability that our weight is less than the legal minimum we would compute

$$P(W < 7.9379|X=5) = P\left(\frac{W - E(W|X=5)}{\sqrt{\mathrm{Var}(W|X=5)}} < \frac{7.9379 - \hat{\beta}_0 - 5\hat{\beta}_1}{\sqrt{SD_5^2 + \frac{SD_5^2}{n_5}}} \right).$$

where everything in the inequality in the right-hand-side is known. Since the random variable on the left-hand-side of this inequality is a standard-normal this expression can be evaluated with the `qnorm` command.

See the `R` script `chap_5_prob_6.R` for the various parts of this problem.

5.7 (π^- data)

For this dataset we can modify the `R` scripts provided with the book to load and perform analysis on the `physics1` dataset. We can evaluate the two linear models with predictors given by $s^{1/2}$ alone and $s^{1/2}$ with s. When we duplicate the commands presented in the book for this example we find that there is about a 7.2 % percent chance that the computed coefficient of $s^{1/2}$ is in fact zero and the observed value is obtained only by chance. The F-statistics for the entire model is 4.296 which has a p-value of 0.072 indicating that there is a 7.2 % chance of getting coefficients this large purly by chance. Fitting a larger power of $s^{1/2}$ does not seem to help this situation. When we include the terms s now the F-statistic falls to 1.918 with a p-value of 0.216 further indicating that there maybe no model of the form

$$E(y|s) = \beta_0 + \beta_1 s^{-1/2} + \beta_2 s\,,$$

that explains this data better than the mean of the data itself. When we look at the first and second order fits to this data with the data itself we obtain the plot in Figure 19. From that plot we see that there are two points in the interior of the domain that don't seem to fit the given curves very well. Heuristically, it is these points that make the linear and quadratic fits not agree well with the data.

See the `R` script `chap_5_prob_7.R` for the various parts of this problem.

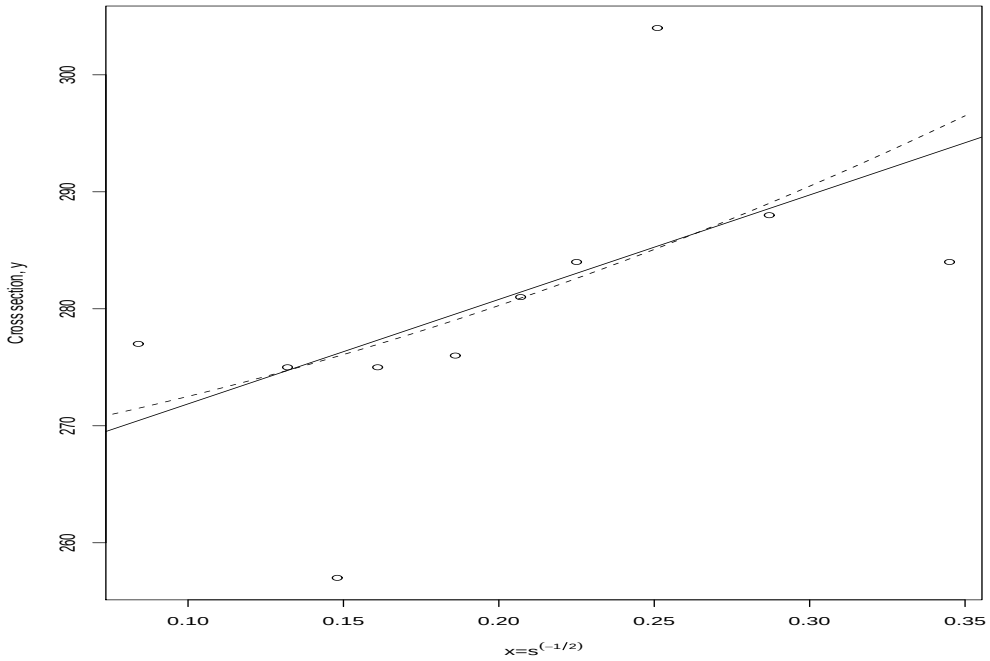

Figure 19: Scatterplot of the strong interaction data for π^- as in Problem 5.7.

Chapter 6 (Polynomials and Factors)

Notes On The Text

Notes on polynomials with several predictors

In this section of the book the claim is made that the `cake` data would not be fit fit with a model that did not include an interaction term. To verify or refute this claim in the `R` script `section_6_1_poly_w_several_predictors.R` we fit the `cake` data to a polynomial model that includes an interaction term

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \,, \tag{52}$$

and then to smaller/simpler model that does *not* have an interaction term of

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 \,. \tag{53}$$

Note we dropped the β_{12} coefficient in Equation 52 in obtain this later model. We then compute the F-statistic and p-value for both fits. We find that the p-value for the more detailed model is given by $5.864\,10^{-5}$ which states that there is strong evidence that the null-hypothesis should be rejected and that this model provides sufficient reduction in the variance. The less specific model given by Equation 53 has a p-value given by $8.913\,10^{-4}$ which still seems quite strong evidence against the null hypothesis. Now in both cases the null hypothesis is the constant average response prediction. If we run the `anova` command where we find

```
> anova(m2,m1)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + I(X1^2) + I(X2^2)
Model 2: Y ~ X1 + X2 + I(X1^2) + I(X2^2) + X1:X2
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1      9 4.2430
2      8 1.4707  1    2.7722 15.079 0.004654 **
```

This result compares the reduction in variance we obtain by using the more complicated model `m1` over the simpler model `m2`. The large value of the F-statistic and the small value of the p-value indicate that in fact this interaction term *does* provide a statistically significant reduction in uncertainty and thus provides additional information.

Notes using the delta method to estimate an extrema

Under the assumptions that $\hat{\theta} \sim N(\theta^*, \sigma^2 D)$ and a linear approximation to $g(\theta)$ of

$$g(\hat{\theta}) \approx g(\theta^*) + \dot{g}(\theta^*)'(\hat{\theta} - \theta^*) \,,$$

we have the *variance* of $g(\hat{\theta})$ given by

$$\begin{aligned} \mathrm{Var}(g(\hat{\theta})) &= \dot{g}(\theta^*)' \mathrm{Var}(\hat{\theta}) \dot{g}(\theta^*) \\ &= \dot{g}(\theta^*)' \left(\sigma^2 D\right) \dot{g}(\theta^*) \,. \end{aligned} \tag{54}$$

Here $\dot{g}(\theta)$ is the gradient of the scalar function $g(\cdot)$. In practice, the factors in the variance expression above are evaluated at $\hat{\theta}$ (rather than the unknown value of θ^*). As an example of the delta method recall that the OLS estimate of the location of a univariate minimum/maximum is given by $g(\hat{\beta}) = -\frac{\hat{\beta}}{2\hat{\beta}_2}$. From this we compute the gradient as

$$\dot{g}(\hat{\theta}) \equiv \begin{bmatrix} \frac{\partial}{\partial \hat{\beta}_0} \\ \frac{\partial}{\partial \hat{\beta}_1} \\ \frac{\partial}{\partial \hat{\beta}_2} \end{bmatrix} \left(-\frac{\hat{\beta}}{2\hat{\beta}_2} \right) = \begin{bmatrix} 0 \\ -\frac{1}{2\hat{\beta}_2} \\ \frac{\hat{\beta}_1}{2\hat{\beta}_2^2} \end{bmatrix} .$$

So that the variance of this function can be computed as

$$\begin{aligned} \mathrm{Var}\left(-\frac{\hat{\beta}}{2\hat{\beta}_2} \right) &= \begin{bmatrix} 0 & -\frac{1}{2\hat{\beta}_2} & \frac{\hat{\beta}_1}{2\hat{\beta}_2^2} \end{bmatrix} \begin{bmatrix} \mathrm{Var}(\hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_2) \\ \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_0) & \mathrm{Var}(\hat{\beta}_1) & \mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ \mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_0) & \mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_1) & \mathrm{Var}(\hat{\beta}_2) \end{bmatrix} \begin{bmatrix} 0 \\ -\frac{1}{2\hat{\beta}_2} \\ \frac{\hat{\beta}_1}{2\hat{\beta}_2^2} \end{bmatrix} \\ &= \begin{bmatrix} 0 & -\frac{1}{2\hat{\beta}_2} & \frac{\hat{\beta}_1}{2\hat{\beta}_2^2} \end{bmatrix} \begin{bmatrix} -\frac{1}{2\hat{\beta}_2}\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) + \frac{\hat{\beta}_1}{2\hat{\beta}_2^2}\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_2) \\ -\frac{1}{2\hat{\beta}_2}\mathrm{Var}(\hat{\beta}_1) + \frac{\hat{\beta}_1}{2\hat{\beta}_2^2}\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) \\ -\frac{1}{2\hat{\beta}_2}\mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_1) + \frac{\hat{\beta}_1}{2\hat{\beta}_2^2}\mathrm{Var}(\hat{\beta}_2) \end{bmatrix} \\ &= \frac{1}{4\hat{\beta}_2^2}\mathrm{Var}(\hat{\beta}_1) - \frac{\hat{\beta}_1}{4\hat{\beta}_2^3}\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) - \frac{\hat{\beta}_1}{4\hat{\beta}_2^3}\mathrm{Cov}(\hat{\beta}_2, \hat{\beta}_1) + \frac{\hat{\beta}_1^2}{4\hat{\beta}_2^4}\mathrm{Var}(\hat{\beta}_2) \\ &= \frac{1}{4\hat{\beta}_2^2}\left(\mathrm{Var}(\hat{\beta}_1) - 2\frac{\hat{\beta}_1}{\hat{\beta}_2}\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2) + \frac{\hat{\beta}_1^2}{\hat{\beta}_2^2}\mathrm{Var}(\hat{\beta}_2) \right) , \end{aligned} \tag{55}$$

which is the books Equation 6.13.

Problem Solutions

6.1 (Cake data)

6.1.1: In the `R` file `chap_6_prob_1.R` we use the `R` function `lm` to estimate the coefficients β in the `cake` model Equation 52. We find a linear model summary given by

```
Call:
lm(formula = 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
```

Every coefficient β appears significant. The "weakest" coefficient appears to be that of β_{12} which has a p-value of 0.004654, still less than the proposed 0.005.

6.2.1: The optimal (X_1, X_2) combination under a model like Equation 52 will require

$$\frac{\partial}{\partial X_1}E(Y|X_1, X_2) = \frac{\partial}{\partial X_2}E(Y|X_1, X_2) = 0\,,$$

when evaluated at the point $(\tilde{X}_1, \tilde{X}_2)$. From the given expression for the `cake` data set we have

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{12}x_1 x_2\,.$$

Thus the required equations to solve become

$$\begin{aligned}\frac{\partial}{\partial X_1}E(Y|X_1, X_2) &= \beta_1 + 2\beta_{11}x_1 + \beta_{12}x_2 = 0 \\ \frac{\partial}{\partial X_2}E(Y|X_1, X_2) &= \beta_2 + 2\beta_{22}x_2 + \beta_{12}x_1 = 0\,.\end{aligned}$$

These two equations are equal to the linear system

$$\begin{aligned}2\beta_{11}x_1 + \beta_{12}x_2 &= -\beta_1 \\ \beta_{12}x_1 + 2\beta_{22}x_2 &= -\beta_2\,.\end{aligned}$$

We can solve these for x_1 and x_2 using Crammer's rule. The determinant, D, of the coefficient system is given by

$$D = \begin{vmatrix} 2\beta_{11} & \beta_{12} \\ \beta_{12} & 2\beta_{22} \end{vmatrix} = 4\beta_{11}\beta_{22} - \beta_{12}^2\,.$$

So that Crammer's rule gives for the desired estimates

$$\begin{aligned}\tilde{X}_1 &= \frac{1}{D}\begin{vmatrix} -\beta_1 & \beta_{12} \\ -\beta_2 & -\beta_{22} \end{vmatrix} = \frac{-2\beta_1\beta_{22} + \beta_2\beta_{12}}{4\beta_{11}\beta_{22} - \beta_{12}^2} \\ \tilde{X}_2 &= \frac{1}{D}\begin{vmatrix} 2\beta_{11} & -\beta_1 \\ \beta_{12} & -\beta_2 \end{vmatrix} = \frac{-2\beta_{11}\beta_2 + \beta_2\beta_{12}}{4\beta_{11}\beta_{22} - \beta_{12}^2}\,.\end{aligned}$$

Now *estimates* of these values would be obtained by evaluating the above expressions at the ordinary least squares estimates $\hat{\beta}$. The standard error of these two expressions will be given by the *delta method* since $\tilde{X}_i$ is a *nonlinear* function of β. Our estimated coefficients $\hat{\beta}$ satisfy

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})\,,$$

where β is the *true* regression coefficients. To compute the standard errors of $\tilde{X}_i$ we need to first compute the gradient of each of these expression with respect to all of the β's in our linear regression. We will do this procedure for $\tilde{X}_1$ only since the derivation will be similar for $\tilde{X}_2$. In the `R` script `chap_6_prob_1.R` we compute these derivatives "by hand" using the `R` command `D` and then compute the square-root of the required inner product in Equation 54 to determine the standard error of $\tilde{X}_1$. We can also very simply do this entire procedure using the `alr3` toolbox function `delta.method`. When we use either of these methods we get

$$\tilde{X}_1 \sim N(179.0288, 6502.26)\,,$$

where 6502.26 is the delta method's estimate of the variance of $\tilde{X}_1$.

6.1.3: To incorporate the block factor terms we would want to add a factor, B, (for block) so that our mean function then becomes

$$E(Y|X_1 = x_1, X_2 = x_2, B = j) = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{11j}x_1^2 + \beta_{22j}x_2^2 + \beta_{12j}x_1x_2\,,$$

for $j = 1, 2$. This would be the most general way of adding a factor an allows for block by term interactions since each affect linear, quadratic, or interaction is allowed to have its own coefficient that depends on the block. A difficulty with this procedure is that since in each block there are only *seven* samples of the required input/output pairs. Since the dimensionality is so small one cannot expect to be able to estimate well these parameters. To hopefully improve on this situation we'll begin with an even simpler model given by

$$E(Y|X_1 = x_1, X_2 = x_2, B = j) = \beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2 + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{12}x_1x_2\,,$$

Thus we hypothesis that the quadratic terms are *not* affected by the block while the linear terms *are*. This might be approximately true if the quadratic terms are a higher order approximation to the predominately linear response. An alternative way to write this later expression is with block dummy variables U_j as

$$E(Y|X_1 = x_1, X_2 = x_2, B = j) = \sum_{j=1}^{2}(\beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2)U_j + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{12}x_1x_2\,,$$

To fit such a model in `R` we would could use the command

```
Y ~ -1 + B + B:X1 + B:X2 + I(X1^2) + I(X2^2) + X1:X2
```

where B has been declared a factor based on the variable `cakes$block`. Then using this model and the model computed above without block interactions we can use the `anova` command to compare the more specific model for significance. We find

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

Model 1: Y ~ X1 + X2 + I(X1^2) + I(X2^2) + X1:X2
Model 2: Y ~ -1 + B + B:X1 + B:X2 + I(X1^2) + I(X2^2) + X1X2
```

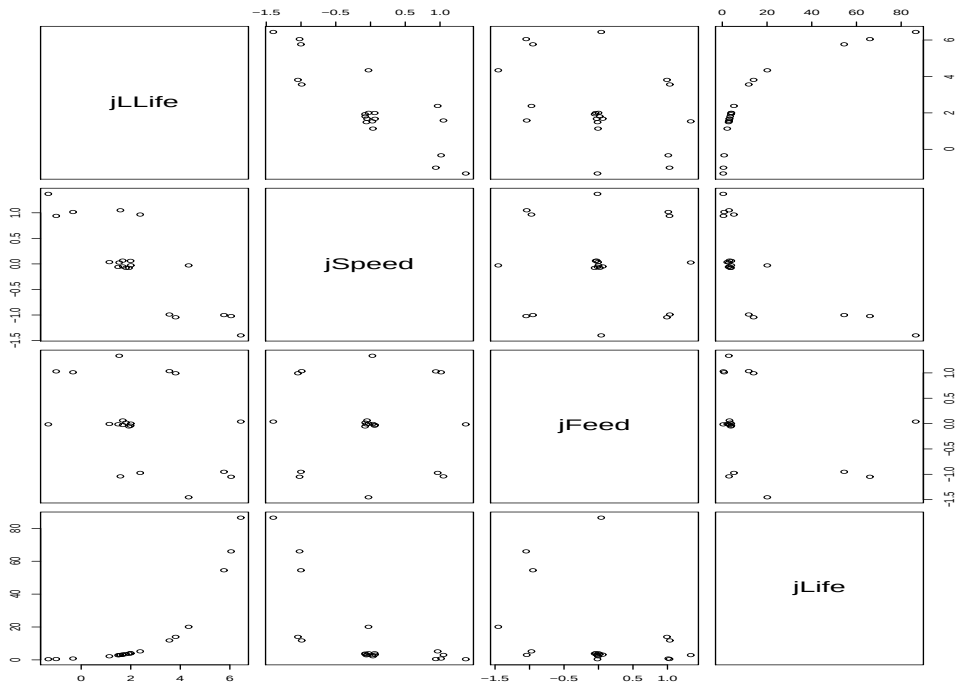

Figure 20: A scatterplot matrix of "jittered" versions of $log(Life)$, $Speed$, $Feed$, and $Life$ from the `lathe1` data set of Problem 6.2.

```
  Res.Df     RSS Df Sum of Sq     F Pr(>F)
1      8 1.47074
2      5 0.48409  3   0.98666 3.397 0.1106
```

The F-statistic and the p-test indicate that the additional complexity in moving to the block estimated model is probably not justified.

6.2 (the lathe data set)

6.2.1: In Figure 20 we plot a scatterplot matrix of the requested variables. From this plot if we view the $Speed$ and $Feed$ variables we see that the experimental design presented varied these two variables in a circle. Taking the logarithm seems to make the projected dependencies of $LLife\ Speed$ and $LLife\ Feed$ look more circular.

6.2.2: We fit a model like Equation 52 (with an interaction) to the variable $log(Life)$. When we look at the summary results from `R` we obtain

```
Call:
lm(formula = LLife ~ Speed + Feed + I(Speed^2) + I(Feed^2) +
    Speed:Feed, data = lathe1)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7141     0.1516  11.307 2.00e-08 ***
Speed        -2.2925     0.1238 -18.520 3.04e-11 ***
Feed         -1.1401     0.1238  -9.210 2.56e-07 ***
I(Speed^2)    0.4156     0.1452   2.863 0.012529 *
I(Feed^2)     0.6038     0.1452   4.159 0.000964 ***
Speed:Feed   -0.1051     0.1516  -0.693 0.499426
---
Residual standard error: 0.4288 on 14 degrees of freedom
Multiple R-Squared: 0.9702,     Adjusted R-squared: 0.9596
F-statistic: 91.24 on 5 and 14 DF,  p-value: 3.551e-10
```

Notice that the interaction term $Speed : Feed$ does not have a very large t-value is not very significant (given the other terms). This indicates that perhaps a simpler model could be used where this interaction term was removed.

6.2.3: We can *test* whether the interaction term provides benefit by constructing a model without it and using a F-test to determine if the reduction in RSS is significantly larger than what we would expect randomly. When we fit this reduced model and compare with the original we get an `anova` result of

```
> anova(m0,m1)
Analysis of Variance Table

Model 1: LLife ~ Speed + Feed + I(Speed^2) + I(Feed^2)
Model 2: LLife ~ Speed + Feed + I(Speed^2) + I(Feed^2) + Speed:Feed
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     15 2.66235
2     14 2.57396  1   0.08839 0.4807 0.4994
```

The p-value above indicates that the addition of the interaction term does not significantly reduce the residual variance and it should probably be left out. Qualitatively we can test this by viewing projections of the regression surface. We can fix values of $Speed$ and vary $Feed$ and then exchange the rolls of the two variables. When do do this for both models we get Figure 21. Visually there is not much difference in the curves, an indication that the interaction term does not provide much of an effect.

6.2.4: This part could be done as the example in the book or using the `alr3` code `delta.method`.

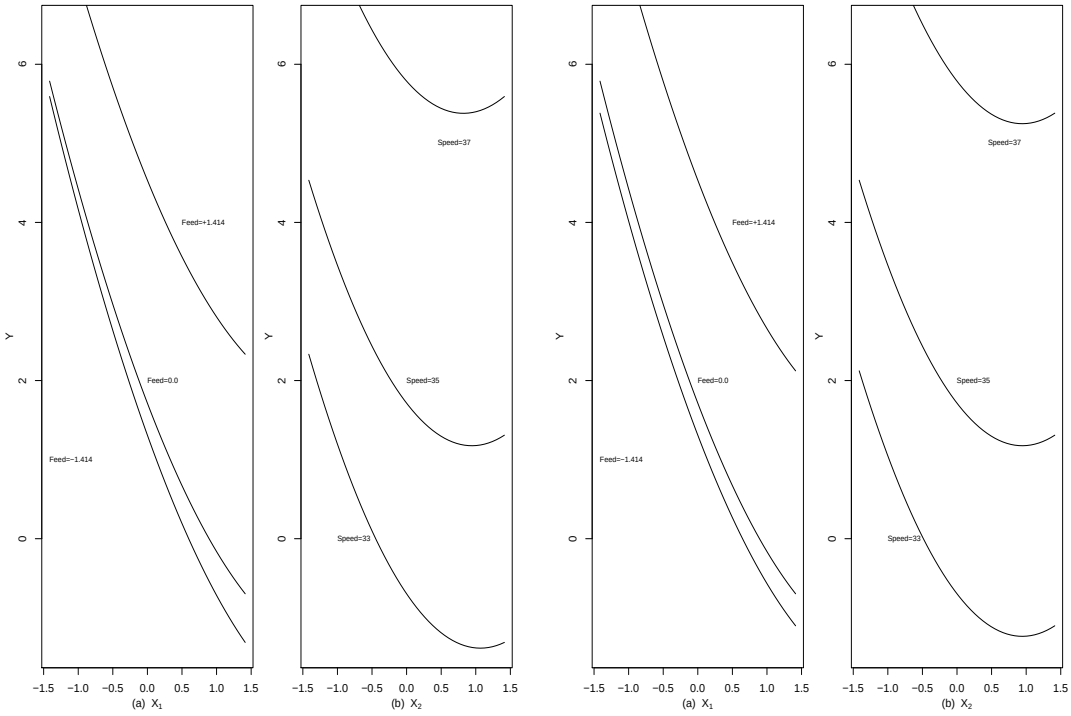

Figure 21: **Left:** The left panel contains plots of $E(log(Life)|Speed, Feed)$, for three fixed values of $Feed$ as a function of $Speed$. The right panel contains plots of $E(log(Life)|Speed, Feed)$, for three fixed values of $Speed$ as a function of $Feed$. Both plots assume an interaction term in the model for computing $E(log(Life)|Speed, Feed)$. **Right:** The same thing but assuming that the model does *not* contain an interaction term.

	Expression: $E(Y\|X=x,F=j)$	R expression
Model 1: most general	$\eta_0+\eta_1 x+\sum_{j=2}^{d}(\eta_{0j}+\eta_{1j}x)U_j$	`Y ~ X + F + F:X`
Model 2: parallel	$\eta_0+\eta_1 x+\sum_{j=2}^{d}\eta_{0j}U_j$	`Y ~ X + F`
Model 3: common intercept	$\eta_0+\eta_1 x+\sum_{j=2}^{d}\eta_{1j}xU_j$	`Y ~ X + F:X`
Model 3': diff	$\beta_0+\sum_{j=1}^{d}\beta_{1j}xU_j$	`Y ~ +1 + F:X`
Model 4: all the same	$\eta_0+\eta_1 x$	`Y ~ X`

Table 1: ANOVA model comparison for the four models for Problem 6.4. Note that most models (all but Model 3') are expressed relative to a common linear fit $\eta_0+\eta_1 x$.

6.4 (the twins data set)

For this problem we want to perform the regression of IQf onto IQb and C, where C is a factor. The most general model of this type is where we assume that each factor has its own parameters and is given by

$$E(IQf|IQb=x,C=j)=\eta_0+\eta_1 x+\sum_{i=2}^{3}(\eta_{0i}+\eta_{1i}x)U_i\,.$$

Various model simplifications happen if we assume that some of the parameters are common among factors. This problem is similar to the example in the book that uses the sleep data where there are a total of four models to consider. To solve this problem we will fit each of these models using the `R` command `lm` and compare the different models using F-statistics based off the most general model (denoted as model # 1) and using the F_l statistic defined by

$$F_l=\left(\frac{RSS_l-RSS_1}{(df_l-df_1)}\right)/\left(\frac{RSS_1}{(df_l-df_1)}\right)\,,$$

as compare to the quantiles of an $F(df_l-df_1,df_1)$ distribution. If the most general model with a residual sum of squares RSS_1 is not sufficiently smaller than the more restrictive model with RSS_l for $l=2,3,4$ the value of F_l (defined above) will be small and the more general model will probability not be needed.

Because the of analysis of a single predictor, X, that maybe affected by several factors, F, is very common and will be used in the problems in Table 1 we present the explicit regression expressions for each of four models that are possible. Because of the different ways to represent the same linear model we will present two forms for the third model. In addition, next to these expressions, we will present the `R` argument to the `lm` command that will compute the various coefficients in the given model. As notation, we have a predictor X of a response Y and d distinct factors denoted by F. The variable U_j is defined as in the book.

When we compute linear regressions on these four model we obtain Table 2. Each F-statistic compares the most general model with the alternative model in question. We see that there is a 60% chance that the reduction in RSS from using the more complicated model is due to chance.

	df	RSS	F	P(>F)
Model 1: most general	21	1317.47		
Model 2: parallel	23	1318.40	$7.42\,10^{-3}$	$9.92\,10^{-1}$
Model 3: common intercept	23	1326.46	$7.16\,10^{-2}$	$9.31\,10^{-1}$
Model 4: all the same	25	1493.53	0.702	0.6

Table 2: ANOVA model comparison for the four models for Problem 6.4.

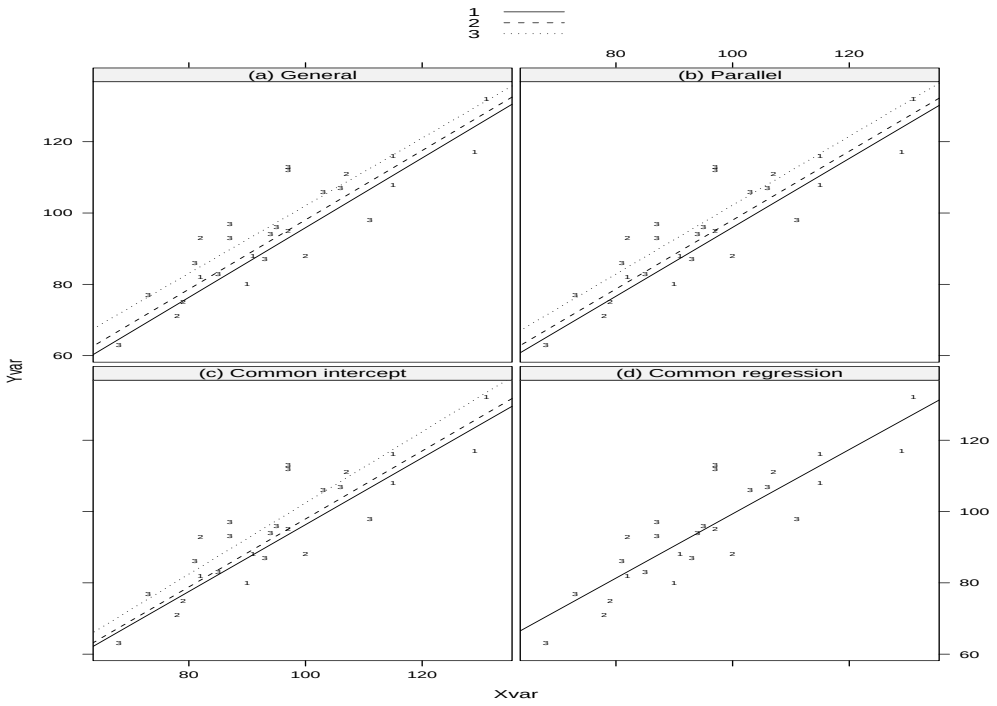

Figure 22: A visual display of the four models from Problem 4. There does not seem to be much variation among the classes in this data set.

	df	RSS	F	P(>F)
Model 1: most general	44	6.004		
Model 2: parallel	45	6.055	0.373	0.544
Model 3: common intercept	45	6.058	0.397	0.532
Model 4: all the same	46	6.091	0.320	0.728

Table 3: ANOVA model comparison for the four models for Problem 6.5.

Finally, a quadrant plot like presented in the book for each of the four models is presented in Figure 22. Visually, this confirms our F-statistics above in that the linear models produced under all variations don't seem significantly different.

This problem is worked in the `R` script `chap_6_prob_4.R`.

6.5 (model variation with temperature)

For this problem we will fit $E(\log(Pressure)|Temp) = \beta_0 + \beta_1 Temp$ for two data sources: Forbes' data and Hooker's data. If we lump the data together in one data frame we can consider the introduction of a factor that corresponds to *which* data set the source data came from. Then under this framework we can use hypothesis testing to determine which of the most general models

$$E(\log(Pressure)|Temp = t, D = j) = \beta_{0j} + \beta_{1j} t\,,$$

is most likely. This problem then is equivalent to example in the book where we compare regression lines within each factor and determine the least general model that explains the data. In the `R` script `chap_6_prob_5.R` we fit the same general models as discussed in Exercise 6.4 on Page 67. When we run that code we produced the anova table shown in Table 3. The fact that the F-statistics are so small and the p-values are so large indicates that there is *no* difference in mean functions.

This is very evident when we graphically visit the four possible mean functions. In Figure 23 we plot these four mean functions.

6.6 (characteristics of $HT18$ and $HT9$ on whether the subject is male or female)

For this problem we will determine if the mean function $E(HT18|HT9 = x, Sex = j)$ depends on in a statistically significant way on the variable Sex. This problem is similar to Problems 6.4 and 6.5 and is worked in the `R` script `chap_6_prob_6.R`.

6.6.1: In Figure 24 we display the scatter plot of $HT18$ versus $HT9$ using the character "m" and "f" for male and female samples. A linear fit seems reasonable.

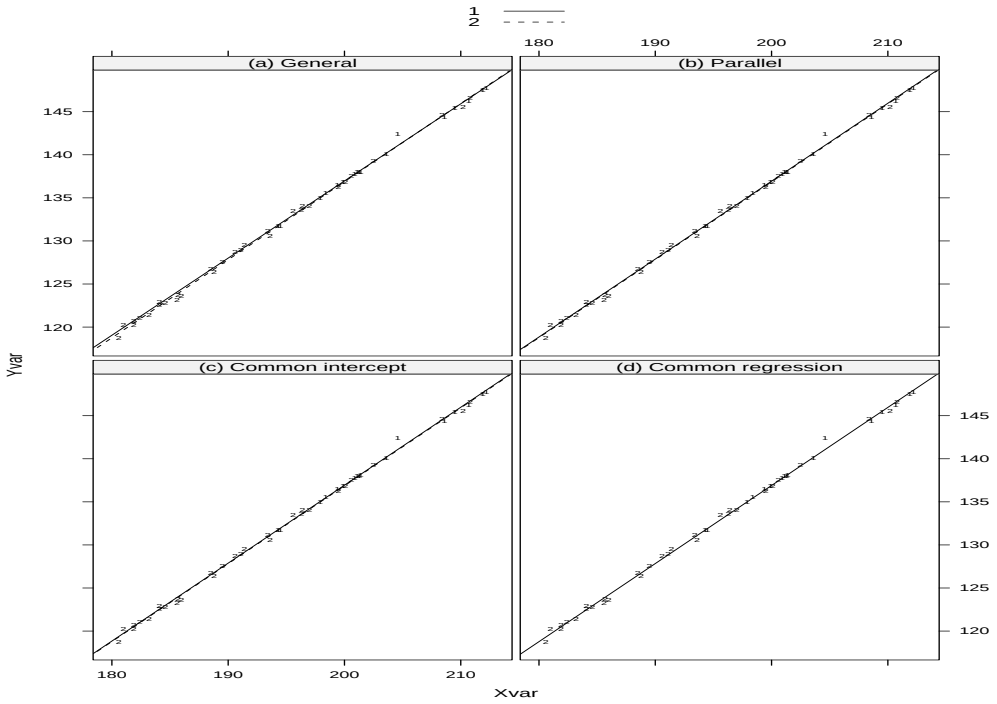

Figure 23: A visual display of the four models from Problem 5. There seems to be almost no variation among the hypothesized two classes in this data set.

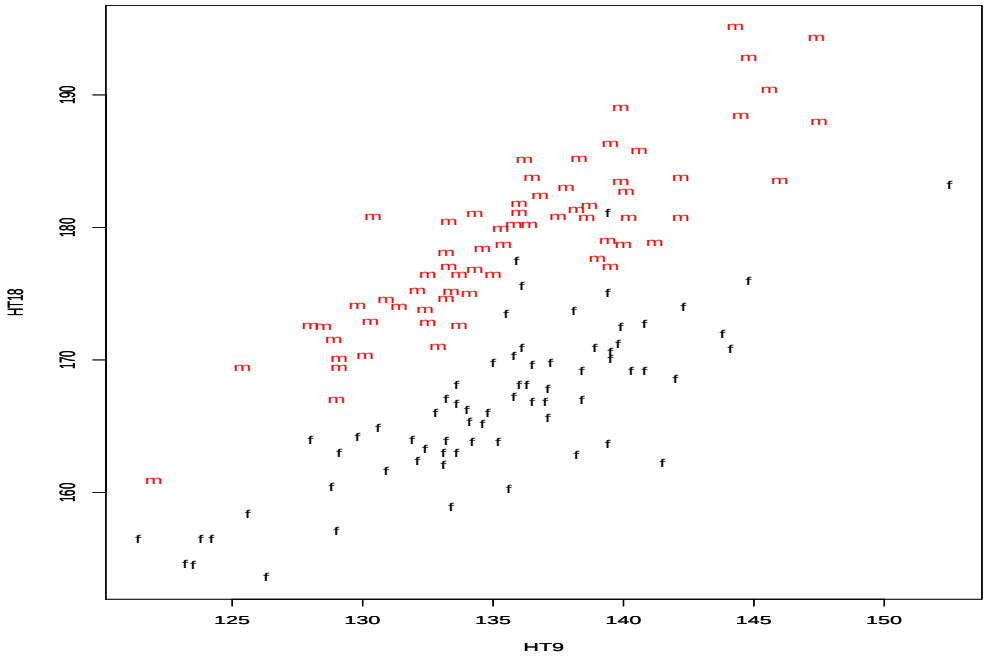

Figure 24: The scatter plot requested by Problem 6. There is an obvious difference between the intercept of the two lines. We can use statistical hypothesis testing to determine the significance of any model slope differences.

	df	RSS	F	P(>F)
Model 1: most general	132	1532		
Model 2: parallel	133	1567	2.964	$8.7\,10^{-2}$
Model 3: common intercept	133	1542	0.8394	0.3612
Model 4: all the same	134	6191	200.6	$9.55\,10^{-41}$

Table 4: ANOVA model comparison for the four models for Problem 6.6.

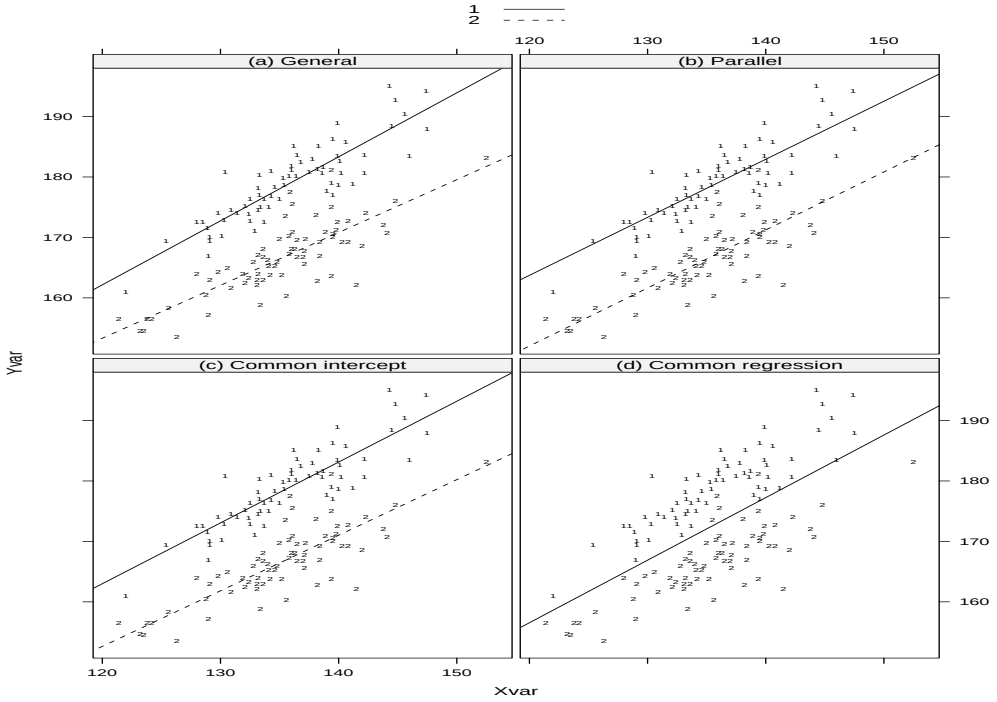

Figure 25: A visual display of the four models from Problem 6.

6.6.2: In Table 4 we produce the ANOVA table that compares the most general model to each of the three remaining restricted models. Each F-statistic is a comparison between the most general model one the each less specific model. There is a 36% chance that the difference in RSS between the most general model and the common intercept model is due to chance. That is the only model that we should consider as a potential simpler candidate model.

This is very evident when we graphically visit the four possible mean functions. In Figure 23 we plot these four mean functions.

6.7 (generalizations to linear models with factors and two predictors)

For an expanded mean function like $E(HT18|HT2 = x_1, HT9 = x_2, Sex = j)$, where we have two continuous predictors the most general model would allow every level to have a different

slope and intercept. This would be expressed using dummy variables mathematically as

$$E(HT18|HT2 = x_1, HT9 = x_2, Sex = j) = \sum_{j=1}^{2}(\beta_{0j} + \beta_{1j}x_1 + \beta_{2j}x_2)U_j\,.$$

Some simplifications of this model would include the pooling (assuming a common value for all factors) of the coefficients β_0, β_1, or β_2. For example, the common mean (the common intercept) model would be given by

$$E(HT18|HT2 = x_1, HT9 = x_2, Sex = j) = \beta_0 + \sum_{j=1}^{2}(\beta_{1j}x_1 + \beta_{2j}x_2)U_j\,.$$

One could also pool samples assuming a common value for β_1. This mean function would look like

$$E(HT18|HT2 = x_1, HT9 = x_2, Sex = j) = \beta_1 x_1 + \sum_{j=1}^{2}(\beta_{0j} + \beta_{2j}x_2)U_j\,.$$

Other variations are also possible.

6.8 (are the regression surfaces parallel for boys and girls)

The suggested hypothesis that we desire to test is to compare the null hypothesis (that there is no "slope" difference between boys and girls) written mathematically as

$$E(HT18|HT2 = x_1, HT9 = x_2, Sex = j) = \eta_0 + \eta_1 x_1 + \eta_2 x_2 + \eta_{02}U_2\,,$$

with the alternative hypothesis that there is a difference. Written mathematically as

$$E(HT18|HT2 = x_1, HT9 = x_2, Sex = j) = \eta_0 + \eta_1 x_1 + \eta_2 x_2 + (\eta_{02} + \eta_{12}x_1 + \eta_{22}x_2)U_2\,.$$

This later model does allow the possibility that they have different intercepts. We fit these two hypothesis using the two `R` command (where `D` has been previously declared a factor

