Solutions Manual for *Regression Analysis by Example*

Richard Charnigo, M.S.

Summer 2001

1 Solutions for Exercises in Chapter 1

1.1
(a) Qualitative
[There are different ways to divide the U.S. (or the world) into geographic regions. One common way is via time zones, which separate the 48 contiguous United States into Pacific, Mountain, Central, and Eastern regions.]
(b) Quantitative
(c) Quantitative
(d) Qualitative
[Some categories might be African-American, Hispanic, Native American, Asian-American, and White.]
(e) Quantitative
(f) Quantitative
(g) Quantitative
(h) Qualitative
[The categories might be Democratic, Republican, and Independent (or Neither); one could also include less prominent third parties as categories.]

1.2
Here are (simplifications of) two real-life examples in which regression has been used as an analytic tool to answer questions; Cf. exercises 5.4 and 5.7.

Example (i)
(a) We might be interested in finding out what factors are related to the differences among states in per capita expenditure on public education.
(b) The response would be per capita expenditure; the predictors could be such things as per capita personal income, proportion of the population under 18, and proportion of the population living in urban areas, and geographic region.
(c) Geographic region is qualitative; all other predictors suggested in part (b) are quantitative.
(d) Presuming that we propose a model that is linear in the parameters, the following adjectives from Table 1.11 apply: Univariate, Multiple, Linear, Analysis

of Covariance.
(e) We might have $Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 R_1 + \gamma_2 R_2 + \cdots + \gamma_k R_k + \epsilon$, where k is the number of geographic regions. Here, Y is per capita expenditure, X_1 is per capita personal income, X_2 is the proportion of the population under 18, X_3 is the proportion of the population living in urban areas, and R_j is 1 for those states in region j (and 0 for those states outside of region j). The parameters are the betas and the gammas.

Example (ii)
(a) We might be interested in finding out what factors affect the popular vote in United States presidential elections.
(b) The response could be the proportion of the two-party vote received by the Democratic candidate; possible predictors include the growth rate of the economy, the inflation rate, and the political party of the incumbent.
(c) The political party of the incumbent is qualitative; the other predictors suggested in part (b) are quantitative.
(d) Presuming that we propose a model that is linear in the parameters, the following adjectives from Table 1.11 apply: Univariate, Multiple, Linear, Analysis of Covariance.
(e) We might have $Y = \beta_1 X_1 + \beta_2 X_2 + \gamma_1 I_1 + \gamma_2 I_2 + \epsilon$. Here, Y is the proportion of the two-party vote received by the Democratic candidate, X_1 is the growth rate of the economy, X_2 is the inflation rate, I_1 is 1 if the incumbent is a Democrat (and 0 otherwise), and I_2 is 1 if the incumbent is a Republican (and 0 otherwise). The parameters are the betas and the gammas.

1.3
(a) Predictor: number of cylinders; Response: gas consumption
We expect gas consumption to be a function of the number of cylinders.
(b) Predictors: SAT scores, grades; Response: college admission
College admission officers base their decisions on SAT scores, grades, and other factors.
(c) Predictor: demand; Response: supply
Generally: If demand for good X is high, manufacturers can increase their revenue by producing more of good X (i.e., raising its supply). If demand is low, manufacturers can reduce cost by shifting their resources elsewhere. [One could also argue that supply should be the predictor and demand thre response, as this is plausible in some cases. In particular, if supply is known to be very low, people may feel pressured into a quick purchase that they might not otherwise decide to make. As examples, consider tickets for sporting events or limited edition collectibles.]
(d) Predictors: net sales, company's assets; Response: return on stocks
The value of a company's stock is related to whether the company's operations are successful (as measured by, e.g., net sales) and to the company's wealth (assets).
(e) Predictors: distance, weather; Response: time to run
The time to run will be a function of the distance and the weather.

(f) Predictors: weight, smoking; Response: lung cancer
Weight and smoking are believed to cause or contribute to health problems.
(g) Predictors: child's gender and age, parents' height and weight; Response: child's height and weight
A child's size is related to his/her gender, age, and genetic makeup (for which parents' height and weight are proxies).

1.4
(a) The decision to admit to college - either a "yes" or a "no" - is qualitative. Similarly, whether a person is a smoker or has cancer is qualitative. Gender is qualitative. Some weather conditions (e.g., "sunny" or "rainy") would be qualitative, but others (e.g., temperature) would be quantitative. All other variables mentioned in problem 1.3 would typically be measured quantitatively.
(b) Whether Linear is an appropriate adjective depends on the mathematical form of the model. However, the following adjectives from Table 1.11 would apply in any case:

(a) Univariate, Simple

(b) Logistic, Multiple

(c) Univariate, Simple (assuming a single measure of demand and a single measure of supply)

(d) Univariate, Multiple

(e) Univariate, Multiple, Analysis of Covariance

(f) Logistic, Multiple, Analysis of Covariance

(g) Multivariate, Multiple, Analysis of Covariance

2 Solutions for Exercises in Chapter 2

2.1
(a) $Var(Y) = \sum_{i=1}^{14}(y_i - \bar{y})^2/13 = 27768.36/13 = 2136.03$; $Var(X) = \sum_{i=1}^{14}(x_i - \bar{x})^2/13 = 114/13 = 8.77$
For parts (b) - (e), we give a proof that the assertion or formula holds in general; the student may also perform a verification for the data in Table 2.6.
(b) $\sum_{i=1}^{n}(y_i - \bar{y}) = \sum_{i=1}^{n} y_i - n\bar{y} = \sum_{i=1}^{n} y_i - n(\sum_{i=1}^{n} y_i/n) = 0$
(c) Let $z_i := (y_i - \bar{y})/s_y$. From part (b), $s_y \sum_{i=1}^{n} z_i = 0$, but the left member of the equation equals $s_y n\bar{z}$, implying that $\bar{z} = 0$. Also, $Var(Z) = \sum_{i=1}^{n}(z_i - \bar{z})^2/(n-1) = \sum_{i=1}^{n} z_i^2/(n-1) = \sum_{i=1}^{n}(y_i - \bar{y})^2/(s_y^2(n-1)) = s_y^2/s_y^2 = 1$, implying that $s_z = 1$.
(d) Equation (2.5) may be obtained from equation (2.6) by substituting the expression for $Cov(Y, X)$ given in equation (2.2). Equation (2.7) may be obtained from equation (2.5) by substituting the expressions for s_y and s_x indicated by equation (2.4), then noting that $(n-1)$ factors cancel.
(e) To see that the equations (2.13) and (2.19) are equivalent, write $Cov(Y, X)/Var(X) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})/(n-1)}{\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)}$ and cancel the factors of $(n-1)$.

2.2
(a) $Cov(Y, X)$ may take any value, but equation (2.8) states that $-1 \leq Cor(Y, X) \leq 1$.
(b) The statement is not true in general; for a counterexample, refer to Table 2.3.
(c) The statement is not true in general; for a counterexample, consider any situation in which $\hat{\beta}_1 = 0$ and, hence, $\hat{Y} = \bar{Y}$ (so that the points in the scatter plot form a vertical configuration).

2.3
(a) Using equation (2.28) with $\beta_1^0 = 15$, we obtain $t_1 = (15.509 - 15)/0.505 = 1.01$. From Table A.2 we obtain -1.78 and 1.78 as the critical values for a two-sided test ($\alpha = .10$, 12 degrees of freedom). Thus we would not reject the null hypothesis.
(b) The test statistic is still 1.01, but the critical value for a one-sided test is 1.36. Thus we would not reject the null hypothesis in favor of the one-sided alternative.
(c) One could use formula (2.29) and Table A.2, but it is easier to look at the regression output in Table 2.9, which provides results for two-sided tests of the coefficients' equality to zero; the p-value of 0.2385 indicates that the null hypothesis would not be rejected for any value of α less than 0.2385.
(d) Here we need to use formula (2.29) with $\beta_0^0 = 5$. We get $t_0 = (4.162 - 5)/3.355 = -0.25$, which is less in magnitude than 1.78. Hence we would not reject the null hypothesis.

2.4
Using equation (2.32) and Table A.2, we obtain $4.162 \pm (3.06)(3.355) = 4.162 \pm 10.266$.

2.5
We prove mathematically that each statement is true in general; the student may also verify that each statement holds for the data in Table 2.5.
(a) $\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = \sum_{i=1}^n (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i) = \sum_{i=1}^n (y_i - \bar{y}) + \hat{\beta}_1 \sum_{i=1}^n (\bar{x} - x_i) = 0$. The first equality follows from equations (2.16) and (2.17); the second equality comes from formula (2.14); and the third equality follows from part (b) of exercise 2.1.
(b) Perhaps the easiest way to prove that (2.25) and (2.31) are equivalent is to begin by using equations (2.19) and (2.46), respectively, to rewrite the numerator and denominator of (2.31). Doing so, we find that $t_1 = \frac{\hat{\beta}_1}{\sqrt{s_y^2(1-R^2)}/\sqrt{s_x^2(n-2)}}$. Since $s.e.(\hat{\beta}_1) = \sqrt{SSE/(n-2)}/\sqrt{(n-2)s_x^2}$ by (2.22) and (2.24), all we need to show is that $\sqrt{s_y^2(1-R^2)} = \sqrt{SSE/(n-2)}$; to do this, we may write $s_y^2 = SST/(n-2)$ and apply formula (2.45).

(c) Since $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$, the scatter plot of Y versus $\hat{Y}$ may be obtained by "stretching" the scatter plot of Y versus X horizontally by a factor of $|\hat{\beta}_1|$ (and reflecting the plot about the vertical axis if $\hat{\beta}_1 < 0$), then "shifting" the resulting plot horizontally by the amount $\hat{\beta}_0$; the relative configuration of points is preserved as long as $\hat{\beta}_1 \neq 0$.
(d) $Cor(Y, \hat{Y}) = Cor(Y, \hat{\beta}_0 + \hat{\beta}_1 X) = Cor(Y, \hat{\beta}_1 X) = sign(\hat{\beta}_1) Cor(Y, X) = sign(Cor(Y, X)) Cor(Y, X) = |Cor(Y, X)| \geq 0$. The critical step was to use formula (2.19) to equate $sign(\hat{\beta}_1)$ and $sign(Cor(Y, X))$.

2.6
(a) From Table 2.6, $Cor(Y, X) = \frac{\sum_{i=1}^{14} (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{14} (y_i - \bar{y})^2 \sum_{i=1}^{14} (x_i - \bar{x})^2}} = \frac{1768.00}{\sqrt{27768.36 \cdot 114}} =$ 0.994. A similar calculation using the data from Table 2.7 yields $Cor(Y, \hat{Y}) =$ 0.994 as well.
(b) This number is actually given to you in Table 2.6, once you recall that $SST = \sum_{i=1}^{14} (y_i - \bar{y})^2$.
(c) Just square the e_i values in Table 2.7 and add them. [Because the entries in Table 2.7 were rounded, the answer will be accurate to three significant digits.] An alternative approach would be to use equations (2.45) and (2.46) in tandem with parts (a) and (b) of this exercise.

2.7
(a) The coefficient estimates agree to two decimal places for all four data sets: $\hat{\beta}_0 = 3.00$ and $\hat{\beta}_1 = 0.50$.
(b) All four correlations are between .8162 and .8165.
(c) All four values of R^2 are between .6662 and .6667.
(d) The t values agree to two decimal places for all four data sets; the common value is 4.24.

2.8
(a) The residuals have the form $y_i - \bar{y}$ because $\hat{y}_i = \hat{\beta}_0 = \bar{y}$.
(b) This follows from part (b) of exercise 2.1.

2.9
(a) $Var(Y) = SSR + SSE = 0.0902$; $Cor(Y, X)$ is the positive square root of R^2 (positive because $\hat{\beta}_1 > 0$), which is 0.631.
(b) The estimated participation rate would be $0.203311 + 0.656040(0.45) =$ 0.4985.
(c) With $\alpha = .05$, $n = 19$, and the information provided in Table 2.10, we may employ formula (2.37) to obtain the 95 percent confidence interval for our prediction in part (b). The result is $0.4985 \pm 2.11(0.0566)\sqrt{1 + \frac{1}{19} + \frac{(.4985 - .5)^2}{18 \cdot Var(X)}} =$ $.4985 \pm .1225$.
(d) We may use the computer output and formula (2.33) to obtain $0.6560 \pm 2.11(0.1961) = 0.6560 \pm 0.4137$ as the 95 percent confidence interval for β_1.

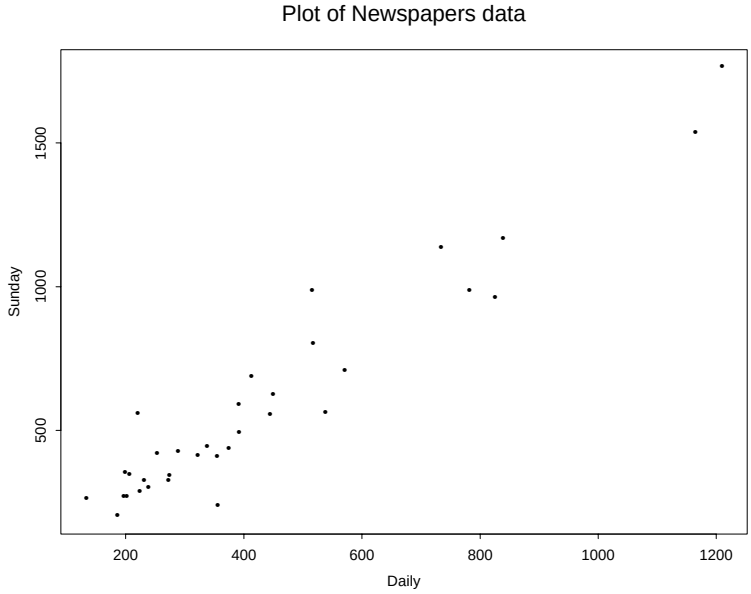
(e) The critical value for the test statistic (see formula (2.28)) is 1.74. However, we can see that the test statistic will be negative without actually computing it; therefore, we may automatically conclude that the null hypothesis will not be rejected.
(f) R^2 would not change because $R^2 = (Cor(Y, X))^2$ and $Cor(Y, X) = Cor(X, Y)$.

2.10
(a) The covariance is 69.41.
(b) The covariance would be multiplied by the square of the centimeters-to-inches conversion factor, 1/2.54; the final result would be 10.76.
(c) The correlation is .7634.
(d) The correlation would not change.
(e) If $X = Y - 5$ exactly, then the correlation would be 1.00.
(f) If one wanted to predict the heights of males' wives, then Y should be the response. If one wanted to predict the heights of females' husbands, then X should be the response. If one merely wants to investigate whether the heights seem to be related, either X or Y may be chosen as the response.
(g) We take Y as the response variable. We find that $\hat{\beta}_1 = 0.6997$ and that the slope estimate has a standard error of 0.0611. The test statistic (see formula (2.25)) is $0.6997/0.0611 = 11.46$, which would result in the rejection of the null hypothesis at any reasonable significance level. [For example, the critical values at level $\alpha = .05$ are ± 2.04.]
(h) Again, we take Y as the response variable. We find that $\hat{\beta}_0 = 16.5079$ and that the intercept estimate has a standard error of 4.1975. The test statistic (see formula (2.30)) is $16.5079/4.1975 = 3.93$, which would result in the rejection of the null hypothesis at any reasonable significance level.
(i) Such a null hypothesis is unreasonable, and may automatically be rejected, for it amounts to hypothesizing that the average height of wives is zero!
(j) The hypothesis that people of similar heights tend to marry each other can be best tested as described in part (k).
(k) Modelling men's heights as $m_i = \mu_m + \varepsilon_{m,i}$ and women's heights as $w_i = \mu_w + \varepsilon_{w,i}$, we get $w_i - m_i = (\mu_w - \mu_m) + (\varepsilon_{w,i} - \varepsilon_{m,i})$. If tall men marry tall women and short men marry short women, then $(\varepsilon_{w,i} - \varepsilon_{m,i})$ will be close to zero, and it will be reasonable to write $w_i = \beta_0 + \beta_1 m_i + \varepsilon_i$, where $\beta_0 = (\mu_w - \mu_m)$, $\beta_1 = 1$, and $\varepsilon_i = (\varepsilon_{w,i} - \varepsilon_{m,i})$. So, we may test the hypothesis that people of similar heights tend to marry each other by fitting the regression model and using formula (2.28) to test whether $\beta_1 = 1$. We obtain $t_1 = (0.6697 - 1)/0.0611 = -5.41$, from which we may conclude that it is not generally true that people who marry each other are of similar heights.

2.11
(a) Many examples are possible. Here is one: Suppose we are interested in the fuel economy of a particular kind of automobile. We may let Y be the distance driven (in miles) and X be the fuel consumption (in gallons). Clearly, Y would

equal zero when X equals zero. Then, in the context of model (2.48), β_1 is a measure of the fuel economy (miles per gallon).
For parts (b) and (c) we consider a fictional data set: Let X take the values $-5, -4, \cdots, 4, 5$ and let $Y := X^2 + 1$.
(b) If we try to fit model (2.48), we will obtain $\hat{\beta}_1 = 0$ so that, in particular, $e_i = y_i$. Each e_i is positive, so their sum is positive.
(c) $SST = 858$ and $SSE = \sum_{i=1}^{11} y_i^2 = 2189$, yielding $R^2 = -1.55$.
(d) Compare the residual mean squares (values of $\hat{\sigma}^2$).