```
NH: HT18 ~ HT2 + HT9 + D
AH: HT18 ~ HT2 + HT9 + D + D:HT2 + D:HT9
```

We can then determine which is the better model using the `anova` command. The output from this is given by

```
> anova(m0,m1)
Analysis of Variance Table

Model 1: HT18 ~ HT2 + HT9 + D
Model 2: HT18 ~ HT2 + HT9 + D + D:HT2 + D:HT9
  Res.Df    RSS  Df Sum of Sq      F  Pr(>F)
1    132 1565.6
2    130 1496.9   2      68.7 2.9831 0.05412 .
```

From which we see that the there is only a 5% chance that this reduction in RSS is due to chance. This indicates that the reduction in RSS *is* significant and we should consider the planes to not be parallel.

This problem is worked in the `R` script `chap_6_prob_8.R`.

6.9 (apple shoots)

6.9.1: To calculate the mean-square expression for pure-error, recall that the pure-error sum of squares is given by

$$\mathrm{SS}_{\mathrm{pe}} = \sum_i (n_i - 1)\mathrm{SD}_i^2 \,, \tag{56}$$

where n_i is the number of repeated experiments with a fixed value of predictor x_i, and SD_i is the unbiased sample standard deviation of these repeated experiments. Also recall that the degrees of freedom of the pure-error is given by

$$\mathrm{df}_{\mathrm{pe}} = \sum_i (n_i - 1) \,. \tag{57}$$

Then an estimate of the pure-error is given by the ratio of these two expressions or

$$\hat{\sigma}^2_{\mathrm{pe}} = \frac{\mathrm{SS}_{\mathrm{pe}}}{\mathrm{df}_{\mathrm{pe}}} \,. \tag{58}$$

For this problem, we will estimate $\hat{\sigma}^2_{\mathrm{pe}}$ for both the long and short shoots from the given data set (using the above formula) and compute another F-statistic in this case given by

$$\frac{\hat{\sigma}^2_{\mathrm{pe,long-shoots}}}{\hat{\sigma}^2_{\mathrm{pe,short-shoots}}} \,,$$

which under the hypothesis that the two distributions have the *same* variance should be distributed as a $F(\mathrm{df}_{\mathrm{pe,long-shoots}}, \mathrm{df}_{\mathrm{pe,short-shoots}})$ distribution. We can use this F-statistics to determine if the assumption of equal pure-error between the shoot types is true. When we compute these two expressions for the short and long shoots we find

$$\frac{\hat{\sigma}^2_{\mathrm{pe,long-shoots}}}{\hat{\sigma}^2_{\mathrm{pe,short-shoots}}} = 1.807 \,,$$

which is approximately two. This is to be compared with the value of the F-distribution with $\mathrm{df}_{\mathrm{pe,long-shoots}} = 167$ and $\mathrm{df}_{\mathrm{pe,short-shoots}} = 292$ degrees of freedom that contains $1 - 0.05 = 0.95$ percent of the density. Using the `R` command `qf( 1-0.05, 167, 292)` we find that the 5% threshold value is 1.2487, which heuristically means that the lack of fit sum-of-squares F-statistic is too large and assumes that the given model does *not* fit and should be rejected because it "does not fit the data". Incidentally, we would get a value of this ratio as large or larger than 1.807 only `1 - pf( F, dfLong, dfShort )` or $5.14\,10^{-6}$ percent of the time.

Our pooled estimate, under the assumption that the long-shoots will have a standard deviation twice that of the short shoots in terms of the pure-error sum of squares will be given

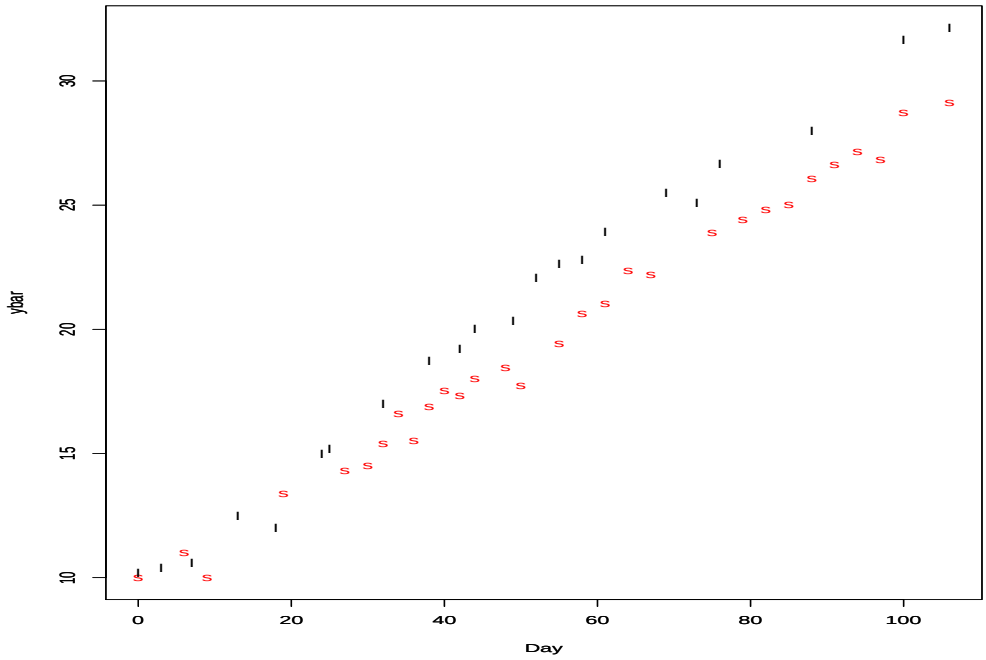

Figure 26: The scatter plot requested by Problem 9. There seems to be an obvious difference between the slopes of the two lines but any differences of intercept are difficult to determine. We can use statistical hypothesis testing to determine the significance of any model slope differences.

by

$$\hat{\sigma}^2_{\text{pooled}} = \frac{\text{SS}_{\text{pe,short-shoots}}}{\text{df}_{\text{pe,short-shoots}}} + \frac{1}{2}\left(\frac{\text{SS}_{\text{pe,long-shoots}}}{\text{df}_{\text{pe,long-shoots}}}\right) \approx 1.61\,,$$

and should be a better estimate than either of the two individual estimates.

As an aside, to determine if our model does not match the data due to a large model specific error we can construct an appropriate F-statistic which in this case is the ratio of the mean square for lack-of-fit to that of the mean-square for pure-error (defined above). This F-statistic is given by

$$F = \frac{\frac{\text{SS}_{\text{lof}}}{\text{df}_{\text{lof}}}}{\frac{\text{SS}_{\text{pe}}}{\text{df}_{\text{pe}}}} = \frac{\frac{\text{RSS}-\text{SS}_{\text{pe}}}{n-p'-\text{df}_{\text{pe}}}}{\frac{\text{SS}_{\text{pe}}}{\text{df}_{\text{pe}}}}\,. \tag{59}$$

This is to be compared to the quantiles of the $F(n-p'-\text{df}_{\text{pe}}, \text{df}_{\text{pe}})$ distribution. If the value of the F-statistic above is "too large" we should reject the null hypothesis that we have the correct model for the given data.

6.9.2: In Figure 26 we present a scatter plot of *ybar* versus *Day*. The long shoots are denoted with the symbol "l" and the short shoots are denoted by the symbol "s". Straight-line mean functions look to be quite plausible.

	df	WRSS	F	P(>F)
Model 1: most general	48	60.55		
Model 2: parallel	49	89.15	22.67	$1.8\,10^{-5}$
Model 3: common intercept	49	62.54	1.58	0.214
Model 4: all the same	50	291.2	91.50	$4.2\,10^{-17}$

Table 5: ANOVA model comparison for the four models for Problem 6.9. Note that the weighted residuals are computed as $\sqrt{w_i}(y_i - \hat{y}_i)$.

6.9.3: In computing the weighted least squared esimates we are told to assume that

$$\mathrm{Var}(\bar{y}_i|Day, \text{short shoots}) = \frac{\sigma^2}{n}\,,$$

and

$$\mathrm{Var}(\bar{y}_i|Day, \text{long shoots}) = \frac{2\sigma^2}{n}\,.$$

Thus we will will write both of these expressions as $\frac{1}{w_i}$ (absorbing σ^2 and n into the same denominator). Under this convention our weights take numerical values given by

$$\frac{1}{w_i} = \begin{cases} n/\hat{\sigma}^2 & \text{short shoots} \\ n/2\hat{\sigma}^2 & \text{long shoots} \end{cases}\,.$$

Where $\hat{\sigma}$ is our pooled estimate computed above.

In table 5 we display an anova table that compares the standard four linear models introduced in this chapter. From that table the difference in weighted residual sum of squares reduction supplied in moving from Model #3 to Model #1 has a twenty-one percent chance of happening by chance. One might conclude that Model #1 is too general and that we can get the same performance by using Model #3. Plots of the four models compared are shown in Figure 27.

This problem is worked in the `R` script `chap_6_prob_9.R`.

6.10 (gothic and romanesque cathedrals)

This problem is to compare various models of the form

$$E(Length|Height = h, D = j) = \beta_{0j} + \beta_{1j}h\,,$$

where D is a factor expressing the type of cathedral. This is similar to Problems 6.4, 6.5, and 6.6 in that Various model simplifications are possible by pooling samples.

6.10.1: In the Figure 28 we present two separate scatter plots of this data set.

6.10.2: In the Table 6 we show the anova table comparing the four candidate models. Since the F-statistics are so low for models #2 and 3 we conclude that the additional complexity

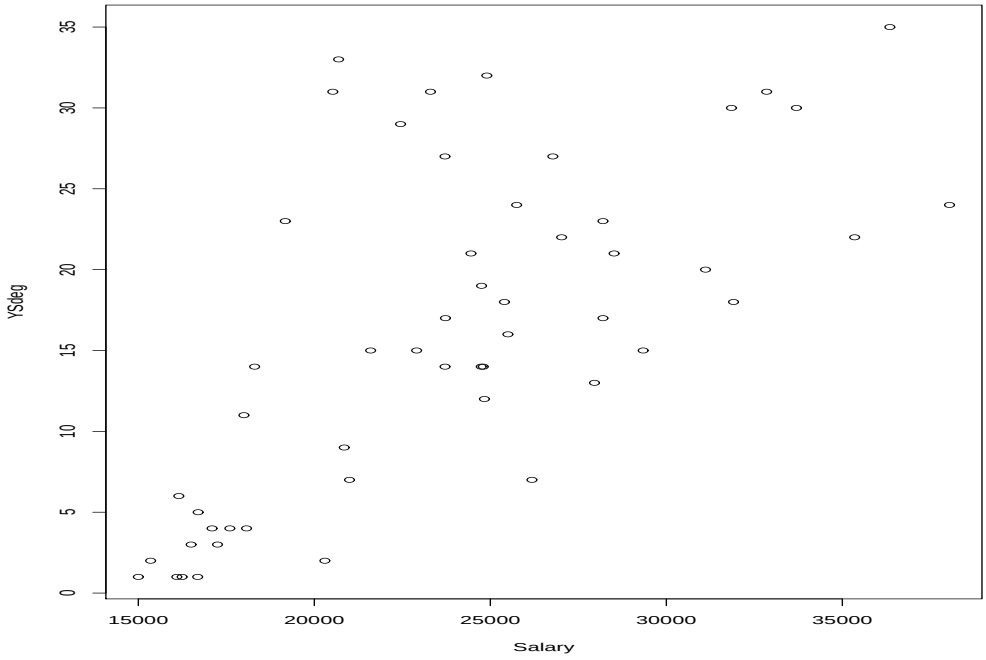

Figure 27: A visual display of the four models from Problem 9.

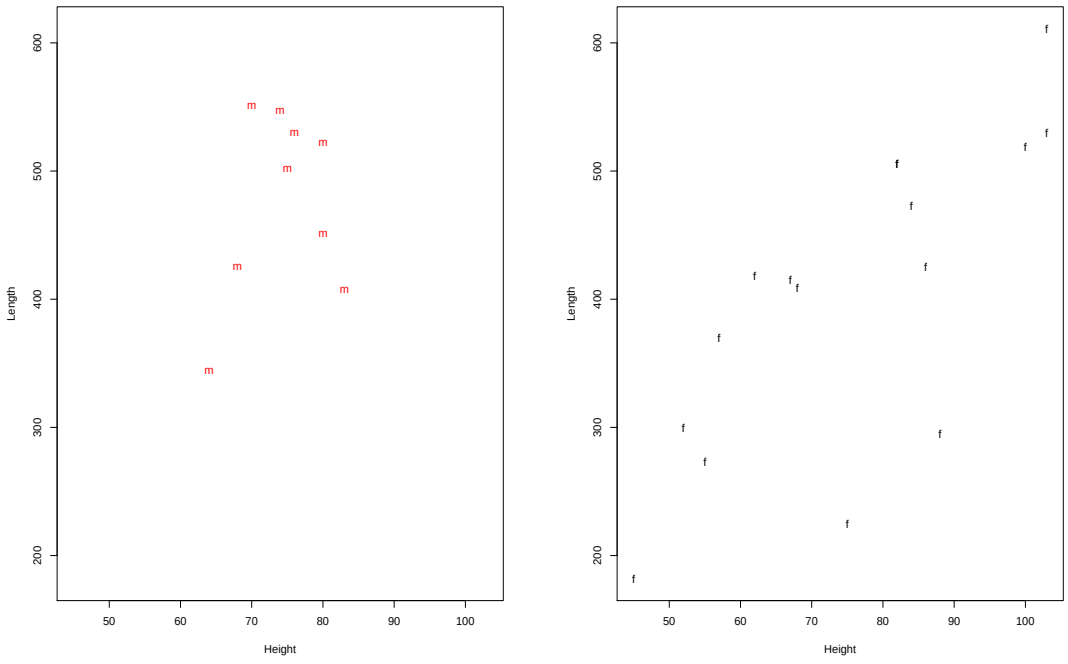

Figure 28: **Left:** Scatter plots of Length vs. Height for Romanesque churches. **Right:** The same but for Gothic churches. The x and y axis are the *same* for both plots so we expect a linear regressions to be quite different for the two different types of churches.

	df	RSS	F	P(>F)
Model 1: most general	21	136200		
Model 2: parallel	22	137100	0.142	0.71
Model 3: common intercept	22	138500	0.3525	0.559
Model 4: all the same	23	171500	2.719	0.089

Table 6: ANOVA model comparison for the four models for Problem 6.10.

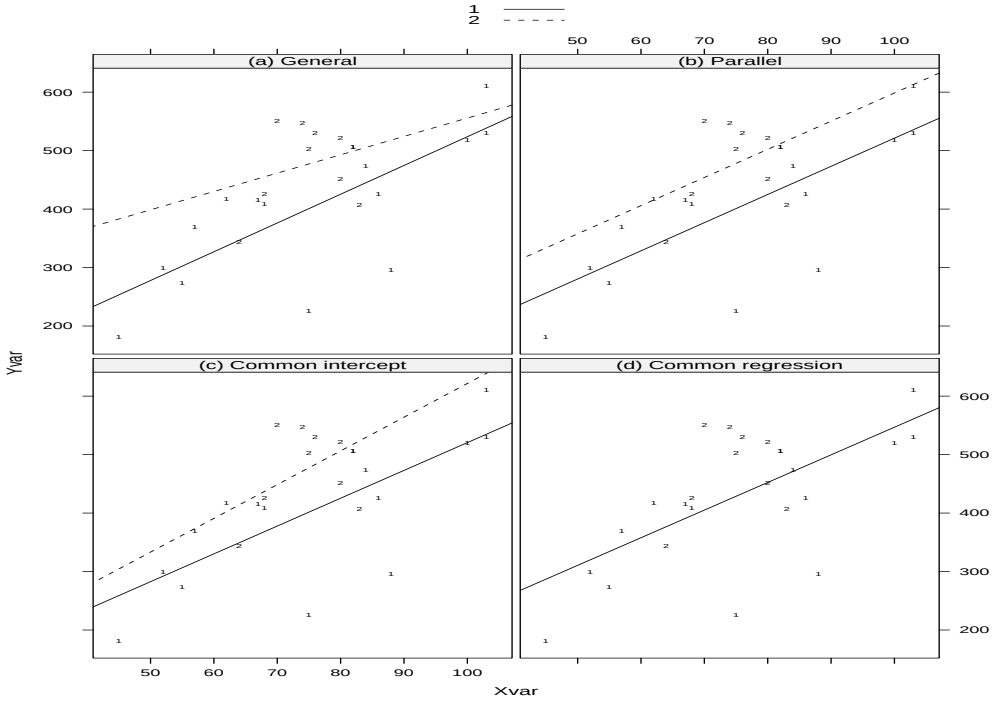

Figure 29: A visual display of the four models from Problem 10. ANOVA considerations would select the parallel slope model.

of model #1 is probably not needed. Given the choice between models #2 and #3 we could take the one with the smaller residual sum of squares, which in this case would be model #2. In Figure 29 we plot these four mean functions.

This problem is worked in the `R` script `chap_6_prob_10.R`.

6.11 (modeling the wind direction in the windmill data)

6.11.1: For this part of the problem we will assume that we should consider the "bin" index as a factor and asses how well the most general regression function where each bin has its own set of coefficients as expressed as

$$E(CSpd|RSpd = x, Bin = j) = \beta_{0j} + \beta_{1j}x \quad \text{for} \quad 1 \leq j \leq 16\,,$$

	df	RSS	F	P(>F)
Model 1: most general	1084	6272.0		
Model 2: parallel	1099	6387.7	1.33	0.1763
Model 3: common intercept	1099	6414.0	1.633	$5.9\,10^{-2}$
Model 4: all the same	1114	6776.0	2.900	$4.01\,10^{-4}$

Table 7: ANOVA model comparison for the four models for Problem 6.11.

compares to the simpler mean functions where no bin restrictions are placed on the β coefficients. When viewed in this way this problem is very similar to several other from this chapter. We present the anova comparison table in Table 7. When we look at that table we see that there is a 17% chance that the difference in RSS between Model #1 and Model #2 is due to chance. This gives reasonable strong evidence that we need the generality provided by Model #1.

6.11.2: We are told to assume that the most general model is appropriate. Then we are given the mean $\bar{x}_{*i}$ of m_i samples drawn from the i-th bin. If the standard errors are indeed independent we can combine the sixteen average estimates by weighting them according to how many samples went into each average. Let $m = \sum_{i=1}^{16} m_i$, then our global estimate of the average wind speed at the candidate site should be taken to be

$$E(CSpd) = \sum_{i=1}^{16} \frac{m_i}{m} E(CSpd|RSPd = \bar{x}_{*i}, B = i)\,.$$

Then the standard error (again assuming independent errors) is given by

$$\mathrm{Var}(CSpd) = \sum_{i=1}^{16} \frac{{m_i}^2}{m^2} \mathrm{Var}(CSpd|RSPd = \bar{x}_{*i}, B = i)\,.$$

These later elements can be computed using the results from Problem 2.13 (see Page 22 Equation 9), where we have

$$\mathrm{Var}(\bar{y}_*) = \frac{\sigma^2}{m} + \sigma^2 \left(\frac{1}{n} + \frac{(\bar{x}_* - \bar{x})}{\mathrm{SXX}} \right)\,.$$

Question: Given the amount of time I wanted to spend on this problem I was unable to determine how to apply the `R` command `predict` with so many factors. If anyone knows how to do this please let me know.

6.12 (land valuation)

Since our goal is to determine if the variable P provides any predictive power over the variable $Value$ we will fit two models to the provided data, one model that includes this variable and another model that does not. We can then use F-statistics to determine if the more complex model reduces the residual sum of squares more than can be attributed to

chance. Thus we will compare the null-hypothesis that *each* year county combination has a different mean value, stated mathematically as

$$E(Value|Year = y, County = c) = \beta_{0y} + \beta_{1c}\,.$$

We can write this mean function in terms of two dummy variables U_i and V_i as

$$E(Value|Year = y, County = c) = \sum_{i=1}^{2} \beta_{0i} U_i + \sum_{i=1}^{4} \beta_{1i} V_i\,.$$

where the coefficient β_{0i} represents the effect of the $Year$ variable and β_{1i} is the same for the county variable. This mean function has $2 \times 4 = 8$ coefficients to be determined. This mean function would be compared to the alternative hypothesis where we include P as an explicit variable

$$E(Value|Year = y, County = c, P = x) = \beta_{0y} + \beta_{1c} + \beta_p x\,. \tag{60}$$

These two mean functions can be compared for significance using the `R` function `anova`. In the time allowed I was not able to figure out how to encode in `R` regression models in the forms specified above. This is not a problem since we can use whatever equivalent representation `R` uses by default to compare mean functions. The anova table for each mean function is

```
> anova(mNH,mAH)
Analysis of Variance Table

Model 1: Value ~ countyFactor + yearFactor
Model 2: Value ~ countyFactor + yearFactor + P
  Res.Df     RSS  Df Sum of Sq      F    Pr(>F)
1    115 1781801
2    114 1423587   1    358214 28.686 4.477e-07 ***
```

From this table we see that the fact that the p-value above is so small (equivalently an F-value so large) is an indication that the addition of the variable P *is* a relevant variable to consider in assessing land values. We can also ask if we can simplify Equation 60 by dropping either the $Year$ factor or the $County$ factor. Appropriate anova tests seem to indicate that the answer is no, these terms do indeed do provide information over that which be measured randomly.

This problem is worked in the `R` script `chap_6_prob_12.R`.

6.13 (sex discrimination)

6.13.1: In Figure 30 we present a scatterplot matrix of the variables $Salary$, $YSdeg$, and $Year$. There does appear to be a increase in salary with both of the variables $YSdeg$ and $Year$.

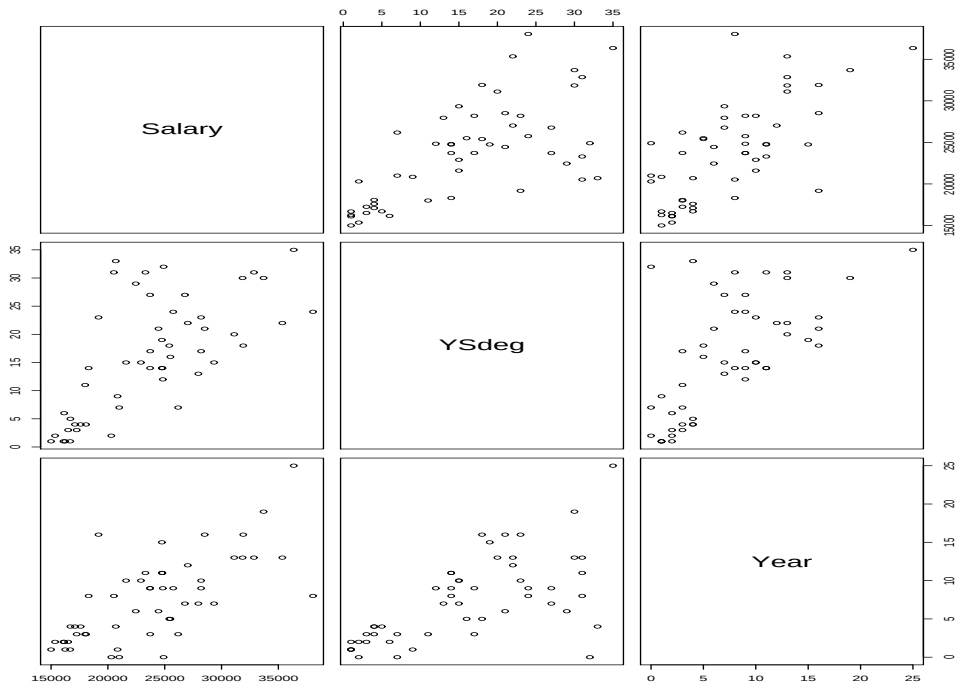

Figure 30: A scatterplot matrix of $Salary$, $YSdeg$, and $Year$ from the `salary` data set of Problem 6.13.

6.13.2: We can take the variable Sex as a factor an compare the more general model where there are two means (one for each factor) with the simpler model with one only mean. If the reduction in the residual sum of squares is significant from model with two means a F-test should determine this. When we compute the `anova` command on these two mean functions we obtain

```
Analysis of Variance Table

Model 1: Salary ~ +1
Model 2: Salary ~ -1 + SF
  Res.Df        RSS Df  Sum of Sq     F Pr(>F)
1     51 1785729858
2     50 1671623638  1  114106220 3.413 0.0706 .
```

When we look at the above table we see that there is only a 7% chance that the this reduction in sum of squares is by chance. This gives strong indication in favor of the hypothesis that there *is* a difference between the two means.

6.13.3: To solve the first part of this problem we generate two models. The first less specific model is based on regressing $Salary$ on the variables $Year$, $Degree$, and $YSdeg$, where $Degree$ is a factor representing the degree. This model is compared with the more general model where in addition to the above variable we add the addition factor $Rank$. The `anova` command comparing these two models gives

```
Analysis of Variance Table

Model 1: Salary ~ Year + HDF + YSdeg
Model 2: Salary ~ Year + HDF + YSdeg + RF
  Res.Df       RSS Df Sum of Sq      F    Pr(>F)
1     48 672102002
2     46 267993336  2 404108665 34.682 6.544e-10 ***
```

The F-statistic comparing these two models is sufficiently large that we must conclude that the factor $Rank$ makes a difference in the determination of $Salary$.

To solve the second part of this problem we consider the model obtained above which uses $Year$, $Degree$, and $YSdeg$ to predict $Salary$ and append to this model a factor representing Sex. This new model is compared to the old one and gives the following `anova` table

```
Model 1: Salary ~ Year + HDF + YSdeg + RF
Model 2: Salary ~ Year + HDF + YSdeg + RF + SF
  Res.Df       RSS Df Sum of Sq     F Pr(>F)
1     46 267993336
2     45 258858365  1   9134971 1.588 0.2141
```

This later table indicates that the difference in residual sum of squares by including this factor has a 21% chance of happening by chance. This gives some indication that Sex is not a helpful factor in determing $Salary$.

This problem is worked in the `R` script `chap_6_prob_13.R`.

6.14 (changing the numerical value of a factor)

We are given a fitted regression function

$$E(Salary|Sex, Year) = 18223 - 571Sex + 741Year + 169Sex \times Year \tag{61}$$

and are asked how this will change when the defining relationship for the factor Sex changes.

6.14.1: In this case we require that the two new definitions of the factor for sex map from their old definitions to new definitions as $(0, 1) \to (2, 1)$. This can be done with the transformation

$$Sex' = -Sex + 2\,.$$

Solving this for Sex we have $Sex = 2 - Sex'$. When we put this expression for Sex into Equation 61 we get

$$E(Salary|Sex', Year) = 17081 + 571Sex' + 1079Year - 169Sex' \times Year\,.$$

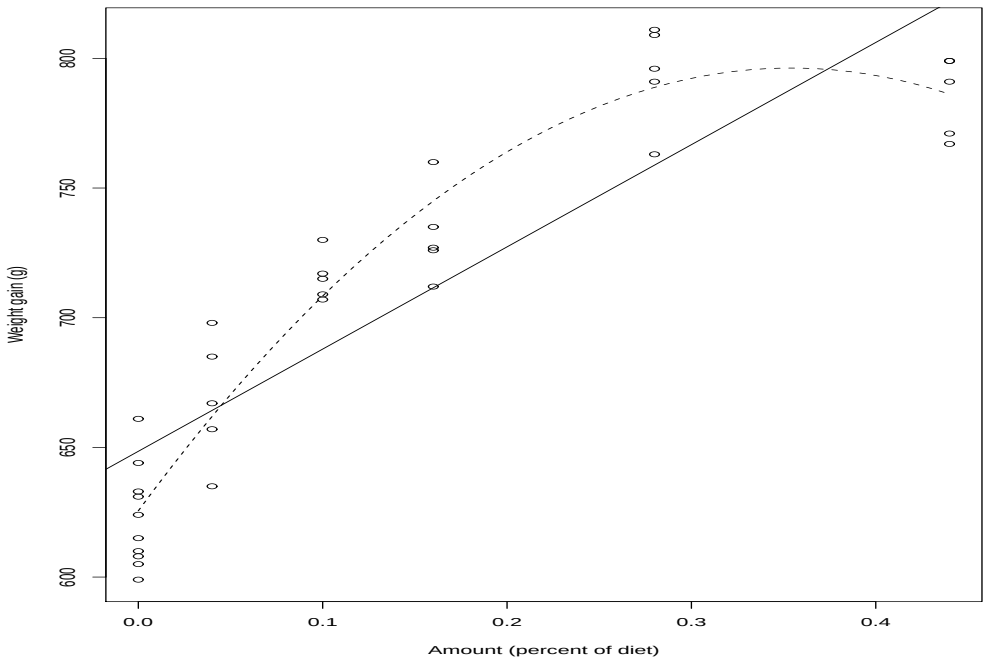

Figure 31: A scatterplot of *Gain* versus amount A for the `turk0` data set of Problem 6.15. In addition, we overlay linear and quadratic fits. The problem explores further which model is appropriate for the data.

6.14.2: If we require that the two new definitions of the factor for sex map from their old definitions to new definitions as $(0,1) \to (-1,1)$. This can be done with the transformation

$$Sex' = 2Sex - 1\,.$$

Solving this for Sex we have $Sex = \frac{1}{2}(Sex' + 1)$. When we put this expression for Sex into Equation 61 we get

$$E(Salary|Sex', Year) = 18508.5 + 285.5Sex' + 825.5Year + 84.5Sex' \times Year\,.$$

6.15 (pens of turkeys)

6.15.1: In Figure 31 we present a scatter plot of *Gain* vs. amount A. Since there are multiple experiments corresponding to each unique value of the *amount* variable we can use the lack-of-fit tests techniques of Chapter 5 under the assumption that the standard deviation is unknown.

6.15.2: Since in this problem we are given several values of the responce *Gain* for each of the possible values of the amount A we can compute the mean of the response along with the sample standard deviation to determine how well the given models (linear and quadratic fit the data). See Problem 6.9 for an example. Specifically, we will use Equation 59 for both

the linear and the quadratic models. For each model we find F-statistics given by

$$F_{\text{linear}} = 18.71063 \quad \text{and} \quad F_{\text{quadratic}} = 1.492030\,.$$

The probability that under the null-hypothesis (our model describes the data) we would obtain F-statistics as large or larger than these is given by

$$\alpha_{\text{linear}} = 1.062082\,10^{-7} \quad \text{and} \quad \alpha_{\text{quadratic}} = 0.2374352\,.$$

This gives an indication that the linear fit is *not* approporiate and that the quadratic fit should be used.

6.15.3: Based on the result of the previous problem the quadratic polynomial fits the data better and would be the prefered model. This is visual observable in Figure 31.

This problem is worked in the `R` script `chap_6_prob_15.R`.

6.16 (estimating the maximum value)

If our *Gain* variable is a quadratic function of amount say

$$E(Gain|Amout = a) = \beta_0 + \beta_1 a + \beta_2 a^2\,,$$

then the extream occurs at $a_M = -\frac{\beta_1}{2\beta_2}$, we can evaluate the standard error of this using the delta method using the command `delta.method` or by using the bootstrap. Using the `alr3` command `delta.method` gives the following:

```
> delta.method(m2,"-b1/(2*b2)")
Functions of parameters:  expression(-b1/(2*b2))
Estimate = 0.3540464 with se = 0.01925134
```

To use the bootstrap we can use the `alr3 R` function `boot.case` to compute bootstrap samples and then take the mean and the standard deviation as estimate of the estimate and standard error. When we do that with $B = 2000$ bootstraped samples we find

```
> print(mean(maxs))
[1] 0.356541
> print(sqrt(var(maxs)))
[1] 0.01851026
```

a similar result.

This problem is worked in the `R` script `chap_6_prob_16.R`.

6.17 (inverse regression in Jevons' coin data)

For a model of the form like computed in Problem 5.6

$$E(\overline{Weight}|Age = a) = \beta_0 + \beta_1 a\,,$$

to derive an estimate of the value of a at which the weight will be the legal minimum of 7.9379 grams we would solve $\beta_0 + \beta_1 a = 7.9379$ for a. When we do this we find an estimate of a given by

$$\hat{a} = \frac{7.9379 - \hat{\beta}_0}{\hat{\beta}_1}\,.$$

We can fit the model for this problem and then use the `alr3` function `delta.method` to compute an estimate and its standard error. We find

```
> delta.method(mwa, "(7.9379 - b0)/b1")
Functions of parameters:  expression((7.9379 - b0)/b1)
Estimate = 2.482909 with se = 0.09657335
```

This problem is worked in the `R` script `chap_6_prob_17.R`.

6.18 (world record running times)

6.18.1: In Figure 32 we present the requested scatter plot.

6.18.2: We could present a sequence of anova table like previous problems but we will simply summarize thier content. The anova values are computed in the `R` script that accompanies this problem. Each anova comparison of the simpler models to the most general model gives a p-value $O(10^{-8})$ which indicates that the simpler models are significantly worse than the more general one at the reduction of the residual sum of squares that they provide. Based on this analysis we should consider the most general model. When we fit separate models we find that the specific coefficients are

$$\begin{aligned} E(Time|Year = y, Sex = m) &= 953.746 - 0.36619y \\ E(Time|Year = y, Sex = f) &= 2309.4247 - 1.03370y\,. \end{aligned}$$

Here "m" stands for male and "f" stands for female. As an intpretation of these number we see that the mean running time for men is less than that of the women. In addition the rate at which the women are decreasing their time is greater than that of the men. Both of these facts are clearly seen in the scatter plot in Figure 32.

6.18.3: To anwser this question requires that we solve for y in the mean regresion function for females. Thus

$$E(Time|Year = y, Sex = f) = \beta_{0f} + \beta_{1f} y = 240\,.$$

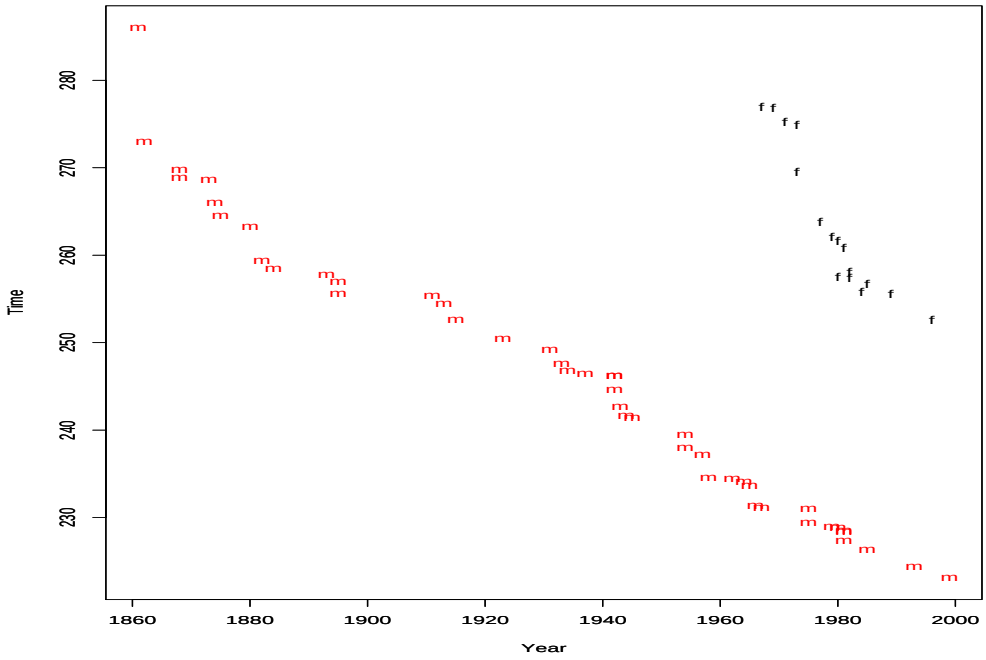

Figure 32: A scatterplot of $Time$ versus $Year$ for men (in red) and women (in black) for the world record times for the one mile run from the `mile` data set of Problem 6.18.

A point estimate of the value for y once we have fit our linear models is given by

$$\hat{y} = \frac{240 - \hat{\beta}_{0f}}{\hat{\beta}_{1f}} .$$

Since this is a nonlinear relationship in terms of the estimated β values we need to use either the bootstrap or the delta method to evaluate its standard error. Using the `alr3` code `delta.method` we get

```
> delta.method(n1, "(240 - b0)/b2")
Functions of parameters:  expression((240 - b0)/b2)
Estimate = 2001.966 with se = 2.357027
```

6.18.4: To anwser this question requires solving for y in

$$E(Time|Year = y, Sex = f) = E(Time|Year = y, Sex = m) ,$$

or

$$\beta_{0f} + \beta_{1f} y = \beta_{0m} + \beta_{1m} y .$$

A point estimate of this value of y once we have fit our two linear models is given by

$$\hat{y} = \frac{\hat{\beta}_{0f} - \hat{\beta}_{0m}}{\hat{\beta}_{1m} - \hat{\beta}_{1f}} .$$

Using the `alr3` code `delta.method` we get

```
> delta.method(n1, "(b0 - b1)/(b3 - b2)")
Functions of parameters:  expression((b0 - b1)/(b3 - b2))
Estimate = 2030.950 with se = 8.16785
```

This problem is worked in the `R` script `chap_6_prob_18.R`.

6.19 (using the delta method to estimate the variance of β_1/β_2)

This is an easy exercise to do when we use the `alr3` code `delta.method`. When use use this we get

```
> delta.method(m1,"b1/b2")
Functions of parameters:  expression(b1/b2)
Estimate = 2.684653 with se = 0.3189858
```

When we look back at the results from Chapter 4 we find that the point estimates agree and the delta method would predict a 95% confidence interval for the ratio of β_1/β_2 of

$$(2.059620, 3.309686)\,.$$

This is a *tigher* interval than what is reported in section 4.6 from the book.

This problem is worked in the `R` script `chap_6_prob_19.R`.

6.23 (using wind to your advantage)

6.23.1: For this problem we choose to fit three linear models to the response $Dist$ using the provided predictors. These three linear models are then compared using the anova techniques from this chapter. Specifically, we choose to fit the following models

$$\begin{aligned} Dist &\sim Velocity + Angle + BallWt + BallDia + Cond \\ Dist &\sim Velocity + Angle + BallWt + BallDia \\ Dist &\sim Velocity + Angle \\ Dist &\sim Velocity + Angle + Cond\,. \end{aligned}$$

Here $Cond$ is a factor that determines the direction of the fans. Since the first model is the most general we compare all models to that one. In Table 8 we present a comparison of the four models. In that table we see table we see that there is a statisticlly significant difference between all of the simpler models. This gives an indication that the placement of the fans *does* contribute to the distance a given ball travels.

This problem is worked in the `R` script `chap_6_prob_23.R`.

	df	RSS	F	P(>F)
Model 1: most general	28	1297		
Model 2: all variables no wind factor	29	1750	9.73	$4.18\,10^{-3}$
Model 3: velocity and angle only	31	2040	5.37	$4.7\,10^{-3}$
Model 4: Model 3 + wind factor	30	1730	4.65	$1.8\,10^{-2}$

Table 8: ANOVA model comparison for the four models for Problem 6.23.

Chapter 7 (Transformations)

Notes On The Text

Notes on the scaled power transformation

In this section of these notes we show the limiting assumption on the scaled power transformation made in the book and given by

$$\lim_{\lambda \to 0} \left(\frac{X^\lambda - 1}{\lambda} \right) = \log(X) \,. \tag{62}$$

To show this recognize that this limit is equal to $\frac{d}{d\lambda}(X^\lambda)\Big|_{\lambda=0}$. To evaluate this derivative let v be defined as $v \equiv X^\lambda$, then we need to evaluate $\frac{dv}{d\lambda}\Big|_{\lambda=0}$. Taking the derivative with respect to λ on both sides of the expression

$$\log(v) = \lambda \log(X) \,,$$

gives

$$\frac{1}{v}\frac{dv}{d\lambda} = \log(X) \,.$$

Using this and solving for $\frac{dv}{d\lambda}$ we find

$$\frac{dv}{d\lambda} = v \log(X) = X^\lambda \log(X) \,.$$

Finally, evaluating this at $\lambda = 0$ gives Equation 62.

Notes on transforming the response only i.e. the inverse fitted value plot

This section of the book indicates that one technique one could use to *find* a mapping to apply to Y that hopefully results in a better OLS fit is the following. First perform transformations on the independent variables X perhaps using the methods discussed in the section on automatic choice of transformation of parameters using the Box-Cox method. Once this has been done and the X variables have been specified we then fit a regression model to Y using these predictors X and compute the predicted values $\hat{Y}$ using the standard formula $\hat{Y} = \hat{\beta}_0 + \hat{\beta}' X$. We then consider the ordered pairs $(Y, \hat{Y})$ and fit a power transformation to the Y variables as discussed in the transformation of the section on transforming the input predictor. That is we look for parameters α_0, α_1 and λ_y such that

$$E[\hat{Y}|Y] = \alpha_0 + \alpha_1' \Psi_S(Y, \lambda_y) \,,$$

has the smallest value of $RSS(\alpha_0, \alpha_1, \lambda_y)$. This value of λ_y is what could then used in the direct regression of $\Psi_S(Y, \lambda_y)$ the predictors X.

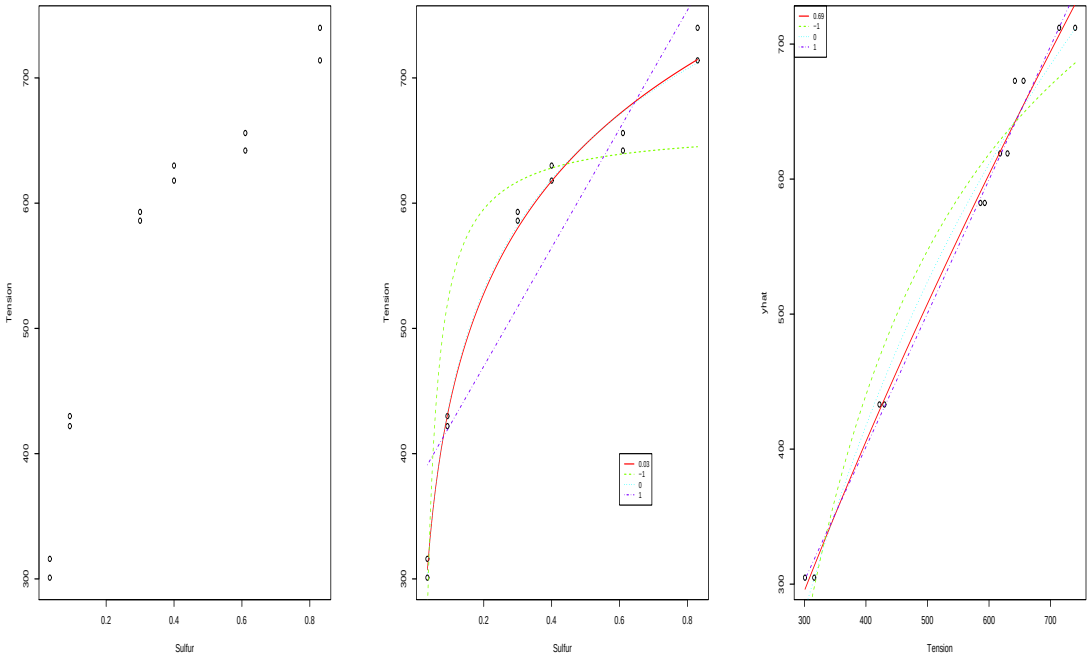

Figure 33: **Left:** A scatter plot of $Tension$ as a function of $Sulfur$. Note the highly non-linear behavior of the variable $Tension$. **Center:** Linear fits of the models $E(Tension|Sulfur) = \beta_0 + \beta_1 Sulfur^{\lambda}$ for three values of λ, specifically $\lambda \in \{-1, 0, +1\}$. **Right:** Linear fits of the models $E(\hat{Y}|Y) = \alpha_0 + \alpha_1 Y^{\lambda}$, where Y in this case is $Tension$ and for various values of λ. The x-axis is the pure variable $Tension$ and the y-axis is the predicted value of $Tension$ using the model found in Part 7.1.2:.

Problem Solutions

7.1 ($Tension$ as a function of $Sulfur$)

7.1.1: See Figure 33 (left) for a scatter plot of $Tension$ as a function of $Sulfur$. The points without any transformation are not very linear, indicating that a transformation if the dependent variable (here $Tension$) is needed.

7.1.2: For this part of the problem we will use the `alr3` command `inv.trans.plot` which produces the inverse transform plot requested when the use specifies the desired values for λ. See the Figure 33 (middle) for the plot produced. This routine also estimates the optimal value of λ which in this case comes very close to $\lambda \approx 0$ or a logarithmic transformation.

7.1.3: For this part of the problem, using the results from earlier we transform the variable $Sulfur$ by taking the logarithm of it. We then look for a model in which to transform Y so that we can better predict $\hat{Y}$. The models we search for are given by

$$E(\hat{Y}|Y) = \alpha_0 + \alpha_1 \psi_S(Y, \lambda)\,,$$

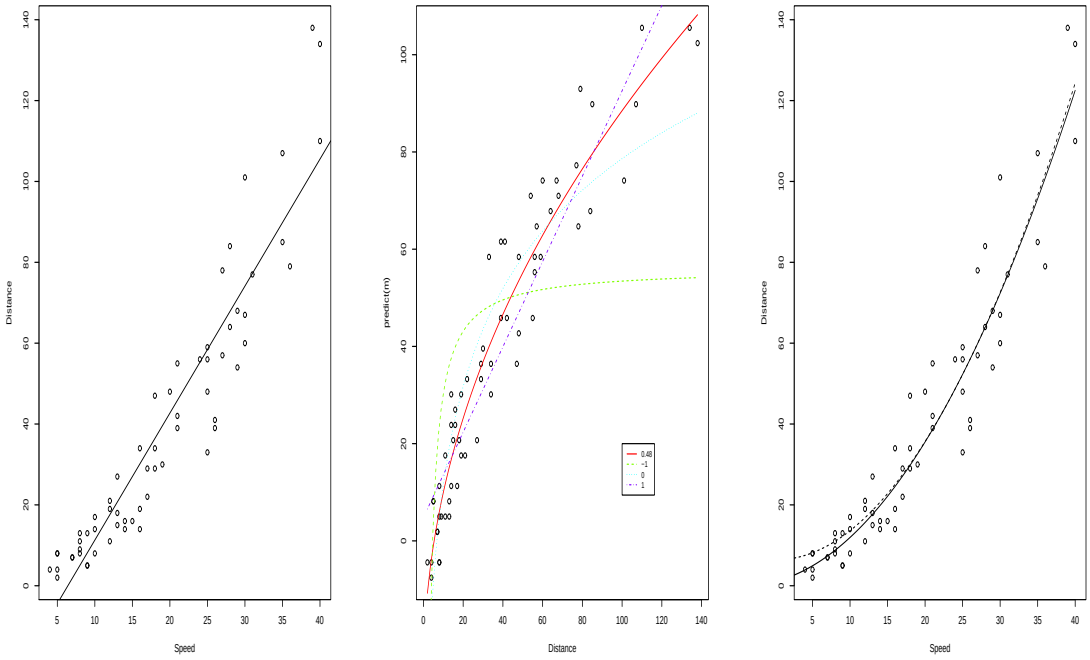

Figure 34: **Left:** A scatter plot of *Distance* as a function of *Speed* for Problem 7.2 with the OLS linear fit of *Distance* vs. *Speed*. **Center:** Power transforms of the response *Distance* for various values of the powers $\lambda \in \{-1, 0, +1\}$ and the optimal λ. This optimal value of λ (the red curve) appears to have the value of $\lambda \approx 0.5$. **Right:** A comparison of the power transformation fit, determined by using the `R` command `inv.trans.plot` and the quadratic model of Hald. The λ power model is the solid line while the quadratic model is the dashed line.

where $\psi_S(Y, \lambda)$ is the "scaled" power transformation. Not we can use the `alr3` command `inverse.responce.plot` to plot to display various scaling transformations of Y (in this case $Tension$) that might result in a better linear fit. When we run this command we obtain the plot presented in Figure 33 (right), where an optimal value of λ is estimated to be 0.68608. One might argue that this optimal value of λ found to close to the value of $\lambda = 1$ representing no transformation. Using the `alr3` command `inv.tran.estimate` we get a standard error of 0.2842 which indicates that the value of 1 is inside two standard deviations of the estimate $\hat{\lambda} = 0.6860$. From this and the fact that the two curves for the optimal λ and the estimated λ shown in Figure 33 (right) look so similar we would decide *not* to use the transformation of the variable Y.

See the `R` function `chap_7_prob_1.R` for code that implements this problem.

7.2 (stopping times as a function of speed)

7.2.1: See Figure 34 (left) for a scatter plot of *Distance* as a function of *Speed*. We also plot the OLS linear fit of the regression of *Distance* onto *Speed*. When we look at this plot

we see that for small and large values of *Speed* the linear fit is not very good at predicting *Distance*.

7.2.2: To find a transformation that will linearize this regression we can use the `alr3` code `inv.tran.plot` to estimate the parameter to use to scale the response *Distance*. When we run this code we see that the optimal value of λ is $\lambda \approx 0.5$. This implies a model given by

$$Distance^{1/2} = \beta_0 + \beta_1 Speed\,.$$

7.2.3: For this part of the problem we need to use weighted least squares to obtain the coefficients that are suggested by the Hald model. The weights, w_i, in weighted least squares are defined as w_i for which

$$\mathrm{Var}(Y|X = x_i) = \frac{\sigma^2}{w_i}\,,$$

so in the case of the Hald model where

$$\mathrm{Var}(Distance|Speed) = \sigma^2 Speed^2 = \frac{\sigma^2}{\left(\frac{1}{Speed}\right)^2}\,,$$

we could take the weights w_i to be

$$w_i = \left(\frac{1}{Speed}\right)^2\,.$$

We perform this weighted least squares fit and then compare it to the power model found in earlier parts of this problem, we obtain Figure 34 (right). The two model are very similar.

See the `R` function `chap_7_prob_2.R` for code that implements this problem.

7.3 (predicting water supply from runoff)

7.3.1: See Figure 35 (left) for a scatter plot matrix of the variables involved in the `water` data set. We next use the `alr3` code `bctrans` which attempts to automatically transform the predictors in such a way that the resulting variables are more closely linearly related. This method works basically as follows. Given a specification of a vector of parameters $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_k)$ (one λ_k for each variable) we can map all of the original predictor data $\mathbf{X}$ to new points $\psi_M(\mathbf{X}, \lambda)$ given by the componentwise mapping

$$\psi_M(\mathbf{X}, \lambda) = (\psi_M(X_1, \lambda_1), \psi_M(X_2, \lambda_2), \ldots, \psi_M(X_k, \lambda_k))\,.$$

Where $\psi_M(Y, \lambda)$ is the *modified* power family transformation given by

$$\begin{aligned} \psi_M(Y, \lambda) &= \psi_S(Y, \lambda) \times \mathrm{gm}(Y)^{1-\lambda} \\ &= \begin{cases} \mathrm{gm}(Y)^{1-\lambda} \times (Y^\lambda - 1)/\lambda & \lambda \neq 0 \\ \mathrm{gm}(Y) \times \log(Y) & \lambda = 0 \end{cases}\,, \end{aligned} \tag{63}$$

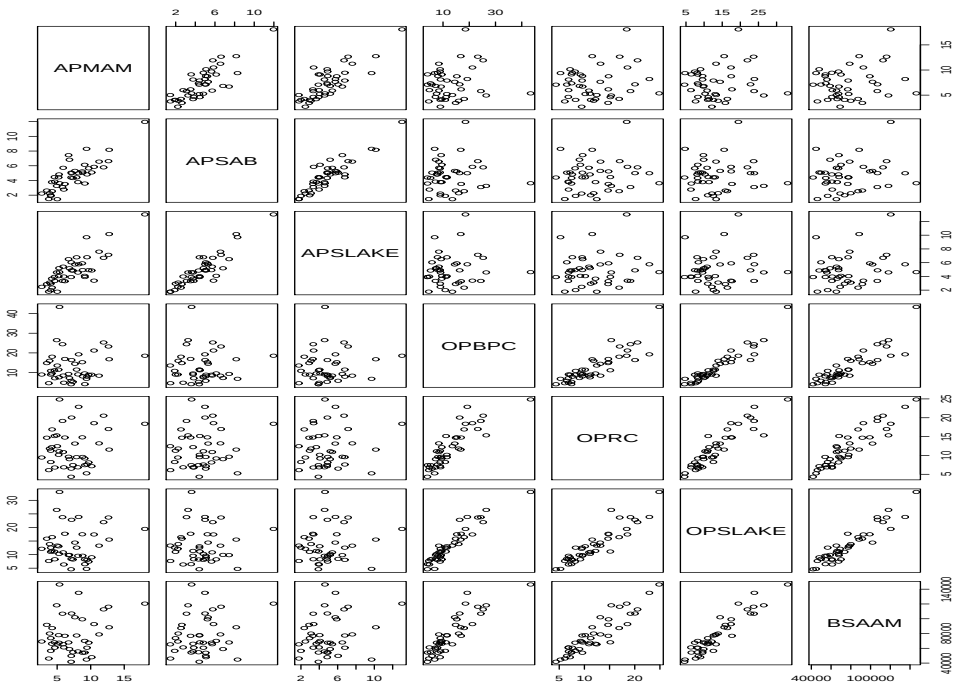

Figure 35: A scatter plot matrix of the variables involved with the `water` data set, before any power transformations. Note that the response $BSAAM$ is located in the lower right corner of this matrix.

and $\text{gm}(Y)$ is the *geometric mean* of the untransformed variable Y. The vector λ is chosen to make the resulting transformed variables as close to linearly related as possible. This means that we select a vector λ such that the transformed sample covariance matrix $\mathbf{V}(\lambda)$ has the smallest logarithm of its determinant. This is conveniently code in the `alr3` function `bctrans`. When we run this routine the resulting output for the `summary` command is

```
> summary(ans)
box.cox Transformations to Multinormality

         Est.Power Std.Err. Wald(Power=0) Wald(Power=1)
APMAM       0.0982   0.2861        0.3434       -3.1522
APSAB       0.3450   0.2032        1.6977       -3.2238
APSLAKE     0.0818   0.2185        0.3741       -4.2020
OPBPC       0.0982   0.1577        0.6227       -5.7180
OPRC        0.2536   0.2445        1.0375       -3.0531
OPSLAKE     0.2534   0.1763        1.4374       -4.2361
                                   LRT df      p.value
LR test, all lambda equal 0  5.452999  6 4.871556e-01
LR test, all lambda equal 1 61.203125  6 2.562905e-11
```

From this output it appears that all of the λ_k estimated are near zero (with a p-value of 0.48) indicating a logarithmic transformations of the independent variables maybe beneficial.

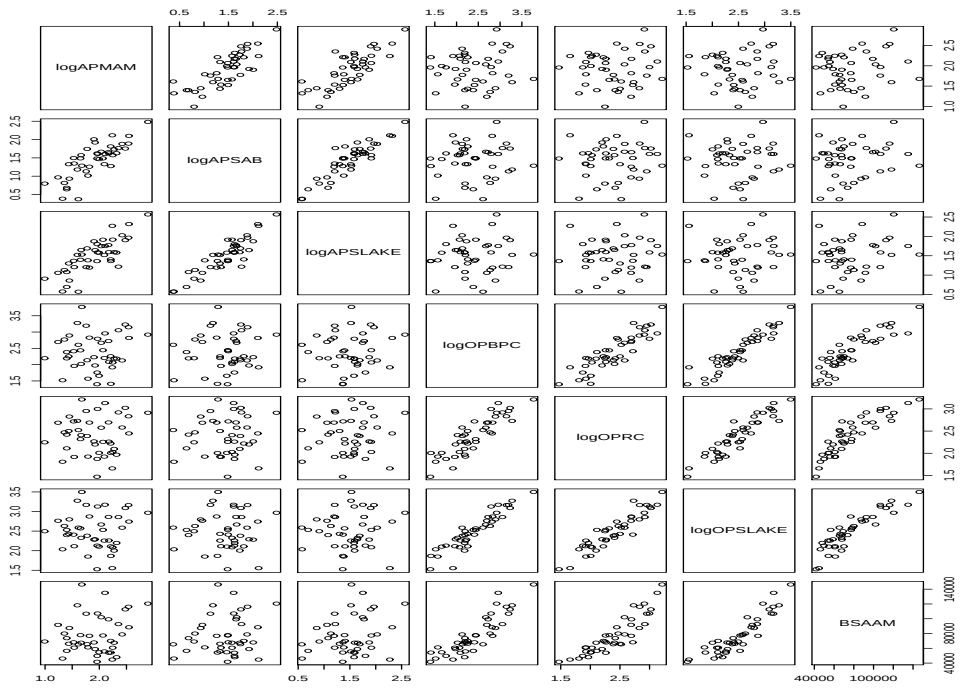

Figure 36: A scatter plot matrix of the variables involved with the `water` data set, after logarithmic transforms of all variables.

A scatter plot matrix of the resulting transformed variables after we do this are shown in Figure 36

7.3.2: Once we have specified a transformation of the predictor variables we can consider possible transformations of the response by considering a inverse fitted value plot. There is a command in the `alr3` code called `inverse.responce.plot` for performing this transformation and also viewing the results of several power transformations of the response graphically. When we use this command we get the plot shown in Figure 37.

7.3.3: We use the standard `R` command `lm` to estimating the transformed coefficients $\hat{\beta}$. Doing this and running the `summary` command we get the following

```
Call:
lm(formula = logBSAAM ~ logAPMAM + logAPSAB + logAPSLAKE + logOPBPC +
    logOPRC + logOPSLAKE, data = waterT)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18671 -0.05264 -0.00693  0.06130  0.17698

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.46675    0.12354  76.626  < 2e-16 ***
```

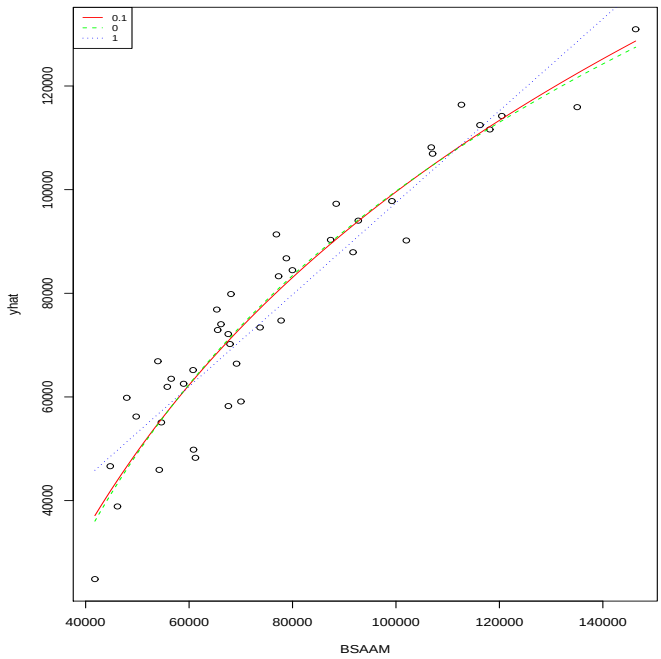

Figure 37: The inverse response plot of $BSAAM$ and the predictions from on a linear model with predictors found using the `bctrans` code. The optimal inverse power transformation resulted in $\hat{\lambda} = 0.1(0.29)$ indicating that a logarithm transformation where $\lambda = 0$ should be considered.