2.12
(a) The plot, shown below, suggests a (positive) linear relationship between daily and Sunday circulation. It is plausible that newspapers with larger daily circulations would also have larger Sunday circulations, and that the basic relationship between the two kinds of circulation would not be too complicated; a linear model therefore seems reasonable.



(b) The fitted line has the equation $\hat{Y} = 13.8356 + 1.3397X$, where X is the daily circulation (in thousands) and Y is the Sunday circulation (in thousands).
(c) The 95 percent confidence interval for β_0 is $13.8356 \pm (2.04)(35.8040)$, which is 13.8356 ± 73.0402. Here, 35.8040 is the standard error of $\hat{\beta}_0$ and 2.04 is close to the 97.5 percentile of a t distribution with 32 degrees of freedom. The 95 percent confidence interval for β_1 is $1.3397 \pm (2.04)(0.0708)$, which is 1.3397 ± 0.1444. Here, 0.0708 is the standard error of $\hat{\beta}_1$. [See formulas (2.32) and (2.33).]
(d) An appropriate procedure is to test the null hypothesis that $\beta_1 = 0$. The critical values for the test statistic (see formula (2.25)) are ± 2.04. The value

we obtain is $1.3397/0.0708 = 18.9348$, which results in rejection of the null hypothesis. We conclude that there is a significant relationship between daily and Sunday circulation.
(e) The proportion of the variability in Sunday circulation accounted for by daily circulation is $R^2 = 0.9181$.
(f) We employ formulas (2.39) and (2.40), using the facts that $\bar{x} = 430.96$, $\sum_{i=1}^{34}(x_i - \bar{x})^2 = 2391669$, $\hat{\sigma} = 109.40$, and $\hat{\mu_0} = 683.69$. We obtain a 95 percent confidence interval of $683.69 \pm (2.04)(109.4)\sqrt{1/34 + 4766.52/2391669}$, which is 683.69 ± 39.55.
(g) Now the relevant formulas are (2.36) and (2.37); this interval will be wider than the preceding one because variability in Y about its conditional mean is taken into account. We get $683.69 \pm (2.04)(109.4)\sqrt{1 + 1/34 + 4766.52/2391669}$, which equals 683.69 ± 226.65.
(h) The 95% prediction interval is $2693.24 \pm (2.04)(109.4)\sqrt{1 + 1/34 + 2461879/2391669}$, which equals 2693.24 ± 320.22. We may note that this interval is more than 41% wider than the interval from part (g). Since none of the newspapers from which the model was fitted had daily circulation close to two million, we are skeptical about the accuracy of this prediction interval.

3 Solutions for Exercises in Chapter 3

3.1
Regressing Y on X_2, we obtain the fitted equation $\hat{Y} = 42.1087 + 0.4239X_2$. Regressing X_1 on X_2, we obtain the fitted equation $\hat{X_1} = 34.3196 + 0.6075X_2$. Finally, fitting the model that relates the two sets of residuals, we obtain $\hat{e}_{Y \cdot X_2} = 0 + 0.7803 e_{X_1 \cdot X_2}$.

3.2
Let X_1 take values $1, 0, 1, 0$ and let X_2 take values $1, 0, -1, 0$;
note that $Cor(X_1, X_2) = 0$. Finally, let Y take values $2, 3, 4, 5$. Then the fitted equation for the model with only X_1 as a predictor is $\hat{Y} = 4.00 - 1.00X_1$, while the fitted equation for the model with both X_1 and X_2 as predictors is $\hat{Y} = 4.00 - 1.00X_1 - 1.00X_2$; note that X_1 has the same coefficient in both models.

3.3
(a) For Model 1, the fitted equation is $\hat{F} = -22.3424 + 1.2605P_1$; for Model 2, the fitted equation is $\hat{F} = -1.8535 + 1.0043P_2$; for Model 3, the fitted equation is $\hat{F} = -14.5005 + 0.4883P_1 + 0.6720P_2$.
(b) The values for the test statistics (see formula (3.23)) are -1.93, -0.25, and -1.57 for Models 1, 2, and 3, respectively. The critical values for rejection of the null hypothesis (at $\alpha = .05$) are ± 2.09. [The 97.5 percentile of a t distribution with 19 or 20 degrees of freedom equals 2.09 to two decimal places.] Therefore, the null hypothesis that $\beta_0 = 0$ cannot be rejected.

(c) P_2 is a slightly better predictor of F than P_1 because its correlation to F is greater.
(d) Model 3 attempts to incorporate more information that might be useful in predicting the final exam score. Indeed, upon fitting the model, we see that the coefficients for both P_1 and P_2 are significant at the .05 level, suggesting that it is helpful to know both P_1 and P_2 when trying to predict F. Using Model 3, then, we obtain a fitted value of 80.7, which rounds to 81.

3.4
To make the requested verifications, we first obtain the results from regressing P_2 on P_1 and vice versa. The fitted equations are $\hat{P_2} = -11.6689 + 1.1490P_1$ and $\hat{P_1} = 25.8981 + 0.6803P_2$.
(a) $0.4883 + 0.6720 \cdot 1.1490 = 1.2604$, which shows that (apart from rounding) $\hat{\beta_1'} = \hat{\beta_1} + \hat{\beta_2}\hat{\alpha_1}$.
(b) $0.6720 + 0.4883 \cdot 0.6803 = 1.0042$, which shows that (apart from rounding) $\hat{\beta_2'} = \hat{\beta_2} + \hat{\beta_1}\hat{\alpha_2}$.
Students who have studied calculus may make an analogy to the chain rule of differentiation: $\frac{dY}{dX_1} = \frac{\partial Y}{\partial X_1} + \frac{\partial Y}{\partial X_2}\frac{dX_2}{dX_1}$.

3.5
Since we know this output is based on a simple regression model fitted from twenty observations, we can immediately fill in $n = 20$ and $d.f.$ values of 1, 18, and 18, respectively. Also, from formula (3.23) it follows that the missing t-test value should be $-23.4325/12.74 = -1.84$ and that the missing Coefficient value should be $8.32 \cdot 0.1528 = 1.27$. From formula (3.42), the missing F-test value must be $(8.32)^2 = 69.22$. Since there is 1 $d.f.$ for Regression, the Mean Square value for Regression is simply 1848.76; the value of the F-test implies that the Mean Square Residual equals $1848.76/69.22 = 26.71$ (and, hence, that the Residual Sum of Squares is $26.71 \cdot 18 = 480.78$). The rest of the table is now easy to fill in: $\hat{\sigma} = \sqrt{26.71} = 5.17$, $R^2 = \frac{1848.76}{1848.76+480.78} = 0.794$, and (see formula (3.19)) $R_a^2 = 1 - \frac{19}{18}(1 - .794) = .783$. To complete the exercise, we are asked to compute $Var(Y)$ and $Var(X_1)$. To find $Var(Y)$, we just add the Regression Sum of Squares and the Residual Sum of Squares, then divide by the total degrees of freedom; we get an answer of $2329.54/19 = 122.61$. We may solve for $Var(X_1)$ by using equation (2.19): $\sqrt{Var(X_1)} = Cor(Y, X) \cdot \sqrt{Var(Y)}/\hat{\beta}_1 = \sqrt{.794} \cdot \sqrt{122.61}/1.27 = 7.77$, whence $Var(X_1) = 60.36$.

3.6
We may immediately fill in $n = 18$ and $d.f.$ values of 1, 16, and 16. From formula (3.23), the missing standard error is $3.43179/0.265 = 12.95$; from formula (3.19), $R_a^2 = 1 - \frac{17}{16}(1 - .716) = .698$. Since $\hat{\sigma} = 7.342$, we know that the Mean Square Residual is 53.90 and, hence, that the Residual Sum of Squares is $53.90 \cdot 16 = 862.4$. The Regression Sum of Squares is then obtained as the solution to the equation $\frac{t}{t+862.4} = 0.716 = R^2$, which is 2174.2. Since there is 1 $d.f.$ for Regression, the Mean Square value for Re-

gression is just 2174.2, from which we find that the missing F-test value must be $2174.2/53.90 = 40.34$. The missing t-test value is the square root of the missing F-test value, 6.35, which implies that the missing Coefficient is $6.35 \cdot .1421 = 0.902$. Finally, we compute $Var(Y) = (2174.2 + 862.4)/17 = 178.62$ and $sqrtVar(X_1) = Cor(Y,X) \cdot \sqrt{Var(Y)}/\hat{\beta}_1 = \sqrt{.716} \cdot \sqrt{178.62}/0.902 = 12.54$, so that $Var(X_1) = 157.2$.

3.7
The 97.5 percentile of a t distribution with 23 degrees of freedom is 2.07, so that the 95 % confidence interval for β_1 is $0.613 \pm (2.07)(0.1610)$ or 0.613 ± 0.333, while the 95 % confidence interval for β_2 is $-0.073 \pm (2.07)(0.1357)$ or -0.073 ± 0.281.

3.8
To reject the null hypothesis in (3.45), we must have evidence that β_1 and β_3 are unequal. But to reject the null in (3.49), we need evidence that β_1 and β_3 are unequal *or* that some of the other coefficients are nonzero. Accordingly, if both tests are equally sensitive to departures from equality of β_1 to β_3, then the probability of a Type I error must be greater for the second test. In other words, if we require the same Type I error probabilities for both tests, the first test will be more sensitive to departures from the equality of β_1 to β_3.

3.9
(a) The residual sum of squares is 1469.575 (on 29 degrees of freedom) for the restricted model, while the residual sum of squares is 1254.649 (on 27 degrees of freedom) for the unrestricted model. Using formula (3.28) with $p = 2$ and $k = 1$, we obtain $F = 2.31$, which is less than 3.35, the 95 percentile of an F distribution with 2, 27 degrees of freedom; the null hypothesis cannot be rejected at the .05 level.
(b) The residual sum of squares is 1410.694 (on 28 degrees of freedom) for the restricted model, while the residual sum of squares is 1224.616 (on 26 degrees of freedom) for the unrestricted model. Using formula (3.28) with $p = 3$ and $k = 2$, we obtain $F = 1.98$, which is less than 3.37, the 95 percentile of an F distribution with 2, 26 degrees of freedom; the null hypothesis cannot be rejected at the .05 level.

3.10
(a) We take Y as the response variable. The residual sum of squares for the model $Y = \beta_0 + \beta_1 X + \epsilon$ is 3303.281 (on 94 degrees of freedom). The residual sum of squares for the model $Y = \epsilon$ is 2586654 (on 96 degrees of freedom). Formula (3.28) with $p = 1$ and $k = 0$ yields $F = 36757$, which clearly indicates that the null hypothesis should be rejected (which we already knew without even performing a formal test; see part (i) of exercise 2.10).

3.11
(a) To see if men are paid more than equally qualified women, we should look

at the coefficient for the Sex variable in Model 1. Since Sex was coded as 1 for male employees and 0 for female employees, a positive coefficient would give evidence that men are paid more. Indeed, the (estimated) coefficient is positive, although it is not statistically significant at any reasonable significance level.
(b) To see if men are less qualified than equally paid women, we should look at the coefficient for the Sex variable in Model 2. Since Sex was coded as 1 for male employees and 0 for female employees, a negative coefficient would give evidence that men are less qualified than equally paid women. However, the (estimated) coefficient is *positive* and has a fairly small p-value.
(c) Questions (a) and (b) are essentially two different ways of asking whether the company is discriminating against women. The above results seem inconsistent insofar as one suggests that perhaps there is discrimination while the other suggests that there is not.
(d) A defense lawyer would prefer to focus attention on Model 2 because it suggests that there is no discrimination against women.

3.12
(a) When we use an F-test to assess the overall fit of a regression, the null hypothesis is that the true values of all of the regression coefficients are zero; the alternative hypothesis is that the true values of some or all of the regression coefficients are nonzero. The formula for the test statistic is given by equation (3.34), and the value of the test statistic for this particular model is given in Table 3.14 as 22.98. The critical value at the 5% significance level is approximately 2.49, so we reject the null hypothesis.
(b) We may test the null hypothesis that the true value of the Experience coefficient is zero against the one-sided alternative hypothesis that the true value of the Experience coefficient is greater than zero. The formula for the relevant test statistic is given by equation (3.23), and the value for this particular model appears in Table 3.14 as 2.16. The critical value at the 5% significance level is approximately 1.66, so we may reject the null hypothesis.
(c) The forecasted salary would be $3526.4 + 1 \cdot 722.5 + 12 \cdot 90.02 + 10 \cdot 1.2690 + 15 \cdot 23.406 = 5692.92$.
(d) The value of the forecasted salary is the same as in part (c), but the standard error is smaller for the average than for an individual.
(e) The forecasted salary would be $3526.4 + 0 \cdot 722.5 + 12 \cdot 90.02 + 10 \cdot 1.2690 + 15 \cdot 23.405 = 4970.42$.

3.13
We use formula (3.28) to test the null hypothesis that the reduced model involving only Education is adequate; note that $p = 4$ and $k = 2$. We get $F = \frac{(38460756 - 22657938)/3}{22657938/88} = 20.46$, which exceeds the critical value of 2.71. We reject the null hypothesis and conclude that some or all of the other predictors are needed.

3.14

The complete model is given by $Sales = \beta_0 + \beta_1 \cdot Age + \beta_2 \cdot HS + \beta_3 \cdot Income + \beta_4 \cdot Black + \beta_5 \cdot Female + \beta_6 \cdot Price$.
(a) The null hypothesis is that $\beta_5 = 0$; the alternative hypothesis is that $\beta_5 \neq 0$. The critical values (level .05) for the test statistic given in formula (3.43) are approximately ± 2.02. The actual value for the test statistic is $-1.0529/5.5610 = -0.1893$; therefore, we cannot reject the null hypothesis.
(b) The null hypothesis is that $\beta_5 = \beta_2 = 0$; the alternative hypothesis is that at least one of the two coefficients is nonzero. The relevant test statistic is given by formula (3.28), in which RM is a model that excludes the Female and HS variables; the critical value is approximately 3.21. The actual value for the test statistic is $\frac{(34959.77-34925.97)/2}{(34925.97)/44} = 0.0213$; therefore, we cannot reject the null hypothesis.
(c) If we use the complete model, the 95% confidence interval for β_3 is given by the formula $\hat{\beta}_3 \pm t_{(44,.025)} \cdot s.e.(\hat{\beta}_3)$ and takes the numerical values of 0.0189 ± 0.0206; in particular, 0 is inside the confidence interval!
(d) If the Income variable is excluded from the model, the R^2 value is 0.2678, indicating that approximately 27 % of the variation in Sales can be accounted for by the remaining five predictors.
(e) If only Price, Age, and Income are included as predictors, the R^2 value is 0.3032, indicating that approximately 30 % of the variation in Sales can be accounted for by these three predictors.
(f) If only Income is included as a predictor, the R^2 value is 0.1063, indicating that a little more than 10 % of the variation in Sales can be accounted for by Income.

3.15
(a) We could adapt the F-test in formula (3.28) for this situation. Specifically, we would have $k = 0$ and (since $\hat{Y} \equiv 0$) $SSE(RM) = \sum_{i=1}^{n} (y_i)^2$. We would then compare the test statistic to the upper α quantile of an F distribution with (p + 1) and (n - (p + 1)) degrees of freedom.
(b) Let X_1 take values $-5, -4, \cdots, 4, 5$, and let $Y := X_1^2 - 9$. Then $SSE(RM) = 869$ and $SSE(FM) = 858$. The test statistic equals $\frac{(869-858)/2}{858/9} = 0.0577$, which is much less than the (level .05) critical value of 4.26. Therefore we do not reject the null hypothesis.
(c) We are not appreciably better off making predictions from the regression model than we are if we make the constant prediction $\hat{Y} \equiv 0$.
(d) One could redefine SST to be $\sum_{i=1}^{n} (y_i)^2$, then use the formula $R^2 = 1 - SSE/SST$.

4 Solutions for Exercises in Chapter 4

4.1
(a) Let Y denote the current milk production, and let X_1 through X_6 denote

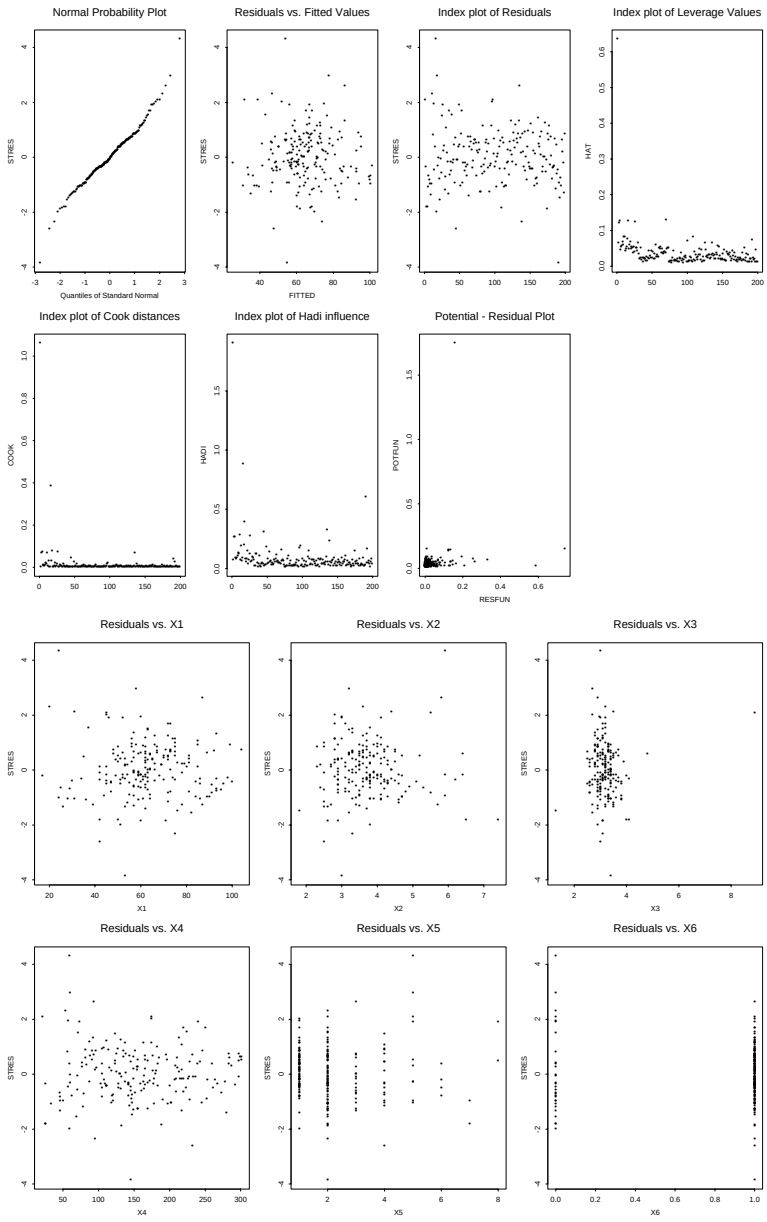
the predictor variables identified in Table 1.1. Below are three figures, each containing a number of diagnostic plots relevant for checking the assumptions underlying the regression of Y on $X1$ through $X6$.