```
logAPMAM    -0.02033    0.06596  -0.308  0.75975
logAPSAB    -0.10303    0.08939  -1.153  0.25667
logAPSLAKE   0.22060    0.08955   2.463  0.01868 *
logOPBPC     0.11135    0.08169   1.363  0.18134
logOPRC      0.36165    0.10926   3.310  0.00213 **
logOPSLAKE   0.18613    0.13141   1.416  0.16524
---

Residual standard error: 0.1017 on 36 degrees of freedom
Multiple R-Squared: 0.9098,     Adjusted R-squared: 0.8948
F-statistic: 60.54 on 6 and 36 DF,  p-value: < 2.2e-16
```

From this we see that the two coefficients that are negative are the values for $log(APMAM)$ and $log(APSAB)$. I would think that negative coefficients would not make much sense. When one looks at the coefficient values and their standard error the estimates of these coefficients are such that their standard error would indicate that these estimates for β are not very reliable and the negative values are probably spurious. In fact from the t-values and the corresponding `P(>|t|)` column we see that we are not certain of their values and we have a relatively large probability of obtaining these values by chance.

7.3.4: We can use the `anova` function to compare the two models suggested in the book. We find that this function call gives

```
> anova(ms,m)
Analysis of Variance Table

Model 1:
  logBSAAM ~ logAPMAM + logAPSAB + logAPSLAKE + I(logOPBPC + logOPRC + logOPSLAKE)
Model 2:
  logBSAAM ~ logAPMAM + logAPSAB + logAPSLAKE +   logOPBPC + logOPRC + logOPSLAKE
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     38 0.40536
2     36 0.37243  2   0.03293 1.5915 0.2176
```

which indicates that if the simpler model were true there is a 21 percent chance of getting a reduction in residual sum of squares this large simply by chance. This is a relatively large percentage value but still probably not enough to justify using the simpler model.

See the `R` function `chap_7_prob_3.R` for code that implements this problem.

7.4 (predicting salary's)

7.4.1: If the variables are given in terms of job class rather than employee one could imagine a situation where there is a great deal of different skill levels among for the same jobs class.

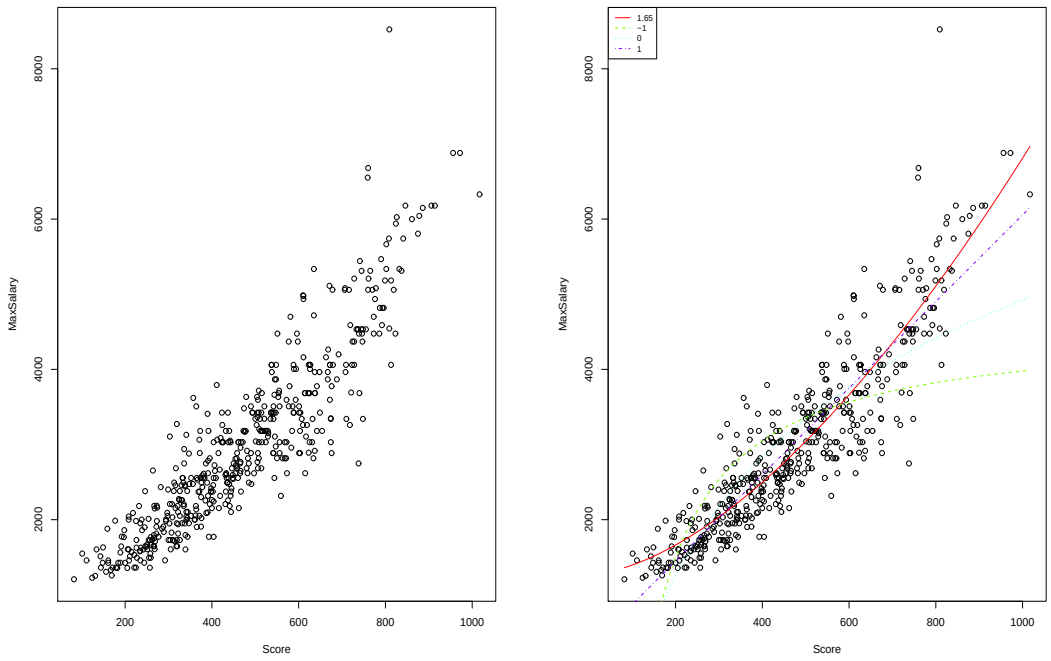

Figure 38: **Left:** The raw scatter plot of $MaxSalary$ as a function of $Score$. Not the curved appearance indicates that a linear model is probably not optimal. **Right:** Linear regressions of $MaxSalary$ as a function of a power transformed value of $Score$. The best fitting line (in red) corresponds to the model $E(MaxSalary|Score) = \beta_0 + \beta_1 Score^{1.645}$.

Not having the data broken down in the correct way would make further analysis difficult.

7.4.2: See Figure 38 (left) for a scatter plot of $MaxSalary$ versus $Score$. The appearance of this looks like it needs a nonlinear transformation to be well approximated by a linear model. To consider nonlinear transformations we first try to perform an power (Box-Cox) transformation of the independent variable $Score$, using the `alr3` function `inv.tran.plot`. The estimated value of λ under this procedure is $\hat{\lambda} \approx 1.645313$. We could then try to find a power transformation for the response $MaxSalary$. When we do this we see that the optimal transformation power $\hat{\lambda} = 0.74$ value seemed close enough to $\lambda = 1$ (meaning no transformation) that it was felt that this was not worth while. The difference between these two curves seems only to matter at larger values of $MaxSalary$. This transformation does appear to give a variance that is constant.

See the `R` function `chap_7_prob_4.R` for code that implements this problem.

7.5 (the price of hamburgers in different cities)

We will perform this exercise with the two cities with the largest value of $BigMac$ *included*. Modifications needed to exclude these two cities should be obvious.

7.3.1: See Figure 39 (left) for a scatter plot of $BigMac$ as a function of $FoodIndex$. We first look for a transformation of $FoodIndex$. Using the `R` command `inv.tran.plot` we obtain an estimate $\hat{\lambda} = -0.7751(0.69)$. Since the standard error of this estimate is so large we will take $\lambda = 0$ and perform a logarithmic transformation of the independent data. We next look for a transformation of the response $BigMac$ that would be beneficial for linear regression. To do this we first fit a linear model of $BigMac$ using the predictor $\log(FoodIndex)$ and then call the `alr3` function `inv.tran.plot`. We present the associated inverse response plot in Figure 39 (right). The optimal power to use to transform the value of $BigMac$ was found to be $-0.34(0.3)$. Again due to the large standard error we will transform the response with a logarithm. The scatter plot that results from the

These total transformations do indeed help the prediction accuracy resulting in a significant increase in the value of R^2 from $R^2 \approx 0.554$ when using the fully transformed model to $R^2 \approx 0.382$ when using the model with only $FoodIndex$ transformed to using the original model with no transformations where $R^2 \approx 0.33$.

7.3.2: See Figure 40 (left) for a scatter plot matrix of the requested variables. We can use the `alr3` command `bctrans` to estimate the optimal powers for the two variables $Rice$ and $Bread$. We find that the `summary` command (excluding the likelihood ratio information) for this command gives

```
> summary(ans)
box.cox Transformations to Multinormality
```

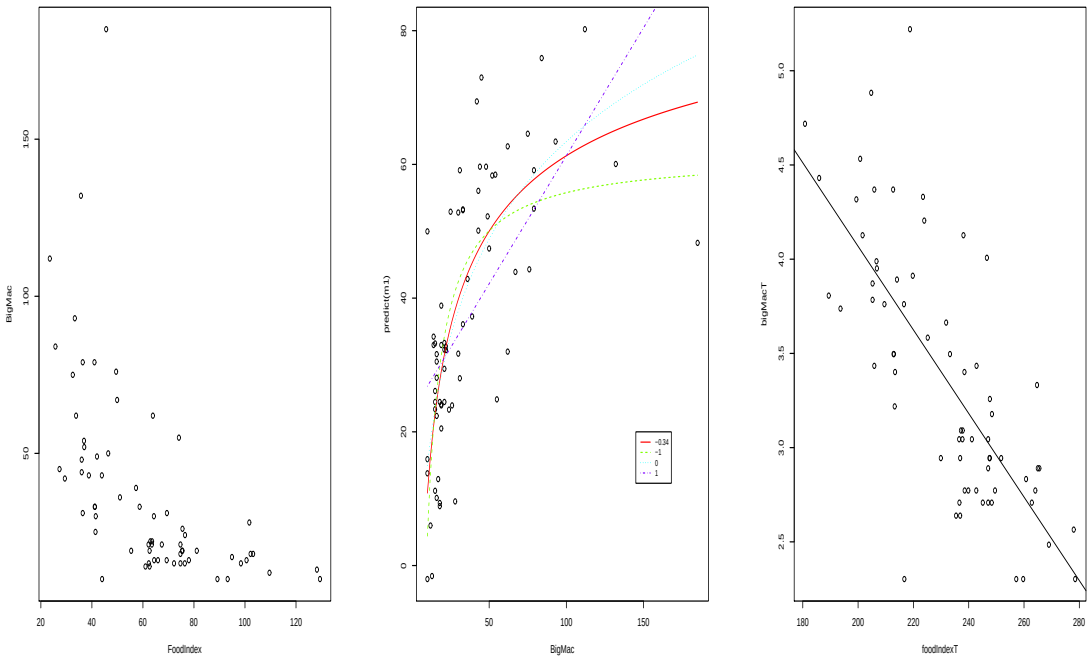

Figure 39: **Left:** The raw scatter plot of $BigMac$ as a function of $FoodIndex$. Note the strong curvature present in this scatter plot. **Center:** The inverse response plot of possible power transformations to optimally regress the linear predictions of $BigMac$ from $\log(FoodIndex)$ onto $BigMac$. **Right:** The scatter plot of $\log(BigMac)$ as a function of $\log(FoodIndex)$. This has a much more "linear" appearance.

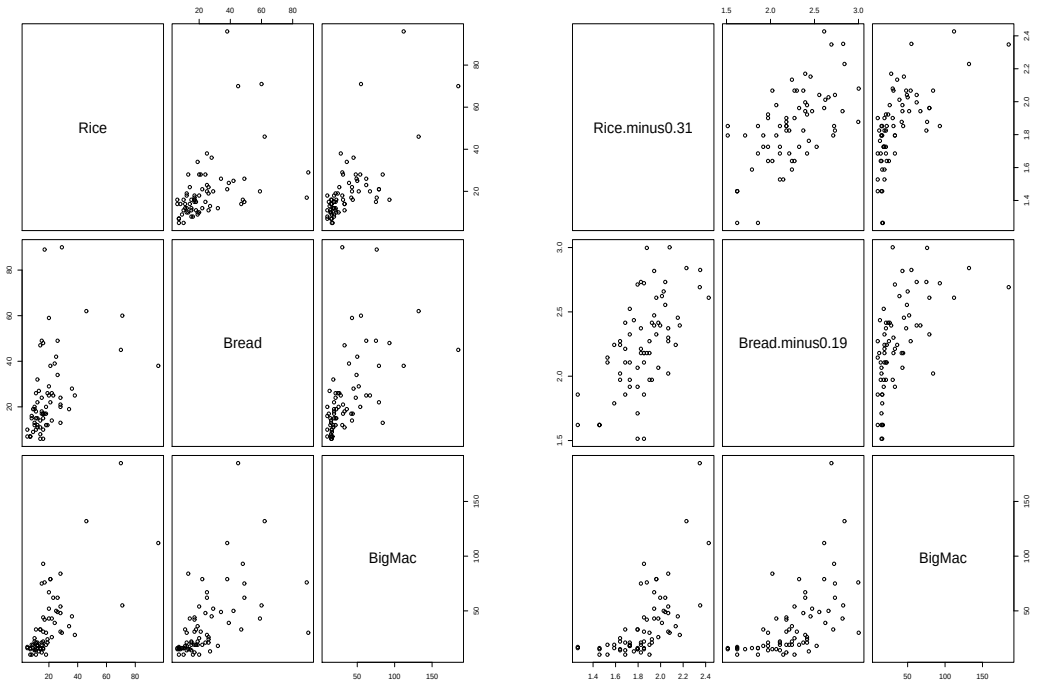

Figure 40: **Left:** The original scatter plot matrix of the variables *Rice*, *Bread*, and *BigMac*. **Right:** The `bctrans` found power transforms of the variables *Rice* and *Bread*. After transformation this data looks much more linearly distributed. Note specifically the *Rice* and *Bread* interactions.

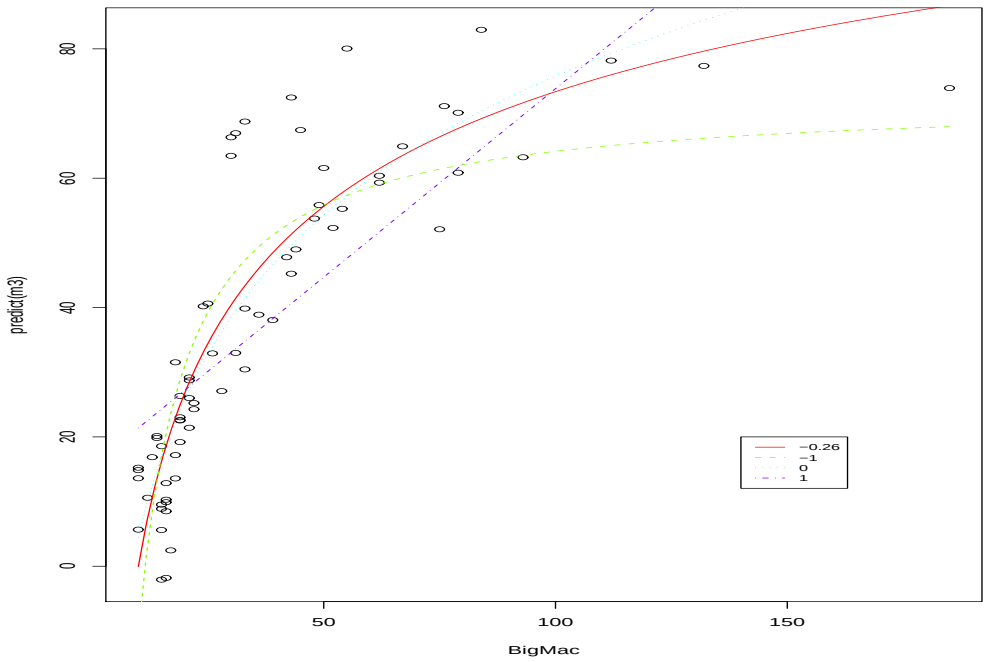

Figure 41: **Left:** The inverse response plot of the $BigMac$ samples after fitting a linear model on the variables $\log(Bread)$, $\log(Bus)$, $\log(TeachGI)$, and $Apt^{0.33}$.

```
      Est.Power Std.Err. Wald(Power=0) Wald(Power=1)
Rice    -0.3136   0.1481       -2.1180       -8.8724
Bread   -0.1939   0.1543       -1.2568       -7.7382
```

From this we see that the variable $Bread$ (due to the large standard error) may indicate a logarithmic transformation for that variable. The variable $Rice$ has a smaller standard error and perhaps the power -0.31 should be used for a transformation. If we accept the values predicted by `bctrans` after transformation the scatter plot matrix looks like Figure 40 (right).

7.3.3: For this part of the problem we will use the command `inv.tran.plot` to find an inverse response plot using the suggested transformation suggested for the independent variables. Note that we have not shown that these hypothesized transformations are optimal. The plotted result from this command is shown in Figure 41 where the optimal power for the response $BigMac$ is found to be $-0.256(0.17)$. Without an power transformation of the variable $BigMac$ the coefficient of determinism for this linear model is $R^2 \approx 0.58$. When we do perform this power transformation we obtain an $R^2 \approx 0.80$.

See the `R` function `chap_7_prob_5.R` for code that implements this problem.

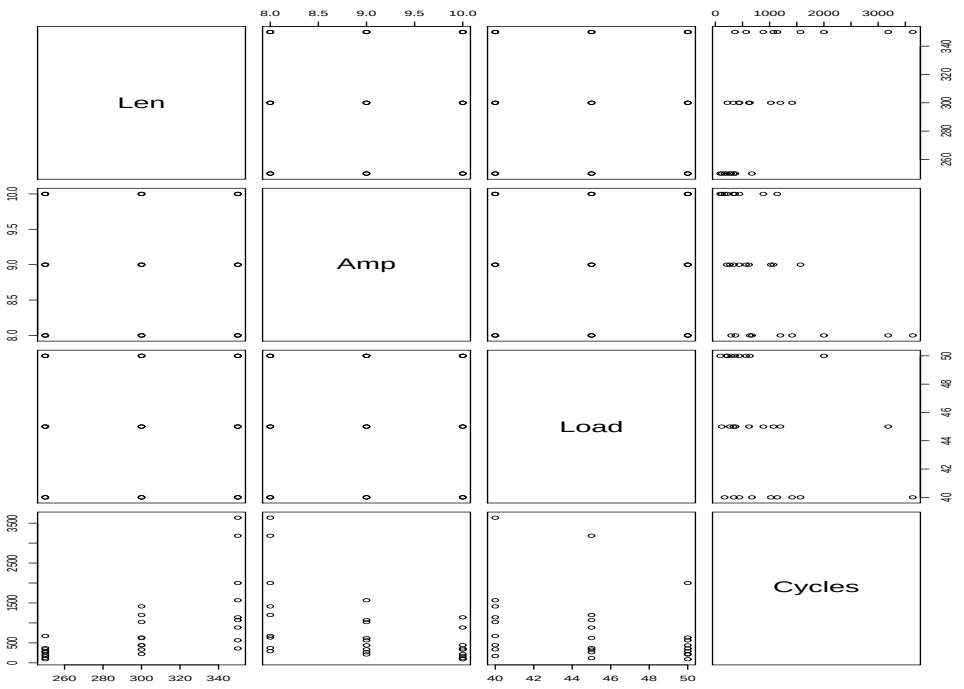

Figure 42: **Left:** The original scatter plot matrix of the variables *Cycles*, *Len*, *Amp*, and *Load* from the `wool` data set before any transformation. Note that the input specifications of *Len*, *Amp*, and *Load* are the result of an experimental design and thus have the given "checkerboard" appearance. The response *Cycles* is located in the lower right corner of this graph.

7.6 (transforming the wool data)

7.6.1: See Figure 42 (left) for a scatter plot matrix of the original variables for this problem.

7.6.2: When we make all three predictors *Len*, *Amp*, and *Load* factors we see that fitting the main effects with interactions second order model of *Cycles* on each of these factors using the regression

```
Cycles ~ lenF + ampF + loadF + lenF:ampF + lenF:loadF + ampF:loadF
```

where $lenF$, $ampF$, and $loadF$ are factors that we get summary statistics of $R^2 = 0.995$ and several highly significant parameter estimates indicating that the above model describes the data well.

7.6.3: When we fit the simpler model consisting of only the main effects

```
Cycles ~ lenF + ampF + loadF
```

we find that the $R^2 = 0.769$, which given the result above indicates that this simpler model is lacking. Using the `R` command `anova` to directly compare the two models indicates that there is a statistically significant reduction in variance in using the more complicated model and that it is to be preferred. We next use the `alr3` command `bctrans` to find an optimal power transformations for each of the independent variables *Len*, *Amp*, and *Load*. The optimal powers are approximately 0.5978 for each variable with standard errors that is much larger than this value. This indicates that a logarithmic transformation maybe appropriate for each independent variable. When we consider the inverse response plot for this data we find that a logarithm transformation of the response maybe helpful. Finally, using all of this information we fit the model

```
log(Cycles) ~ log(Len) + log(Amp) + log(Load)
```

and find a coefficient of determinism estimated at $R^2 = 0.966$ with all parameter estimates (but the intercept) well estimated. This indicates that nonlinear transformations can sometimes remove the need for interactions and the book calls this a *removable nonadditivity*.

See the `R` function `chap_7_prob_6.R` for code that implements this problem.

7.7 (justifying the transformation of $Miles$ in the Fuel data set)

One justification for transforming $Miles$ is provided by the *log* rule which states that if a variables range is more than one order of magnitude then replacing the variable by its loga-

rithm is likely to be helpful. For this variable the range of $Miles$ before and transformation is [1534300767] which satisfies the above criterion.

7.8 (predicting fertility)

7.8.1: For this problem we use the `alr3` command `bctrans` but with the "family" option set to "yeo.johnson", since that will allow for the fact that the variable $Change$ is negative. When we use this command we get a summary that looks like

```
> summary(ans)
yeo.johnson Transformations to Multinormality

           Est.Power Std.Err. Wald(Power=0) Wald(Power=1)
Change        0.9681   0.1193        8.1147       -0.2676
PPgdp        -0.1226   0.0506       -2.4224      -22.1812
Frate         1.1173   0.1776        6.2912        0.6607
Pop           0.0508   0.0346        1.4699      -27.4638
Fertility    -0.3099   0.1821       -1.7020       -7.1948
Purban        0.9629   0.1608        5.9896       -0.2309
```

Based on the estimated values and the standard errors we might consider the following values for the vector λ

$$(1, 0, 1, 0, 0, 1)\,.$$

For example, the first 1 in the vector above indicates that there should be no transformation for the variable $Change$ while the second 0 indicates that there should be a "yeo.johnson" logarithmic transformation of the variable $PPgdp$ etc.

7.8.2: We next look for a nonlinear transformation of the response $ModernC$. To do that we fit a linear model of $ModernC$ using the transformed variables found above and use the `inv.tran.plot` to determine a possible power transformation. The result from this is an estimated power of 0.793 with a standard error of 0.194. Given this result we accept this power and fit a linear model of the form

$$\begin{aligned} ModernC^{0.793} &= \beta_0 + \beta_1 Change + \beta_2 \log(PPgdp) + \beta_3 Frate + \beta_4 \log(Pop) \\ &+ \beta_5 \log(Fertility) + \beta_6 Purban\,. \end{aligned}$$

See the `R` function `chap_7_prob_8.R` for code that implements this problem.

Chapter 8 (Regression Diagnostics: Residuals)

Notes On The Text

It may help in reading this chapter to realized that much of the *practical* content of this chapter is about determining and testing the specific form that a given data set should have for a variance function. For most of the book we have assumed that

$$\mathrm{Var}(Y|X=x) = \sigma^2 \,.$$

This assumption may or may not be true for data set one may consider. The topics of the "score" test allow one to hypothesis different models for the variance function and to determine if the given data supports such a model. An example variance model of this type would be one that depended linearly on one of the predictors

$$\mathrm{Var}(Y|X=x) = \sigma^2 x \,.$$

The problems 8.3 and 8.4 in this chapter deal with such issues.

Notes on the residuals

In this section of these notes we derive some very simple properties of the residuals. The residuals $\hat{e}$ are defined as $\hat{e} = (I-H)Y$ so the variance of $\hat{e}$ is given by

$$\begin{aligned} \mathrm{Var}(\hat{e}) &= \mathrm{Var}(\hat{e}\hat{e}') \\ &= \mathrm{Var}((I-H)YY'(I-H)') \\ &= (I-H)\mathrm{Var}(YY')(I-H) \,. \end{aligned}$$

Since $H' = H$. To finish this evaluation we next need to compute $\mathrm{Var}(YY')$. Using the assumed model for Y of $Y = X\beta + e$ we see that

$$\mathrm{Var}(YY') = \mathrm{Var}((X\beta+e)(X\beta+e)') = \mathrm{Var}(ee') = \sigma^2 I \,.$$

Which gives that

$$\mathrm{Var}(\hat{e}) = \sigma^2(I-H)(I-H) = \sigma^2(I-2H+H^2) \,.$$

Next consider H^2. We see that

$$H^2 = X(X'X)^{-1}X'X(X'X)^{-1}X' = H \,.$$

Thus we finally get that

$$\mathrm{Var}(\hat{e}) = \sigma^2(I-H) \,, \tag{64}$$

as claimed in the books equation 8.5.

Notes on the hat Matrix H

In this subsection we derive some of the results quoted in the book in this section. First note that HX can be simplified as

$$HX = X(X'X)^{-1}X'X = X\,.$$

From this it follows that $(I-H)X = 0$. Now the covariance between the residuals $\hat{e}$ and the predictions $\hat{Y}$ or $\mathrm{Cov}(\hat{e},\hat{Y})$ can be computed as

$$\begin{aligned}\mathrm{Cov}(\hat{e},\hat{Y}) &= \mathrm{Cov}((I-H)Y, HY)\\ &= (I-H)\mathrm{Cov}(Y, HY)\\ &= (I-H)\mathrm{Cov}(Y,Y)H'\\ &= (I-H)\sigma^2 IH'\\ &= \sigma^2(I-H)H' = \sigma^2(I-H)H = 0\,,\end{aligned}$$

verifying the result presented in the book.

Now the expression for h_{ij} can be obtained from e_i and e_j using the inner product $e_i'He_j$. We can compute this expression in other ways by looking at the definition of H. We find

$$\begin{aligned}h_{ij} &= e_i'X(X'X)^{-1}X'e_j\\ &= (X'e_i)'(X'X)^{-1}(X'e_j)\,.\end{aligned}$$

Note that $X'e_i$ is the i-th column of X' which is the ith predictors from our data set i.e. x_i. Thus we see that

$$h_{ij} = x_i'(X'X)^{-1}x_j = h_{ji}\,, \tag{65}$$

which is again claimed in the book.

To evaluate the sum $\sum_{i=1}^n h_{ii}$ use the above expression to get

$$\begin{aligned}\sum_{i=1}^n h_{ii} &= \sum_{i=1}^n x_i'(X'X)^{-1}x_i\\ &= x_1'(X'X)^{-1}x_1 + x_2'(X'X)^{-1}x_2 + \cdots + x_n'(X'X)^{-1}x_n\\ &= \mathrm{tr}\left(\begin{bmatrix} x_1' \\ x_2' \\ \vdots \\ x_n' \end{bmatrix}(X'X)^{-1}\begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}\right)\\ &= \mathrm{tr}(X(X'X)^{-1}X') = \mathrm{tr}(X'X(X'X)^{-1}) = \mathrm{tr}(I) = p'\,,\end{aligned}$$

or equation 8.8. Since when the linear model includes a constant we have that $\hat{e}'\mathbf{1} = 0$, we can use this relationship to derive an expression involving the components of the hat matrix H. Using the fact that $\hat{e} = Y - \hat{Y}$ in that expression we get

$$\hat{e}'\mathbf{1} = Y'(I-H)\mathbf{1} = 0\,.$$

This last expression in tern implies that $(I-H)\mathbf{1}=0$, or $\mathbf{1}=H\mathbf{1}$. The ith equation in this later system is given by

$$1=\sum_{j=1}^{n} h_{ij}\,, \tag{66}$$

or the books equation 8.9.

When h_{ii} gets closer to 1 we can show that the prediction at x_i or $\hat{y}_i$ limits to y_i the measured value. To see this consider the ith equation from $\hat{Y}=HY$ which is

$$\hat{y}_i=\sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i+\sum_{j\neq i} h_{ij}y_j\,. \tag{67}$$

Then as $h_{ii}\to 1$ we see that since $\sum_{i=1}^{n} h_{ij}=1$ (derived in Equation 66) that the sum of h_{ij} (without the term h_{ii}) limits to

$$\sum_{i\neq j} h_{ij}\to 0\,. \tag{68}$$

Then using Equation 67 we have

$$\begin{aligned} |\hat{y}_i-h_{ii}y_i| &\leq |\sum_{j\neq i} h_{ij}y_j| \leq \sum_{i\neq j}|h_{ij}||y_j| = \sum_{i\neq j} h_{ij}|y_j| \\ &\leq \max_j(|y_j|)\sum_{i\neq j} h_{ij}\,. \end{aligned}$$

This later expression goes to zero by Equation 68 and we have that $\hat{y}_i\to y_i$. Note that we can drop the absolute value on h_{ii} in the above derivation because h_{ii} is positive by the fact that $h_{ii}\geq\frac{1}{n}$.

Notes on non-constant variance

In the book when considering the possible variance models for the `sniffer` data set the comment is made that since the two pressure variables $TankPres$ and $GasPres$ are very linearly related we might not want to use both of them in the mean function. One of the reasons for this is that the variance of our *estimates* of the coefficients in the linear regression, β_k, become much worse as the predictors get more and more correlated. If we desire or require accurate values for β_k then excluding one of the predictors in our linear model might be a good idea.

Problem Solutions

8.1 (working with the hat matrix H)

8.1.1: For this part of the problem see the results on Page 105.

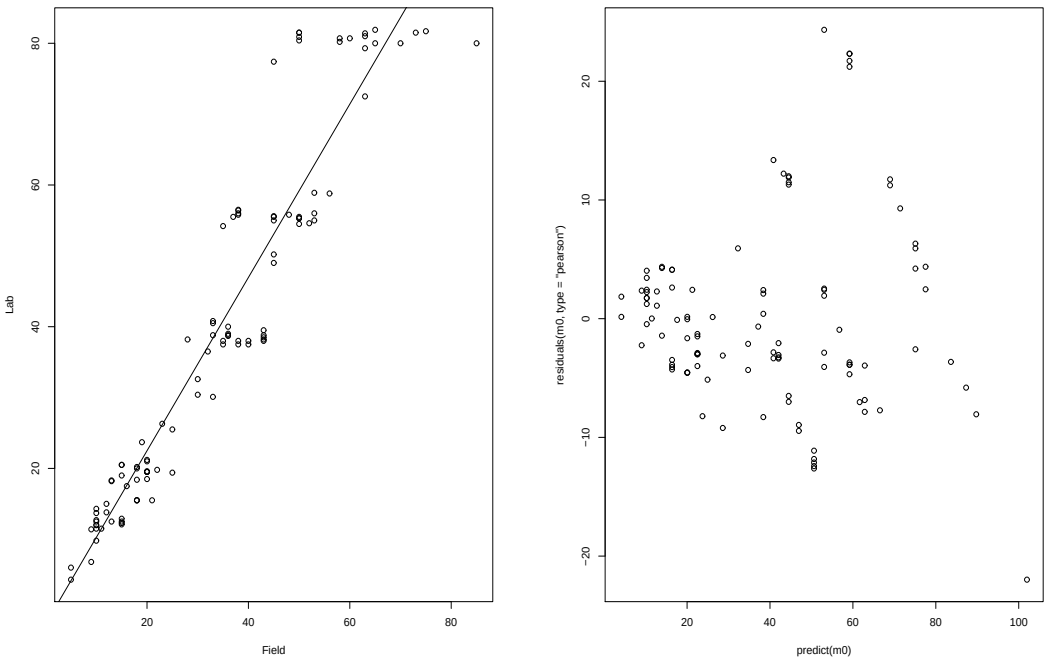

Figure 43: **Left:** A scatter plot of *Lab* vs. *Field* for the `pipeline` data set and the OLS linear fit. Note that the variance of the scatter points seems to change as *Field* increases. **Right:** The residual plot for the `pipeline` data set. Again notice the non-constant variance.

8.2 (removing the linear trend in *Fuel* vs. $\hat{y}$)

The plot of the response *Fuel* as a function of the fitted values when the linear trend is removed is equivalent to a plot of the residuals vs. the fitted values $\hat{y}$, which is Figure 8.5e.

8.3 (defects in the Alaska oil pipeline)

8.3.1: See Figure 43 (left) for a scatter plot of *Lab* vs. *Field*. Linear regression looks like it would work reasonable well but the variance of the fit does not look to be constant.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pipeline$Field	1	59.17	59.17	18.52	0.0000
Residuals	105	335.43	3.19		

Table 9: The ANOVA table for the score test for the non-constant variance model $\text{Var}(Lab|Field = f) = \sigma^2 f$ for Problem 8.3.2. The large F-value indicates that this is a better variance model than a constant.

8.3.2: In Figure 43 (right) we have presented a residual plot. A residual plot is a scatterplot of the Pearson residuals $\hat{e}_i$ vs. the fitted values $\hat{y}_i$. We next consider the score test where

we consider a variance model that is linear in the variable $Field$. Using $Field$ to predict the scaled squared residuals has an ANOVA table is given in Table 9. From Table 9 the score test for non-constant variance computes the value $S = (1/2)SSreg = (1/2)59.17 = 29.58$, which is to be compared with the chi-squared distribution with one degree of freedom. This has a very small p-value indicating that this data almost certainly has a non-constant variance and that this variance model is better than a constant.

8.3.3: For this subproblem we use weighted least squares to attempt to better predict the observed variance for the given data and then repeat the score test. We obtain the ANOVA table for this second test given in Table 10. Again the score tests indicates that there is strong evidence for a variance different than the one specified.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(1/Field)	1	49.97	49.97	11.38	0.0010
Residuals	105	461.07	4.39		

Table 10: The ANOVA table for the score test corresponding to the non-constant variance model $\mathrm{Var}(Lab|Field = f) = \sigma^2/f$ for Problem 8.3.3.

8.3.4: For this subproblem we use weighted least squares to attempt to better predict the observed by using a variance model of $\mathrm{Var}(Lab|Field) = \sigma^2/Field^2$. The ANOVA table for this model is given in Table 11. The score tests in this case gives a p-value of 0.01 the largest seen under any of the variance models. This is clearly the best model specified.

See the `R` function `chap_8_prob_3.R` for code that implements this problem.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
I(1/Field^2)	1	12.11	12.11	0.91	0.3431
Residuals	105	1402.26	13.35		

Table 11: The ANOVA table for the score test corresponding to the non-constant variance model $\mathrm{Var}(Lab|Field = f) = \sigma^2/f^2$ for Problem 8.3.4.

8.4 (determining a variance model for the `stopping` data set)

In Figure 44 we display the residuals of a linear fit of $Speed$ vs. $Distance$ under the assumption of constant variance for various possible predictors. Specifically, we consider possible models for the variance given by $Speed$, $Speed^2$, and $\hat{y}$. We next use the score test to determine which of the possible predictors given maybe best to use in predicting the variance. The score tests for the variance residual models is given in Table 12. These results indicate that the variances for this problem are almost certainly *not* constant. The book discusses a test using the χ^2 distribution to determine if the *nested* variance models reduces the variance significaltly relative to smaller less general models. That would be used here to determine if the addition of $Speed^2$ was a useful modification. The S values in Table 12 give an indication that this variable is helpful.

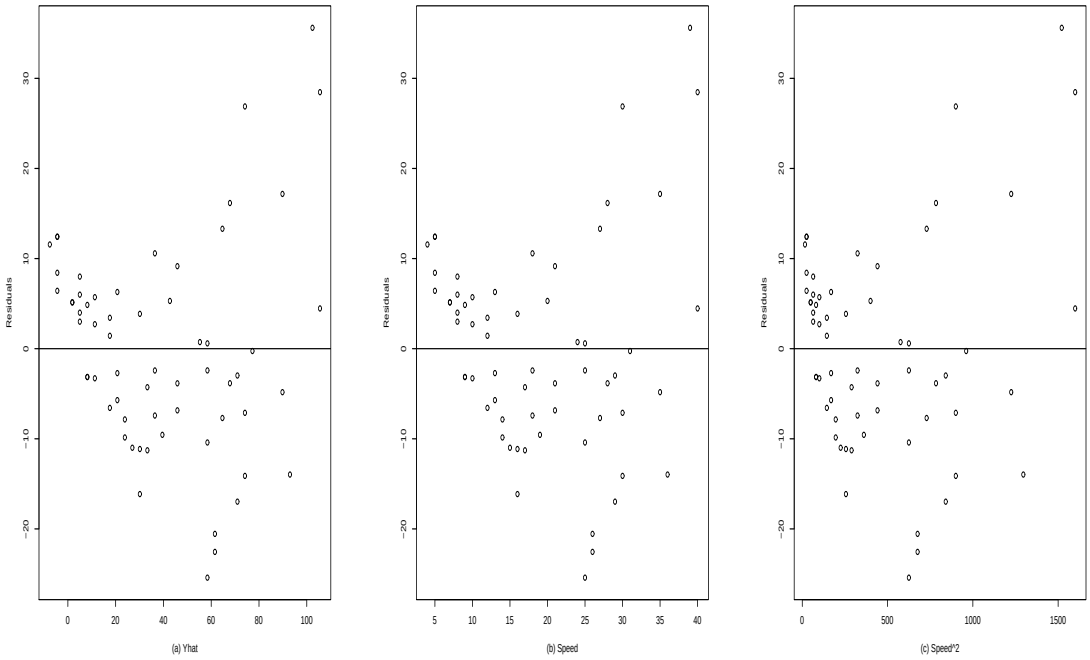

Figure 44: Three residual plots for the `stopping` data set. **Left:** A scatter plot of the residuals vs. the fitted values. **Middle:** A scatter plot of the residuals vs. *Speed*. **Right:** A scatter plot of the residuals vs. $Speed^2$.

	df	S	p
~Speed	1.00	20.49	0.00
~Speed + I(Speed^2)	2.00	27.44	0.00
~I(Speed^2)	1.00	25.11	0.00
~fitted.values	1.00	20.49	0.00

Table 12: The ANOVA table for the score test corresponding to different variance models for Problem 8.4.

See the `R` function `chap_8_prob_4.R` for code that implements this problem.

8.5 (leverages in the simple regression model $E[Y|X = x] = \beta_0 + \beta_1 x$)

8.5.1: As demonstrated in Chapter 3 of this book, when we consider the simple linear regression in matrix terms we have

$$(X'X)^{-1} = \frac{1}{SXX} \begin{bmatrix} \frac{1}{n}\sum x_k^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix},$$

with $\bar{x} = \frac{1}{n}\sum x_k$ and SXX given by Equation 104. Using the expression for h_{ij} given by Equation 65 above we see that

$$\begin{aligned} h_{ij} &= x_i'(X'X)^{-1}x_j \\ &= \begin{bmatrix} 1 & x_i \end{bmatrix} \left(\frac{1}{SXX} \begin{bmatrix} \frac{1}{n}\sum x_k^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \right) \begin{bmatrix} 1 \\ x_j \end{bmatrix} \\ &= \frac{1}{SXX} \begin{bmatrix} 1 & x_i \end{bmatrix} \begin{bmatrix} \frac{1}{n}\sum x_k^2 - x_j\bar{x} \\ -\bar{x} + x_j \end{bmatrix} \\ &= \frac{1}{SXX}\left(\frac{1}{n}\sum x_k^2 - (x_i + x_j)\bar{x} + x_i x_j \right) \end{aligned} \tag{69}$$

Thus the leverage h_{ii} is given by

$$h_{ii} = \frac{1}{SXX}\left(\frac{1}{n}\sum x_k^2 - 2x_i\bar{x} + x_i^2 \right) . \tag{70}$$

8.5.3: For this we want $h_{ii} = 1$ which using Equation 70 this requires

$$x_i^2 - 2\bar{x}x_i + \frac{1}{n}\sum x_k^2 - SXX = 0 .$$

In this later expression $\bar{x}$, SXX, and $\sum x_k^2$ all have the variable x_i in them. One would need to write the above expression as an expression in x_i. For example one would need to express

$$\bar{x} = \frac{1}{n}x_i + \frac{1}{n}\sum_{k;k\neq i} x_k ,$$

and the same for the other expressions. One would then obtain a quadratic equation for the variable x_i that could be solved using the quadratic equation to given the value of x_i for which $h_{ii} = 1$.

8.6 (an expression for h_{ii} in terms of the QR factorization of X)

Factoring X using the QR factorization as $X = QR$ the hat matrix H becomes

$$\begin{aligned} H &= X(X'X)^{-1}X' \\ &= QR(R'Q'QR)^{-1}R'Q' \\ &= QR(R'R)^{-1}R'Q' \\ &= QRR^{-1}R'^{-1}R'Q' \\ &= QQ' . \end{aligned}$$

Using this expression we can write h_{ij} as

$$h_{ij} = e_i'He_j = (e_i'Q)(Q'e_j) = (Q'e_i)'(Q'e_j) ,$$

where e_i is a vector of all zeros with a single one at the ith location. Since $Q'e_i$ is the ith column of Q' it is also the ith row of Q. Defining the ith row of Q as q_i we see that

$$h_{ij} = q_i'q_j .$$

8.7 (the values of h_{ij} for a special regression)

8.7.1: For this particular vector $U = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ the fitted values of the regression of U on X are given by HU. Since U has a single 1 in the first component we have that $\hat{U} = HU = \begin{bmatrix} h_{11} \\ h_{21} \\ \vdots \\ h_{n1} \end{bmatrix}$. Since H is symmetric this is equivalent to a vector with components h_{1j} for $j = 1, \ldots, n$.

8.7.2: The vector of residuals $\hat{e}$ is defined as $U - \hat{U}$ which by the way that U is defined and the above result gives

$$\begin{aligned} \hat{e}_1 &= 1 - h_{11} \\ \hat{e}_j &= 0 - h_{1j} = -h_{1j} \quad \text{for} \quad j > 1\,. \end{aligned}$$

8.8 (orthogonal matrices)

To show that H and $I - H$ are orthogonal see the results on Page 105. To consider the slope of the regression of $\hat{e}$ on $\hat{Y}$ recall that these two expressions are given by

$$\begin{aligned} \hat{Y} &= HY \\ \hat{e} &= Y - \hat{Y} = Y - HY = (I - H)Y\,. \end{aligned}$$

The slope of the regression of $\hat{e}$ onto $\hat{Y}$ is the estimated slope where the x-variable is $\hat{Y}$ and the y-variable is $\hat{e}$. This is given by Equation 110 or

$$\hat{\beta}_1 = \frac{S\hat{Y}\hat{e}}{S\hat{Y}\hat{Y}}\,.$$

Now since an intercept is included in the regression the mean of $\hat{e}$ is zero so using Equation 106 to evaluate $S\hat{Y}\hat{e}$ we see that

$$\begin{aligned} S\hat{Y}\hat{e} &= \sum \hat{y}_i \hat{e}_i = \hat{Y}'\hat{e} \\ &= (HY)'(I - H)Y \\ &= Y'H'(I - H)Y \\ &= Y'H(I - H)Y = 0\,, \end{aligned}$$

since H is symmetric and $I - H$ are orthogonal.

8.9 (the hat matrix with weighted errors)

Let $W^{1/2}$ be the $n \times n$ diagonal matrix with elements $\sqrt{w_i}$ and $W^{-1/2}$ its corresponding inverse. Then define $\hat{Y} = W^{1/2}Y$. Assuming the suggested model for Y is true that is

$Y = X\beta + e$, for some value of β we see that $\hat{Y}$ has the following linear model

$$\hat{Y} = W^{1/2}Y = W^{1/2}X\beta + W^{1/2}e\,.$$

The error term in the regression of $\hat{Y}$ onto $W^{1/2}X$ then has a variance given by

$$\begin{aligned}\mathrm{Var}(W^{1/2}e) &= W^{1/2}\mathrm{Var}(ee')W^{1/2} \\ &= W^{1/2}(\sigma^2 W^{-1})W^{1/2} \\ &= \sigma^2 I\,.\end{aligned}$$

Thus we can apply OLS regression on the variable $W^{1/2}X$. The hat matrix H for these variables is given by

$$\begin{aligned}H &= (W^{1/2}X)((W^{1/2}X)'(W^{1/2}X))^{-1}(W^{1/2}X)' \\ &= (W^{1/2}X)(X'W^{1/2}W^{1/2}X)^{-1}X'W^{1/2} \\ &= W^{1/2}X(X'WX)^{-1}X'W^{1/2}\,,\end{aligned}$$

the desired expression.

8.10 (residual plots of the California water data set)

The mean function described in Problem 7.3.3 is given by

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

We can use the `alr3` command `residual.plots` to generate the residual plots for the above model. This command also generates tests for curvature which are given by