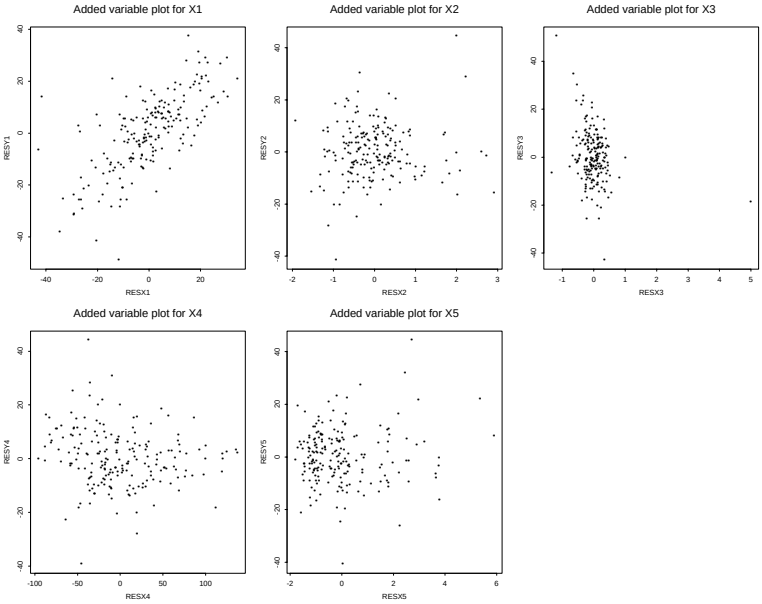
The first figure contains a normal probability plot of the internally studentized residuals; a plot of the internally studentized residuals against the fitted values from the regression; index plots of the internally studentized residuals, leverage values, Cook's distance measure, and Hadi's influence measure; and a potential-residual plot. Observation 1 has extremely high values for the Cook and Hadi influence measures, while observation 16 has moderately high values for these measures. The leverage plot and potential-residual plot indicate that observation 1 has a very high leverage; observation 16 is a regression outlier, and observation 190 is a regression outlier with sufficiently low leverage that it is not a point of great influence.

The second figure contains plots of the internally studentized residuals against each of the predictors; these plots are unremarkable.

The third figure contains the added variable plots for each quantitative predictor. Here we see something very interesting: The impression we have of the relationship between Y and X_3 (adjusting for the other predictor variables) depends very much on whether observation 1 is included.

In practice, we would question observation 1, and perhaps observation 16 as well. Specifically, we are suspicious that there has been an error in measurement or transcription for observation 1 because its protein value is nearly twice as high as that of any other in the data set.

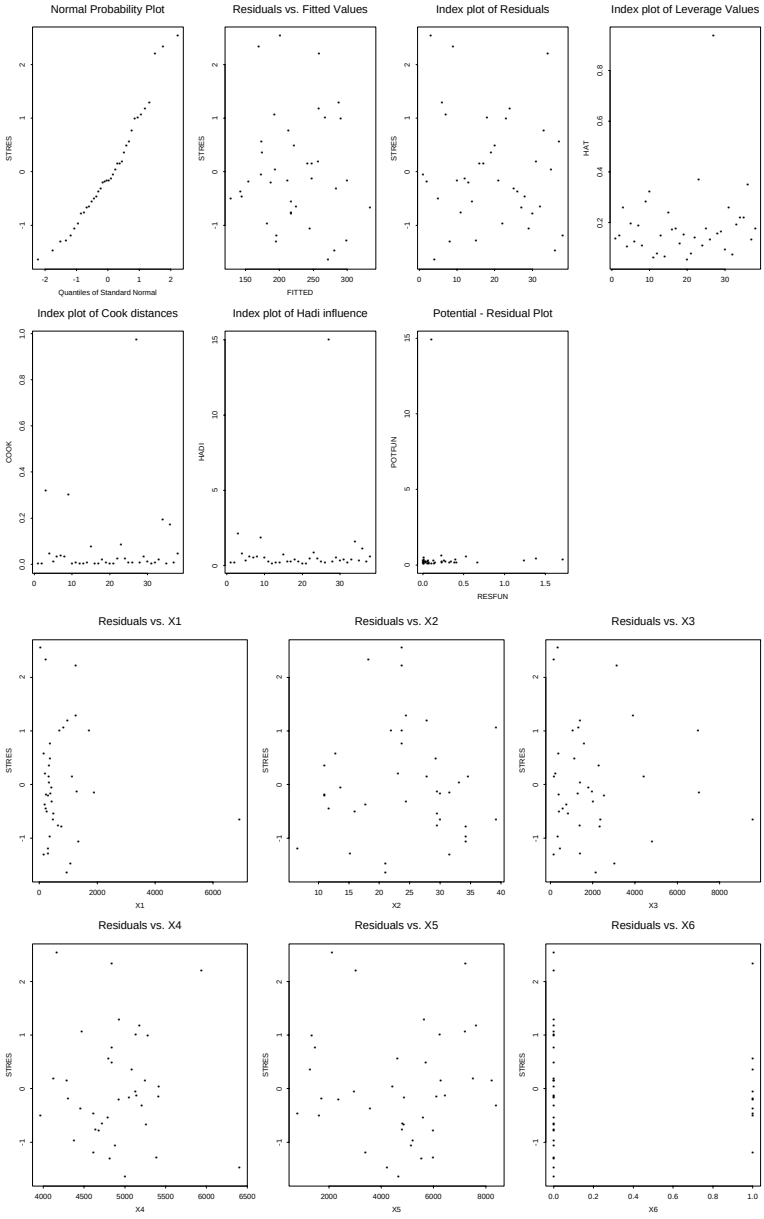

Normal Probability Plot
STRES
Quantiles of Standard Normal
Residuals vs. Fitted Values
STRES
FITTED
Index plot of Residuals
STRES
Index plot of Leverage Values
HAT
Index plot of Cook distances
COOK
Index plot of Hadi influence
HADI
Potential - Residual Plot
POTRUN
RESFUN
Residuals vs. X1
STRES
X1
Residuals vs. X2
STRES
X2
Residuals vs. X3
STRES
X3
Residuals vs. X4
STRES
X4
Residuals vs. X5
STRES
X5
Residuals vs. X6
STRES
X6

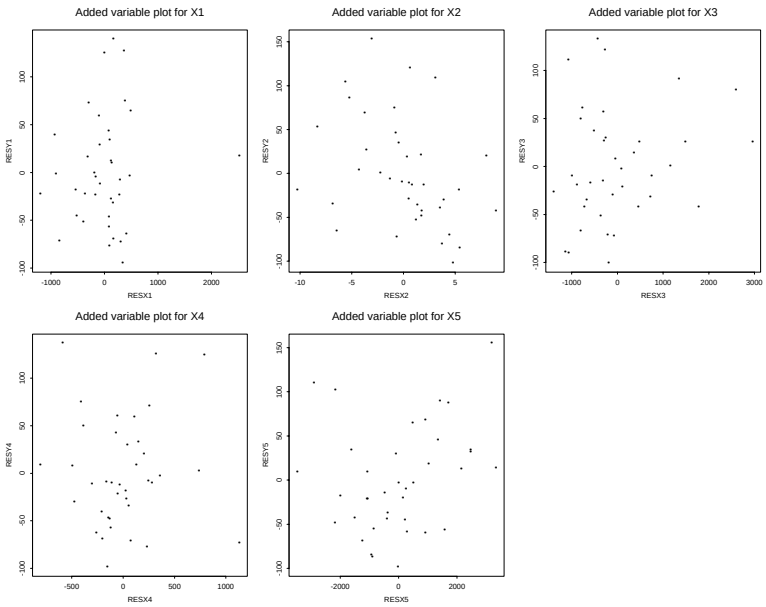


(b) Let Y denote the cost of living, and let X_1 through X_6 denote the other variables indicated in Table 1.2. For computational stability, we will take X_3 to be the population in thousands.

The three figures below are analogous to the ones produced for part (a).

We see that observation 27 (corresponding to New York) has high influence; it is a leverage point, mainly because the population density of New York is so much greater than that of the other cities. Observations 3, 9, and 34 are regression outliers but do not appear to be influential.

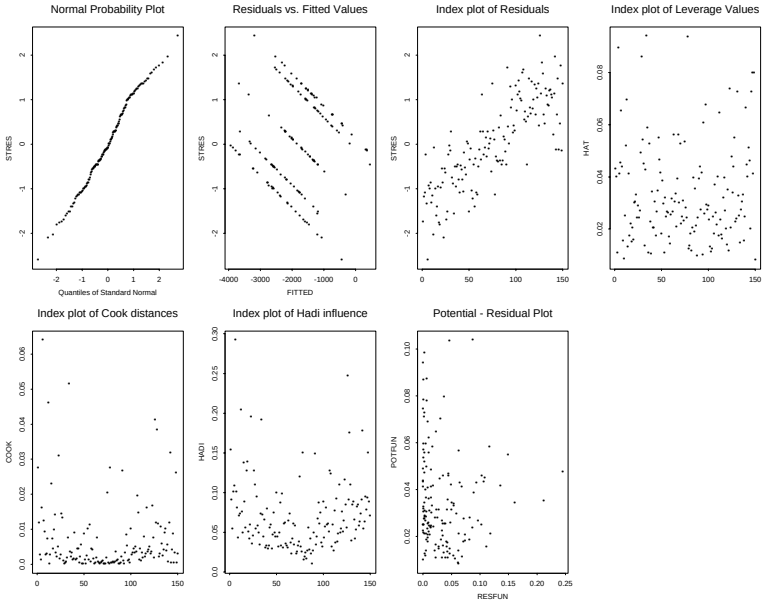

Normal Probability Plot
STRES
Quantiles of Standard Normal
Residuals vs. Fitted Values
STRES
FITTED
Index plot of Residuals
STRES
Index plot of Leverage Values
HAT
Index plot of Cook distances
COOK
Index plot of Hadi influence
HADI
Potential - Residual Plot
POTRUN
RESFUN
Residuals vs. X1
STRES
X1
Residuals vs. X2
STRES
X2
Residuals vs. X3
STRES
X3
Residuals vs. X4
STRES
X4
Residuals vs. X5
STRES
X5
Residuals vs. X6
STRES
X6



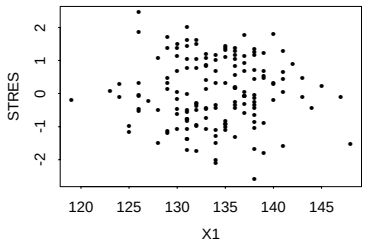
(c) Let Y denote the year, and let X_1 through X_4 denote the other variables indicated in Table 1.4.

We use the same plots as in parts (a) and (b).

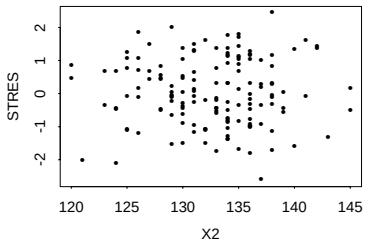
The dramatic positive pattern in the index plot of residuals indicates that the older skulls tend to have their ages underestimated by the model, while the newer skulls tend to have their ages overestimated. The pattern of lines in the residuals versus fitted values plot arises because there are only five values of Y in the data set; moreover, because there are only five values of Y, we know that equation (4.1), which postulates that Y is a linear combination of the predictor variables plus a normally distributed error term, is not an accurate representation of the structure of the data.



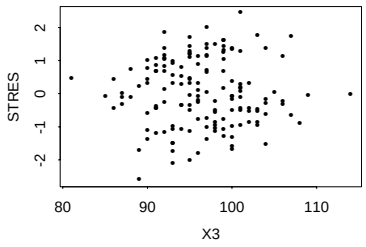
Residuals vs. X1



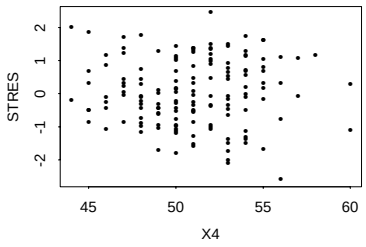
Residuals vs. X2

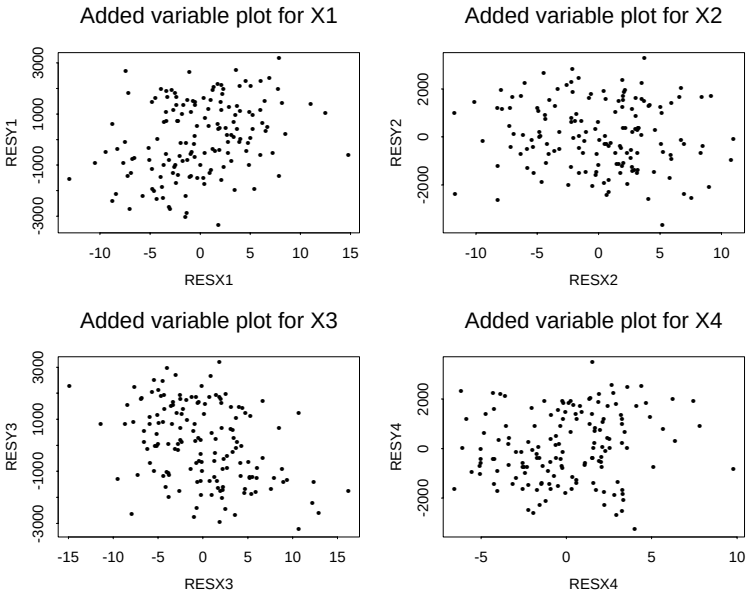


Residuals vs. X3



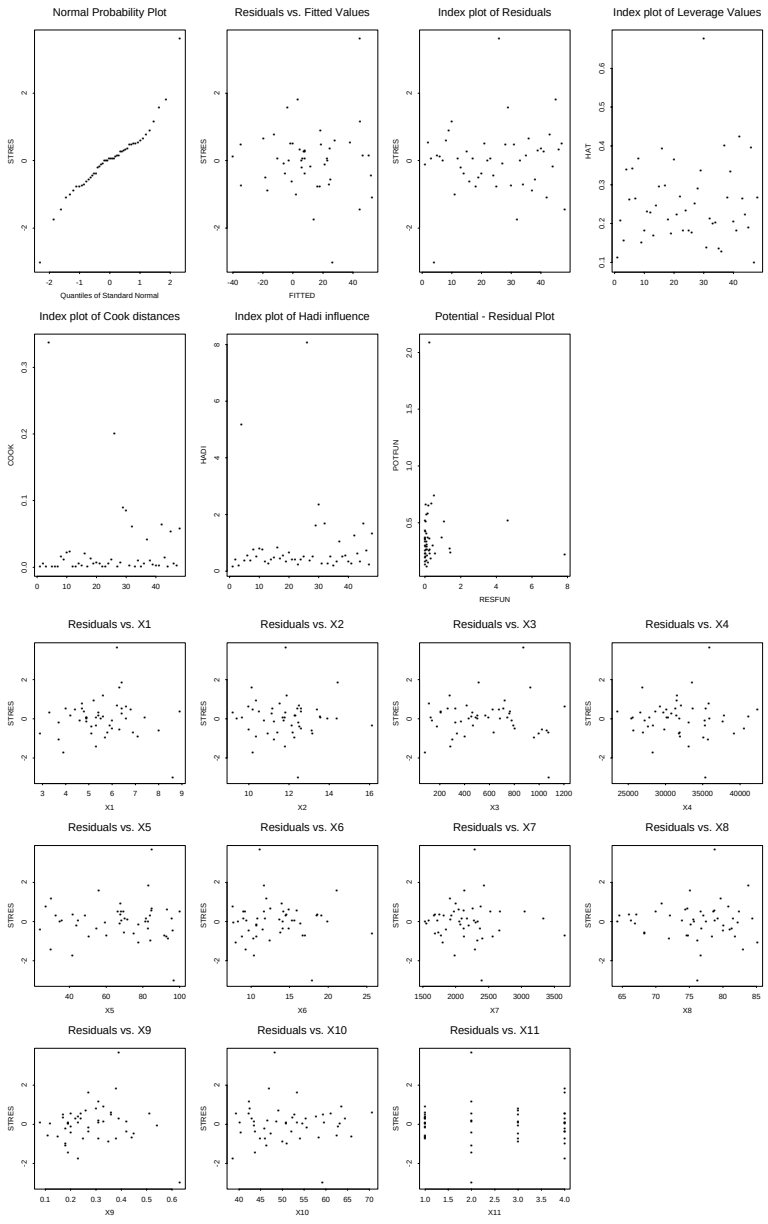
Residuals vs. X4

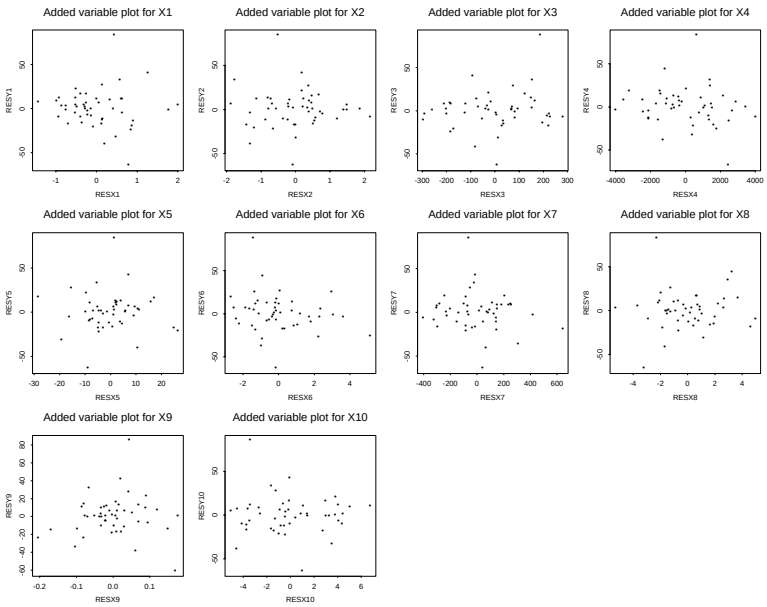




(d) Let Y denote the net domestic immigration, and let X_1 through X_{11} denote the other variables indicated in Table 1.5. For X_{11} we will code the regions as follows: South = 1, West = 2, Northeast = 3, Midwest = 4.

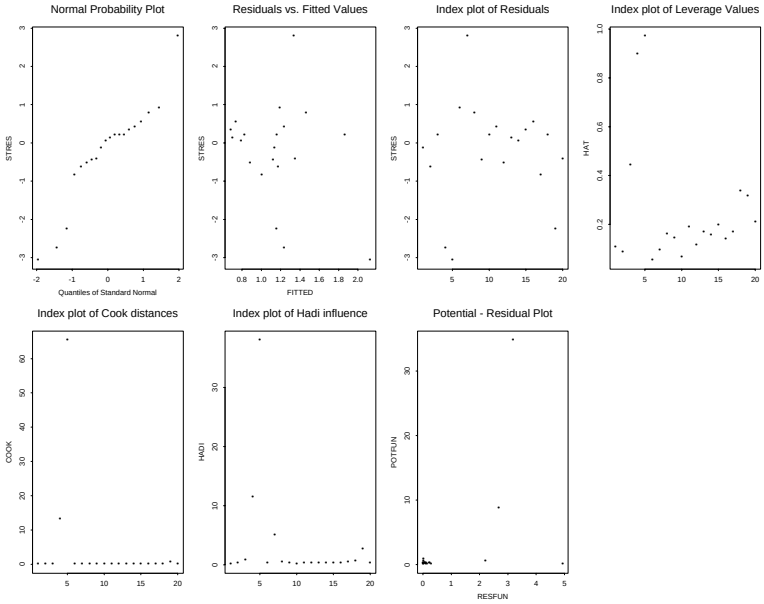
Observations 4 (California) and 26 (Nevada) are regression outliers, and are influential; observation 30 (New York) has high leverage but does not seem to be influential. Homogeneity of the error variance across different values of the predcitors (especially X_8 and X_{10}) is questionable. The impressions we have about the relationship between Y and some of the predictors (especially X_{10}) are affected by observations 4 and 26.

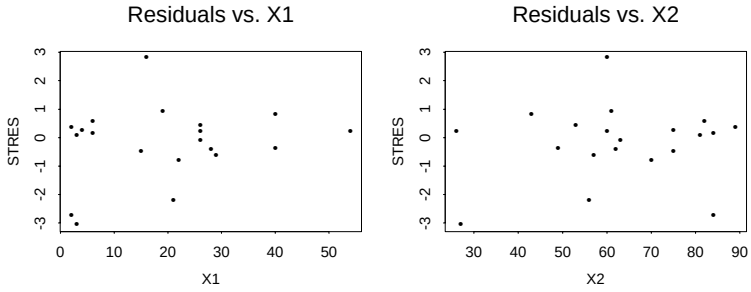

Normal Probability Plot
STRES
Quantiles of Standard Normal
Residuals vs. Fitted Values
STRES
FITTED
Index plot of Residuals
STRES
Index plot of Leverage Values
HAT
Index plot of Cook distances
COOK
Index plot of Hadi influence
HADI
Potential - Residual Plot
POTRUN
RESFUN
Residuals vs. X1
STRES
X1
Residuals vs. X2
STRES
X2
Residuals vs. X3
STRES
X3
Residuals vs. X4
STRES
X4
Residuals vs. X5
STRES
X5
Residuals vs. X6
STRES
X6
Residuals vs. X7
STRES
X7
Residuals vs. X8
STRES
X8
Residuals vs. X9
STRES
X9
Residuals vs. X10
STRES
X10
Residuals vs. X11
STRES
X11

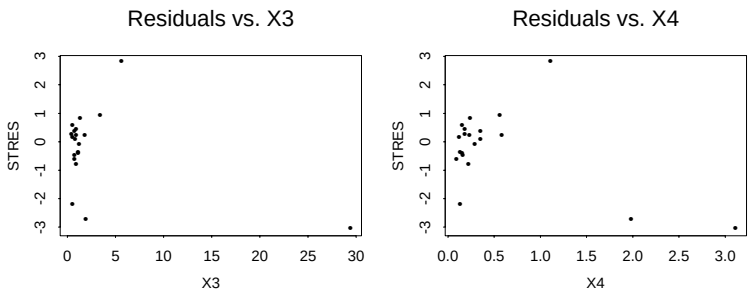


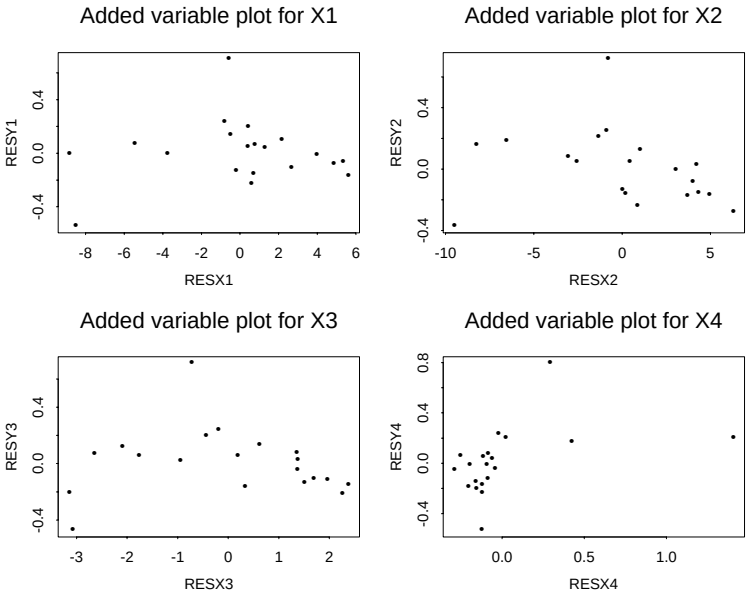
(e) Let Y denote the mean nitrogen concentration, and let X_1 through X_4 be as indicated in Table 1.8.

Observation 5 has extremely high values of the Cook and Hadi influence measures, and observation 4 has moderately high values. Observations 4, 5, 7, and 19 are all regression outliers. The latter two have extremely low leverages and so are not as infuential as the former two; observation 5 has a very large leverage value because of its extreme X_3 value, and observation 4 has a moderate leverage value. Homogeneity of the error variance across X_1 is questionable. Observation 4, in particular, affects our impression of how Y and X_4 are related. Since X_1 through X_4 are percentages that must sum to 100 or less, we may have a collinearity problem.


Normal Probability Plot
STRES
Quantiles of Standard Normal
Residuals vs. Fitted Values
STRES
FITTED
Index plot of Residuals
STRES
Index plot of Leverage Values
HAT
Index plot of Cook distances
COOK
Index plot of Hadi influence
HADI
Potential - Residual Plot
POTRUN
RESFUN


Residuals vs. X1
STRES
X1
Residuals vs. X2
STRES
X2


Residuals vs. X3
STRES
X3
Residuals vs. X4
STRES
X4



4.2
Many different data sets could be examined. For part (a), diagnostic plots of the kind used in exercise 4.1 should be employed to check whether the usual assumptions are valid. If the assumptions seem valid, then, for part (b), the question of interest can be answered straightforwardly based on the regression results.

4.3
Let Y denote Minutes, and let X_1 denote the number of Units.
(a) The regression equation is $\hat{Y} = 37.21 + 9.97X_1$.
(b) A plot of the internally studentized residuals against the fitted values (or against X_1) reveals that the relationship between Y and X_1 is not linear; a quadratic model ($Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon$) would be more appropriate.

4.4
Observation 11 is a point of high leverage that is not a regression outlier; observation 18 is a regression outlier that has a low leverage value; observation 7 is a moderate regression outlier with a moderately high leverage value. We would expect observation 7 to have the most influence on the least squares fit.

4.5
(a) One could use scatterplots of the internally studentized residuals against the predictors and/or residual plus component plots. An assumption of linearity would be invalid if we had a picture like Figure 4.4(a); if the assumption were valid, we would expect to get a picture without any such curvature.

(b) If the data are presented in some kind of chronological order, one could use an index plot of the standardized residuals to test the assumption of independence. Such an assumption would be invalid if there were a pattern indicating correlated residuals (e.g., something like a sine curve); if the assumption were valid, no such pattern should be discernible.
(c) The internally studentized residuals could be plotted against the fitted values and/or against the predictors. An assumption of constant variance would be invalid if we had a picture like Figure 4.4(b); if the assumption were valid, the spread of the residuals would not increase based on the fitted values or predictors.
(d) If the errors are normally distributed, then this assumption is equivalent to the one in part (b) and may be tested in the same way.
(e) One could use a normal probability plot of the internally studentized residuals to check the normality assumption. If the pattern formed by the points is close to a straight line, then the normality assumption is tenable. If the pattern formed by the points has curvature (e.g., looks like a cubic function), then the normality assumption is not reasonable. If a couple points at the corners of the plot stand far out from the line formed by the majority of the points, then these couple points are regression outliers; nonnormality may or may not be an issue.
(f) Index plots of Cook's distance and Hadi's influence measures could be used to detect influential points. In either of these pictures, points standing out from the others would be flagged as being influential on the least squares results. If the assumption of equally influential observations were valid, none of the points should stand out from the others in these plots.

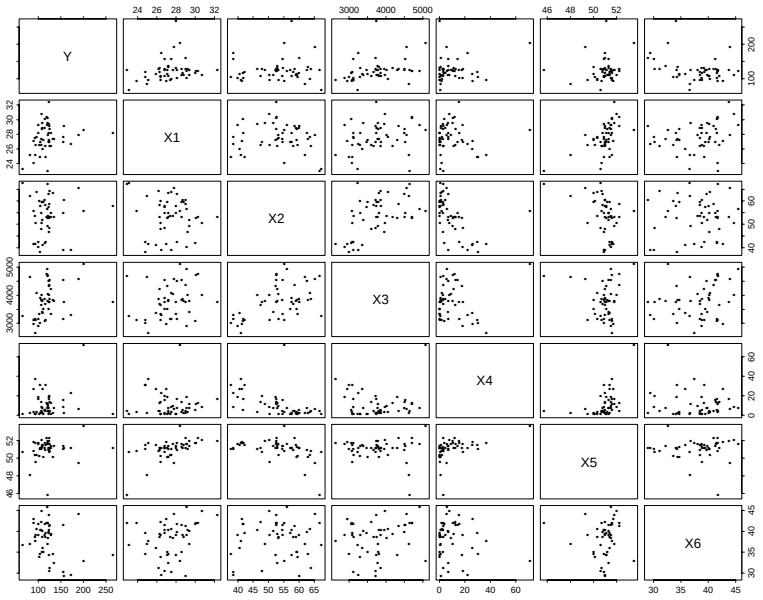
4.6
1. The scatter plot of Y versus a predictor would be useful to check the assumption that Y is linearly related to the predictor. Figure 4.1(a) depicts a situation where the assumption seems reasonable, while Figure 4.1(b) shows a situation where it is not.
2. The scatter plot matrix of the predictor variables would be useful to check the assumption that the predictors are linearly independent. A collinearity problem would be indicated if one or more of the plots had a nearly straight line pattern. In the absence of such an indication or of knowledge suggesting that the assumption is unreasonable (e.g., that percentages must sum to 100), the assumption of linear independence may provisionally be adopted and then checked later, after fitting the model, using the methods presented in Chapter 9.
3. This plot may be used to check the assumption that the errors are normally distributed. If the pattern formed by the points is close to a straight line, then the normality assumption is tenable. If the pattern formed by the points has a perceptible curvature (e.g., looks like a cubic or cube root function), then the normality assumption is not reasonable. If a couple points at the corners of the plot stand far out from the line formed by the majority of the points, then these couple points are regression outliers; nonnormality may or may not be an

issue.
4. A plot of the residuals against the fitted values can be used to check the assumptions of linearity and homoscedasticity. Patterns like those in Figure 4.4 indicate violations of these assumptions. If the assumptions are correct, then there should be no discernible pattern in the plot of residuals against fitted values.
5. The potential-residual plot may be used to check the assumption that none of the observations is unduly influential on the least squares fit. Isolated points that stand out from the majority of the points, vertically and/or horizontally, may have undue influence on the least squares fit. An absence of such points in the potential-residual plot suggests that the assumption of equally influential observations is reasonable.
6. The index plot of Cook's distances may be used to check the assumption that none of the observations is unduly influential on the least squares fit. Points with unusually large Cook's distances are suspect.
7. The index plot of Hadi's influence measure is another tool for checking the assumption that none of the observations is unduly influential on the least squares fit. Points with unusually large values of Hadi's influence measure are suspect.

4.7
(a) A priori, we might expect sales to be negatively related to price and positively related to income, based on economic considerations. We would expect sales to be positively related to median age, since a higher median age suggests that the proportion of the population old enough to smoke legally is larger. We would expect sales to be negatively related to the percentage completing high school, as high school students are educated about the dangers of smoking. It is not clear what, if any, relationships we should expect sales to have with the two demographic variables.
(b) Let Y denote Sales, and let X_1 through X_6 denote the predictor variables. The correlation matrix of $Y, X_1, \cdots, X_6$ is:

$$\begin{pmatrix} 1.000 & 0.227 & 0.067 & 0.326 & 0.190 & 0.146 & -0.301 \\ 0.227 & 1.000 & -0.099 & 0.257 & -0.040 & 0.553 & 0.248 \\ 0.067 & -0.099 & 1.000 & 0.534 & -0.502 & -0.417 & 0.057 \\ 0.326 & 0.257 & 0.534 & 1.000 & 0.017 & -0.069 & 0.215 \\ 0.190 & -0.040 & -0.502 & 0.017 & 1.000 & 0.451 & -0.148 \\ 0.146 & 0.553 & -0.417 & -0.069 & 0.451 & 1.000 & 0.022 \\ -0.301 & 0.248 & 0.057 & 0.215 & -0.148 & 0.022 & 1.000 \end{pmatrix}$$

And here is the scatter plot matrix of these same variables:



(c) Upon inspection of the scatter plot matrix, we see that there are some very unusual observations that must be having large effects on the correlation coefficients. The unusual observations are number 30 (New Hampshire, whose Y value is more than 65 units greater than the second largest Y value), number 9 (District of Columbia, whose values of X_4 and X_5 are very high), and number 2 (Alaska, whose value of X_5 is very low). None of the correlation coefficients involving X_4 or X_5 should be taken at face value, since the scatter plot suggests that they are potentially misleading. For a dramatic example, note that the correlation between X_3 and X_4 is -0.292 when the District of Columbia is excluded!

(d) There does not seem a relationship, negative or otherwise, between Y (Sales) and X_2 (education), as speculated in part (a).

(e) The regression equation is $\hat{Y} = 103.3448 + 4.5205X_1 - 0.0616X_2 + 0.0189X_3 + 0.3575X_4 - 1.0529X_5 - 3.2549X_6$. Again, there does not seem to be a relationship between Y and X_2.

(f) The regression coefficients measure the relationships between Y and the various predictors, adjusting for the other predictors; the pairwise correlation coefficients do not adjust for relationships with other variables. The most noticeable difference we see here is that the regression coefficients for X_2 and X_5 are negative, while the pairwise correlations of Y with these variables were positive. [However, these two regression coefficients are not even close to being significantly different from zero.]