```
> residual.plots( m0 )
            Test stat   Pr(>|t|)
logAPMAM     0.4499390 0.65552893
logAPSAB    -0.4647128 0.64501524
logAPSLAKE -0.8524521 0.39975903
logOPBPC     1.3848392 0.17486642
logOPRC      0.8386546 0.40735461
logOPSLAKE   1.6295066 0.11217455
Tukey test   1.8386288 0.06596981
```

Based on this result only the "Tukey test" seems to be somewhat (although not overly so) significant. This means that there is some chance that there is a dependence of the variance on the value of the response y and perhaps variance stabalization would help. Using the `alr3` command `mmps` show graphically that the linear fit seems to be a good one.

See the `R` function `chap_8_prob_10.R` for code that implements this problem.

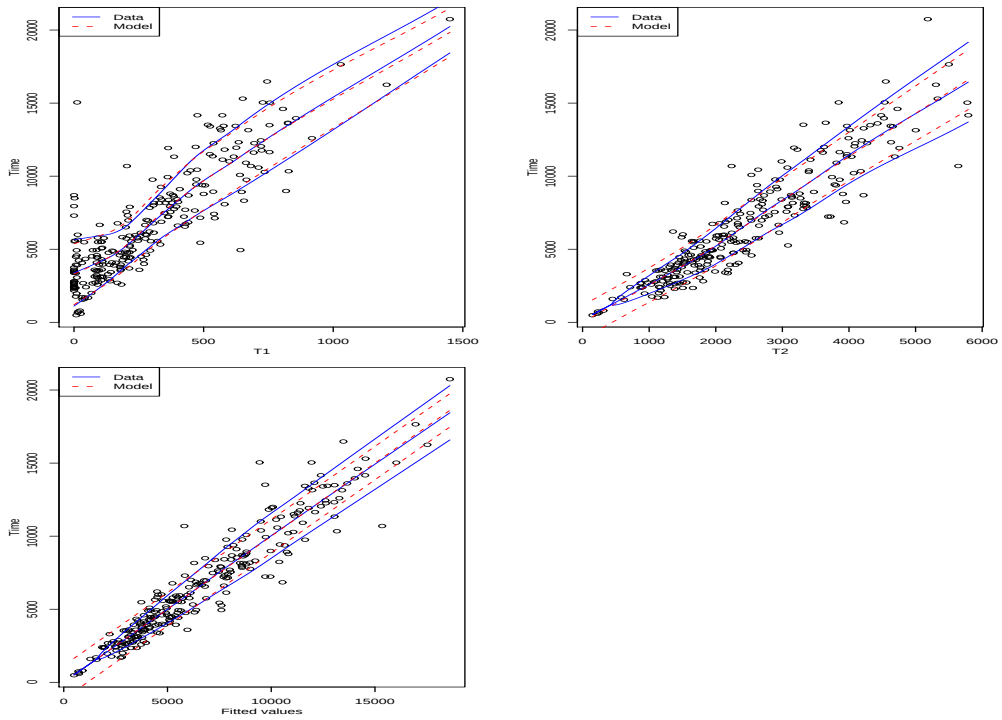

Figure 45: Plots from the `mmps` command for the `transaction` data set.

8.11 (marginal model plots of the transaction data)

This data set is discussed on page 102 in Section 4.6 in the book where the following model for $Time$ is proposed

$$E(Time|T_1, T_2) = \beta_0 + \beta_1 T_1 + \beta_2 T_2 \, .$$

We fit this model and then use the `alr3` command `mmps` to assess visually the mean and variance specifiations. When we run this command we get the results shown in Figure 45. The loess curves and the linear curves are relativly close indicating that the fit is a good one.

See the `R` function `chap_8_prob_11.R` for code that implements this problem.

8.12 (modeling crustacean zooplankton)

For this problem we are given the `lakes` data set which has a large number of features and are asked to perform a modeling study of this data. Since there were quite a few variable as a first step I computed the correlation matrix of all predictors and the single response *Species*. To simplify the complexity of the problem I then choose to predict *Species* based on the three predictors that are most correlated with it. This resulted in the variables (and correlations values) of

```
Dist        Elev        Area
```

```
0.35328247 0.40441612 0.74810087
```

The range of the variables *Dist* and *Elev* span three orders of magintude indicating that a power (i.e. logarithm) transformation maybe helpful. I choose to use the `alr3` command `bctrans` to search for optimal power transformations. The edited summary of running that command is given by

```
box.cox Transformations to Multinormality

     Est.Power Std.Err. Wald(Power=0) Wald(Power=1)
Elev    0.1863   0.0725        2.5689      -11.2229
Dist   -0.3597   0.1432       -2.5124       -9.4963
Area   -0.0292   0.0328       -0.8910      -31.3975
```

From this summary power transforamtion of zero (or logarithmic transformation) for all the independent variables are reasonable approxmations. We also find using the `alr3` command `inv.trans.plot` that *no* transformation of the response seems to help the linear regression. A scatter plot matrix of all three terms and the response is given in Figure 46 (left). From there we see that the variable $log(Elev)$ does not seem to predict *Species* vary well. In addition, when we fit a linear model using these three terms the coefficient of $log(Elev)$ has a p-value of 0.32 indicating that its value is not well known. Based on this evidence I decided to drop this term from our model. Next we plotted to study if the assumption of constant variance is violated. Using the `alr3` command `plot.residuals` we obtain the plot in Figure 46 (right) with summary statistics given by

```
           Test stat   Pr(>|t|)
logDist    -1.042883 0.30458247
logArea     1.911697 0.06463105
Tukey test  1.758213 0.07871135
```

Since these p-values are relativly large (at least they are not zero) we might conclude that the constant variance specification is reasonably sound. If we were to include a term in the variance because the $log(Area)$ coefficient has the smallest p-value we should begin with a variance model like

$$\mathrm{Var}(Species|log(Dist) = d, log(Area) = a) = \sigma^2 a\,.$$

Finally, using the command `mmps` command we can visually compare the least squares fit with the provided loess fits. Unfortunatly, the two do not seem to agree very well for the mean function we have specified. This indicates that we should revist the choice of predictors and try to get a better match.

See the `R` function `chap_8_prob_12.R` for code that implements this problem.

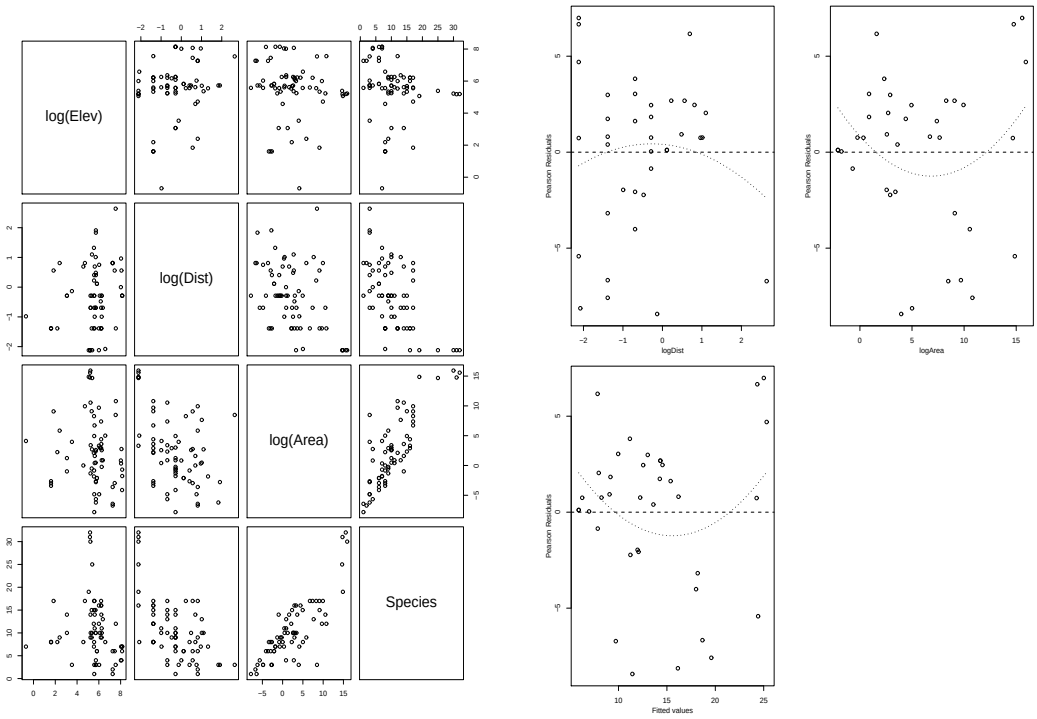

Figure 46: **Left:** A scatter plot matrix of three possible predictors in the `lakes` data set. **Right:** Residual plots corresponding to the model $E(Species|log(Dist) = d, log(Aread) = a) = \beta_0 + \beta_1 d + \beta_2 a$.

Chapter 9 (Outliers and Influence)

Notes On The Text

Notes on an outlier test

We will derive an alternative expression for t-tests used to determine in the mean shift outlier model if each given point x_i is an outlier. The t-tests for each sample x_i involve computing

$$t_i = \frac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1 + x_i'(X_{(i)}'X_{(i)})^{-1}x_i}}\,. \tag{71}$$

We will write t_i in terms of the standardized residuals r_i which are given by

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}\,. \tag{72}$$

To do this recall that $\hat{y}_{i(i)} = x_i'\hat{\beta}_{(i)}$ so from Problem 9.4

$$y_i - \hat{y}_{i(i)} = y_i - x_i'\hat{\beta}_{(i)} = \frac{\hat{e}_i}{1-h_{ii}}\,.$$

From Equation 121 we have that the inner product of the matrix $X_{(i)}'x_{(i)}$ with x_i is given by

$$\begin{aligned} x_i'(X_{(i)}'X_{(i)})^{-1}x_i &= h_{ii} + \frac{1}{1-h_{ii}}x_i'(X'X)^{-1}x_ix_i'(X'X)^{-1}x_i \\ &= h_{ii} + \frac{h_{ii}^2}{1-h_{ii}} = \frac{h_{ii}}{1-h_{ii}}\,. \end{aligned}$$

Using these expressions t_i is given by

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1+\frac{h_{ii}}{1-h_{ii}}}}\left(\frac{1}{1-h_{ii}}\right) = \frac{\hat{e}_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}\,,$$

which is one of the equations in the books 9.4. Transforming this expression into one involving the standardized residual r_i and using the books A.39 of

$$\hat{\sigma}_{(i)}^2 = \hat{\sigma}^2\left(\frac{n-p'-1}{n-p'-r_i^2}\right)^{-1}\,,$$

we have

$$t_i = \frac{\hat{\sigma}}{\hat{\sigma}_{(i)}}r_i = \left(\frac{n-p'-1}{n-p'-r_i^2}\right)^{1/2}r_i\,, \tag{73}$$

which is the second expression in the books equation 9.4. The benefit of this expression over that given in Equation 71 is that without any modification this later expression would require computing $(X_{(i)}'X_{(i)})^{-1}$ for each of $i = 1, \cdots, n$ which could be computationally costly.

Notes on Cook's distance

Cook's distance D_i is defined as

$$D_i = \frac{(\hat{\beta}_{(i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(i)} - \hat{\beta})}{p'\hat{\sigma}^2} \,. \tag{74}$$

By moving the X in the "sandwiched" expression $X'X$ to each of the beta vectors in the inner product we can write D_i as

$$D_i = \frac{(X(\hat{\beta}_{(i)} - \hat{\beta}))'(X(\hat{\beta}_{(i)} - \hat{\beta}))}{p'\hat{\sigma}^2} = \frac{(\hat{Y}_{(i)} - \hat{Y})'(\hat{Y}_{(i)} - \hat{Y})}{p'\hat{\sigma}^2} \,,$$

which is the books equation 9.7.

Notes on computing D_i

Starting with the definition of Cook's distance D_i given by Equation 74 and using Equation 125 we can write D_i as

$$\begin{aligned} D_i &= \frac{1}{p'\hat{\sigma}^2} \left(\frac{\hat{e}_i (X'X)^{-1} x_i}{1 - h_{ii}} \right)' (X'X) \left(\frac{\hat{e}_i (X'X)^{-1} x_i}{1 - h_{ii}} \right) \\ &= \frac{\hat{e}_i^2}{p'\hat{\sigma}^2 (1 - h_{ii})^2} x_i' (X'X)^{-1} (X'X) (X'X)^{-1} x_i = \frac{\hat{e}_i^2 h_{ii}}{p'\hat{\sigma}^2 (1 - h_{ii})^2} \,. \end{aligned}$$

Next using the definition of the standardized residual, r_i, given in Equation 72 to write the $\hat{e}_i$ above in terms of r_i we get

$$D_i = \frac{r_i^2 h_{ii}}{p'(1 - h_{ii})} \,, \tag{75}$$

which is the books equation 9.8.

Problem Solutions

9.1 (examples computing r_i, D_i, and t_i)

To solve this problem we will first compute r_i using Equation 72. Next using the values of r_i in Equation 75 we will compute D_i. Finally, we will use Equation 73 to compute t_i. To test each case to be an outlier we will compare the values of t_i to the quantiles of a t-distribution with $n - p' - 1$ degrees of freedom. When we do this we get the following p-values for each of the given cases

```
[1] 0.434861199 0.392019262 0.007963954 0.003330443
```

Thus the last two points (the ones with the largest $\hat{e}_i$ values) are candidates for outliers. The influence of each sample on the estimated coefficients is determined by looking at the Cook distances D_i. The values of D_i computed for this data are given by

```
[1] 1.1250000 0.4499736 0.4500000 0.3689939
```

which make the argument that the most influential point is the *first* one even though it does not have a seemingly large value of the residual $\hat{e}_1$.

See the `R` function `chap_9_prob_1.R` for code that implements this problem.

9.2 (examples computing r_i, D_i, and t_i)

From the given specification of the mean we see that $p = 4$ so that $p' = 5$. Since the degrees of freedom df of the residuals is given by $df = n - p'$ since we are told $df = 46$ we find $n = df + p' = 51$. We will follow the same specifications as in Problem 9.1 and find that the values we find for t_i are given by

```
[1] -3.1927822 -2.4376317 -1.8147106  3.2465847 -0.9962917
```

indicating that the first (Alaska) and the second from the last (Wyoming) are candidates for outliers. The values we find for Cooks distance D_i are given by

```
[1] 0.5846591 0.2074525 0.1627659 0.1601094 0.1408527
```

Indicating that the most influential measurement is the first (Alaska again). Depending on the use of this model maybe the data point represented by Alaska should be removed.

See the `R` function `chap_9_prob_2.R` for code that implements this problem.

9.3 (proving the case deletion matrix inverse lemma)

This problem is worked in the Appendix on Page 161.

9.4 (deriving the PRESS or predicted residual)

Using Equation 125 we can write the *predicted* residual or PRESS as

$$y_i - x_i'\hat{\beta}_{(i)} = y_i - x_i'\left(\hat{\beta} - \frac{(X'X)^{-1}x_i\hat{e}_i}{1-h_{ii}}\right)$$

$$= \hat{e}_i + \frac{h_{ii}\hat{e}_i}{1-h_{ii}} = \frac{\hat{e}_i}{1-h_{ii}}\,,$$

the requested expression. Note we have used the fact that $\hat{e}_i \equiv y_i - x_i'\beta$.

9.5 (deriving the expression for D_i)

See the notes in this text given on Page 117 for this derivation.

9.6 (deriving the expression for D_i^*)

Warning: There seems to be an error in the following derivation, since it does not match the final result from the text. I'm not sure where the error in this derivation might be. If anyone finds anything incorrect with this argument please contact me.

We will define D_i^* as

$$D_i^* = \frac{1}{p\hat{\sigma}^2}(\hat{\beta}^*_{(i)} - \beta^*)'(\mathcal{X}'\mathcal{X})(\hat{\beta}^*_{(i)} - \beta^*)\,. \tag{76}$$

To simplify this we will start with Equation 125 for the *full* vector $\hat{\beta}$ but written as

$$(X'X)(\hat{\beta}^*_{(i)} - \beta^*) = -\frac{\hat{e}_i}{1-h_{ii}}x_i\,.$$

Using this expression we will follow the steps we performed on Page 25 of these notes. To do that we define $\Delta\hat{\beta}$ as

$$\Delta\hat{\beta} = \hat{\beta}_{(i)} - \beta\,,$$

with two parts $\Delta\hat{\beta}_0$ and $\Delta\hat{\beta}^*$ just as we had done before when we split β into two pieces. We get

$$\begin{bmatrix} n & n\bar{x}' \\ n\bar{x} & V'V \end{bmatrix}\begin{bmatrix} \Delta\hat{\beta}_0 \\ \Delta\hat{\beta}^* \end{bmatrix} = -\frac{\hat{e}_i}{1-h_{ii}}\begin{bmatrix} 1 \\ x_i^* \end{bmatrix}\,.$$

We will multiply the first equation above by $1/n$, write out explicitly the first equation in terms of $\Delta\hat{\beta}_0$ and $\Delta\hat{\beta}^*$ and then solve for $\Delta\hat{\beta}_0$, which is then put into the second equation. This gives a single equation for $\Delta\hat{\beta}^*$ which is

$$(V'V - n\bar{x}\bar{x}')\Delta\hat{\beta}^* = (\bar{x} - x_i^*)\frac{1}{1-h_{ii}}\hat{e}_i\,.$$

as in Equation 19 we find that the coefficient of $\Delta\hat{\beta}^*$ simplifies to $\mathcal{X}'\mathcal{X}$. We finally get

$$\Delta\hat{\beta}^* = (\mathcal{X}'\mathcal{X})^{-1}(\bar{x} - x_i^*)\frac{\hat{e}_i}{1-h_{ii}}\,.$$

Using this expression we can evaluate D_i^* to get

$$\begin{aligned} D_i^* &= \frac{1}{p\hat{\sigma}^2}\left((\mathcal{X}'\mathcal{X})^{-1}(\bar{x} - x_i^*)\frac{\hat{e}_i}{1-h_{ii}}\right)'(\mathcal{X}'\mathcal{X})\left((\mathcal{X}'\mathcal{X})^{-1}(\bar{x} - x_i^*)\frac{\hat{e}_i}{1-h_{ii}}\right) \\ &= \frac{\hat{e}_i^2}{(1-h_{ii})^2}\frac{1}{p\hat{\sigma}^2}(\bar{x} - x_i^*)'(\mathcal{X}'\mathcal{X})^{-1}(\bar{x} - x_i^*)\,. \end{aligned}$$

Using the fact that r_i and $\hat{e}_i$ are related by

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \Rightarrow \hat{e}_i^2 = \hat{\sigma}^2(1-h_{ii})r_i^2\,,$$

we can write D_i^* as

$$, D_i^* = \frac{{r_i}^2}{p(1-h_{ii})}(\bar{x}-x_i^*)'(\mathcal{X}'\mathcal{X})^{-1}(\bar{x}-x_i^*)\,.$$

To finish this derivation recall the books equation 8.11 or

$$h_{ii} = \frac{1}{n} + (x_i^*-\bar{x})'(\mathcal{X}'\mathcal{X})^{-1}(x_i^*-\bar{x})\,. \tag{77}$$

Using this we can replace the expression $(x_i^*-\bar{x})'(\mathcal{X}'\mathcal{X})^{-1}(x_i^*-\bar{x})$ with $h_{ii}-\frac{1}{n}$ to get

$$D_i^* = \frac{{r_i}^2}{p}\left(\frac{h_{ii}-1/n}{1-h_{ii}}\right)\,.$$

This is different than the books result in that the denominator does not have the $1/n$ term. If anyone sees an error in this calculation please email me.

9.8 (elections in Florida)

In Figure 47 (left) we display the scatter plot of the number of votes for *Buchanan* vs. *Bush* found in the `florida` data set. Notice the potential outlier at the top of this graph. We next compute the values of t_i for each sample and find that the largest one is located at the 50-th location with values

```
> florida[spot,]
       County   Gore   Bush Buchanan
50 PALM BEACH 268945 152846     3407
```

The p-value for this element using the Bonferroni bound is zero indicating that there is "no" chance that this residual is this large by chance. We thus conclude that this point is an outlier. If we next look for the next largest value of t_i (in search for another outliers) we find its value is given by -3.280 at the spot containing the data

```
> florida[spot,]
   County   Gore   Bush Buchanan
13   DADE 328702 289456      561
```

The Bonferroni p-value expression computed with `n*2*(1-pt(t,n-pprime-1))` has a value greater than one indicating that we should truncate its value to 1. Thus we conclude that this point is *not* a candidate for an outlier.

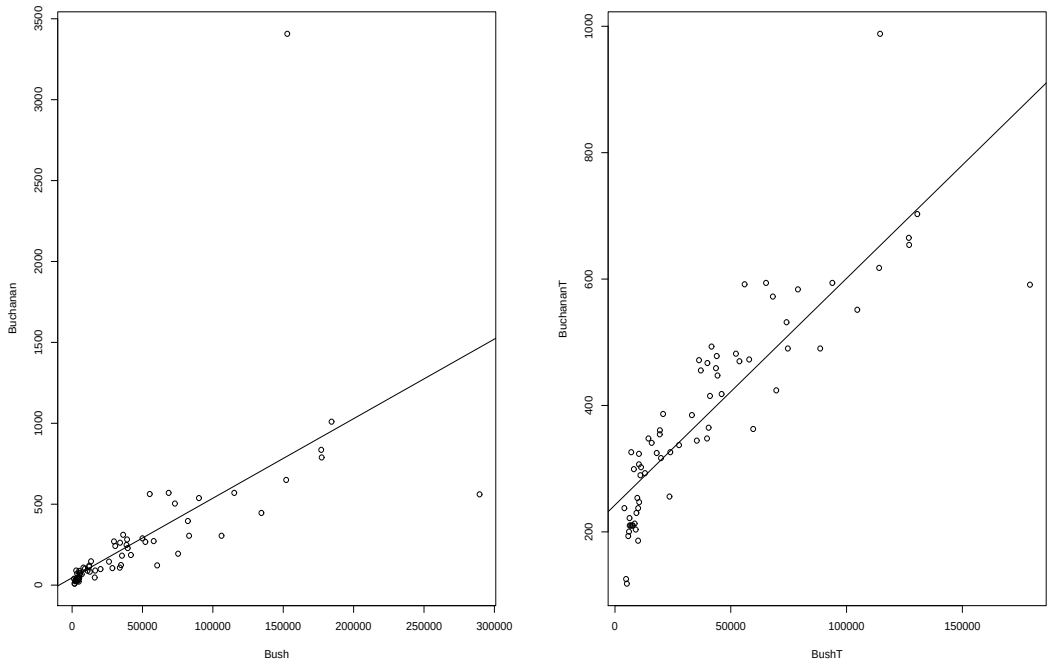

Figure 47: **Left:** A scatter plot of *Buchanan* and *Bush* votes (unmodified) from the `florida` data set and the corresponding ordinary least squares fit. **Right:** A scatter plot of the *transformed Buchanan* and *Bush* data and the corresponding ordinary least squares fit.

We next try to transform the independent and dependent variables using the techniques from this book. When we use the `alr3` command `inv.tran.estimate` we find a possible scaled power transformation for the *Bush* variable. This command gives

```
      lambda           se          RSS
7.016553e-01 2.556051e-01 7.934565e+06
```

and visually the fitted mean function using the power 0.7 looks very similar to the same thing under no transformation (using the value of 1.0). In addition, the standard error of the above estimate indicates that the value of 0.7 may not be sufficiently different than 1.0. In any case we will accept the value of 0.7 as valid and perform this power transform on the raw *Bush* vote data. Next we look for a transformation of the dependent variable *Buchanan*. The scaled power transformation 0.23 seems to be a good fit. When we perform these transformation we get the plot shown in Figure 47 (right). We can run the same tests as performed earlier. We find the same two points are possibly outliers and again find that "Palm Beach" is very likely an outlier but that "Dade" is most likely not. This is the same conclusion reached earlier.

See the `R` function `chap_9_prob_8.R` for code that implements this problem.

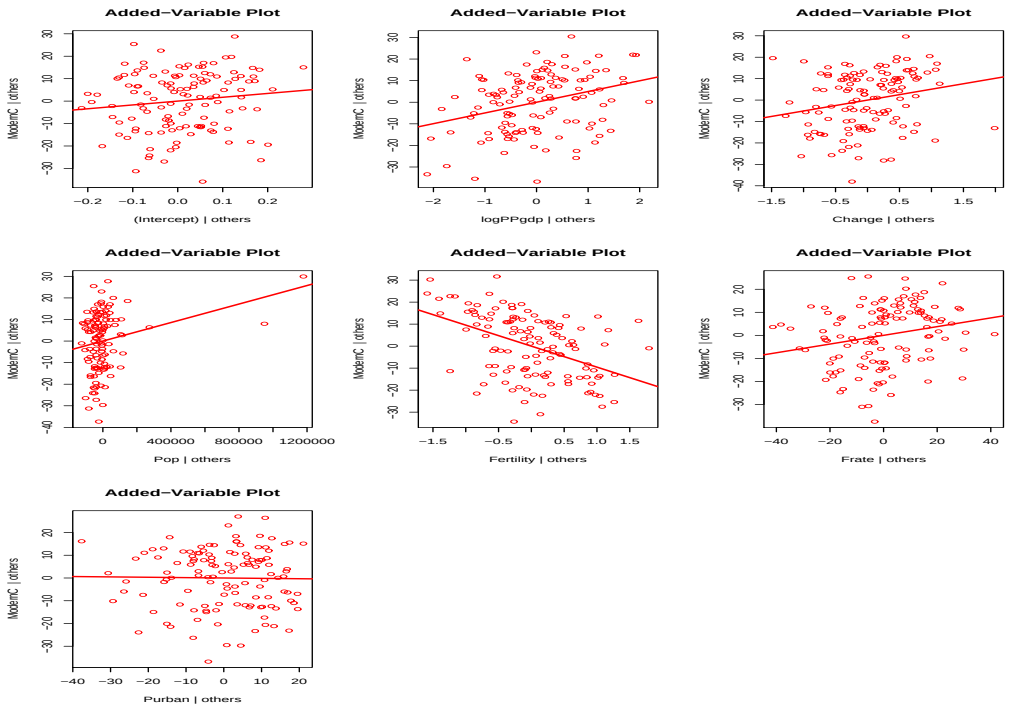

Figure 48: An added variable plot for the `UN3` data set from the `avp` command in the `car` package. Note that the slope of the line in the added variable plot for the variable *Pop* seems to be heavily determined by the two points to the far right of that plot.

9.9 (outliers in the united nations data set)

9.9.1: In Figure 48 we present the added variable plots (AVP) for the regression suggested. We see several things from these plots, from the added variable plot corresponding to the variable *Pop*, we see that the two points with the largest value of *Pop*

	Locality	ModernC	Change	PPgdp	Frate	Pop	Fertility	Purban
25	China	83	0.73	918	73	1304196	1.83	37
50	India	43	1.51	467	34	1065462	3.01	28

seem to have a large influence on their β coefficient since the slope of the linear fit presented there is strongly dependent on their two values. Without these two points the added variable plot would resemble a null plot indicating no dependence on the variable *Pop*. The second thing to notice is that the AVP for the variable *Purban* is a null plot indicating that this variable gives no information (when we have already included the others). This information is also present in the p-value for this coefficient when we fit a linear model to *ModernC* using all the coefficients (its p-value turns out to be 0.87 indicating no rejection of the null-hypothesis).

9.9.2: We can use the command `rstudent` to look for outliers and the Bonferroni inequity to get significance levels for each point to be an outlier. This test gives that *none* of the

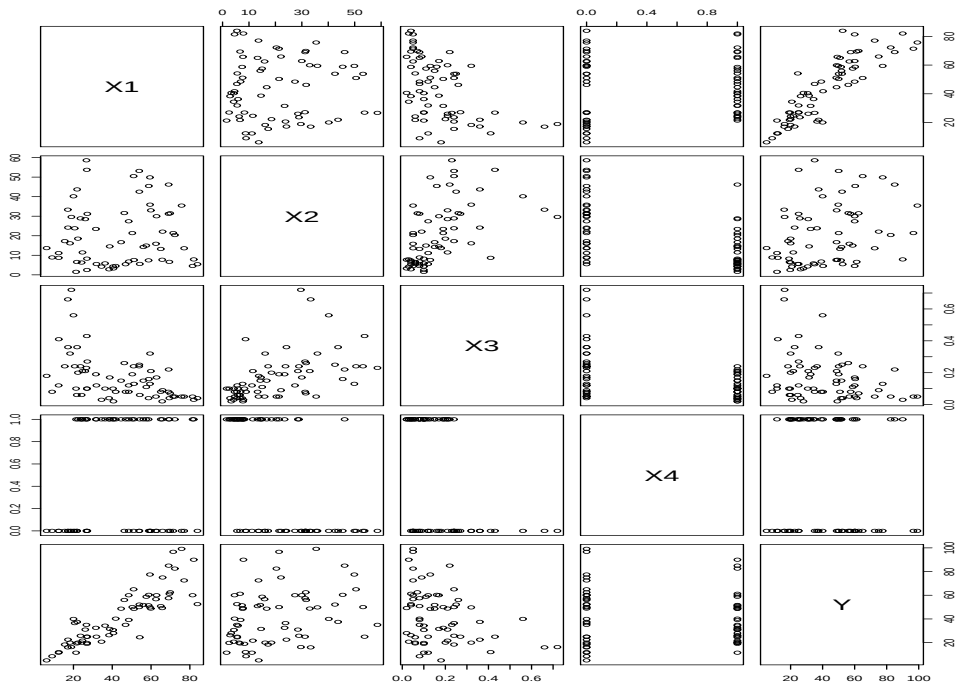

Figure 49: A scatter plot matrix for all variables in Problem 9.10.

points in this data set should be classified as an outlier.

See the `R` function `chap_9_prob_9.R` for code that implements this problem.

9.10 (land/rent agricultural data)

For this problem we are asked to determine a linear model for Y, the average rent per acre planted to alfalfa in counties in Minnesota in terms of the variables X_i for $i = 1, \ldots, 4$. If all independent variables were deemed equally important one might initially attempt a model of the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \,.$$

There are some problems with this initial model however. The first is that the variable X_4 represents whether or not the field was required to be "limed" (the compound lime applied to the field) in order to grow alfalfa. This variable is necessarily Boolean and should be properly represented as a factor. Using X_4 as a factor, the most general of linear model is

$$E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = j) = \beta_{0j} + \beta_{1j} x_1 + \beta_{2j} x_2 + \beta_{3j} x_3 \,, \tag{78}$$

for $j = 0, 1$ indicating the liming requirement. In this model depending on the value of X_4, the mean function for Y can be entirely different. One might expect that the presence or absence of a *requirement* for liming would not materially affect the average price of land Y and because of that the above model is too general and a simpler model ignoring the feature

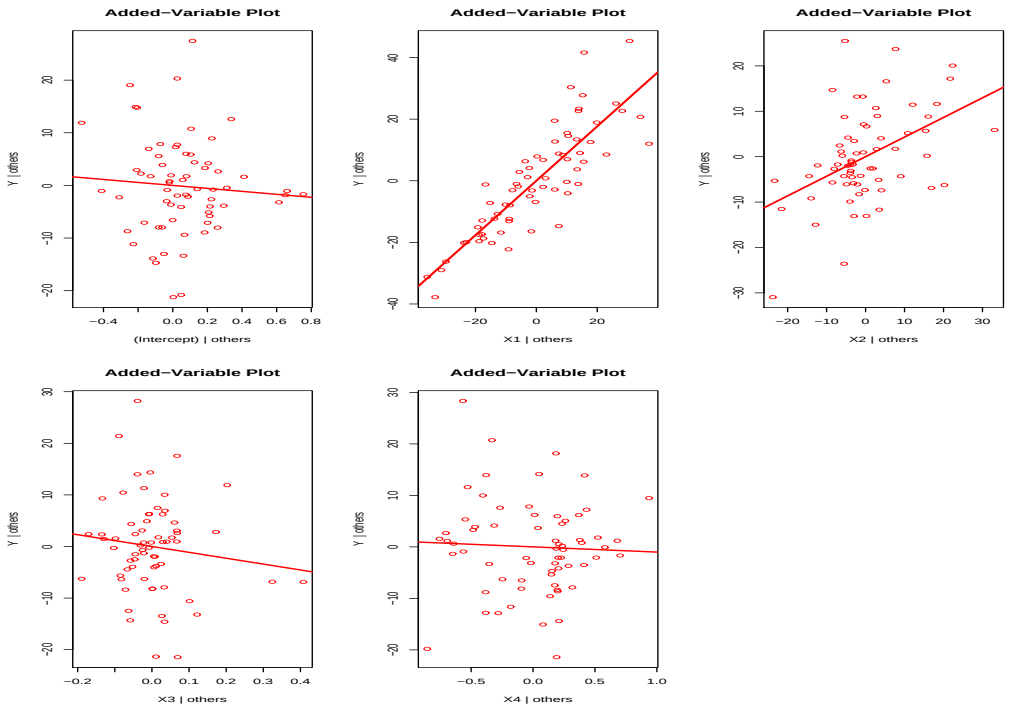

Figure 50: The added variable plot for the variables under the full model in Problem 9.10.

X_4 would be preferred. Dropping the term X_4 one such model would be

$$E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \,. \tag{79}$$

We can test that this model is not sufficiently different in predictive power than the more general model in Equation 78 using the `anova` function. As a second observation we would expect that the variables X_2 and X_3 would be highly correlated since X_2 is the density of dairy cows and X_3 is the proportion of farmland used for pasture (which would include land used for cows). Thus the added variable plot of either X_2 or X_3 in the full model should show an almost horizontal line indicating that given all but either of these two variables the other is determined. All possible added variable plots for the model 78 are shown in Figure 50. There we see that the line in the X_3 added variable plot is indeed nearly horizontal with the exception caused by the two potential outliers at the far right of the plot. Where these outliers found to be incorrect in some way the slope of this line would change dramatically. In addition, the added variable plot of X_4 is basically a null plot adding argument to our hypothesis that X_4 is not an important variable for predicting Y.

When we compute the models given by Equation 78 and 79 the `anova` command gives