(g) New Hampshire doesn't fit in with the rest of the observations because its sales level is so much higher. [One might suspect that there is a transcription error, but actually this is not the case: in 1970, cigarettes were not taxed in New

Hampshire, so some people from neighboring states bought their cigarettes in New Hampshire.] Also, the District of Columbia and Alaska are quite different demographically from the 48 contiguous states. [In fact, cigarettes were not taxed in these locations either.] Thus, including New Hampshire, the District of Columbia, and Alaska in the construction of a regression model may yield results that do not accurately describe the phenomenon of cigarette sales in (most of) the 48 contiguous states; therefore, conclusions reached based on such a regression model are questionable.

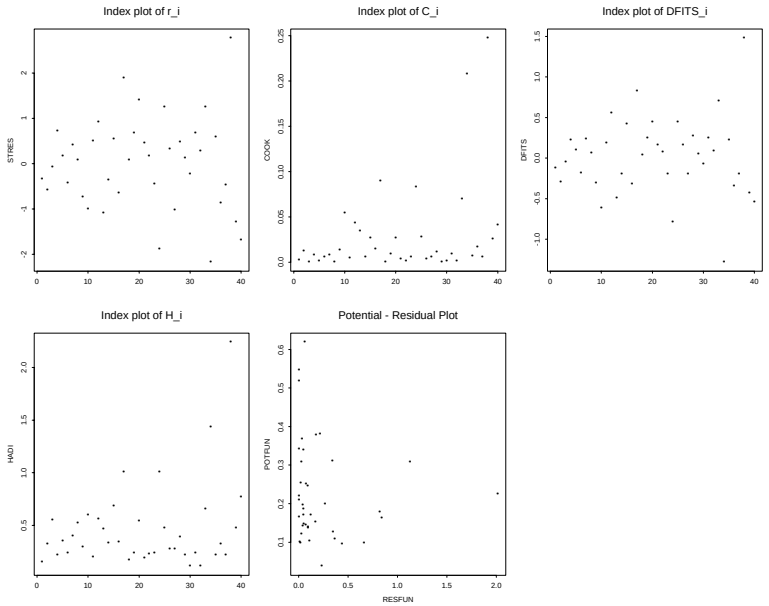
4.8
The full data set is needed to do this exercise.

4.9
We will prove statement (a) mathematically, and we will perform a numerical check on statement (b) with the unusual observation $i = 30$ removed.
(a) Take the expression for $\hat{\beta}_0$ near the bottom of page 55, multiply both sides by n, and rearrange to get $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p)$. The terms being summed on the right hand side are, by definition, the fitted values, so it follows that $\sum_{i=1}^{n} (y_i - \hat{y_i}) = 0$ (i.e., the sum of the residuals is zero).
(b) For the model constructed using all 51 observations, $\hat{\sigma}^2 = 793.772$, $p = 6$, and $r_i^2 = 23.945$. Thus, the right hand side of equation (4.26) is equal to $793.772[\frac{44-23.945}{43}] = 370.212$. On the other hand, when we remove observation 30, we also obtain 370.212 as the value of $\hat{\sigma}^2_{(i)}$.

4.10
Observation 1 is a leverage point. Observation 2 is a regression outlier with moderate leverage. Observation 2 has a heavy influence on the least squares results.

4.11
We will perform the requested verifications for observation 33 of the Scottish Hiles Races Data.
(a) Fitting Model (4.27), we obtain $\hat{Time} = -539.4829 + 373.0727 \cdot Distance + 0.6629 \cdot Climb$; fitting Model (4.28), we obtain $\hat{Time} = -498.3380 + 369.5481 \cdot Distance + 0.6436 \cdot Climb + 714.8072 \cdot U_{33}$. The t value for β_3 in Model (4.28) is 0.7335. The 33rd externally studentized residual in Model (4.27) may be computed from formula (4.15), given that $r_{33} = 0.7389$, $n = 35$, and $p = 2$; as anticipated, it comes out to be 0.7335.
(b) The residual sum of squares for Model (4.28) is 24386804, while the residual sum of squares for Model (4.27) is 24810082. Thus the F statistic is $\frac{(24810082-24386804)/1}{24386804/31} = 0.5381$, whose square root is 0.7335. [Also, see the second remark on page 69.]
(c) We obtain $\hat{Time} = -498.3380 + 369.5481 \cdot Distance + 0.6436 \cdot Climb$.
(d) Indeed, the estimates of β_0, β_1, and β_2 are the same whether we omit the 33rd observation or give it its own indicator variable.

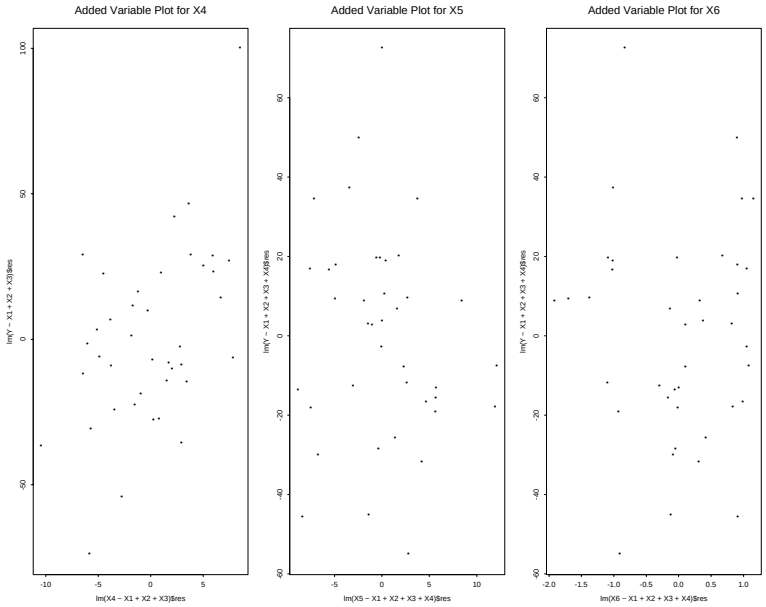
4.12
(a) Based on a plot of the internally studentized residuals versus fitted values (not shown), the assumption of homogeneous error variances is questionable. There are also some unusual points, described in part (d) below.
(b) To compute $DFITS_i$, use formulas (4.15) and (4.23). The other items may be obtained straightforwardly.
(c) The index plots of r_i, C_i, $DFITS_i$, and H_i, as well as the Potential-Residual plot, are given in the figure directly below.



(d) Observations 3, 8, and 15 have the highest leverages, although they do not really stand out from the other observations in this respect. Observation 38 is a regression outlier, and observation 34 is a mild regression outlier. The index plots of the Cook and Hadi influence measures suggest that observations 38 and 34 are influential.

4.13
(a) The added variable plot for X_4 (see below) suggests that, qualitatively, there is a positive linear relationship between Y and X_4 (adjusting for the first three predictors). The removal of observation 38 would change the slope of the line that would be fit to the points in the added variable plot, but the qualitative conclusion is unaffected. Based on the added variable plot and the associated t-test value of 3.35, we decide that X_4 should be added to the model.
(b) The added variable plot for X_5 does not suggest that there is a linear relationship between Y and X_5 (adjusting for the first four predictors); based on this, and the associated t-test value of -0.75, we decide not to add X_5 to our

model.
(c) The added variable plot for X_6 does not suggest that there is a linear relationship between Y and X_6 (adjusting for the first four predictors); based on this, and the associated t-test value of -0.18, we decide not to add X_6 to our model.



(d) Of the models we have considered, the best one is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$. One should, of course, perform some diagnostic checks before giving this model final approval. It turns out that, as with the model based on all six predictors variables, two of the points (observations 38 and 34) are influential.

4.14
We verify the assertions numerically using the data in Table 4.8.
(a) $\sum_{i=1}^{n} u_i v_i = 35089.34$ and $\sum_{i=1}^{n} v_i^2 = 17394.3$, so that their quotient is 2.0173. This is also the value of $\hat{\beta}_3$.
(b) $\hat{\sigma} = 31.632$, so that $\hat{\sigma}/\sqrt{\sum_{i=1}^{n} v_i^2} = 0.2398$. This is also the value of the standard error of $\hat{\beta}_3$.

5 Solutions for Exercises in Chapter 5

5.1
(a) Observation 11, being a regression outlier with fairly high leverage, is influential. Observation 19 is also influential, but to a lesser extent.

(b) The residual sum of squares is 40.32. If Z were dropped, the residual sum of squares would be 45.57. Thus, the F-value for testing $H_0 : \gamma = 0$ is $\frac{(45.57-40.32)/1}{(40.32)/17} = 2.21$. The critical value at the .05 level is 4.45, so we do not reject the null hypothesis.
(c) The t-value is 1.49. The critical values at the .05 level are ± 2.11, so we do not reject the null hypothesis.
(d) The F-value and the critical value for the F statistic are the squares of the t-value and the critical values for the t statistic. Thus, the tests we used in parts (b) and (c) are equivalent. [See also the second remark at the bottom of page 69.]

5.2
(a) Observation 11, being a regression outlier with high leverage, is influential.
(b) The residual sum of squares is 34.71. If the interaction variable were dropped, the residual sum of squares would be 45.57. Thus, the F-value for testing $H_0 : \delta = 0$ is $\frac{(45.57-34.71)/1}{(34.71)/17} = 5.32$, which is greater than the critical value ($\alpha = .05$) of 4.45. So, we would reject the null hypothesis.
(c) The t-value is 2.31, which is greater in magnitude than 2.11, so we would reject the null hypothesis.
(d) The F-value and the critical value for the F statistic are the squares of the t-value and the critical values for the t statistic. Thus, the tests we used in parts (b) and (c) are equivalent. [See also the second remark at the bottom of page 69.]

5.3
One might begin by considering four competing models. We use the notation provided on page 142, along with I_1, I_2, and I_3 to denote interactions between personal disposable income and the seasonal indicator variables.

(i) $S_t = \beta_0 + \beta_1 PDI_t + \varepsilon_t$
(ii) $S_t = \beta_0 + \beta_1 PDI_t + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + \varepsilon_t$
(iii) $S_t = \beta_0 + \beta_1 PDI_t + \delta_1 I_1 + \delta_2 I_2 + \delta_3 I_3 + \varepsilon_t$
(iv) $S_t = \beta_0 + \beta_1 PDI_t + \gamma_1 Z_1 + \gamma_2 Z_2 + \gamma_3 Z_3 + \delta_1 I_1 + \delta_2 I_2 + \delta_3 I_3 + \varepsilon_t$

The residual sums of squares from fitting these four models are 346.43, 47.24, 57.39, and 47.18, respectively. One may use F-tests to see that models (ii), (iii), and (iv) are substantially better than model (i), which does not take into account seasonality. On the other hand, model (iv) is very complicated and not any better than model (ii). Given these results, it would seem that model (ii) is the best choice of the four. But, a close inspection of the results from fitting model (ii) suggests that, really, only one indicator variable (representing the spring and summer seasons) is needed. Letting $Z := Z_2 + Z_3$, we consider the following, which we will call model (v):

(v) $S_t = \beta_0 + \beta_1 PDI_t + \gamma Z + \varepsilon_t$.

The residual sum of squares for model (v) is 47.86, so that this model is about as good as either model (ii) or (iv) and easier to interpret. The regression equation, $\hat{S}_t = 15.0045 + 0.1987PDI_t - 5.4643Z$, suggests that people prefer to buy ski equipment in the fall and winter, and that there is a positive relationship between expenditure on this leisure activity and disposable income.

Of course, one should perform diagnostic checks before accepting model (v) as a final product. In particular, since the data were collected over time, the assumption of independent errors needs to be verified. It turns out that this assumption is reasonable.

5.4

One might begin by looking at the model proposed on page 144: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 T_1 + \gamma_2 T_2 + \delta_1 T_1 \cdot X_1 + \delta_2 T_1 \cdot X_2 + \delta_3 T_1 \cdot X_3 + \alpha_1 T_2 \cdot X_1 + \alpha_2 T_2 \cdot X_2 + \alpha_3 T_2 \cdot X_3 + \varepsilon$. Including the terms with γ_1 and γ_2 allows us to accommodate the possibility that the intercept should vary by year; including the terms with δ_j and α_j allows us to accommodate the possibility that the coefficient for X_j should vary by year.

Noting that observations 11 and 12 from 1960 should have Y values of 84, we fit the model and obtain $\hat{Y} = -556.57 + 0.0724X_1 + 1.5521X_2 - 0.0043X_3 + 545.16T_1 + 267.39T_2 - 0.0275T_1 \cdot X_1 - 1.4858T_1 \cdot X_2 - 0.0247T_1 \cdot X_3 + 0.0085T_2 \cdot X_1 - 0.7336T_2 \cdot X_2 - 0.0995T_2 \cdot X_3$. Diagnostic plots indicate mild heteroscedasticity (the spread of the residuals increases somewhat as the fitted values increase) and that observation 149 is highly influential. Observation 149 corresponds to Alaska in the year 1975; the oil crisis of the mid-1970's affected Alaska's economy, so it makes sense to exclude observation 149. Refitting the model, we get the regression equation $\hat{Y} = -277.58 + 0.0483X_1 + 0.8869X_2 + 0.0668X_3 + 266.17T_1 - 11.60T_2 - 0.0034T_1 \cdot X_1 - 0.8207T_1 \cdot X_2 - 0.0957T_1 \cdot X_3 + 0.0326T_2 \cdot X_1 - 0.0685T_2 \cdot X_2 - 0.1706T_2 \cdot X_3$. Apart from mild heteroscedasticity, there do not seem to be any problems with this model.

However, we may ask whether all of the terms are really needed. The coefficients for T_2, $T_1 \cdot X_1$, and $T_2 \cdot X_2$ are (individually) not significant, with p-values greater than 0.80. However, all of the other coefficients have p-values less than 0.10. The F-value associated with eliminating the three terms named above is 0.44 (compared to the critical value of 2.67), so it is reasonable to eliminate these three terms. When we do, we get $\hat{Y} = -275.64 + 0.0510X_1 + 0.8292X_2 + 0.0724X_3 + 240.50T_1 - 0.7469T_1 \cdot X_2 - 0.1092T_1 \cdot X_3 + 0.0254T_2 \cdot X_1 - 0.1799T_2 \cdot X_3$. Each of the coefficients is statistically significant at the .05 level, and all but the coefficient for X_3 are significant at the .01 level. Apart from mild heteroscedasticity, the diagnostic plots for this reduced model look good. The reduced model tells us that the relationships between Y and the predictors did change over time.

5.5

(a) F_j equals 1 for observations in group j and 0 otherwise, for j = 1, 2, 3.
(b) We obtain $\hat{Y} = 29.80 + 6.80F_1 + 0.10F_2 + 5.10F_3$.
(c) The null hypothesis is that $\mu_1 = \mu_2 = \mu_3 = 0$. The regression output gives a value of 5.14 for the F-test, which is much greater than the level .05 critical value of 2.87. So, we reject the null hypothesis.
(d) The null hypothesis is that $\mu_1 = \mu_2 = \mu_3$. We employ the usual F-test to compare the full model against the reduced model $Y = \mu_0 + \mu \cdot (F_1 + F_2 + F_3) + \varepsilon$. The residual sum of squares from the reduced model is 1088.4. The F-value is $\frac{(1088.4 - 845.8)/2}{845.8/36} = 5.16$, which is greater than the level .05 critical value of 3.26. We reject the null hypothesis.
(e) From the model we fit in part (b), it seems that fertilizer 1 is the best. [Our answer to part (d) indicates that, indeed, there is some difference among the fertilizers.]

5.6
(a) Either weight or height could be taken as the response variable; either choice would permit us to study the relationship between these two variables.
(b) We first consider the model $Weight = \beta_0 + \beta_1 \cdot Height$; age is not used as a predictor for the reason given in part (d). The regression equation is $\hat{Weight} = -85.5 + 3.35 \cdot Height$. One finds that the standardized residuals for the women tend to be lower than for the men, suggesting that gender needs to be incorporated into our model; that is, a single equation is not adequate to describe the relationship between height and weight.
(c) Let us consider three models:

(i) $Weight = \beta_0 + \beta_1 \cdot Height + \gamma \cdot Sex + \delta \cdot (Height \cdot Sex) + \varepsilon$
(ii) $Weight = \beta_0 + \beta_1 \cdot Height + \gamma \cdot Sex + \varepsilon$
(iii) $Weight = \beta_0 + \beta_1 \cdot Height + \delta \cdot (Height \cdot Sex) + \varepsilon$

Using F-tests, we find that models (ii) and (iii) are about as good as model (i). In other words, it is not necessary to adjust both the slope and the intercept according to gender; adjusting one or the other seems sufficient. If model (ii) is chosen, the regression equation relating weight to height is $\hat{Weight} = 42.93 + 1.62 \cdot Height - 26.07 \cdot Sex$. For model (ii), the standardized residuals do not tend to be lower for women; however, the standardized residuals are less spread out for the women than for the men.
(d) All of the people in the sample (except two) are almost the same age; that is, we don't have enough variation in age in our sample for age to be useful as a predictor. Moreover, if we did try to include age in our model, the two observations corresponding to the slightly older students would become leverage points and might exert undue influence on the least squares results.

5.7
(a) The regression equation corresponding to the initial model (5.11) is $\hat{V} = 0.5112 - 0.0201 \cdot I + 0.0546 \cdot D + 0.0134 \cdot W + 0.0097 \cdot (G \cdot I) - 0.0007 \cdot P - 0.0052 \cdot N$.

Only the coefficients for D and $(G \cdot I)$ are statistically significant at the .05 level. In particular, we do not seem to need I if D is already in the model because I and D are very similar; including either one in the model allows us to take into account the advantage to the Democratic candidate (resp., Republican candidate) of having a Democratic incumbent (resp., Republican incumbent) in office.
(b) Yes, based on the very low p-value (0.0001) as well as the knowledge that a strong economy will usually help a Democratic candidate (resp., Republican candidate) if a Democratic incumbent (resp., Republican incumbent) is in office. However, it may make more sense to have $G \cdot D$ in the model than $G \cdot I$ if we discard I.
(c) There are at least three things that we might try to change about model (5.11). First, we do not need to include both D and I, as stated in part (a). Second, while there may be truth to the idea that stability of leadership is desired during and after a major war, we do not seem to have enough data (three observations for which $W = 1$) to quantify this phenomenon in our model; thus, we might be better off dropping W. Third, since good economic performance should not help a Democratic candidate if the incumbent is a Republican, it does not make sense to have terms in the model involving P and N without including the corresponding interactions with D or with I.

The three considerations above lead us to consider the following alternative model: $V = \beta_0 + \beta_1 \cdot D + \beta_2 \cdot G + \beta_3 \cdot (G \cdot D) + \beta_4 \cdot P + \beta_5 \cdot (P \cdot D) + \beta_6 \cdot N + \beta_7 \cdot (N \cdot D) + \varepsilon$. [One could also use I instead of D; we chose D over I because its simple correlation to V is somewhat stronger.] Fitting the alternative model, we obtain the regression equation $\hat{V} = 0.5251 - 0.0070 \cdot D + 0.0028 \cdot G + 0.0102 \cdot (G \cdot D) - 0.0022 \cdot P - 0.0023 \cdot (P \cdot D) - 0.0088 \cdot N + 0.0081 \cdot (N \cdot D)$. Only the $G \cdot D$, N, and $N \cdot D$ terms are statistically significant at the .05 level. If we reduce the alternative model to $V = \beta_0 + \beta_1 \cdot G + \beta_2 \cdot (G \cdot D) + \beta_3 \cdot N + \beta_4 \cdot (N \cdot D)$, we obtain the regression equation $\hat{V} = 0.5109 + 0.0041 \cdot G + 0.0114 \cdot (G \cdot D) - 0.0083 \cdot N + 0.0053 \cdot (N \cdot D)$; each of the coefficients in the reduced model is statistically significant, and an F-test indicates that the reduction is permissible.

For this reduced model, observations 2, 5, and 6 are influential; the corresponding elections were held in years when the economy was much more volatile than it has been in modern times. Perhaps the older observations, some of which took place during world wars and turbulent economic times, should not be receiving the same weight as the more recent observations.

In any case, the terms in the reduced model may be interpreted as follows: The terms involving $G \cdot D$ and $N \cdot D$ are intended to account for the fact that a candidate will benefit from good economic leadership by his party. On the other hand, the *degree* to which a candidate benefits from such leadership may be different for Republicans and Democrats; the terms involving G and N allow such a difference to be accounted for.

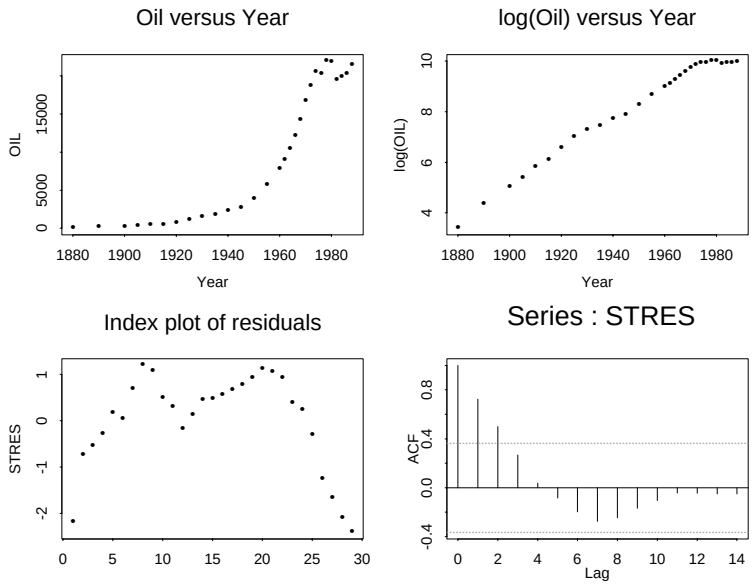
6 Solutions for Exercises in Chapter 6

6.1
(a) The regression equation is $\hat{R} = 7.6041 + 0.3527P$. The fit is poor, as evidenced by the low R^2 value of 0.1263 and the unsatisfactory diagnostic plots which show that observation 23 (True Story) is both an extreme leverage point and a severe regression outlier. It is difficult to say much else because observation 23 has such great influence. If we delete observation 23 and refit the linear model, we obtain $\hat{R} = -2.5726 + 1.5617P$; the R^2 rises to 0.6458, but the plot of standardized residuals versus fitted values plainly shows that there is a heteroscedasticity problem.
(b) A scatter plot of $\log(R)$ versus $\log(P)$ shows (apart from a few outliers) a basically linear relationship between $\log(R)$ and $\log(P)$ without an obvious heteroscedasticity problem. Fitting a regression model accordingly, we obtain $\widehat{\log(R)} = -0.0847 + 1.0259\log(P)$; the R^2 value is 0.5162, and the only problem suggested by the diagnostic plots is that observations 15 and 22 are leverage points and regression outliers.
(c) If we remove observations 15 and 22, we get $\widehat{\log(R)} = -1.2134 + 1.5278\log(P)$; R^2 is 0.7910, and the diagnostic plots are satisfactory.

6.2
(a) The table from the web site organizes the data into three columns.
(b) We obtain $\hat{W} = -9.0566 - 1.1055V + 1.4187T$, and, indeed, the residual plots show that the linear model is inadequate. Specifically, the linear model seems to fit the data poorly when V and T are simultaneously high or simultaneously low. [These points stand out in the index plot of standardized residuals as well as in the index plots of the Cook and Hadi influence measures.]
(c) The plot of standardized residuals against V has strong curvature, indicating that the relationship between W and V is not linear.
(d) The plot of standardized residuals against T does not show curvature, but it does suggest that the error variances are not homogeneous across the range of T.
(e) We obtain $\hat{W} = 48.9094 + 1.6506V + 1.4187T - 26.6231\sqrt{V}$. The diagnostic plots still do not look good; points with low values of V do not seem to be handled well even with the extra term in the model. After replacing the 0.817 by 0.0817, we find that formula (6.19) produces values that are quite close to the W values given in Table 6.16. We see that the essential difference between (6.18) and (6.19) is that the latter incorporates interaction between T and V.
(f) The agreement between the values produced by (6.19) and the W values in Table 6.16 is weakest for the points with low temperatures; if one knew the form of (6.19) but not the coefficients, one might wish to downweight some of the points with low temperatures in order to get precise estimates of the coefficients. However, since we have the coefficients and since (6.19) is an operational definition rather than a working model, there is no practical way in which (6.19) can be improved upon.

6.3
(a) We obtain the regression equation $\hat{Y} = 0.0438 - 0.0821 \cdot I + 0.2222 \cdot D + 0.0503 \cdot W + 0.0394 \cdot (G \cdot I) - 0.0030 \cdot P - 0.0207 \cdot N$.
(b) The two models are not greatly different because Y and V are almost linearly related over the range of V observed within the data set. However, the diagnostic plots for the model with Y as the response do appear slightly better in the sense that the observations' influences seem more comparable.
(c) $Y = \log(V/(1-V)) \iff \exp(Y) = V/(1-V) \iff \exp(Y)/(1+\exp(Y)) = V$; thus, $f(x) = \exp(x)/(1+\exp(x))$.

6.4
(a) The pattern in the scatter plot of OIL versus Year (first panel in the figure below) is obviously nonlinear.
(b) The pattern in the scatter plot of log(OIL) versus Year (second panel) is basically linear until 1972.
(c) The regression equation is $\log(\hat{O}IL) = -111.8753 + 0.0616 \cdot Year$. The R^2 value is 0.9834, but the model does not fit the data well at the beginning and end of the time span under consideration.
(d) The index plot of the standardized residuals (third panel) shows a sinusoidal pattern, suggesting that the assumption of independent errors is unreasonable. The plot of the autocorrelation function (fourth panel) shows that, indeed, residuals for successive observations are very strongly correlated.



6.5

(a) No, the data do not follow a linear time trend. The (absolute) decreases in price are much larger in the earlier years.
(b) The model $P_t = P_0 \exp(\beta t)\epsilon_t'$ (multiplicative error) is equivalent to the model $\log(P_t) = \alpha + \beta t + \epsilon_t$. For the latter, the regression equation comes out as $\widehat{\log(P_t)} = 3.5069 - 0.5605t$. Although R^2 is 0.9791, this model does not fit well at the beginning of the time span, and there is a problem with correlated errors.
(c) Let I denote the requested indicator variable. The resulting regression equation is $\log(P_t) = 2.7336 - 0.2679t + 1.4769I - 0.3776I \cdot t$. The R^2 value is 0.9972, and adding the indicator variable seems to have remedied the problem of correlated errors. The indicator and interaction variables allow us to adjust the intercept and slope when we pass from the earlier years (1988-1991) to the later years (1992-1998). Specifically, the intercept and slope are estimated to be 2.7336 and -0.2679 for the earlier years but are estimated to be 4.2105 and -0.6455 for the later years.

7 Solutions for Exercises in Chapter 7

7.1
Note that observations 11 and 12 (corresponding to Indiana and Illinois) should have Y values of 84. Fitting the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$, we get the regression equation $\hat{Y} = -11.40 + 0.0449X_1 + 0.0662X_2 - 0.0290X_3$. Observations 30 and 32 (Kentucky and Alabama) are flagged as being somewhat influential, but there is not a compelling reason to remove these observations. There are indications of heteroscedasticity across fitted values and across regions. But when we look at the plot of standardized residuals by region, the most striking problem is that all of the standardized residuals are negative for states in the first region! This suggests that there is a regional effect on the expected value of Y. So, before we try to do anything about the heteroscedasticity, we should refine our model by including regional indicator variables.

When we include regional indicator variables, we obtain $\hat{Y} = 41.42 + 0.0382X_1 + 0.0026X_2 - 0.0237X_3 - 33.08I_1 - 17.14I_2 - 22.26I_3$. The plot of standardized residuals against fitted values for this refined model shows that the residuals tend to be more spread out for larger fitted values than for smaller ones; however, the severity of the heteroscedasticity is difficult to assess, as observations 30 and 33 have moderately large standardized residuals to go along with their rather small fitted values. Looking at a plot of the standardized residuals by region, we do find that the residuals' spreads are different in each of the four regions. We compute that $\hat{\sigma}_1^2 = 53.45$, $\hat{\sigma}_2^2 = 133.30$, $\hat{\sigma}_3^2 = 49.20$, and $\hat{\sigma}_4^2 = 184.59$; each $\hat{\sigma}_j^2$ is an estimate of regional error variance, obtained by summing the squared residuals for the region and dividing by the number of states in the region.

To do weighted least squares, we define $c_j^2 := \hat{\sigma}_j^2 / (\sum_{i=1}^n e_i^2/n)$, which yields

$c_1^2 = 0.507$, $c_2^2 = 1.265$, $c_3^2 = 0.467$, and $c_4^2 = 1.752$. The weighted least squares regression equation based these c_j^2 values is $\hat{Y} = 49.05 + 0.0380X_1 - 0.0152X_2 - 0.0233X_3 - 33.98I_1 - 17.58I_2 - 22.06I_3$. The diagnostic plots from the weighted least squares model suggest that we have adequately treated the heteroscedasticity problem. [For weighted least squares diagnostic plots, one should divide the standardized residuals, as they are usually computed, by the appropriate c_j values.]

7.2
Trying out $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$ as an initial model, we obtain the regression equation $\hat{Y} = -289.18 + 0.0809X_1 + 0.8184X_2 - 0.1038X_3$. Diagnostic plots show that observations 6 and 49 are influential; the former (Connecticut) is a regression outlier, while the latter (Alaska) is a leverage point. Of the two, observation 49 is the more troublesome. In particular, observation 49 has the highest Y value (by nearly 100 units!), the highest X_2 value, and one of the highest X_1 values. Because Alaska is very different from the other states, we will exclude it from the remainder of our analysis.

Refitting our original model without Alaska, we get $\hat{Y} = -244.73 + 0.0746X_1 + 0.7146X_2 - 0.0852X_3$. The plot of standardized residuals versus fitted values indicates heteroscedasticity, and the plot of standardized residuals by region shows that the residuals for states in the third region are less spread out. Also, the latter plot shows that the residuals tend to be more positive for states in the fourth region, prompting us to refine the model (as in exercise 7.1) by introducing regional indicator variables before we attempt to deal with the heteroscedasticity. The regression equation for the refined model is $\hat{Y} = -101.43 + 0.0712X_1 + 0.4392X_2 - 0.1030X_3 - 29.18I_1 - 27.25I_2 - 31.46I_3$; indications of heteroscedasticity are present.

We compute (as in exercise 7.1) that $\hat{\sigma}_1^2 = 800.82$, $\hat{\sigma}_2^2 = 789.75$, $\hat{\sigma}_3^2 = 229.94$, and $\hat{\sigma}_4^2 = 546.18$; these values yield $c_1^2 = 1.458$, $c_2^2 = 1.438$, $c_3^2 = 0.419$, and $c_4^2 = 0.994$. The weighted least squares regression equation is $\hat{Y} = -151.19 + 0.0821X_1 + 0.4820X_2 - 0.1054X_3 - 31.39I_1 - 27.52I_2 - 26.06I_3$. Diagnostic plots indicate that the heteroscedasticity problem has been mitigated.

7.3
The figure below contains index plots of the leverages, standardized residuals, Cook's distances, and DFIT values. Only observation 49 (Alaska) is influential, as seen in the index plots of Cook's distances and DFIT values. Observation 49 has the highest leverage, and observation 44 (Utah) has the second-highest.

7.4

As in section 7.4, Alaska is excluded because it is quite different from the other states. When we include regional indicator variables, we get the following regression equation based on the other 49 states: $\hat{Y} = -150.72 + 0.0436X_1 + 0.6570X_2 + 0.0481X_3 - 17.32I_1 - 21.48I_2 - 29.73I_3$. There is still a heteroscedasticity problem after we have included the regional indicator variables, as can be seen in plots of the standardized residuals against the fitted values and by region. We remark that the estimated coefficients for X_1, X_2, and X_3 are similar qualitatively but not exactly equal to the WLS coefficients given in Table 7.7. The estimated coefficients for the indicator variables are (individually) not statistically significant at the .05 level.

Putting aside the issue of heteroscedasticity for the moment, we are asked to test the hypothesis that I_1, I_2, and I_3 are not needed in the present model; the F-value is $\frac{(57699.76-52781.62)/3}{52781.62/42} = 1.30$, which falls well short of the 2.83 required for rejection of the hypothesis. Thus, the decision in section 7.4 to construct a weighted least squares model directly from $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$ (i.e., without first incorporating regional indicator variables) seems justified.

7.5

The Y values for observations 11 and 12 should be 84. Noting this, we obtain the following regression equation, incorporating regional indicator variables: $\hat{Y} = 41.42 + 0.0382X_1 + 0.0026X_2 - 0.0237X_3 - 33.08I_1 - 17.14I_2 - 22.26I_3$. There appears to be some heteroscedasticity across fitted values, although the extent of the heteroscedasticity is hard to assess because of observations 30 and 33. The

standardized residuals are spread out more for states in the second and fourth regions than in the first and third. We remark that the estimated coefficients of all three indicator variables are (individually) statistically significant. The 1960 analogue to the WLS model presented in section 7.4 has regression equation $\hat{Y} = 17.43 + 0.0406X_1 + 0.0104X_2 - 0.0294X_3$; the estimated coefficients of X_1, X_2, and X_3 in the present model are similar qualitatively but not identical.

Putting aside the issue of heteroscedasticity for the moment, we are asked to test the hypothesis that I_1, I_2, and I_3 are not needed in the present model; the F-value is $\frac{(11332.31-5267.476)/3}{5267.476/43} = 16.50$, which is considerably larger than the 2.82 required for rejection of the hypothesis. We conclude that it would be wise to include regional indicator variables as part of a model for the Table 5.12 data (weighted least squares or otherwise) because, heteroscedasticity aside, there are differences in the expected value of Y across the regions; see the solution to exercise 7.1.

8 Solutions for Exercises in Chapter 8

8.1
(a) The Durbin-Watson statistic is $\frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{0.0002377}{0.0003829} = 0.621$. This is much less than 1.29, the lower critical value ($\alpha = .05$) given in Table A.6. We would conclude that the errors have a positive autocorrelation.
(b) There are, in order, five negative residuals, eleven positive, five negative, one positive, one negative, and two positive. Under randomness, the expected number of runs would be 11.32 and the standard deviation would be 2.41. The observed number of runs is 6, which is more than two standard deviations less than the expected number under randomness. Again, we would conclude that the errors have a positive autocorrelation.

8.2
(a) The Durbin-Watson statistic is $\frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{0.3459}{1.778} = 0.195$. This is considerably less than 1.34, the lower critical value ($\alpha = .05$) given in Table A.6. We would conclude that the errors have a positive autocorrelation.
(b) There are, in order, four negative residuals, seven positive, one negative, twelve positive, and five negative. Under randomness, the expected number of runs would be 12.10 and the standard deviation would be 2.38. The observed number of runs is 5, which is nearly three standard deviations less than the expected number under randomness. Again, we would conclude that the errors have a positive autocorrelation.