```
> anova(m1,m0)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3 + x4Factor + X1:x4Factor + X2:x4Factor + X3:x4Factor
Model 2: Y ~ X1 + X2 + X3
```

```
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     59 4941.0
2     63 5385.7 -4    -444.8 1.3278 0.2702
```

This indicates that there is a 27% chance that the reduction in RSS due to the addition complexity in model 78 is due to chance. This is not small enough to warrant the complexity and we drop X_4 from.

The question as to whether the process of liming results in an increase in the value of Y might be answered by considering the model

$$E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = j) = \beta_{0j} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \,, \tag{80}$$

where now only the *intercept* β_{0j} depends on the factor X_4. When we compare this model to that of model 79 we get an `anova` table given by

```
> anova(m2,m0)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3 + x4Factor
Model 2: Y ~ X1 + X2 + X3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     62 5374.8
2     63 5385.7 -1     -10.9 0.1261 0.7237
```

The large value of `Pr(>F)` indicates that we should not consider this model further.

We are left considering model 79. The added variable plot for this model is shown in Figure 51 (left). We note that there appear to be two possible outliers with vary large values of X_3 that contribute to the non-zero estimate of the coefficient β_3. When we look at possible outliers we find that the "most likely" candidate for an outlier is *not* one of the two points suggested in the above added variable plots. The point with the largest value of t_i has a probability of begin an outlier given by the Bonferroni inequality of 0.08 which is not overwhelming evidence. We conclude that we don't need to remove these points and refit. We can still determine if the value of X_3 significantly helps predict the value of Y. Again using the `anova` command we find that including the term X_3 is not needed and we come to the model

$$E(Y|X_1 = x_1, X_2 = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \,. \tag{81}$$

Note this is like deleting the feature X_2 or X_3 that has the smaller t-test value when we fit the linear model in Equation 79. The added variable plot for the model 81 is shown in Figure 51 (right).

To determine if rent is higher in areas where there is a high density of dairy cows means we would like to determine if $\beta_2 > 0$. Since the `R summary` command gives the following for this two term (and an intercept model)

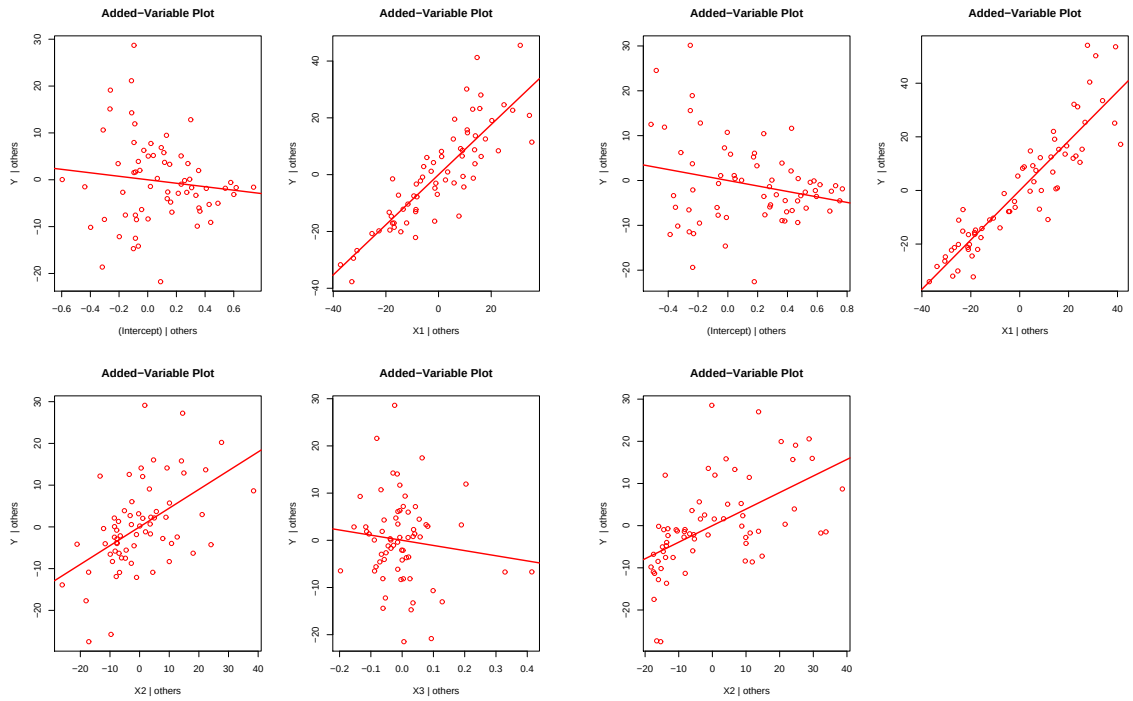

Figure 51: **Left:** The added variable plot (AVP) for the variables under the full model in Problem 9.10 but without the factor variable X_4. The almost horizontal line in the AVP for X_3 indicates that perhaps the value of X_3 is not needed given the others. **Right:** The added variable plot for the model given by Equation 81.

```
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
     Min       1Q   Median       3Q      Max
-21.4827  -5.8720   0.3321   4.3855  28.6007

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.11433    2.96123  -2.065    0.043 *
X1           0.92137    0.05382  17.121  < 2e-16 ***
X2           0.39255    0.07422   5.289 1.59e-06 ***
---

Residual standard error: 9.236 on 64 degrees of freedom
Multiple R-Squared: 0.8379,     Adjusted R-squared: 0.8328
F-statistic: 165.3 on 2 and 64 DF,  p-value: < 2.2e-16
```

we can be reasonably sure that the value of $\beta_2 \approx 0.39(0.07)$ is positive. This could be tested with a statistical test and a p-value computed.

See the `R` function `chap_9_prob_10.R` for code that implements this problem.

9.11 (cloud seeding)

In Figure 52 we present a scatter plot matrix for the `cloud` data set. At the outset we can ignore the variable D since the days index of the experiment should not be an input to this model. The correlation matrix for this data set shows that we expect that S (suitability of seeding), E (echo motion or type of cloud category), C (percent of cloud cover), and P (prewetness) are the variable that are most correlated with $Rain$ (in that order). Note that at the outset the variables A and $Rain$ do not seem very correlated.

We begin this problem by trying to predict whether the value of A, a factor variable indicating whether seeding was performed had any type of effect on the response variable $Rain$. We expect that whether or not seeding was performed the amount of rain would depend on the other explanatory variables: S, E, C and P. A very general model would be

$$E(Rain|S = s, E = i, C = c, P = p, A = j) = \beta_{0ij} + \beta_{1ij}s + \beta_{3ij}c + \beta_{4ij}p\,,$$

where in the above E *and* A are taken to be factors. The above model does not include any direct *interaction* terms between E and A. This seems to be reasonable as A (whether to seed or not) was chosen randomly by flipping a coin and the type of cloud E would be independent of the flip obtained. Rather than deal with the complexity of two factors that would be required to work with the above model lets instead consider a simpler model where E is not taken as a factor but is instead taken to be a continuous variable. We are then led

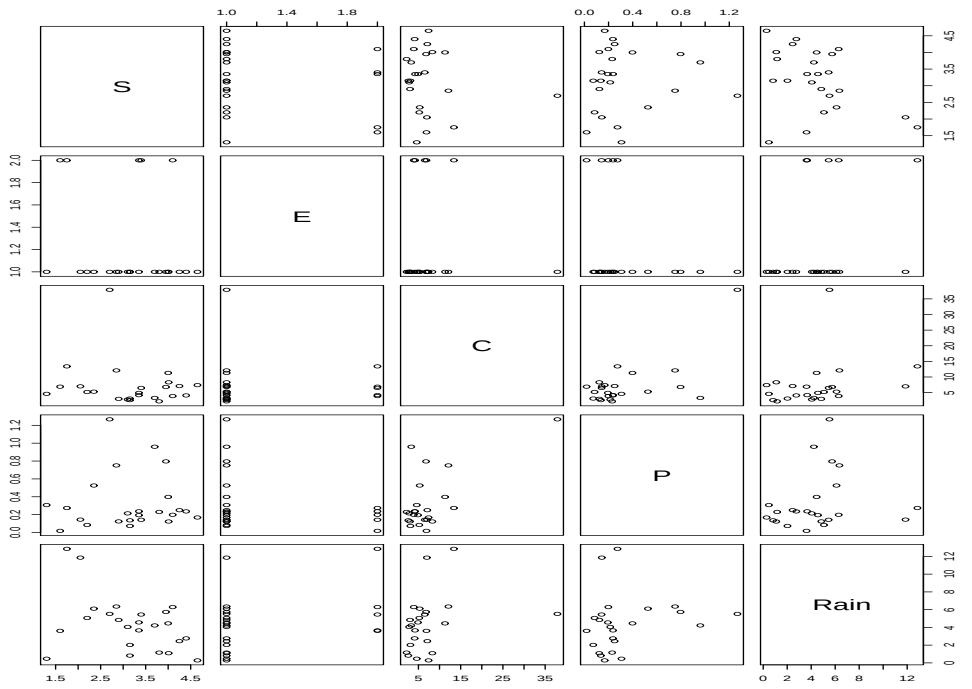

Figure 52: The scatter plot matrix for the variables in the `cloud` data set.

to consider the model

$$E(Rain|S=s,E=e,C=c,P=p,A=j)=\beta_{0j}+\beta_{1j}s+\beta_{2j}e+\beta_{3j}c+\beta_{4j}p\,.$$

Starting with this larger model we effectively perform "backwards selection" by hand removing terms that when removed don't result in a "statistically significant" increse in RSS. After this is done the model that seems approporiate is

$$E(Rain|S=s,C=c,A=j)=\beta_{0j}+\beta_{1j}s+\beta_{3j}c\,. \tag{82}$$

Thus we have dropped the features P (prewetness) and E echo motion from consideration. The added variable plot for the model above is presented in Figure 53. As an addition reference we note that this problem is also discussed in detail in the book [1].

See the `R` function `chap_9_prob_11.R` for code that implements this problem.

9.12 (health plans)

For this problem we want to predict $COST$ based on possible predictors GS (percentage of generic substituions), RI (restrictiveness index), F (percentage female members), AGE (average members age), $RXPM$ (average number of predictions per year), $COPAY$ (average member copay), and MM (member months). In Figure 54 we present a scatter plot matrix of the variables for this problem. As an additional piece of information from the correlation

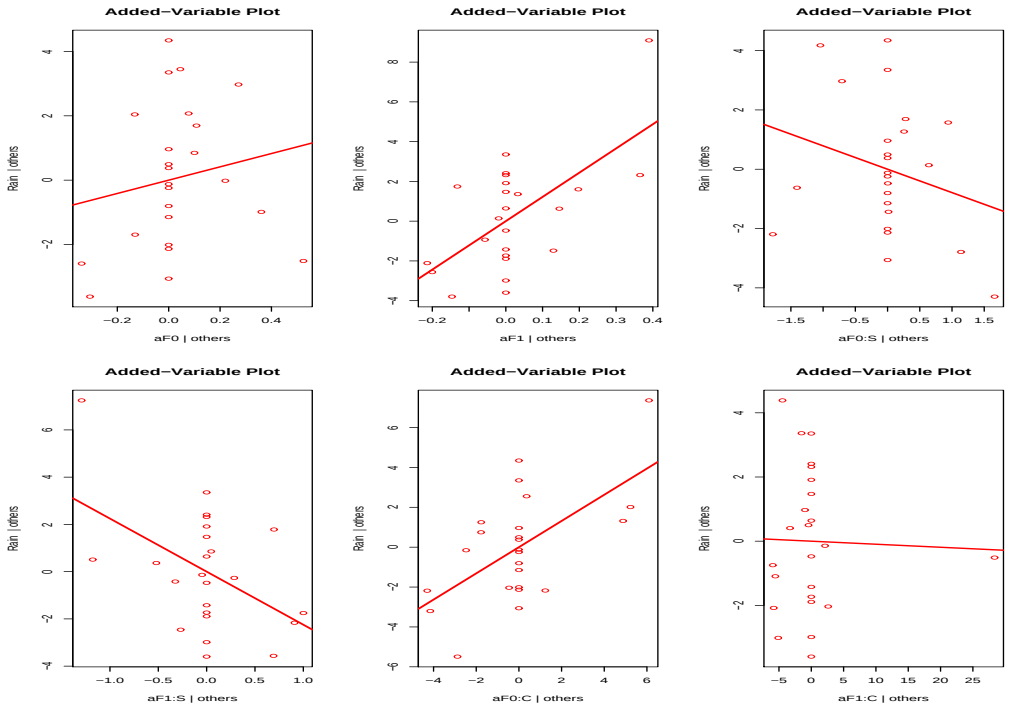

Figure 53: The added variable plot for the model given by Equation 82.

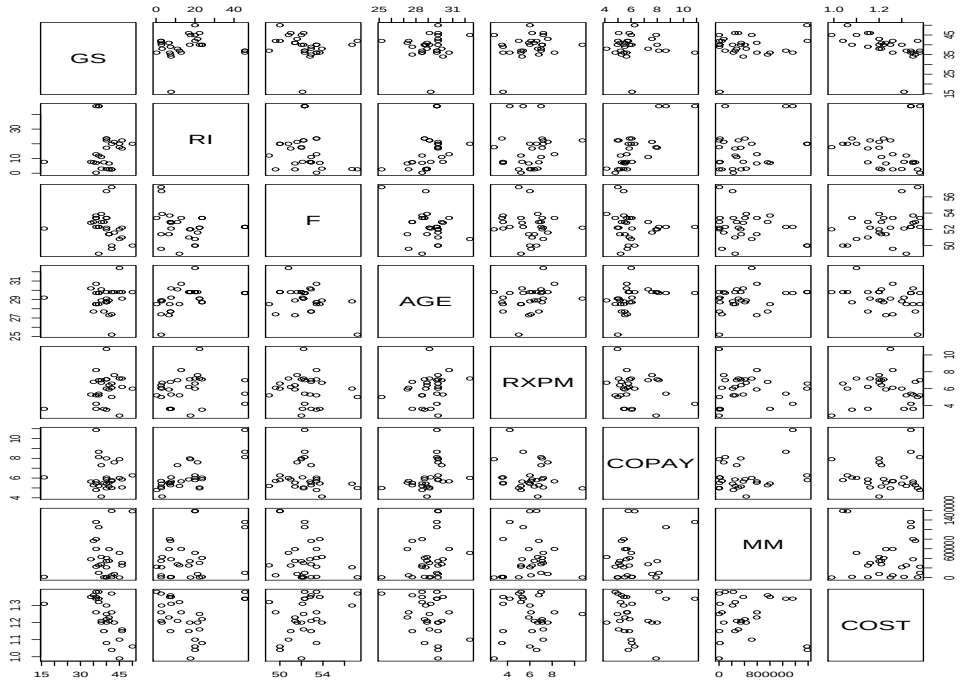

Figure 54: A scatter plot matrix for the data in the `drugcost` data set.

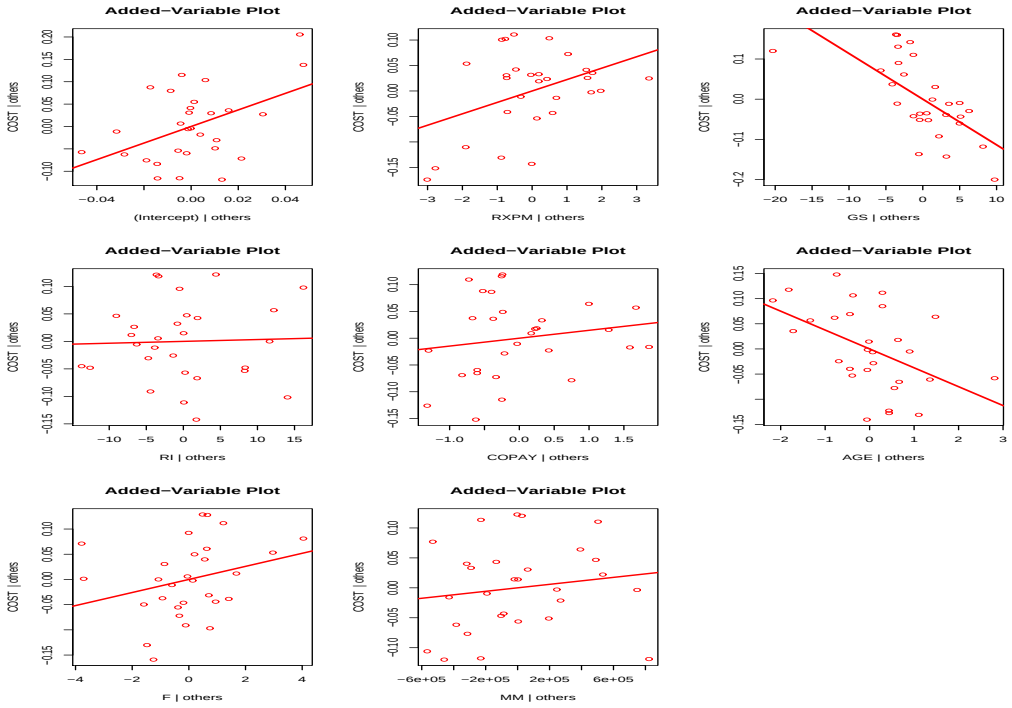

Figure 55: The added variable plot for the full model.

matrix we expect that the variables GS, AGE, and F are the most predictive of the value of $COST$.

We want to stastically test if the coefficients of GS and RI are negative indicating that using more GS and RI will reduce drug costs. The main question to ask is from the two factors GS and RI which is more important at lowering the value of $COST$. When we look at the various models we can drop several of the terms because their inclusion does not result in a significant reduction in RSS. After performing the "backwards selction" by hand we finally end with a model

$$E(COST|RXPM = r, GS = g, AGE = a) = \beta_0 + \beta_1 r + \beta_2 g + \beta_3 a\,. \tag{83}$$

The fact that the variable RI can be dropped without affecting the predictive power of the model too much indicated that this parameter in fact does not affect $COST$ very much. The added variable plot for the full model is given in Figure 55. In that plot we see that the variable RI has an almost horizontal line indicating that its coefficient is not very informative given the other variables. Finally, when we look at the `summary` command for the model Equation 83 we get that

```
Call:
lm(formula = COST ~ RXPM + GS + AGE)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-0.14956 -0.04003  0.00120  0.05936  0.12197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.52905    0.36930   6.848 3.54e-07 ***
RXPM         0.01813    0.01003   1.807 0.082749 .
GS          -0.01159    0.00277  -4.185 0.000307 ***
AGE         -0.03263    0.01262  -2.586 0.015940 *
---

Residual standard error: 0.0831 on 25 degrees of freedom
Multiple R-Squared: 0.4908,     Adjusted R-squared: 0.4297
F-statistic: 8.032 on 3 and 25 DF,  p-value: 0.0006474
```

This indicates that the coefficient of the term GS is $\beta_2 = -0.011(0.002)$ indicating that if one were to *increase* the value of GS this would result in a *decrease* in the value of $COST$. This model also paradoxily indicates that if we increase the value of AGE the plans cost also goes down.

See the `R` function `chap_9_prob_12.R` for code that implements this problem.

Chapter 10 (Variable Selection)

Notes On The Text

Notes on approximate collinearity when $p > 2$

After defining the term approximate collinearity of the predictors X_j in the case when $p > 2$ it can be helpful to have an computational definition to use in searching for it. To determine if a set of predictors is approximate collinear we can perform p regressions by regressing X_i on the set of all X's excluding X_i. That is we regress

$$X_i \sim X \setminus \{X_i\} ,$$

for $i = 1, 2, \ldots, p$. For each regression we compute the coefficient of determination, R^2, and since this depends on the ith predictor X_i that is the target of the regression these values can be indexed with i as R_i^2. Then consider the largest of all these R_i^2 values

$$R_{\max}^2 = \max_i (R_i^2) ,$$

if $R_{\max}^2$ is close to 1 then we declare that the variable X_i are approximately collinear.

Notes on Computationally Intensive Criteria

In this section the book mentions two cross-validation the procedures. The first is where the total data set is split into two parts: a construction set and a validation set. The second was denoted as computing *predicted residuals*. This second procedure might be better understood in that it appears to be a form of *leave-one-out* cross validation, where we hold out a single sample $x_{\mathcal{C}i}$, fit the regression coefficients using all of the other data points to get $\hat{\beta}_{\mathcal{C}(i)}$ and then determine how well this linear model predicts the response of $x_{\mathcal{C}i}$ via computing the square error $(y_i - x'_{\mathcal{C}i}\hat{\beta}_{\mathcal{C}(i)})^2$. How well this candidate set $\mathcal{C}$ does at predicting y can be estimated adding up the above square error when we hold out each sample one at a time or

$$\sum_{i=1}^{n} (y_i - x'_{\mathcal{C}i}\hat{\beta}_{\mathcal{C}(i)})^2 .$$

The above expression is defined as the predicted residuals or $PRESS$. Up to this point this procedure could be applied to any model selection procedure, but we may have a significant amount of computation to do to obtain the leave-one-out models. What makes this cross-validation procedure different is that it can be shown that when we are considering *linear regression* the above expression for $PRESS$ can be determined from expressions that we can compute when we do a global fit to calculate $\hat{\beta}_{\mathcal{C}}$ keeping all samples (i.e. not holding out any samples). The expression for $PRESS$ above then becomes

$$PRESS = \sum_{i=1}^{n} \left(\frac{\hat{e}_{\mathcal{C}i}}{1 - h_{\mathcal{C}ii}} \right)^2 ,$$

where $\hat{e}_{\mathcal{C}i}$ and $h_{\mathcal{C}ii}$ are the residual and the leverage for the ith case. Thus by doing a fit over all samples we can evaluate the $PRESS$ and not have to refit our model n times as would need to be done with a very simple implementation of hold-one-out cross validation. Some modeling techniques don't have this refitting problem. For example selecting the k to use in k nearest neighbor regression does not suffer from this since it is relatively easy to evaluate each model (change k) when one point is "held out".

Notes on Computational Methods

We note that stepwise methods like forward selection and backwards selection, using the information criteria suggested in the book, are not really equivalent when one of the variable is a factor (or equivalently a group of variables). The reason for this is that when a factor enters or exits a fit we must *also* change the value of $p_{\mathcal{C}}$ in addition to the value of $RSS_{\mathcal{C}}$. The different coefficients of the term $p_{\mathcal{C}}$ in the different information criterion i.e. AIC, BIC or Mallow's C_p can result in different predictions for the subset to use in minimizing the information criterion. Thus when factors are *not* considered the three information criterion AIC, BIC, and Mallow's C_p will all select the equivalent subsets. When one of the variables is a factor this may no longer be true.

Problem Solutions

10.1 (correlated features make feature selection more difficult)

In this problem we duplicate the example in the book that demonstrates that in linear regressions on correlated variables is can be difficult to determine which variables are active and which are inactive when the sample size is small and the variables are correlated. To implement this problem we change the random seed and rerun the `R` code that was used to generate the example from the book. We see the same behavior discussed in the book. The most interesting result being that when we consider strongly correlated variables the *variance* of the estimate of $\hat{\beta}_i$ are so much larger than in the uncorrelated case. This results in smaller t-values that would be expected and more conclusions on the significance of coefficients $\hat{\beta}_i$. As an example, when on runs the `R` script below one see that if we have a small sample size and strongly correlated variables (represented by the model `m2`) the predictions of the linear models `y ~ x1 + x4` and `y ~ x3 + x4` both have statistically significant estimates for their β coefficients even though the data as generated did not explicitly involve these variables. This is not much of an practical issue since in either case the variables `x3` and `x4` could be used to predict `y`.

See the `R` file `chap_10_prob_1.R` for an implementation of the problem.

10.2 (backwards elimination (BE) and forward selection (FS))

Warning: The code for this problem seemed to produce different results depending on the version of `R` on which it is run. The results shown here are for `R version 2.6.0 (2007-10-03)`, where backwards selection results in keeping all three terms. If the code is run on the `R` version `R version 2.10.0 (2009-10-26)` the AIC for the full three term model was estimated at -285.77 and backwards selection removed the variable `X3` before finishing. The general conclusion that backwards elimination and forward selection can yield different subsets is still a valid conclusion however.

For this problem we first perform backwards elimination (BE) on the `mantel` data set and then second perform forward selection (FE) on the same data set. When we perform backward elimination starting with the regression of Y on all of the other variables the `R` function `step` quickly tells us that the optimal subset (under any of the subset selection criterion AIC, BIC, or C_p) is the one containing *all* of the terms. This set has an AIC of -315.23 and the estimated coefficients are given by

```
Call:
lm(formula = Y ~ X1 + X2 + X3, data = mantel)

Coefficients:
(Intercept)           X1           X2           X3
 -1.000e+03    1.000e+00    1.000e+00    4.404e-15
```

From which we notice that the coefficient of `X3` is effectively zero when compared to the magnitude of the others.

When one runs forward selection on the the other hand all three criterion function select a regression with only *one* (the X_3) term. This model has an AIC given by -0.31 and has coefficients given by

```
Call:
lm(formula = Y ~ X3, data = mantel)

Coefficients:
(Intercept)           X3
     0.7975       0.6947
```

This is another example where determining the true active set is difficult and the two model selection techniques give different answers. We can tell that this is the case where the sample size is very small (only five measurement) and where the input variables are strongly correlated. Displaying the correlation matrix we see that it is given by

```
> cor(as.matrix(mantel)[,c(2,3,4)])
```

```
           X1         X2         X3
X1  1.0000000 -0.9999887  0.6858141
X2 -0.9999887  1.0000000 -0.6826107
X3  0.6858141 -0.6826107  1.0000000
```

from which we see that indeed the input variables are very strongly correlated and we expect that determining the active variables will be difficult. The fact that several of the variables like `X1` and `X2` are so correlated means the variance of their estimates will be particularly poor. See Problem 10.6 on page 137 and Equations 89 and 92 that demonstrate how correlation among factors affects the variance of the estimate for β.

See the `R` file `chap_10_prob_2.R` for an implementation of the problem.

10.3 (backwards elimination (BE) on the highway data)

For this problem we perform backwards elimination (BE) on the highway data set. To do this we use the `R` command `step`. The results of this study can be found by running the `R` script file `chap_10_prob_3.R`. When we run that command we start with a "full" model and sequentially remove measurements taking the measurement to remove that results in the smallest AIC. This procedure continues to remove features until the smallest model is obtained (in this case that is a model with only the feature $logLen$). The routine stops when there is no reduction in AIC by removing a feature. The final model produced by backward selection in this case is given by

```
logRate ~ logLen + logADT + logSigs1 + Slim + Hwy
```

and has an AIC value of -74.71. When we look at the books result from forward selection gives the model

```
logRate ~ logLen + Slim + logTrks + Hwy + logSigs1
```

with an AIC of -73.03. These two models are the same in the number of terms 5 but differ in that the first has the variable $logADT$ while the second has the variable $logTrks$.

10.4 (optimal subset selection for $HT18$ as a function of younger measurements)

For this problem we will apply subset selection to the $HT18$ data from the Berkeley Guidance Study using variables produced during for younger ages. To save time we don't perform transformations of the variables but consider them in their raw form. We consider *all* possible variables that we could use to predict the value of $HT18$ and then fit a linear model on all

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	44.3997	16.2947	2.72	0.0085
WT2	0.5303	0.3228	1.64	0.1057
HT2	-0.3030	0.1764	-1.72	0.0910
WT9	-0.0533	0.1996	-0.27	0.7903
HT9	1.2510	0.1145	10.92	0.0000
LG9	-0.6149	0.4746	-1.30	0.2002
ST9	0.0396	0.0338	1.17	0.2449

Table 13: The fitted coefficients $\hat{\beta}$ from the largest model for predicting the variable $HT18$ in Problem 10.4.

of this data. When we do that and using the `xtable` command we obtain Table 13. In that table we see that the only variable that is know with a very strong certainty is $HT9$. We expect that this will be used the optimal subset from the set of variables