8.3
(a) To suspect the presence of an autocorrelation problem would be eminently reasonable. However, when one actually fits model (5.11) to the data, there is

no indication that the errors are positively autocorrelated. [For example, the Durbin-Watson statistic is 2.20.]
(b) In situations where there is a positive autocorrelation, adding a time trend predictor variable could help. To see why, consider an extreme case: Suppose the first several residuals from the original model are negative and the last several are positive (i.e., the original model is over-predicting the response early in time and under-predicting it later in time); one can see, in this case, that including the time trend predictor variable would allow the prediction of the response to be adjusted downward for the early years and upward for the later years, thereby improving upon the original model which had the autocorrelation problem.

8.4
(a) The regression equation for the linear trend model is $\hat{DJIA} = 5212 + 4.024 \cdot Day$. However, the linear trend model is not adequate, as there are strong time dependencies in the residuals. [For example, the Durbin-Watson statistic is a miniscule 0.0559!]
(b) The regression equation is $\hat{DJIA}_{(t)} = 36.90 + 0.9945 \cdot DJIA_{(t-1)}$. This model appears to be quite good: there is no indication of an autocorrelation problem, and $R^2 = 0.9856$.
(c) We obtain $\log(\hat{DJIA}) = 8.5627 + 0.00069 \cdot Day$ and $\log(\hat{DJIA}_{(t)}) = 0.0621 + 0.9929 \cdot \log(DJIA_{(t-1)})$. The basic conclusions are the same as those in parts (a) and (b).

8.5
(a) We use the model from exercise 8.4(b) to get the following regression equation for the data from the first 130 days: $\hat{DJIA}_{(t)} = 239.97 + 0.9573 \cdot DJIA_{(t-1)}$. The residual mean square is 1699.9.
(b) The prediction errors for the first 15 trading days of July are 76.8, -4.9, -13.1, 3.5, -111.3, -38.7, 28.1, 20.2, -83.8, -14.2, -165.7, -2.3, 7.0, 76.9, and -44.0.
(c) The average of the squared prediction errors over the first 15 trading days of July is 4253.1, which is a lot larger than the residual mean square.
(d) The average of the squared prediction errors over the last 132 trading days of the year is 2427.7, which is somewhat larger than the residual mean square.
(e) Two useful remarks can be made here. First, the residual mean square is generally an optimistic measure of model fit insofar as it will typically be smaller than the mean square error of prediction for future observations. Second, for this particular model, the mean square error of prediction for the trading days in July was large because there was a brief downturn in the economy in July; this can be seen in a plot of the $DJIA$ values.

8.6
(a) The t-value for testing $H_0 : \beta_0 = 0$ (individually) is 0.85, much less in magnitude than the level .05 critical values of ± 1.97. The t-value for testing $H_0 : \beta_1 = 1$ (individually) is -0.73, again much less in magnitude than the

critical values. However, the appropriate way to proceed is to test the hypotheses simultaneously. In the context of formula (3.28), the reduced model is $DJIA_{(t)} = DJIA_{(t-1)} + \epsilon$. We obtain an F-value of $\frac{[472756.5-464981.5]/2}{464981.5/259} = 2.17$, which falls short of the level .05 critical value of 3.03. So, we do not have evidence against the adequacy of the reduced model $DJIA_{(t)} = DJIA_{(t-1)} + \epsilon$.
(b) One may use an index plot, a normal probability plot, and a plot of the autocorrelation function to see that there are no serious contraindications to the properties suggested by the random walk theory (for either $DJIA$ or its logarithm).
(c) One should, in particular, repeat exercises 8.4(b), 8.6(a), and 8.6(b) for the new data.

9 Solutions for Exercises in Chapter 9

9.1
If we run the regression without A_t, we get $\hat{S}_t = 10.51 + 22.79E_t + 3.70P_t - 0.769A_{t-1} - 0.969P_{t-1}$; the associated variance inflation factors are 1.066, 1.427, 1.481, and 1.512, respectively, suggesting that the collinearity problem has been eliminated.

If we eliminate P_t, we get $\hat{S}_t = 28.88 + 23.09E_t - 3.84A_t - 4.27A_{t-1} - 4.78P_{t-1}$; the associated variance inflation factors are 1.062, 1.575, 1.374, and 1.774, respectively, suggesting that the collinearity problem has been eliminated.

If we run the regression without A_{t-1}, we get $\hat{S}_t = 5.76 + 22.71E_t + 1.15A_t + 4.61P_t + 0.0314P_{t-1}$; the variance inflation factors are 1.069, 2.111, 1.775, and 2.113, respectively, suggesting that the collinearity problem has been mitigated.

If we eliminate P_{t-1}, we get $\hat{S}_t = 5.38 + 22.81E_t + 1.17A_t + 4.65P_t + 0.159A_{t-1}$; the variance inflation factors are 1.058, 1.283, 1.365, and 1.258, respectively, suggesting that the collinearity problem has been eliminated.

9.2
(a) Here is the correlation matrix for the eleven predictor variables:

$$\begin{pmatrix} 1.00 & 0.94 & 0.99 & -0.35 & -0.67 & 0.64 & -0.77 & 0.87 & 0.80 & 0.95 & 0.82 \\ 0.94 & 1.00 & 0.96 & -0.29 & -0.55 & 0.76 & -0.63 & 0.80 & 0.71 & 0.89 & 0.71 \\ 0.99 & 0.96 & 1.00 & -0.33 & -0.67 & 0.65 & -0.75 & 0.86 & 0.79 & 0.94 & 0.80 \\ -0.35 & -0.29 & -0.33 & 1.00 & 0.41 & 0.04 & 0.56 & -0.30 & -0.38 & -0.36 & -0.44 \\ -0.67 & -0.55 & -0.67 & 0.41 & 1.00 & -0.22 & 0.87 & -0.56 & -0.45 & -0.58 & -0.76 \\ 0.64 & 0.76 & 0.65 & 0.04 & -0.22 & 1.00 & -0.28 & 0.42 & 0.30 & 0.52 & 0.40 \\ -0.77 & -0.63 & -0.75 & 0.56 & 0.87 & -0.28 & 1.00 & -0.66 & -0.66 & -0.71 & -0.85 \\ 0.87 & 0.80 & 0.86 & -0.30 & -0.56 & 0.42 & -0.66 & 1.00 & 0.88 & 0.96 & 0.68 \\ 0.80 & 0.71 & 0.79 & -0.38 & -0.45 & 0.30 & -0.66 & 0.88 & 1.00 & 0.90 & 0.63 \\ 0.95 & 0.89 & 0.94 & -0.36 & -0.58 & 0.52 & -0.71 & 0.96 & 0.90 & 1.00 & 0.75 \\ 0.82 & 0.71 & 0.80 & -0.44 & -0.76 & 0.40 & -0.85 & 0.68 & 0.63 & 0.75 & 1.00 \end{pmatrix}$$

Based on the correlation matrix and pairwise plots of the predictors (not shown here), it is clear that multicollinearity is present. In particular, X_1, X_2, and X_3 are very strongly related to each other, while X_{10} is strongly related to both X_1 and X_8.

(b) The eigenvalues are 7.703, 1.403, 0.7734, 0.5771, 0.2115, 0.1419, 0.09514, 0.05009, 0.03327, 0.008418, and 0.003497; the condition number is 46.93, so we conclude (as in part (a)) that multicollinearity is present. Eigenvectors corresponding to the smallest two eigenvalues are
$(-0.291, 0.291, -0.466, 0.051, -0.086, -0.005, -0.056, -0.294, -0.055, 0.714, 0.017)'$
and
$(0.618, 0.259, -0.682, 0.013, -0.045, -0.060, 0.049, 0.091, 0.053, -0.260, -0.010)'$;
in part (c), we will refer to these eigenvectors as v_{10} and v_{11}, respectively.

(c) The first, second, third, eighth, and tenth components of v_{10} are greater in absolute value than 0.29, while all the other components are less in absolute value than 0.09. The first, second, third, and tenth components of v_{11} are greater in absolute value than 0.25, while all the other components are less in absolute value than 0.10. Based on these results, the variables most involved in the multicollinearity are X_1, X_2, X_3, X_8, and X_{10}.

(d) The regression equation is $\hat{Y} = 17.77 - 0.0779X_1 - 0.0734X_2 + 0.1211X_3 + 1.329X_4 + 5.976X_5 + 0.3042X_6 - 3.199X_7 + 0.1854X_8 - 0.3991X_9 - 0.0052X_{10} + 0.5987X_{11}$. The variance inflation factors for the eleven predictor variables are found to be 128.2, 43.86, 161.3, 2.06, 7.78, 5.33, 11.74, 20.58, 9.42, 85.47, and 5.14. The predictors most affected are the same five that we identified in part (c): X_1, X_2, X_3, X_8, and X_{10}.

9.3

(a) The maximum number of terms allowable in a linear regression model is n, the number of observations. To see why this is true, consider the special case where there is one predictor. Then you are trying to find the slope and intercept of the line in two-dimensional space that most accurately represents the relationship between the response variable and the predictor. If you have only one observation, there will be infinitely many lines that go through the corresponding point in the plane; *two* distinct points in a plane uniquely determine a line going through both points, and more than two distinct points determine

a most accurate line in the sense of least squares. More generally, when there are p predictors, you will need at least $p + 1$ observations. [However, from a practical point of view, one would like to have many more observations than predictors, so that the regression coefficients can be estimated well.]

Answers for the next three parts will vary depending on what interaction terms have been chosen. Here is one possible solution:
(b) Suppose that we include I, D, W, G, N, P, and the year as covariates, along with the interactions of I, D, and the year with each of the three economic variables. This makes a total of 16 predictors, so that there will be 17 terms in the regression model for V. When this model is actually fit, there is only one t-value bigger in magnitude than 0.51, suggesting the presence of multicollinearity. When we try to compute the eigenvalues of the correlation matrix, two of the eigenvalues are reported to be negative! This means that these two eigenvalues are small even in comparison to the numerical error involved in their computation (which we usually ignore). So, we cannot even compute the condition number (although we know that it is very, very large). All of the predictors have large variance inflation factors. The *smallest* is 15.75, for W; several are in the hundreds and thousands.
(c) The correlations of the economic variables with the respective time-interaction variables are all nearly equal to 1, so we first eliminate the time-interaction terms. We can then obtain the correlation matrix for the remaining 13 predictors. The condition number is 61.04; by looking at the eigenvector corresponding to the smallest eigenvalue of the correlation matrix, we see that much of the remaining collinearity problem lies with D, I, and their interactions with N. Because of this, and for the reasons given in the solution to exercise 5.7, we may want to eliminate I and the interaction terms involving I. When we do so, the multicollinearity problem is considerably lessened, as the correlation matrix for the remaining 9 predictors has a condition number of 8.19.
(d) The regression equation for the model with the remaining 9 predictors is $\hat{V} = -1.2643 + 0.0022G - 0.0057N - 0.0006P - 0.0007D - 0.0110W + 0.0009(Year) + 0.0122(D \cdot G) + 0.0066(D \cdot N) - 0.0014(D \cdot P)$. The highest variance inflation factor is 11.71, for D, so the remaining multicollinearity is not severe. If we wanted to go one step further, we could eliminate D; then the condition number of the predictors' correlation matrix would fall to 4.36, and the highest variance inflation factor would be 4.32. For a final model, of course, one should eliminate the irrelevant variables (even if they do not contribute to a multicollinearity problem) and check the adequacy of the diagnostic plots.

10 Solutions for Exercises in Chapter 10

10.1
(a) We obtain $\hat{\theta_1} = 0.0463$, $\hat{\theta_2} = -1.0137$, $\hat{\theta_3} = -0.5375$, $\hat{\theta_4} = -0.2047$, $\hat{\theta_5} = -0.1012$, and $\hat{\theta_6} = 2.4797$. Just by looking at the data set, we can tell that

X_1, X_2, X_5, and X_6 are very highly correlated; they all follow a trend of steady increase. We conclude that there is likely to be a multicollinearity problem; we will investigate further beginning in part (d).

(b) From equation (9.18), we get $\hat{\beta}_1 = 1.5062$, $\hat{\beta}_2 = -0.0358$, $\hat{\beta}_3 = -2.0202$, $\hat{\beta}_4 = -1.0332$, $\hat{\beta}_5 = -0.0511$, $\hat{\beta}_6 = 1829.15$, and $\hat{\beta}_0 = -3.4823 \times 10^6$.

(c) The same answers as in part (b) will be obtained (up to differences in rounding).

(d) The pairwise correlation matrix of the predictors is:

$$\begin{pmatrix} 1.000 & 0.992 & 0.621 & 0.465 & 0.979 & 0.991 \\ 0.992 & 1.000 & 0.604 & 0.446 & 0.991 & 0.995 \\ 0.621 & 0.604 & 1.000 & -0.177 & 0.687 & 0.668 \\ 0.465 & 0.446 & -0.177 & 1.000 & 0.364 & 0.417 \\ 0.979 & 0.991 & 0.687 & 0.364 & 1.000 & 0.994 \\ 0.991 & 0.995 & 0.668 & 0.417 & 0.994 & 1.000 \end{pmatrix}$$

This correlation matrix and the corresponding scatter plot matrix (not shown here) indicate very strong linear relationships among X_1, X_2, X_5, and X_6.

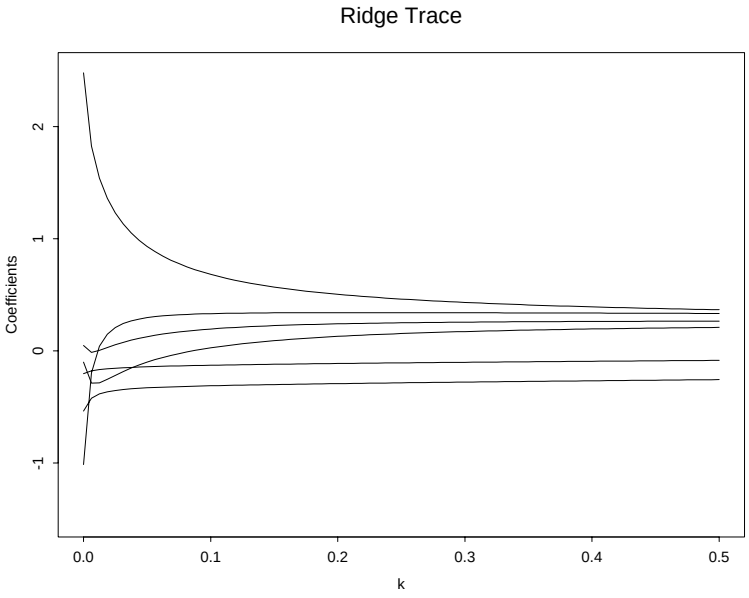
(e) The principal components are:

$C_1 = 0.462\tilde{X}_1 + 0.462\tilde{X}_2 + 0.321\tilde{X}_3 + 0.202\tilde{X}_4 + 0.462\tilde{X}_5 + 0.465\tilde{X}_6$,
$C_2 = 0.058\tilde{X}_1 + 0.053\tilde{X}_2 - 0.596\tilde{X}_3 + 0.798\tilde{X}_4 - 0.046\tilde{X}_5 + 0.001\tilde{X}_6$,
$C_3 = -0.149\tilde{X}_1 - 0.278\tilde{X}_2 + 0.728\tilde{X}_3 + 0.562\tilde{X}_4 - 0.196\tilde{X}_5 - 0.128\tilde{X}_6$,
$C_4 = 0.793\tilde{X}_1 - 0.122\tilde{X}_2 + 0.008\tilde{X}_3 - 0.077\tilde{X}_4 - 0.590\tilde{X}_5 - 0.052\tilde{X}_6$,
$C_5 = 0.338\tilde{X}_1 - 0.150\tilde{X}_2 + 0.009\tilde{X}_3 + 0.024\tilde{X}_4 + 0.549\tilde{X}_5 - 0.750\tilde{X}_6$,
$C_6 = 0.135\tilde{X}_1 - 0.818\tilde{X}_2 - 0.107\tilde{X}_3 - 0.018\tilde{X}_4 + 0.312\tilde{X}_5 + 0.450\tilde{X}_6$.

The corresponding variances (i.e., the the eigenvalues of the correlation matrix) are 4.6034, 1.1753, 0.2034, 0.0149, 0.0026, and 0.0004, respectively. It appears that there are three sets of multicollinearity in the data, corresponding to C_4, C_5, and C_6: X_1, X_2, and X_5 are prominent in C_4; X_1, X_2, X_5, and X_6 are prominent in C_5; X_1, X_2, X_3, X_5, and X_6 are prominent in C_6. The condition number of the correlation matrix is 110.5.

(f) We will retain C_1, C_2, and C_3; the estimated regression coefficients for the principal components are 0.4457, 0.1116, and -0.5297, respectively. From these we obtain $\hat{\theta}_1$ as a sum of products: $0.462 \cdot 0.4457 + 0.058 \cdot 0.1116 + -0.149 \cdot -0.5297 = 0.2913$. Similarly, $\hat{\theta}_2 = 0.3591$, $\hat{\theta}_3 = -0.3091$, $\hat{\theta}_4 = -0.1186$, $\hat{\theta}_5 = 0.3046$, and $\hat{\theta}_6 = 0.2752$. In turn, we find that $\hat{\beta}_1 = 9.4800$, $\hat{\beta}_2 = 0.0127$, $\hat{\beta}_3 = -1.1617$, $\hat{\beta}_4 = -0.5985$, $\hat{\beta}_5 = 0.1538$, $\hat{\beta}_6 = 203.00$, and $\hat{\beta}_0 = -3.5881 \times 10^5$.