```
WT2, HT2, WT9, HT9, LG9, ST9
```

Next we use forward selection (starting at the constant model) to derive the optimal model subset using the `R` command `step`. When we do that we see that the first variable added is $HT9$ which was to be predicted from the t-value found for this variable from the full model. Only one more variable is added $LG9$ to form the optimal (under forward selection) regression `HT18~HT9+LG9`, a model which has an AIC given by 149.76.

When we use backwards selection we find the model `HT18~WT2+HT2+HT9+LG9` with a AIC of 149.87. Note that the two variables $HT9$ and $LG9$ found under forward selection are also found to be informative under backwards selection.

See the `R` file `chap_10_prob_4.R` for an implementation of the problem.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.9683	2.1096	-2.36	0.0336
BOD	-0.0000	0.0012	-0.04	0.9719
TKN	0.0030	0.0029	1.03	0.3188
TS	0.0003	0.0002	1.67	0.1179
TVS	0.0183	0.0323	0.56	0.5814
COD	0.0003	0.0002	1.92	0.0751

Table 14: The fitted coefficients $\hat{\beta}$ from the largest model for predicting the variable $logO2UP$ in Problem 10.5.

10.5 (subset selection of the predictors of $O2UP$)

For this problem we will try to predict the response $O2UP$ given the variables suggested in this problem: BOD, TKN, TS, TVS, COD. We begin by plotting a scatter plot matrix of all the variables without any transformations. We see that the range of the $O2UP$ variable covers three orders of magnitude from 0.3 to 36.0 and does not appear to be linearly correlated with any of the variables very strongly. Because the range of $O2UP$ spans several orders of magnitude this might suggest a logarithmic transformation. When we perform that transformation we see that several of the variables (like BOD, TS, TVS, and COD) appear to be linearly related.

If we consider the linear model with a response $logO2UP$ that uses *all* possible predictors we find a Table 14. This table indicates that *none* of the beta coefficients (given all of the others) are known with great certainty. This and the fact that the sample size is so small 20 points we expect to get considerable benefit from performing subset selection.

We begin with forward selection on this problem. The `R` function `step` predicts a model with two predictors given by `logO2UP~TS+COD` and having an AIC given by -18.92. Backwards selection in this case gives exactly the same model and AIC value. The estimated parameters from this model are given in Table 15. From their t-values and associated probabilities we see that these variable are much more accurately known.

See the `R` file `chap_10_prob_4.R` for an implementation of the problem.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-3.1547	0.4533	-6.96	0.0000
TS	0.0003	0.0001	2.72	0.0146
COD	0.0003	0.0001	2.66	0.0165

Table 15: The fitted coefficients $\hat{\beta}$ from the model fit on TS and COD that predicts the variable $logO2UP$ (see Problem 10.5).

10.6 (deriving the variance inflation factor)

For this problem we will first derive the requested result in the situation when we are considering adding a single additional predictor (say X_2) to the simple linear regression $E(Y|X_1) = \beta_0 + \beta_1 X_1$, since this derivation is easier to follow and shorter. In a second part of this problem we present the derivation in the general case where we regress Y on k predictors $X_1, X_2, \cdots, X_k$ and then add another predictor X_{k+1}. This second derivation is very similar to the first derivation which is simpler to understand and follow. This second derivation can be skipped on first reading since it is more involved and is somewhat more complicated

The Simple Case: Recall that in an added variable plot for simple linear regression we

begin with a linear regression of Y on the single predictor X_1

$$E(Y|X_1 = x_1) = \hat{\beta}_0 + \hat{\beta}_1 x_1 \,, \tag{84}$$

and the corresponding residuals for this model. We then proceed to add the 2nd predictor X_2. The added variable plot is obtained by first performing a linear regression of X_2 onto the previous X_1 predictor obtaining a linear model

$$E(X_2|X_1 = x_1) = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 \,, \tag{85}$$

and the corresponding residuals for this model. We then plot the residuals of Equation 84 as a function of the residuals for the model Equation 85. This gives an indication of the information that the variable X_2 contains (and that is not already contained in the variable X_1) that can be used to predict the residuals of the model Equation 84 (the missing information needed in explaining Y and not contained in the variable X_1). In addition, the estimated *slope* in the added variable plot is the slope that will enter Equation 84 as $\hat{\beta}_2$ when we add the variable X_2.

With this background we will use our one-dimensional regression formulas from the appendix corresponding to simple linear regression applied to the variables from in the added variable plots. To do this we will let the variable V denote the *residuals* of the regression of X_2 onto X_1, and let the variable U denote the residuals of the regression of Y onto X_1. For notational simplicity in what follows we will also denote the variable X_2 by W. In this notation, the added variable plot is a plot of V versus U and the estimated slope coefficient $\hat{\beta}_1$ in a simple linear regression formulation is equivalent to the coefficient $\hat{\beta}_2$ that would enter the model in Equation 84 if we added add the variable X_2. From the appendix the estimate of β_1 is given by

$$\hat{\beta}_1 = \frac{SUV}{SUU} \,,$$

which has a variance given by

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SUU} \,.$$

To evaluate this later variance we need to evaluate SUU. From the definition of SUU we have that

$$SUU = \sum_{j=1}^{n} (u_j - \overline{u})^2 = \sum_{j=1}^{n} {u_j}^2 \,,$$

in which we have used the fact that $\overline{u} = 0$ since the regression of X_2 (or W) onto X_1 includes a constant term (see Equation 85). Recalling that u are the residuals of the model given by Equation 85 we have that

$$u_j = w_j - \tilde{\beta}_0 - \tilde{\beta}_1 x_{j1} \,.$$

So the expression for SUU becomes

$$SUU = \sum_{j=1}^{n} \left(w_j - \tilde{\beta}_0 - \tilde{\beta}_1 x_{j1} \right)^2 \,.$$

Recalling Equation 109 of $\tilde{\beta}_0 = \overline{w} - \tilde{\beta}_1 \bar{x}_1$ by replacing the value of $\tilde{\beta}_0$ in the above we can write SUU as

$$SUU = \sum_{j=1}^{n} \left(w_j - \overline{w} + \tilde{\beta}_1 \bar{x}_1 - \tilde{\beta}_1 x_{j1} \right)^2$$

$$
\begin{aligned}
&= \sum_{j=1}^{n} \Big(w_j - \overline{w} - \tilde{\beta}_1 (x_{j1} - \bar{x}_1) \Big)^2 \\
&= \sum_{j=1}^{n} \Big[(w_j - \overline{w})^2 - 2\tilde{\beta}_1 (w_j - \overline{w})(x_{j1} - \bar{x}_1) + \tilde{\beta}_1^2 (x_{j1} - \bar{x}_1)^2 \Big] \\
&= SWW - 2\tilde{\beta}_1 \sum_{j=1}^{n} (w_j - \overline{w})(x_{j1} - \bar{x}_1) + \tilde{\beta}_1^2 \sum_{j=1}^{n} (x_{j1} - \bar{x}_1)^2 \\
&= SWW - 2\tilde{\beta}_1 SX_1W + \tilde{\beta}_1^2 SX_1X_1 \,. \qquad (86)
\end{aligned}
$$

Since $\tilde{\beta}_1 = \frac{SX_1W}{SX_1X_1}$ the above simplifies and we get

$$
\begin{aligned}
SUU &= SWW - 2SX_1W \left(\frac{SX_1W}{SX_1X_1} \right) + \left(\frac{SX_1W}{SX_1X_1} \right)^2 SX_1X_1 \\
&= SWW - \frac{(SX_1W)^2}{SX_1X_1} = SWW \left(1 - \frac{(SX_1W)^2}{(SWW)(SX_1X_1)} \right) \\
&= SWW(1 - r_{12}^2) \,,
\end{aligned}
$$

which is the books equation 10.4 and ends the derivation of the simple case.

The General Case: Recall that in an added variable plot we begin with a linear regression of Y on k predictors $X_1, X_2, \cdots, X_k$ as

$$
E(Y|X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \,, \qquad (87)
$$

and the corresponding residuals for this model. We then proceed to add the $k+1$st predictor X_{k+1}. The added variable plot is obtained by first performing a linear regression of X_{k+1} onto the previous k predictors obtaining a linear model

$$
E(X_{k+1}|X_1 = x_1, X_2 = x_2, \cdots, X_k = x_k) = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \cdots + \tilde{\beta}_k x_k \,, \qquad (88)
$$

and the corresponding residuals for this model. We then plot the residuals of Equation 87 as a function of the residuals for the model Equation 88. This gives an indication of the information that the variable X_{k+1} contains (and that is not already contained in the variables $X_1, X_2, \cdots, X_k$) that can be used to predict the residuals of the model Equation 87 (the missing information needed in explaining Y and not contained in the variables $X_1, X_2, \cdots, X_k$). In addition, the estimated *slope* in the added variable plot is the slope that will enter Equation 87 as $\hat{\beta}_{k+1}$ when we add the variable X_{k+1}.

With this background we will use our one-dimensional regression formulas from the appendix corresponding to simple linear regression applied to the variables from in the added variable plots. To do this we will let the variable V denote the *residuals* of the regression of X_{k+1} onto $X_1, X_2, \cdots, X_k$, and let the variable U denote the residuals of the regression of Y onto $X_1, X_2, \cdots, X_k$. For notational simplicity in what follows we will also denote the variable X_{k+1} by W. In this notation, the added variable plot is a plot of V versus U and the estimated slope coefficient $\hat{\beta}_1$ in a simple linear regression formulation is equivalent to the coefficient $\hat{\beta}_{k+1}$ that would enter the model in Equation 87 if we added add the variable X_{k+1}. From the appendix the estimate of β_1 is given by

$$
\hat{\beta}_1 = \frac{SUV}{SUU} \,,
$$

which has a variance given by

$$\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SUU}\,. \tag{89}$$

To evaluate this later variance we need to evaluate SUU. From the definition of SUU we have that

$$SUU = \sum_{j=1}^{n}(u_j - \overline{u})^2 = \sum_{j=1}^{n} {u_j}^2\,,$$

in which we have used the fact that $\overline{u} = 0$ since the regression of X_{k+1} (or W) onto $X_1, X_2, \cdots, X_k$ includes a constant term (see Equation 88). Recalling that u are the residuals of the model given by Equation 88 we have that

$$u_j = w_j - \tilde{\beta}_0 - \sum_{l=1}^{k} \tilde{\beta}_l x_{jl}\,.$$

So the expression for SUU becomes

$$SUU = \sum_{j=1}^{n}\left(w_j - \tilde{\beta}_0 - \sum_{l=1}^{k}\tilde{\beta}_l x_{jl}\right)^2\,.$$

Recalling Equation 18 or

$$\tilde{\beta}_0 = \overline{w} - \tilde{\beta}'\bar{\mathbf{x}} = \overline{w} - \sum_{l=1}^{k}\tilde{\beta}_l \bar{x}_l\,,$$

where $\bar{\mathbf{x}}$ is a vector containing the means of the predictors X_l. Using this expression for $\tilde{\beta}_0$ we can write SUU as

$$\begin{aligned}
SUU &= \sum_{j=1}^{n}\left(w_j - \overline{w} + \sum_{l=1}^{k}\tilde{\beta}_l\bar{x}_l - \sum_{l=1}^{k}\tilde{\beta}_l x_{jl}\right)^2 \\
&= \sum_{j=1}^{n}\left(w_j - \overline{w} - \sum_{l=1}^{k}\tilde{\beta}_l(x_{jl} - \bar{x}_l)\right)^2 \\
&= \sum_{j=1}^{n}\left[(w_j - \overline{w})^2 - 2(w_j - \overline{w})\left(\sum_{l=1}^{k}\tilde{\beta}_l(x_{jl} - \bar{x}_l)\right) + \left(\sum_{l=1}^{k}\tilde{\beta}_l(x_{jl} - \bar{x}_l)\right)^2\right] \\
&= SWW - 2\sum_{j=1}^{n}(w_j - \overline{w})\left(\sum_{l=1}^{k}\tilde{\beta}_l(x_{jl} - \bar{x}_l)\right) + \sum_{j=1}^{n}\left(\sum_{l=1}^{k}\tilde{\beta}_l(x_{jl} - \bar{x}_l)\right)^2\,.
\end{aligned} \tag{90}$$

We will now simplify the last sum in the above. Expanding the square of the sum over $\tilde{\beta}_l$ we can write it as

$$\sum_{j=1}^{n}\sum_{l=1}^{k}\sum_{m=1}^{k}\tilde{\beta}_l\tilde{\beta}_m(x_{jl} - \bar{x}_l)(x_{jm} - \bar{x}_m) = \sum_{l=1}^{k}\tilde{\beta}_l\left(\sum_{m=1}^{k}\sum_{j=1}^{n}\tilde{\beta}_m(x_{jl} - \bar{x}_l)(x_{jm} - \bar{x}_m)\right)\,. \tag{91}$$

To this last expression we will apply some of the results derived earlier. Recalling Equation 26 we see that the expression in parenthesis above is equal to $((\mathcal{X}'\mathcal{X})\tilde{\beta})_l$, which is equal to the l-th component of $\mathcal{X}'\mathcal{Y}$ and can be written as the sum in Equation 28. In this case this inner sum is specifically given by

$$\sum_{j=1}^{n}(x_{jl} - \bar{x}_l)(w_j - \bar{w})\,.$$

Using this result, the above triple sum in Equation 91 becomes a double sum given by

$$\sum_{l=1}^{k} \tilde{\beta}_l \sum_{j=1}^{n} (x_{jl} - \bar{x}_l)(w_j - \bar{w}) \,,$$

which can then be combined with the second term in Equation 90. Combining these two terms we get for SUU the following

$$\begin{aligned} SUU &= SWW - \sum_{l=1}^{k} \left(\tilde{\beta}_l \sum_{j=1}^{n} (x_{jl} - \bar{x}_l)(w_j - \bar{w}) \right) \\ &= SWW \left(1 - \frac{\sum_{j=1}^{n} \sum_{l=1}^{k} \tilde{\beta}_l (x_{jl} - \bar{x}_l)(w_j - \bar{w})}{SWW} \right) . \end{aligned}$$

A few more transformations and we will have the result we seek. Considering the inner summation in the fraction above we can write this as two parts and using ideas like on Page 5 of these notes we see that

$$\sum_{l=1}^{k} \tilde{\beta}_l (x_{jl} - \bar{x}_l) = (\hat{w}_j - \tilde{\beta}_0) - \sum_{l=1}^{k} \tilde{\beta}_l \bar{x}_l = (\hat{w}_j - \tilde{\beta}_0) - (\bar{w} - \tilde{\beta}_0) = \hat{w}_j - \bar{w} \,.$$

Using this we then have that the expression for SUU becomes

$$\begin{aligned} SUU &= SWW \left(1 - \frac{\sum_{j=1}^{n} (\hat{w}_j - \bar{w})(w_j - \bar{w})}{SWW} \right) \\ &= SWW \left(1 - R_w^2 \right) , \end{aligned} \tag{92}$$

where R_w is the multiple correlation coefficient or the correlation of the fitted values $\hat{w}_j$ and the true values w_j. This is the desired expression.

10.7 (feature selection on the Galapagos Island)

For this problem we want to predict some measure of the given islands diversity. To solve the problem I choose to consider the ratio of the number of endemic species to the total number of species or ES/NS. Then to use such a model for a given novel island we would simple count up the number of species found on that island and then multiply by the result of the regression model to get the number of endemic species on that island.

We begin our analysis by a scatter plot matrix of ES/NS considered with all of the variables: *Area*, *Anear*, *Dist*, *DistSC*, and *Elevation*. Several of these variable we might expect to be irrelevant (like *DistSC*) but lets see if features selection demonstrates this fact. This scatter plot matrix is plotted in Figure 56 (left). From that plot we see that several of the variables (like *Area*, *Anear*, and *Elevation*) have relatively large dynamic ranges. These can perhaps be better modeled by taking logarithms of their values in the given data set. In addition, by experimentation it appears that if we take the logarithm of the fraction ES/NS we get more "linear" looking data in the subsequent scatter plots. These variables are plotted in a

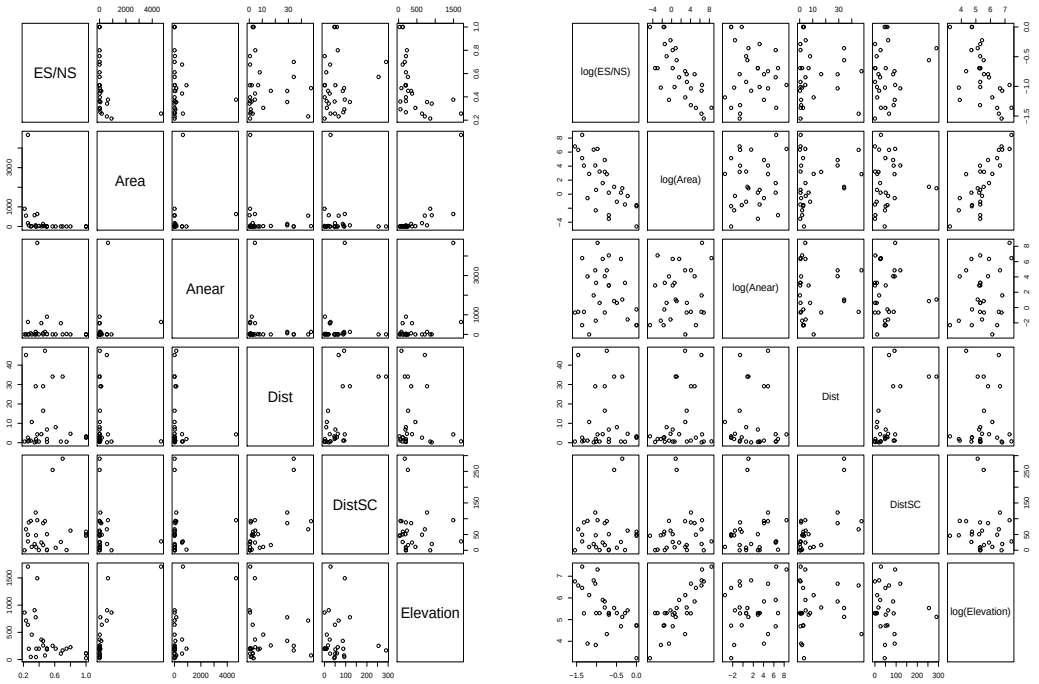

Figure 56: **Left:** A scatter plot matrix of the variables ES/NS, $Area$, $Anear$, $Dist$, $DistSC$, and $Elevation$. **Right:**A scatter plot matrix of the variables $log(ES/NS)$, $log(Area)$, $log(Anear)$, $Dist$, $DistSC$, and $Elevation$.

scatter plot matrix in Figure 56 (right) and these will be used to derive a linear model to help predict diversity.

Next we will use forward and backwards selection to determine which variable are the most helpful in predicting this function of NE/NS. When we run forward selection we find the first variable added is $logArea$ and the second variable select is $DistSC$. Giving a final model of

$$E(log(NE/NS)|logArea = a, DistSC = b) = \beta_0 + \beta_1 a + \beta_2 b\,,$$

which has an AIC value of -65.44. When we do backwards elimination we find the same final model. It is a bit surprising that the value of $DistSC$ was found to be so predictive. This maybe an example of a *lurking* variable that is unknown but that is strongly correlated with the variable $DistSC$.

See the `R` file `chap_10_prob_7.R` for an implementation of the problem.

10.8 (conditions under which $E(Y|X_{\mathcal{C}})$ will be linear)

We are told to assume that $X_{\mathcal{C}}$ will *not* include all of the active terms in $X_{\mathcal{A}}$. Lets break down $X_{\mathcal{A}}$ into two sets of terms $X_1 \equiv X_{\mathrm{C}}$ the terms we *do* include in our set of predictors and $X_2 \equiv X_{\neg\mathcal{C}}$ the terms we do not include in our set of predictors but that are in X_{A}. Then

we are told

$$E(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1' x_1 + \beta_2' x_2 \,, \tag{93}$$

We want to determine conditions such that $E(Y|X_1 = x_1)$ is also a pure *linear* function of x_1. To do this first recall the conditional expectation formula from the appendix which states

$$E(Y) = E[E[Y|X = x]] \,. \tag{94}$$

In the above context, we want to use this expression as

$$E[Y|X_1 = x_1] = E_{X_2}\left[E[Y|X_1 = x_1, X_2 = x_2]|X_1 = x_1\right] \,, \tag{95}$$

where we have conditioned on $X_1 = x_1$ in *all* expectations and the outer expectation is taken with respect to the variables in X_2. Now our hypothesis given in Equation 93 means that the expectation we should be considering in Equation 95 is given by

$$\begin{aligned} E[Y|X_1 = x_1] &= E_{X_2}\left[\beta_0 + \beta_1' x_1 + \beta_2' x_2 | X_1 = x_1\right] \\ &= \beta_0 + \beta_1' x_1 + \beta_2' E[X_2|X_1 = x_1] \,. \end{aligned}$$

Thus if the expectation $E[X_2|X_1 = x_1]$ is linear i.e. can be expressed as

$$E[X_2|X_1 = x_1] = \tilde{\beta}_0 + \tilde{\beta}_1' x_1 \,,$$

then our regression $E[Y|X_1 = x_1]$ will also be linear.

Chapter 11 (Nonlinear Regression)

Notes On The Text

Notes on estimation for nonlinear mean functions

The book describes the *score vector* $u_i(\theta^*)$, and gives a formula for it but it can sometimes be helpful when performing the Gauss-Newton iterates to visualize this as a *new* measurement vector similar to how x_i is defined. Its definition has it going into the design matrix in exactly the same way that the vector x_i does. The ith score vector $u_i(\theta^*)$ in vector form is defined as

$$u_i(\theta^*) = \begin{bmatrix} \frac{\partial m}{\partial \theta_1}(x_i, \theta) \\ \frac{\partial m}{\partial \theta_2}(x_i, \theta) \\ \vdots \\ \frac{\partial m}{\partial \theta_k}(x_i, \theta) \end{bmatrix} . \tag{96}$$

On each iteration of the Gauss-Newton iterates these vectors is stacked horizontally into a matrix $U(\theta^*)$ and ordinary least squares performed using this design matrix. If our mean function $m(x, \theta)$ is actually linear (as opposed to nonlinear) then $m(x, \theta) = x'\theta$ and

$$\frac{\partial m(x_i, \theta)}{\partial \theta_k} = x_{ik} \, ,$$

of the kth component of the ith vector. This gives a score vector $u_i(\theta^*) = x_i$ and the design matrices $U(\theta^*)$ is then the same as earlier matrix X. Note that this is an *iterative* way to obtain a sequence of coefficients $\hat{\beta}_i$ that should converge to the ordinary least squares solution $\hat{\beta} = (X^T X)^{-1} X^T Y$.

If we perform a linearization about the true solution θ^* then we have that in the large sample case that our approximate estimate $\hat{\theta}$ can be expressed in terms of θ^* as

$$\hat{\theta} = \theta^* + (U(\theta^*)' W U(\theta^*))^{-1} U(\theta^*)' W e \, . \tag{97}$$

The only random factor in this expression is e and so the variance of our estimate $\hat{\theta}$, can be computed using the fact that if $y = Az$ then the variance of y in terms of the variance of z is given by

$$\mathrm{Var}(y) = A \mathrm{Var}(z) A' \, . \tag{98}$$

Using this we can compute the large sample variance of $\hat{\theta}$ as

$$\mathrm{Var}(\hat{\theta}) = (U(\theta^*)' W U(\theta^*))^{-1} U(\theta^*)' W \mathrm{Var}(e) W U(\theta^*) (U(\theta^*)' W U(\theta^*))^{-1} \, .$$

Where we have used the fact that $W' = W$ as the weight matrix is diagonal. Since $\mathrm{Var}(e) = \sigma^2 W^{-1}$ the above simplifies to

$$\mathrm{Var}(\hat{\theta}) = \sigma^2 (U(\theta^*)' W U(\theta^*))^{-1} \, , \tag{99}$$

which is the books equation 11.14 when σ^2 is approximated with $\hat{\sigma}^2 = \frac{\mathrm{RSS}(\hat{\theta})}{n-k}$.

Problem Solutions

11.1 (a nonlinear model for the sleep data)

11.1.1: The suggested mean function

$$E(Y|X=x, G=j) = \beta_0 + \beta_{1j}(x-\gamma)\,,$$

is nonlinear because of the terms $\beta_{1j}\gamma$ in the above. This mean function has all intercept values (β_0) the same when $x=\gamma$. The model for each group can have a different slope out of this point (β_{1j}).

11.1.2: To fit the given model to the `sleep` data set we need to create indicator variables to denote the factor/class membership for each of the data points. We can introduce decision variables D_i for $i = 1, 2, \cdots, 5$ to construct the mean function

$$\begin{aligned} E(TS|\log_2(BodyWt) = x, D=j) &= \beta_0 + \sum_{j=1}^{5} D_j\beta_{1j}(x-\gamma) \\ &= \beta_0 + \left(\sum_{j=1}^{5} D_i\beta_{1j}\right)(x-\gamma)\,. \end{aligned}$$

This model can be fit much in the same way in which the "source of methionine" example was done in the text. To use the `R` function `nls` we need a starting value to begin the search for the parameters. For this problem the parameter vector θ in terms of its component unknown is

$$\theta' = \begin{bmatrix} \beta_0 & \beta_{11} & \beta_{12} & \beta_{13} & \beta_{14} & \beta_{15} & \gamma \end{bmatrix}\,.$$

To get initial values for the unknowns in θ we will take $\gamma = 0$ and fit the *linear* model

$$E(TS|\log_2(BodyWt) = x, D=j) = \beta_0 + \left(\sum_{j=1}^{5} D_i\beta_{1j}\right)x\,.$$

When we do this using the `R` command `lm` we obtain values of θ given by

```
> coefficients(m1)
(Intercept)       lB:D1       lB:D2       lB:D3       lB:D4       lB:D5
 11.6258529  -0.2004724  -0.4110485  -0.6463494  -0.4446115  -1.1495892
```

Using these parameters as the starting point in the full nonlinear model in the call to `nls` we find that the `summary` command to the results of the model gives

```
> summary(m2)

Formula:
```

```
  TS ~ beta0
     + (D1 * beta11 + D2 * beta12 + D3 * beta13 +
        D4 * beta14 + D5 * beta15) * (lB + gamma)

Parameters:
       Estimate Std. Error t value Pr(>|t|)
beta0   49.3719   192.6534   0.256 0.798771
beta11  -0.4091     0.1785  -2.291 0.026099 *
beta12  -0.4365     0.1160  -3.762 0.000436 ***
beta13  -0.4504     0.1331  -3.383 0.001385 **
beta14  -0.4518     0.1326  -3.408 0.001285 **
beta15  -0.4890     0.2689  -1.818 0.074916 .
gamma   86.7476   440.1284   0.197 0.844536
---
Residual standard error: 3.375 on 51 degrees of freedom
```

A potential difficulty with this fit is that the estimated values for $\gamma \approx 86.7$ and $\beta_0 \approx 49.3$ are significantly outside of the supplied range of $\log_2(BodyWt)$ and TS respectively.

See the `R` file `chap_11_prob_1.R` for an implementation of this problem.

11.2 (a nonlinear model for fish growth)

11.2.1: See Figure 57 (left) for a scatter plot of this data.

11.2.2: For this part of this problem we want to fit the von Bertalanffy model to this data. From the scatter plot we take $L_\infty = 1.05 * max(Length) = 197.4$. Given the data samples of L and the assumed nonlinear model

$$E(Length|Age = t) = L_\infty(1 - \exp(-K(t - t_0))) \,, \tag{100}$$

we solve for the expression $-K(t - t_0)$ and find

$$-K(t - t_0) = \log\left(1 - \frac{L}{L_\infty}\right) \,. \tag{101}$$

We next fit a linear model in t to the right hand side of the above. That is we look for a model involving t of the form $\log\left(1 - \frac{L}{L_\infty}\right) = \beta_0 + \beta_1 t$. If we can fit such a model and determine estimates of β_0 and β_1 then we see that from Equation 101 this would mean that

$$\beta_1 = -K \quad \text{and} \quad \beta_0 = +Kt_0 \,.$$

Solving these two equations for K and t_0, which are needed for the nonlinear least squares fit we see that

$$K = -\beta_0 \quad \text{and} \quad t_0 = \frac{\beta_0}{K} = -\frac{\beta_0}{\beta_1} \,.$$

See Figure 57 (left) for the nonlinear model plotted onto the data points.

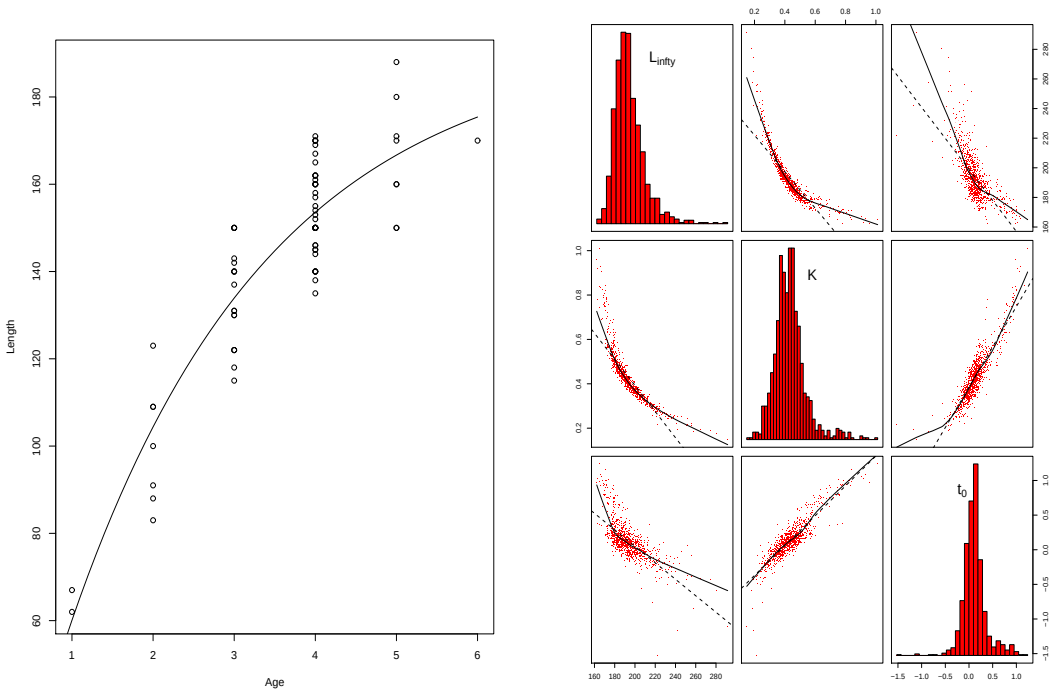

Figure 57: **Left:** A scatter plot of the variables *Length* vs. *Age* and the nonlinear von Bertalanffy model fit to the given data in problem 11.2. **Right:** A scatter plot matrix of various parameters L_∞, K, and t_0 that are be obtained from nonlinear regression on 999 bootstrap samples.

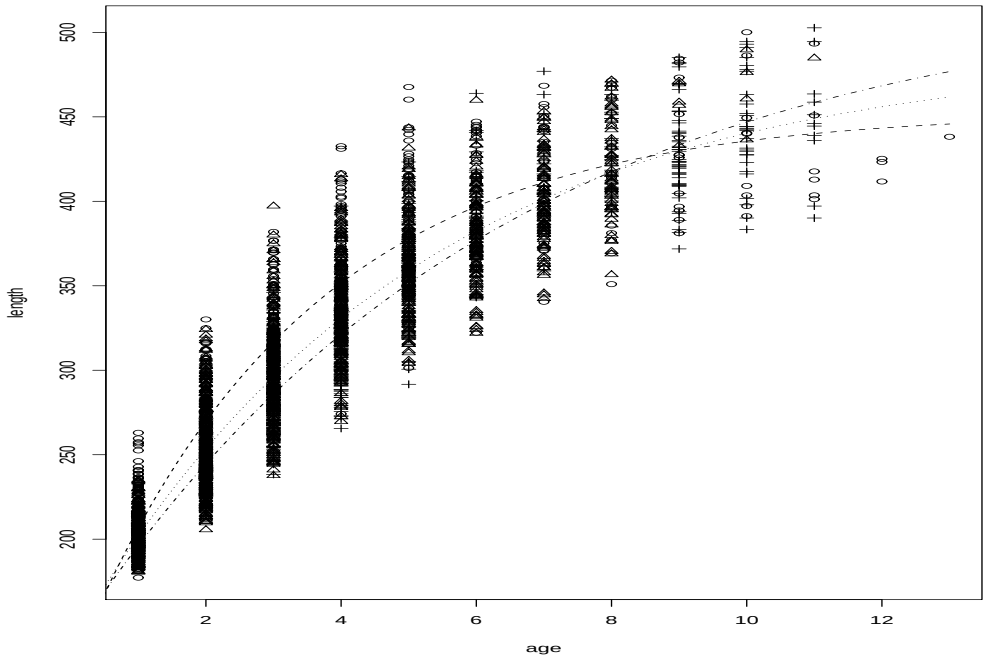

Figure 58: A scatter plot of the variables *length* vs. *age* for the `walleye` data set. Also plotted are the most general nonlinear von Bertalanffy model fit to the given data (one curve for every period).

11.2.3: See Figure 57 (right) for the scatterplot matrix of the parameter estimates found when fitting 999 bootstrap nonlinear models. Notice that the parameter t_0 seems reasonably Gaussian while the distribution of the other two parameters L_∞ and K appear to be skewed to the right. The large sample least squared parameter estimates compared to the bootstrapped samples are shown in in the following table:

```
> round(cbind(n1.ls.summary,n1.boot.summary),2)
      LInfinity    K    t0 LInfinity    K    t0
Mean     192.81 0.41  0.08    193.69 0.43  0.12
SD        13.08 0.09  0.24     15.50 0.11  0.26
2.5%     167.17 0.23 -0.39    171.25 0.25 -0.30
97.5%    218.45 0.58  0.55    232.93 0.75  0.82
```

This table shows the hypothesis above that the parameters L_∞ and K appear to be skewed to the right

See the `R` file `chap_11_prob_2.R` for an implementation of this problem.

11.3 (fitting a growth model to the `walleye` data set)

See the Figure 58 for a scatter plot of the `walleye` data set. If we consider the assumption that there are three different models (one for each period) then the *most general model* will have

$$E(Length|Age = t, P_1, P_2, P_3)$$

given by

$$P_1 L_1(1 - \exp(-K_1(t - t_{01}))) + P_2 L_2(1 - \exp(-K_2(t - t_{02}))) + P_3 L_3(1 - \exp(-K_3(t - t_{03}))) .$$

In addition to this model there are several simplifications that are less general but may fit the data just as well. We will consider some of the possible models, like the *common intercept model* where all periods share the same value of L_∞ and our expectation is equal to

$$L\left[1 - P_1 \exp(-K_1(t - t_{01})) - P_2 \exp(-K_2(t - t_{02})) - P_3 \exp(-K_3(t - t_{03}))\right] ,$$

the *common intercept-rate model* where our expectation is equal to

$$L\left[1 - \exp(-K(t - P_1 t_{01} - P_2 t_{02} - P_3 t_{03}))\right] ,$$

the *common intercept-origin model* where our expectation is equal to

$$L\left[1 - P_1 \exp(-K_1(t - t_0)) - P_2 \exp(-K_2(t - t_0)) - P_3 \exp(-K_3(t - t_0))\right] ,$$

and finally the model where there is *no* difference in parameters among the periods or Equation 100. We can compare these models using the `R` function `anova`. We find that when we compare the most general model to the most specific model that the ANOVA summary statistics are

```
> anova(m0,m1)
Analysis of Variance Table

Model 1: length ~ L * (1 - exp(-K * (age - t0)))
Model 2: length ~ (P1 * (L1 * (1 - exp(-K1 * (age - t01)))) +
                   P2 * (L2 * (1 - exp(-K2 * (age - t02)))) +
                   P3 * (L3 * (1 - exp(-K3 * (age - t03)))))
  Res.Df Res.Sum Sq Df Sum Sq F value    Pr(>F)
1   3195    2211448
2   3189    1963513  6 247935  67.113 < 2.2e-16 ***
```

indicating that the more specific model *does* result in a significant decrease in RSS.

See the `R` file `chap_11_prob_3.R` for an implementation of this problem.

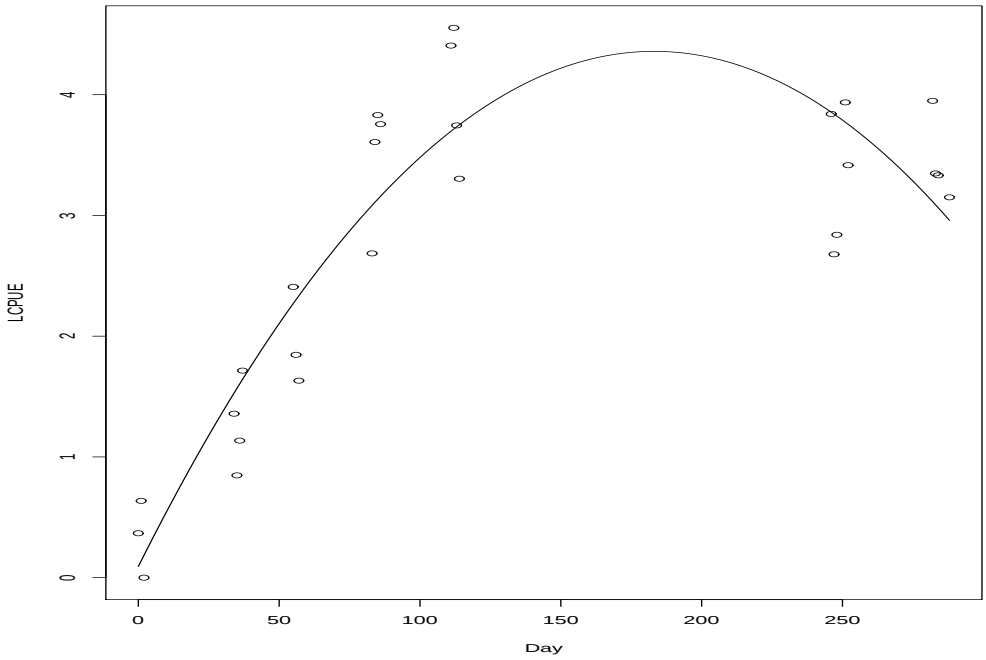

Figure 59: A scatter plot of the variables $LCPUE$ vs. Day for the `swan96` data set and the fitted curve for the estimated quadratic model.

11.4 (a quadratic polynomial as a nonlinear model)

11.4.1: See Figure 59 for the requested plots.

11.4.2: When the model for $LCPUE$ is given by

$$E(LPUE|Day = x) = \beta_0 + \beta_1 x + \beta_2 x^2 \,,$$

the x location where the maximum is obtained is given by $x_m = -\beta_1/(2\beta_2)$ and we can use the delta method to compute the variance of this statistic. When we do that with the `alr3` command `delta.method` we find

```
> delta.method(m0,"-b1/(2*b2)")
Functions of parameters:  expression(-b1/(2*b2))
Estimate = 183.1104 with se = 5.961452
```

We can check this result by using the bootstrap and find similar results.

11.4.3: When the model for $LCPUE$ is parameterized by the nonlinear regression

$$E(LPUE|Day = x) = \theta_1 - 2\theta_2\theta_3 x + \theta_3 x^2 \,,$$

then the x location where the maximum is obtained is given by $x_m = -(-2\theta_2\theta_3)/(2\theta_3) = \theta_2$. Thus we can use the command `nls` to try and fit this nonlinear model and observe the

estimate and large sample variance of the parameter θ_2 from that procedure. We use the estimates from the linear model above to compute starting values for the parameters in θ. We find the estimate and variance of the parameter $\hat{\theta}_2$ given by the corresponding row in the `summary` call or

```
Parameters:
      Estimate Std. Error t value Pr(>|t|)
th2  1.831e+02  5.961e+00  30.716  < 2e-16 ***
```

These results are very similar to the ones we obtained earlier.

See the `R` file `chap_11_prob_4.R` for an implementation of this problem.

11.5 (selecting the transformation using nonlinear regression)

I think the mean function for this problem is supposed to read

$$\begin{aligned} E(\log(Rate)|X_1 = x_1, X_2 = x_2, \Lambda_1 = \lambda_1, \Lambda_2 = \lambda_2) &= \beta_0 + \beta_1 \psi_S(x_1, \lambda_1) + \beta_2 \psi_S(x_2, \lambda_2) \\ &= \beta_0 + \beta_1 \left(\frac{x_1^{\lambda_1} - 1}{\lambda_1} \right) + \beta_2 \left(\frac{x_1^{\lambda_2} - 1}{\lambda_2} \right) , \end{aligned}$$

from which we need to compute starting values for the parameters we need to estimate. To do this we set $\lambda_i = 1$ and then estimate the following model

$$E(\log(Rate)|X_1 = x_1, X_2 = x_2, \Lambda_1 = \lambda_1, \Lambda_2 = \lambda_2) = \beta_0 + \beta_1(x_1 - 1) + \beta_2(x_2 - 1) ,$$

using least squares. When we do that we get the following starting value for β_i

```
> c(b00,b10,b20)
[1]  2.477882358 -0.046937134 -0.005468399
```

We then use these estimated coefficients β_i in the nonlinear mean function and obtain the following

```
Parameters:
     Estimate Std. Error t value Pr(>|t|)
b0     5.0927     1.7023   2.992  0.00514 **
b1    -1.6723     1.9696  -0.849  0.40180
b2    -0.5655     0.6918  -0.817  0.41935
lam1  -0.3524     0.5444  -0.647  0.52176
lam2  -0.6927     0.8793  -0.788  0.43626
```

One thing to note that that the sign of the two estimates of λ_i above are *negative* and the values are different from what is estimated using the `bctrans` function. One could perhaps get different values for λ_i by using different starting values. One could even use the output from `bctran` as a starting value for the parameters in the `nls` command.

See the `R` file `chap_11_prob_5.R` for an implementation of this problem.

11.6 (fitting partial one-dimensional models (POD))

11.6.1: Recall that a POD model has the following functional form

$$E(Y|X = x, F = j) = \eta_{0j} + \eta_{1j}(x'\beta)\,,$$

for vector predictors X and a factor F. When we expand the expression above we see that it is a nonlinear model because of the products terms $\eta\beta$. In the case of the Australian athletes data where our model is given by

$$\begin{aligned} E(LBM|Sex, Ht, Wt, RCC) &= \beta_0 + \beta_1 Sex + \beta_2 Ht + \beta_3 Wt + \beta_4 RCC \\ &+ \eta_0 Sex + \eta_1 Sex \times (\beta_2 Ht + \beta_3 Wt + \beta_4 RCC)\,, \end{aligned}$$

one very simple method one could use to get starting values needed when fitting this with nonlinear least squares is to take $\eta_0 = \eta_1 = 0$ and then fit the model

$$E(LBM|Sex, Ht, Wt, RCC) = \beta_0 + \beta_1 Sex + \beta_2 Ht + \beta_3 Wt + \beta_4 RCC\,.$$

using ordinary least squares.

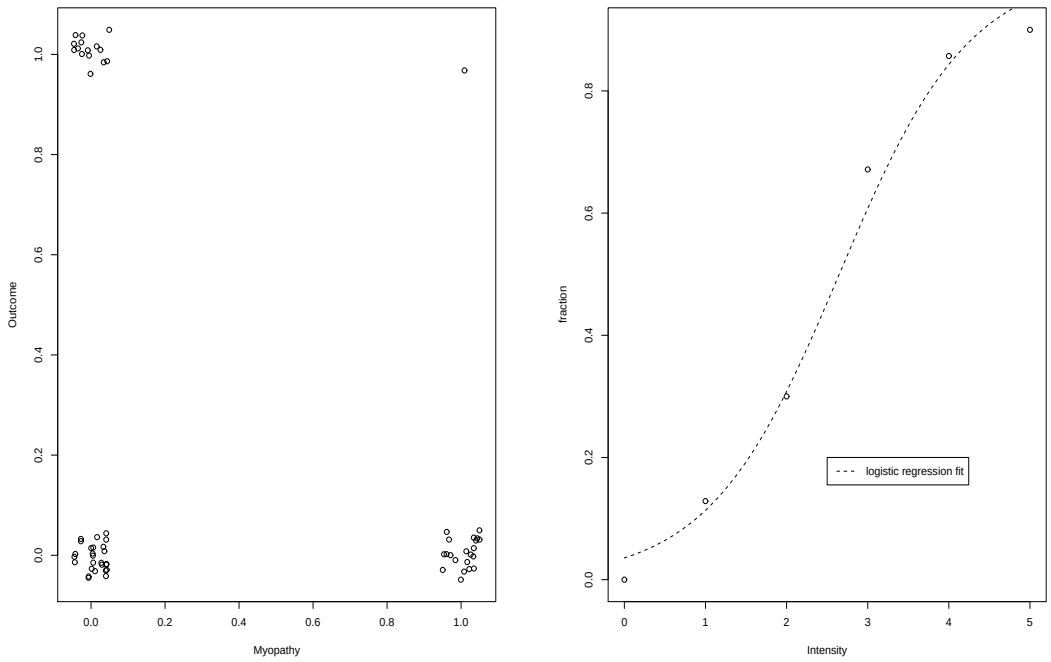

Figure 60: **Left:** A scatter plot of the jittered variables $Myopathy$ vs. $Outcome$ for the data given in problem 12.1. All records with a `NA` have been removed. **Right:** A scatter plot of the variables $Intensity$ vs. $fraction = Y/m$ for the data given in problem 12.3.

Chapter 12 (Logistic Regression)

Problem Solutions

12.1 (Downer data)

12.1.1: See Figure 60 (left) for a scatter plot of the two variables $Myopathy$ and $Outcome$. This plot seems to indicate that when $Myopathy \approx 1$ few cows survive (most values of $Outcome$ are near 0) but when $Myopathy \approx 0$ fewer cows . The fraction of cows that survive under the two cases are given by

$$E(Surviving|Myopathy = 0) = 0.238 \quad \text{and} \quad E(Surviving|Myopathy = 1) = 0.0158\,.$$

12.1.2: We next fit a logistic regression model to this data. The `R` summary command gives

```
glm(formula = Outcome ~ Myopathy, family = binomial(link = "logit"),
    data = downer)

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

```
(Intercept)  -0.5500     0.3242  -1.696   0.0898 .
Myopathy     -2.4945     1.0736  -2.324   0.0201 *
```

To compute a 95% confidence interval for the β coefficient of $Myopathy$, as noted in the book we should use the quantiles of the standard normal distribution rather than the t-distribution. Using the above estimate and standard errors we find the required confidence interval given by

$$-0.390 \leq \hat{\beta}_1 \leq -4.598\,.$$

This is a rather wide range but indicates that $\beta_1 < 0$ indicating that the more $Myopathy$ present in a cow results in an increased chance of death (lower chance that $Outcome = 1$). To determine the estimated probability of survival under the two values of the variable $Myopathy$ we use the `R` function `predict` and the previously determined logistic regression model. When we do this for $Myopathy = 0$ we obtain (we also display the sample estimate derived from the data directly)

$$E(Y|Myopathy = 0) = 0.3658 \quad \text{vs.} \quad 0.2380\,,$$

for the observed survival fraction. When $Myopathy = 1$ in the same way we obtain

$$E(Y|Myopathy = 1) = 0.0454 \quad \text{vs.} \quad 0.0158\,.$$

12.1.3: For some reason that I don't understand I was not able to get `R` to download the `sm` package in the same way in which I was for the `alr3` package. The package was stated as not found in the CRAN archive. A search on the CRAN web site seem to have reference to this package however.

12.1.4: When we fit a logistic model using the `R` function `glm` we find estimated coefficients now given by

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.9099     1.6961   1.716   0.0862 .
logCK        -0.5222     0.2260  -2.311   0.0209 *
```

From which we see that the sign of the β coefficient of $log(CK)$ is negative indicating that larger values of CK indicate less chance of survival.

See the `R` file `chap_12_prob_1.R` for an implementation of this problem.

12.3 (electric shocks)

In Figure 60 (right) we see the requested scatter plot of $Intensity$ vs. $fraction = Y/m$. We see that as $Intensity$ increases we have a steady rise in the value of $fraction$. When we fit a logistic regression model to this data we find fitted coefficients $\hat{\beta}$ given by

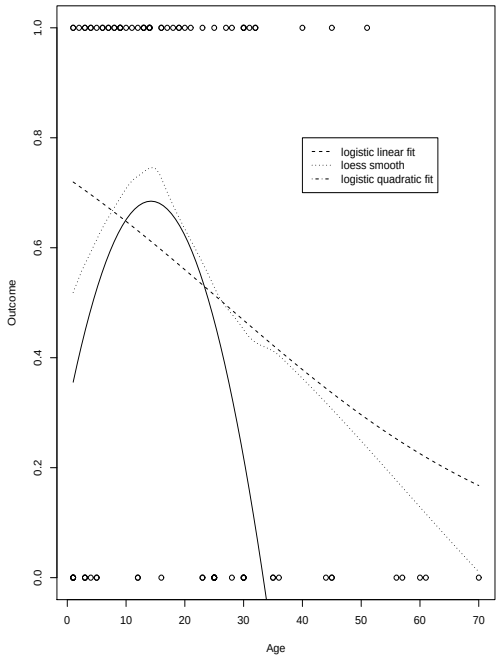

Figure 61: A scatter plot of the variables *Outcome* vs. *Age* for the data given in problem 12.4 along with a logistic regression fit.

```
glm(formula = lRT ~ Intensity, family = binomial(), data = shocks)

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 ***
```

The fact that the z value of the coefficient for *Intensity* is so small means that the found coefficient is significant and unlikely to be zero. This indicates that the probability of response is not independent of intensity. This is consistent with the observed scatter plot. The fact that the sign of the *Intensity* coefficient is positive indicates that as we increase *Intensity* we increase the probability of "mouth movement" increases.

See the `R` file `chap_12_prob_3.R` for an implementation of this problem.

12.4 (the Donner party)

12.4.1: From the given data we have $N_M = 53$ males and $N_F = 35$ in the Donner party and the sample survival rates of each party is given by $s_M = 0.452$ and $s_F = 0.714$.

12.4.2: In Figure 61 we present a scatter plot of *Outcome* vs. *Age* (along with some other

curves). From the scatter plot we see that in general Age seems to increase the probability of death. We can verify this conjecture by fitting a logistic regression model to the variable $Outcome$. We find coefficients of the logistic regression fit given by

```
glm(formula = Outcome ~ Age, family = binomial(), data = donner)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.97917    0.37460   2.614  0.00895 **
Age         -0.03689    0.01493  -2.471  0.01346 *
```

The fact that the estimated coefficient for Age is negative adds evidence to the argument above.

12.4.3: In Figure 61 we overlay on the scatterplot of the raw data a loess smooth and the logistic regression fits given by

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 Age\,,$$

and

$$\log\left(\frac{\theta(x)}{1-\theta(x)}\right) = \beta_0 + \beta_1 Age + \beta_2^2 Age^2\,.$$

In comparing the linear model with the loess fit we see that for values of $Age > 50$ the loess fit and the logistic regression line don't match well. In addition, for small values of Age (near zero) and for values of $Age \approx 15$ the logistic fit does not match the data well. This mismatch of the logistic regression fit and the data may indicate that adolescents may have been more likely to survive. When we compare the quadratic model with the data we see that it fits better for $Age \approx 15$ but performs much poorly for $Age35$ where the quadratic model predicts a zero probability of survival.

See the `R` file `chap_12_prob_4.R` for an implementation of this problem.

12.5 (counterfeit banknotes)

12.5.1: In Figure 62 we present a scatter plot matrix for the variables represented in this problem. Several variables look attractive for this classification problem, but there seems to be a great deal of correlation among the variables that makes using *all* of them problematic. When we fit a logistic model over all terms we get very uncertain estimates of the β parameters.

See the `R` file `chap_12_prob_5.R` for an implementation of this problem.

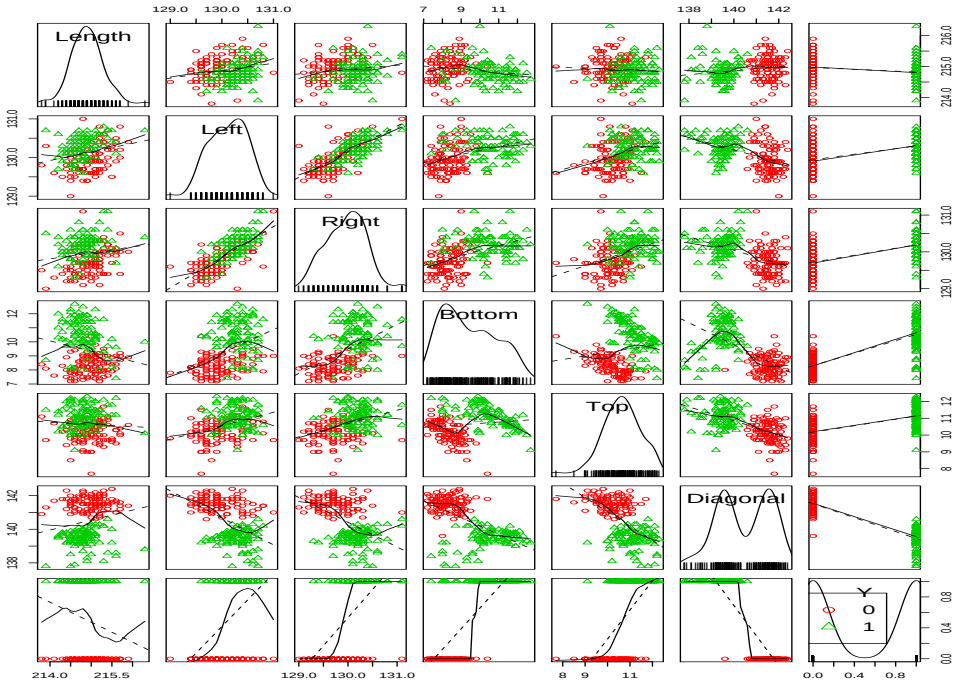

Figure 62: A scatter plot matrix of all the variables for the data given in problem 12.5. It looks like several variables are quite good at predicting forgeries.

Appendix

Notes Least Squares for Simple Regression (A.3):

Here we derive many of the expressions given in this section of the appendix and some that are only stated but not proven. To begin we recall, equations A.7 for simple regression given by

$$\begin{aligned} \beta_0 n + \beta_1 \sum x_i &= \sum y_i \quad (102) \\ \beta_0 \sum x_i + \beta_1 \sum x_i^2 &= \sum x_i y_i \,. \quad (103) \end{aligned}$$

Using the facts that $\sum x_i = n\bar{x}$, $\sum y_i = n\bar{y}$ with

$$\begin{aligned} \text{SXX} &= \sum (x_i - \bar{x})^2 = \sum x_i (x_i - \bar{x}) = \sum x_i^2 - n\bar{x}^2 \quad (104) \\ \text{SYY} &= \sum (y_i - \bar{y})^2 = \sum y_i (y_i - \bar{y}) = \sum y_i^2 - n\bar{y}^2 \quad (105) \\ \text{SXY} &= \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i (y_i - \bar{y}) = \sum y_i (x_i - \bar{x}) = \sum x_i y_i - n\bar{x}\bar{y} \,, \quad (106) \end{aligned}$$

the solutions of Equations 102 and 103 would then be solved for β_0 and β_1. These solutions are denoted $\hat{\beta}_0$ and $\hat{\beta}_1$ where the "hat" notation reminds us that they are *estimates* of the true population parameters, which are written without hats. When we perform the substitution of the summations of $\sum x_i$, $\sum y_i$ and $\sum x_i y_i$ from Equations 104, 105 and 106 into Equations 102 and 103 we get the system

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 \bar{x} &= \bar{y} \quad (107) \\ \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \frac{1}{n} (\text{SXX} + n\bar{x}^2) &= \frac{1}{n} \text{SXY} + \bar{x}\bar{y} \,. \quad (108) \end{aligned}$$

Solving for $\hat{\beta}_0$ in Equation 107 we find

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \,. \quad (109)$$

When we put this value into Equation 108 we find

$$\hat{\beta}_1 = \frac{\text{SXY}}{\text{SXX}} \,, \quad (110)$$

which are the books equations A.9.

Notes on Means and Variances of Least Squares Estimates (A.4)

We next would like to derive expressions for the expected values and variances of the estimators derived above. Since we will be computing expectations conditional on *knowing* the values of X we will write many of expressions we derive in terms of only the y_i variables since if we know X then these are the only variables that are random (under the assumed

model $y_i = \beta_0 + \beta_1 x_i + e_i$) and this simplifies notation some. With this motivation note that $\hat{\beta}_1$ can be written as

$$\begin{aligned}\hat{\beta}_1 &= \frac{\text{SXY}}{\text{SXX}} = \frac{1}{\text{SXX}}\sum(x_i-\bar{x})(y_i-\bar{y}) \\ &= \frac{1}{\text{SXX}}\left(\sum y_i(x_i-\bar{x}) - \bar{x}\sum(x_i-\bar{x})\right) \\ &= \sum\left(\frac{x_i-\bar{x}}{\text{SXX}}\right)y_i = \sum c_i y_i\,,\end{aligned}$$

if we introduce the definition $c_i = \frac{x_i-\bar{x}}{\text{SXX}}$. Note we have used the fact that $\sum(x_i-\bar{x}) = 0$. Here we have lumped all of the x dependence into the variables c_i which makes computing expectations holding x constant easier. We can now compute the bias of the estimator more easily since the c_i's are then constant. We find

$$\begin{aligned}E(\hat{\beta}_1|X) &= E\left(\sum c_i y_i|X=x_i\right) = \sum c_i E(y_i|X=x_i) \\ &= \sum c_i(\beta_0+\beta_1 x_i) = \beta_0\sum c_i + \beta_1\sum c_i x_i\,.\end{aligned}$$

To finish this calculation we need to compute the above two sums. We find

$$\begin{aligned}\sum c_i &= \frac{1}{SXX}\sum(x_i-\bar{x}) = 0 \\ \sum c_i x_i &= \frac{1}{SXX}\left(\sum x_i^2 - \bar{x}\sum x_i\right) \\ &= \frac{1}{SXX}\left(\sum x_i^2 - n\bar{x}^2\right) = 1\,,\end{aligned}$$

so that when we use these results we find that $E(\hat{\beta}_1|X) = \beta_1$, showing that the estimate for β_1 is unbiased. To compute the variance of the estimator $\hat{\beta}_1$ we have

$$\begin{aligned}\text{Var}(\hat{\beta}_1|X) &= \text{Var}(\sum c_i x_i|X=x_i) \\ &= \sum c_i^2\text{Var}(y_i|X=x_i) = \sum c_i^2\sigma^2 = \sigma^2\sum c_i^2 \\ &= \frac{\sigma^2}{SXX^2}\sum(x_i-\bar{x})^2 \\ &= \frac{\sigma^2}{SXX^2}SXX = \frac{\sigma^2}{SXX}\,. \qquad (111)\end{aligned}$$

For the estimate $\hat{\beta}_0$ we can compute its bias as

$$\begin{aligned}E(\hat{\beta}_0|X) &= E(\bar{y}-\hat{\beta}_1\bar{x}|X) = E\left(\frac{1}{n}\sum y_i - \hat{\beta}_1\bar{x}|X\right) \\ &= \frac{1}{n}\sum E(y_i|X) - E(\hat{\beta}_1|X)\bar{x} \\ &= \frac{1}{n}\sum(\beta_0+\beta_1 x_i) - \beta_1\bar{x} \\ &= \beta_0+\beta_1\bar{x}-\beta_1\bar{x} = \beta_0\,,\end{aligned}$$

showing that as claimed in the book the estimate $\hat{\beta}_0$ is also unbiased. The variance of this estimate is given by

$$\begin{aligned}\text{Var}(\hat{\beta}_0) &= \text{Var}(\bar{y}-\hat{\beta}_1\bar{x}|X) \\ &= \text{Var}(\bar{y}|X) + \bar{x}^2\text{Var}(\hat{\beta}_1|X) - 2\bar{x}\text{Cov}(\bar{y},\hat{\beta}_1|X)\,.\end{aligned}$$

So we need to compute several things to evaluate this. We begin with the covariance calculation

$$\begin{aligned}\mathrm{Cov}(\bar{y},\hat{\beta}_1|X) &= \mathrm{Cov}(\frac{1}{n}\sum y_i, \sum c_i y_i|X) \\ &= \frac{1}{n}\sum_i\sum_j c_j \mathrm{Cov}(y_i,y_j|X) \\ &= \frac{1}{n}\sum_i c_j \mathrm{Var}(y_i|X)\,,\end{aligned}$$

since $\mathrm{Cov}(y_i,y_j)=0$ if $i\neq j$ under the assumption that each y_i is independent (given x). Continuing our calculation above we have

$$\mathrm{Cov}(\bar{y},\hat{\beta}_1|X) = \frac{1}{n}\sigma^2\sum c_i = 0\,. \tag{112}$$

Next we evaluate the two variances $\mathrm{Var}(\bar{y}|X)$ and $\mathrm{Var}(\hat{\beta}_1|X)$ to get

$$\begin{aligned}\mathrm{Var}(\bar{y}|X) &= \frac{1}{n^2}\sum \mathrm{Var}(y_i|X) = \frac{\sigma^2}{n^2}n = \frac{\sigma^2}{n} \\ \mathrm{Var}(\hat{\beta}_1|X) &= \mathrm{Var}(\sum c_i y_i|X) \\ &= \sigma^2\sum c_i^2 = \sigma^2\frac{1}{\mathrm{SXX}^2}\sum(x_i-\bar{x})^2 = \frac{\sigma^2}{\mathrm{SXX}}\,.\end{aligned} \tag{113}$$

Thus combining these expressions we find

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\bar{x}^2}{\mathrm{SXX}}\sigma^2 = \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\mathrm{SXX}}\right)\,. \tag{114}$$

Computing next the covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$ we find

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1|X) &= \mathrm{Cov}(\bar{y}-\hat{\beta}_1\bar{x},\hat{\beta}_1|X) \\ &= \mathrm{Cov}(\bar{y},\hat{\beta}_1|X) - \bar{x}\mathrm{Cov}(\hat{\beta}_1,\hat{\beta}_1|X) \\ &= 0 - \bar{x}\mathrm{Var}(\hat{\beta}_1|X) \\ &= -\frac{\bar{x}\sigma^2}{\mathrm{SXX}}\,.\end{aligned} \tag{115}$$

For the variances of a *fitted* value $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$, using many of the results above we have

$$\begin{aligned}\mathrm{Var}(\hat{y}|X) &= \mathrm{Var}(\hat{\beta}_0+\hat{\beta}_1 x|X) \\ &= \mathrm{Var}(\hat{\beta}_0|X) + x^2\mathrm{Var}(\hat{\beta}_1|X) + 2x\mathrm{Cov}(\hat{\beta}_0,\hat{\beta}_1|X) \\ &= \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\mathrm{SXX}}\right) + x^2\left(\frac{\sigma^2}{\mathrm{SXX}}\right) - \frac{2x\bar{x}\sigma^2}{\mathrm{SXX}} \\ &= \sigma^2\left(\frac{1}{n}+\frac{1}{\mathrm{SXX}}(\bar{x}^2-2x\bar{x}+x^2)\right) \\ &= \sigma^2\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\mathrm{SXX}}\right)\,,\end{aligned} \tag{116}$$

which is equation A.11 in the book. Now $\hat{y}$ is called a *fitted* value since it is computed based/using the fitted values $\hat{\beta}_0$ and $\hat{\beta}_1$. That is $\hat{\beta}_0$ and $\hat{\beta}_1$ are determined (or fit) from the give data.

The variance of a *predicted* value $\tilde{y}$ will depend on *both* the errors in the estimated coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ *and* the natural unexplainable variation present in our model. What we mean by that last statement is that for a linear model of the type

$$y_i = \beta_0 + \beta_1 x_i + e_i \,, \tag{117}$$

the error component e_i has a variance of σ^2. Note this last variance is missing when we are talking about the variance of *fitted* values (see above). Thus for predicted values we need to add another σ^2 to Equation 116 to get

$$\mathrm{Var}(\tilde{y}|X) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\mathrm{SXX}} \right) . \tag{118}$$

Notes on Least Squares Using Matrices (A.8)

In this section of the appendix the book shows that

$$\hat{\beta}' \mathbf{X}' \mathbf{X} \hat{\beta} = \mathbf{Y}' \mathbf{X} \hat{\beta} \,. \tag{119}$$

Using this we can derive an expression for $\mathrm{RSS}(\cdot)$ evaluated at the least squares estimate $\hat{\beta}$. Using some of the results from the book in this section we have

$$\begin{aligned} \mathrm{RSS} &\equiv \mathrm{RSS}(\hat{\beta}) = \mathbf{Y}'\mathbf{Y} + \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} - 2\mathbf{Y}'\mathbf{X}\hat{\beta} \\ &= \mathbf{Y}'\mathbf{Y} + \mathbf{Y}'\mathbf{X}\hat{\beta} - 2\mathbf{Y}'\mathbf{X}\hat{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\hat{\beta} \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} \,, \end{aligned}$$

where the last equation is obtained from the one before it by using Equation 119. We can *also* write $\hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}$ as

$$\hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} = (\mathbf{X}\hat{\beta})'(\mathbf{X}\hat{\beta}) = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} \,.$$

This last result show that we can write $\mathrm{RSS}(\hat{\beta})$ as

$$\mathrm{RSS} = \mathbf{Y}'\mathbf{Y} - \hat{\mathbf{Y}}'\hat{\mathbf{Y}} \,. \tag{120}$$

Notes on Case Deletion in Linear Regression (A.12)

In this subsection of these notes we will derive many of the results presented in this section of the appendix. Our first goal will be to prove the case deletion inverse identity which is given by

$$(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'_i\mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}} \,. \tag{121}$$

To do this we will begin by proving two related identities that we will use in this proof. The first is

$$\mathbf{X}_{(i)}'\mathbf{X}_{(i)} = \mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i' \,. \tag{122}$$

To do this lets the left-hand-side of Equation 122 in terms of the sample vectors $\mathbf{x}_i$

$$\begin{aligned}
\mathbf{X}_{(i)}'\mathbf{X}_{(i)} &= \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_{i-1}' \\ \mathbf{x}_{i+1}' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}' \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_{i-1}' \\ \mathbf{x}_{i+1}' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{i-1} & \mathbf{x}_{i+1} & \cdots & \mathbf{x}_n \end{bmatrix} \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_{i-1}' \\ \mathbf{x}_{i+1}' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix} \\
&= \sum_{i=1;k\neq i}^{n} \mathbf{x}_k\mathbf{x}_k' = \sum_{k=1}^{n} \mathbf{x}_k\mathbf{x}_k' - \mathbf{x}_i\mathbf{x}_i' = \mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i' \,,
\end{aligned}$$

which is the desired expression. The second identity we will show is the analogous expression for $\mathbf{X}_{(i)}'\mathbf{Y}_{(i)}$ and is given by

$$\mathbf{X}_{(i)}'\mathbf{Y}_{(i)} = \mathbf{X}'\mathbf{Y} - \mathbf{x}_i y_i \,. \tag{123}$$

The proof of this is done in exactly the same way as in the proof of Equation 122, but is somewhat easier to understand since $\mathbf{Y}_{(i)}$ is vector and not a matrix. Again consider the left-hand-side of Equation 123 in terms of the sample vectors $\mathbf{x}_i$ and response y_i. We have

$$\begin{aligned}
\mathbf{X}_{(i)}'\mathbf{Y}_{(i)} &= \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_{i-1}' \\ \mathbf{x}_{i+1}' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}' \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{i-1} & \mathbf{x}_{i+1} & \cdots & \mathbf{x}_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{i-1} \\ y_{i+1} \\ \vdots \\ y_n \end{bmatrix} \\
&= \sum_{i=1;k\neq i}^{n} y_k\mathbf{x}_k = \sum_{k=1}^{n} y_k\mathbf{x}_k - y_i\mathbf{x}_i = \mathbf{X}'\mathbf{Y} - y_i\mathbf{x}_i \,,
\end{aligned}$$

completing this derivation. Next we will derive Equation 121. From Equation 122 we have

$$(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1} = (\mathbf{X}'\mathbf{X} - \mathbf{x}_i\mathbf{x}_i')^{-1} \,.$$

To evaluate the right-hand-side of the above we will use the *Sherman-Morrison-Woodbury* formula

$$(\mathbf{A} + UV')^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}U(I + V'\mathbf{A}^{-1}U)^{-1}V'\mathbf{A}^{-1} \,, \tag{124}$$

with $U = -\mathbf{x}_i$, and $V = \mathbf{x}_i$. This then gives

$$(\mathbf{A} - \mathbf{x}_i\mathbf{x}_i')^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{x}_i(1 + \mathbf{x}_i'\mathbf{A}^{-1}\mathbf{x}_i)^{-1}\mathbf{x}_i'\mathbf{A}^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{x}_i\mathbf{x}_i'\mathbf{A}^{-1}}{1 + \mathbf{x}_i'\mathbf{A}^{-1}\mathbf{x}_i} \,.$$

If we take $\mathbf{A} = \mathbf{X}'\mathbf{X}$ and recognize that $\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \equiv h_{ii}$ we have

$$(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1} = (\mathbf{X}'\mathbf{X})^{-1} + \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'\mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}}{1 - h_{ii}} \,,$$

or Equation 121 the books equation A.37. Next consider how we can use Equation 121 to derive an expression for $\hat{\beta}_{(i)}$ the estimated least squares regression coefficients *excluding* the sample $\mathbf{x}_i$. Since we can write $\hat{\beta}_{(i)}$ as

$$\hat{\beta}_{(i)} = (\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)}\,,$$

we can post multiply $(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}$ by $\mathbf{X}'_{(i)}\mathbf{Y}_{(i)}$ and use Equation 121 to get

$$\begin{aligned}
(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)} + \frac{1}{1+h_{ii}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)} \\
&+ \frac{1}{1+h_{ii}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)} \\
&= \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\left[-\mathbf{X}'\mathbf{Y} + \mathbf{X}'_{(i)}\mathbf{Y}_{(i)} + \frac{1}{1+h_{ii}}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)}\right]\,.
\end{aligned}$$

Next using Equation 123 to replace $\mathbf{X}'_{(i)}\mathbf{Y}_{(i)}$ in terms of $\mathbf{X}'\mathbf{Y}$ the above becomes

$$\begin{aligned}
(\mathbf{X}'_{(i)}\mathbf{X}_{(i)})^{-1}\mathbf{X}'_{(i)}\mathbf{Y}_{(i)} &= \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\left[-\mathbf{x}_i y_i + \frac{1}{1+h_{ii}}\mathbf{x}_i\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y} - \mathbf{x}_i y_i)\right] \\
&= \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\left[-y_i + \frac{1}{1+h_{ii}}\mathbf{x}_i'(\hat{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i y_i)\right] \\
&= \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\left[-y_i + \frac{1}{1+h_{ii}}\hat{y}_i - \frac{1}{1-h_{ii}}h_{ii}y_i\right]\,,
\end{aligned}$$

when we recall that $h_{ii} \equiv \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i$ and $\hat{y}_i = \mathbf{x}_i'\hat{\beta}$. Combining the first and third terms above and using $\hat{e}_i = y_i - \hat{y}_i$ we get

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\hat{e}_i}{1-h_{ii}}\,, \tag{125}$$

which is the book's equation A.38.

References

[1] R. D. Cook and S. Weisberg. *Residuals and Influence in Regression*. Chapman and Hall, 1982.