(g) Here is the ridge trace, from which we can see that the coefficients for the standardized predictors are pretty stable once $k \approx 0.3$.



Taking $k := 0.3$, we have $\hat{\theta_1} = 0.2554$, $\hat{\theta_2} = 0.3390$, $\hat{\theta_3} = -0.2795$, $\hat{\theta_4} = -0.1027$, $\hat{\theta_5} = 0.1714$, and $\hat{\theta_6} = 0.4314$. Accordingly, $\hat{\beta_1} = 8.3117$, $\hat{\beta_2} = 0.0120$, $\hat{\beta_3} = -1.0504$, $\hat{\beta_4} = -0.5183$, $\hat{\beta_5} = 0.0865$, $\hat{\beta_6} = 318.23$, and $\hat{\beta_0} = -5.7522 \times 10^5$.
(h) The principal components and ridge results are qualitatively similar, and both are quite different from the ordinary least squares results (in which, for instance, the estimated regression coefficients for X_2 and X_5 are negative). Either the principal components or ridge approach would be preferrable to the ordinary least squares approach because of the multicollinearity among the predictors.

10.2
(a) The estimated regression coefficients for the standardized predictors are $\hat{\theta_1} = 0.6065$, $\hat{\theta_2} = 0.5277$, $\hat{\theta_3} = 0.0434$, and $\hat{\theta_4} = -0.1603$. Just by looking at the data set, we can see that X_2 and X_4 are negatively correlated, as are X_1 and X_3. We conclude that there is likely to be a multicollinearity problem; we will investigate further beginning in part (d).
(b) From equation (9.18), we get $\hat{\beta_1} = 1.5511$, $\hat{\beta_2} = 0.5102$, $\hat{\beta_3} = 0.1019$, $\hat{\beta_4} = -0.1441$, and $\hat{\beta_0} = 62.405$.
(c) The same answers as in part (b) will be obtained (up to differences in rounding).
(d) The pairwise correlation matrix of the predictors is:

$$\begin{pmatrix} 1.000 & 0.229 & -0.824 & -0.245 \\ 0.229 & 1.000 & -0.139 & -0.973 \\ -0.824 & -0.139 & 1.000 & 0.030 \\ -0.245 & -0.973 & 0.030 & 1.000 \end{pmatrix}$$

This correlation matrix and the corresponding scatter plot matrix (not shown here) indicate a strong negative relationship between X_2 and X_4 and a moderately strong negative relationship between X_1 and X_3, but the latter depends somewhat on observation 10.

(e) The principal components are:

$C_1 = -0.476\tilde{X}_1 - 0.564\tilde{X}_2 + 0.394\tilde{X}_3 + 0.548\tilde{X}_4$,

$C_2 = 0.509\tilde{X}_1 - 0.414\tilde{X}_2 - 0.605\tilde{X}_3 + 0.451\tilde{X}_4$,

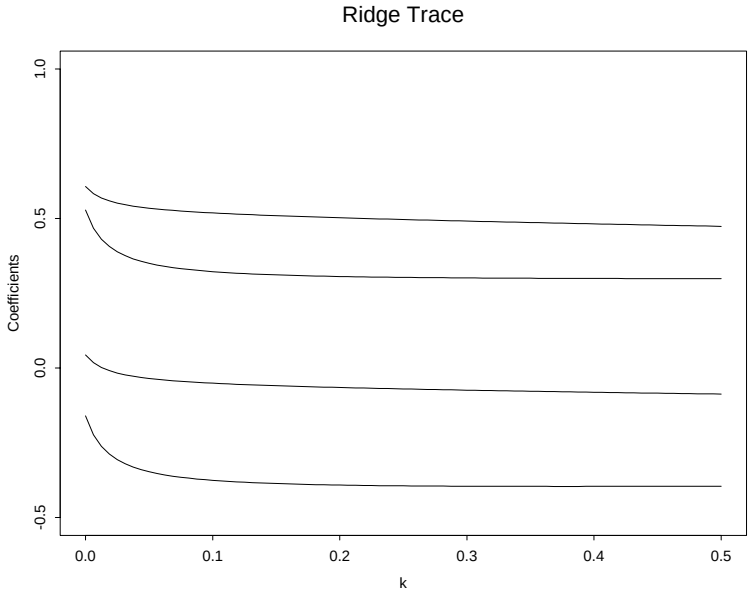
$C_3 = -0.676\tilde{X}_1 + 0.314\tilde{X}_2 - 0.638\tilde{X}_3 + 0.195\tilde{X}_4$,

$C_4 = 0.241\tilde{X}_1 + 0.642\tilde{X}_2 + 0.268\tilde{X}_3 + 0.677\tilde{X}_4$,

The corresponding variances (i.e., the the eigenvalues of the correlation matrix) are 2.2357, 1.5761, 0.1866, and 0.0016, respectively. It appears that there is one set of multicollinearity in the data, corresponding to C_4, in which all four predictors are involved. The condition number of the correlation matrix is 37.11.

(f) We will retain C_1, C_2, and C_3; the estimated regression coefficients for the principal components are -0.6570, -0.0083, and -0.3028, respectively. From these we obtain $\hat{\theta_1}$ as a sum of products: $-0.476 \cdot -0.6570 + 0.509 \cdot -0.0083 + -0.676 \cdot -0.3028 = 0.5132$. Similarly, $\hat{\theta_2} = 0.2789$, $\hat{\theta_3} = -0.0607$, and $\hat{\theta_4} = -0.4228$. In turn, we find that $\hat{\beta_1} = 1.3125$, $\hat{\beta_2} = 0.2696$, $\hat{\beta_3} = -0.1426$, $\hat{\beta_4} = -0.3800$, and $\hat{\beta_0} = 85.726$.

(g) Here is the ridge trace, from which we can see that the coefficients for the standardized predictors are pretty stable once $k \approx 0.2$.



Taking $k := 0.2$, we have $\hat{\theta_1} = 0.5022$, $\hat{\theta_2} = 0.3058$, $\hat{\theta_3} = -0.0655$, and $\hat{\theta_4} = -0.3920$. Accordingly, $\hat{\beta_1} = 1.2843$, $\hat{\beta_2} = 0.2956$, $\hat{\beta_3} = -0.1538$, $\hat{\beta_4} = -0.3523$, and $\hat{\beta_0} = 83.985$.

(h) The principal components and ridge results are qualitatively similar. Both are somewhat different from the ordinary least squares results. Either the principal components or ridge approach would be preferrable to the ordinary least squares approach because of the multicollinearity among the predictors.

10.3
No. In these data sets, it is the first principal component that is most strongly related to Y: for the Longley data set, 91.4% of the variation in Y is captured by C_1; for the Hald data set, 96.5% of the variation in Y is captured by C_1. Moreover, we do not have problems with outliers in these data sets.

11 Solutions for Exercises in Chapter 11

11.1
One may complete this exercise using software that performs regression calculations directly from the correlation matrix of the variables. For example, in all three parts, one may find that variables 9, 6, 2, 14, and 1 are selected (in that order) via the method of forward selection (using the first stopping criterion on page 296).

11.2
Using the information in Tables 11.11 and 11.13, along with formula (9.18), we obtain the estimated regression equation $\hat{Y} = \hat{\beta}_0 + \sum_{j=1}^{15} \hat{\beta}_j X_j$, where $\hat{\beta}_1 = 1.907$, $\hat{\beta}_2 = -1.945$, $\hat{\beta}_3 = -3.097$, $\hat{\beta}_4 = -9.076$, $\hat{\beta}_5 = -103.1$, $\hat{\beta}_6 = -17.05$, $\hat{\beta}_7 = -0.628$, $\hat{\beta}_8 = 0.0036$, $\hat{\beta}_9 = 4.463$, $\hat{\beta_{10}} = -0.189$, $\hat{\beta_{11}} = -0.150$, $\hat{\beta_{12}} = -0.662$, $\hat{\beta_{13}} = 1.320$, $\hat{\beta_{14}} = 0.0883$, $\hat{\beta_{15}} = 0.104$, and $\hat{\beta}_0 = 1748$.

11.3
Consider the following subset of variables: FTP, $UEMP$, LIC, GR, HE, and WE. Proceeding via backward elimination (with stopping rule number 1 on page 296), we select $UEMP$, LIC, and WE as our predictors. On the other hand, proceeding via forward selection (with stopping rule number 1 on page 296), we select FTP, $UEMP$, LIC, HE, and WE as our predictors.

11.4
The correlation matrix of the eleven predictors has only five eigenvalues greater than 0.030, and only three of those are greater than 0.101. In view of this evidence of multicollinearity, it would not be wise to directly apply forward selection or backward elimination to this set of predictors. Rather, we will follow the procedure introduced in section 11.13. Constructing a ridge trace for all eleven predictors (not shown), we find that FTP, $UEMP$, M, GR, and $CLEAR$ can be eliminated according to the first criterion on page 303. [Such a determination is somewhat subjective, of course.] Reconstructing the ridge trace without these five variables, we find that $NMAN$ can also be eliminated

according to the first criterion on page 303. Continuing, we find that we can remove G (second criterion), then W (second criterion again), and finally WE (third criterion). So, in the end, we are modeling H based only on LIC and HE. We obtain the regression equation $\hat{H} = -35.84 + 0.0215 \cdot LIC + 12.52 \cdot HE$. For this model, $R^2 = 0.9759$, and the p-values for both predictors are less than 0.0001. However, diagnostic plots reveal that the observation from 1968 is highly influential: it turns out that one of our predictors had an abnormally high value in 1968, while the other had an unusually low value. Hence, we may wish to downweight or omit the observation from 1968 in estimating the regression coefficients for the model we have selected.

11.5
(a) No. As one may suspect from looking at Table 11.17, there will be a (moderate) multicollinearity problem if we include both X_6 and X_7 as predictors: the VIF for X_6 is 11.71, while the VIF for X_7 is 9.72. If we remove X_6, the multicollinearity problem is mitigated; however, it is still not desirable to include all of the remaining predictors in the regression model because some of them offer little explanatory power (in particular, X_3, X_4, X_8, and X_9 have t-values less than one in magnitude).
(b) Upon fitting the proposed model, we find that the t-values for X_6 and X_8 are -0.352 and 0.077. Generally speaking, we are not content with a model in which predictors have t-values this small.
(c) The proposed model is actually not bad: $R^2 = 0.7637$. It turns out that if we were to use forward selection (X_6 excluded from consideration) with the first stopping criterion on page 296, we would select only X_1. Also, if we were to use backward elimination (X_6 excluded from consideration) with the first stopping criterion on page 296, we would again select only X_1. However, under each of the variable selection schemes, X_2 comes close to meeting the inclusion criterion. Thus, either a model with X_1 alone or a model with both X_1 and X_2 would be a reasonable choice.

11.6
(a) No, there would be a multicollinearity problem. For instance, the condition number for the correlation matrix of the eleven predictors is 46.93. [See the solution to exercise 9.2 for a more detailed examination of the multicollinearity.]
(b) Models (c), (e), and (f) are affected by multicollinearity, and model (d) does not really improve on model (b) (the t-value for X_2 is -0.867). Thus, among these six models, either model (a) or model (b) would be preferred. One other reasonable subset choice would be X_{10} and X_5, although the associated R^2 value (0.7568) is less than that of model (a) (0.7601).
(c) The four plots (not presented here) show that Y tends to decrease as the predictors increase but that the rate of decrease gets smaller as the predictors get larger.
(d) The four plots (not presented here) suggest that the relationships between the predictors and W are more linear than the relationships between the predictors and Y.

(e) Models (c), (e), and (f) are affected by multicollinearity. But model (d) does improve on model (b) when W is taken as the response (the t-value for X_2 is 2.170). Among the six models, either model (a) or model (d) would be preferred. One other reasonable subset choice would be X_1 and X_5, which yields a higher R^2 value (0.7775) than model (d) (0.7576).
(f) We obtain $\hat{Y} = -11.37 + 566.00X_{13}$, with $R^2 = 0.8921$. This model does a better job of predicting gasoline consumption than any of the other models that have been considered.
(g) A good report would include: an introduction, in which the variables of interest are described and in which the reason for conducting the analysis is given; a discussion of the process leading to the final model (or models), including the techniques and/or rationale for variable selection; the results of the statistical diagnostic checks that were performed on the final model; a physical interpretation of the final model in the context of the phenomenon being studied; and, if possible, some assessment of the model's capability for making predictions in the future.

11.7
(a) Suppose that we start with I, D, W, G, N, P, and the year as covariates, along with the interactions of I, D, and the year with each of the three economic variables. If we use backward elimination to select variables (with the first stopping criterion on page 296), we eliminate (in order): $Year \cdot P$; $D \cdot N$; $D \cdot G$; $I \cdot P$; $D \cdot P$; P; $Year$; N; and W. The regression equation based on the remaining variables is $\hat{V} = 0.5054 - 0.0736I + 0.0680D - 0.3248G + 0.000167 \cdot Year \cdot G - 0.00000387 \cdot Year \cdot N + 0.00955 \cdot I \cdot G + 0.00855 \cdot I \cdot N$. We note that the model chosen by backward elimination has a severe multicollinearity problem, which reinforces the warning (see page 302) that subset selection needs to be done more carefully when the data are multicollinear. In such situations, instead of relying on forward selection or backward elimination, we can employ the method described in section 11.13. When we do so here, we decide (somewhat subjectively, of course) to eliminate variables as follows: $Year$, W, and $Year \cdot P$ (based on the first criterion given on page 303); P and $D \cdot G$ (first criterion); G, N, $Year \cdot G$, $Year \cdot N$, and $D \cdot P$ (second criterion); I (second criterion); D and $I \cdot N$ (first criterion). We are left with $D \cdot N$, $I \cdot G$, and $I \cdot P$; the regression equation is $\hat{V} = 0.4774 + 0.00772 \cdot D \cdot N + 0.00705 \cdot I \cdot G - 0.00345 \cdot I \cdot P$.
(b) Suppose that we start only with D, W, G, N, P, the year, and the interactions of D with the three economic variables. Since we have eliminated I, its interactions, and the time interactions from consideration, there is not a severe multicollinearity problem and we can attempt subset selection via forward selection and/or backward elimination. If we use forward selection with the first stopping rule on page 296, we choose (in order): $D \cdot G$; $D \cdot N$; and $Year$. The regression equation for the model chosen via forward selection is $\hat{V} = -2.1456 + 0.0122 \cdot D \cdot G + 0.00528 \cdot D \cdot N + 0.00133 \cdot Year$. If we use backward elimination with the first stopping rule on page 296, we drop (in order): D; P; $D \cdot P$; W; N; and G. So, we get the same result from backward elimination as we did from forward selection.

(c) The model obtained from backward elimination in part (a) is unsuitable because of the associated multicollinearity problem. The other model obtained in part (a) and the model obtained in part (b) both have comparable explanatory power on three predictors.
(d) Using the second model obtained in part (a) with the values $I = 1$, $D = 0$, $G = 3.5$, $P = 1.7$, and $N = 8$, the value of V predicted for the 2000 election is 0.496. Predictions for 2004 and 2008 can only be made if the future values of the covariates are guessed.
(e) Predictions for 2004 and 2008 based on guessed values of the covariates are likely to be less accurate than the prediction for 2000 based on known values of the covariates.
(f) The Democratic share of the two-party vote in 2000 was 0.503.

11.8
(a) The condition number for the correlation matrix of the predictors is 3.80, indicating that multicollinearity is not a concern. So, we may select variables via forward selection and/or backward elimination. If we use backward elimination, HS and $Female$ are dropped first. If we rigidly apply the first stopping rule on page 296, $Black$ and Age will also be dropped; however the t-value for Age (1.89) is fairly close to the level 0.05 critical value (2.01), suggesting that perhaps we should retain Age. If we do, then the regression equation is $\hat{Sales} = 64.25 + 4.16 \cdot Age + 0.0193 \cdot Income - 3.40 \cdot Price$. Diagnostic plots for the model show that observation 30 (New Hampshire) is influential. This is because the per capita sales figure for New Hampshire is much higher than that of any other state, making New Hampshire a regression outlier. [As mentioned in the solution to exercise 4.7, the high sales figure has been attributed to the non-taxation of cigarettes in New Hampshire.] If we remove observation 30 and *then* start with the backward elimination, only HS and $Female$ are dropped, and the regression equation is $\hat{Sales} = 41.58 + 3.39 \cdot Age + 0.0184 \cdot Income + 0.543 \cdot Black - 2.38 \cdot Price$. Diagnostic plots show that observation 29 (Nevada) is a moderately influential regression outlier and that observation 9 (D.C.) is a leverage point. To arrive at a final model, one must eventually decide whether and how to incorporate New Hampshire, Nevada, and D.C. (perhaps also Alaska and Hawaii) into the model-building process; we have already seen that such decisions can affect not only the regression coefficients but also the variables that are selected through a process like backward elimination.
(b) A thorough report would include: an introduction, in which the variables of interest are described and in which the reason for conducting the analysis is given; a discussion of the process leading to the final model (or models), including the techniques and/or rationale for variable selection as well as the decisions made about whether and how to include certain observations; the results of the statistical diagnostic checks that were performed on the final model; a social or economic interpretation of the final model in the context of the phenomenon being studied; and, if possible, some assessment of the model's capability for making predictions in the future.

12 Solutions for Exercises in Chapter 12

12.1
After the three unusual observations are deleted, we obtain $\hat{\beta}_0 = -59.25$, $\hat{\beta}_1 = 1.77$, $\hat{\beta}_2 = 2.08$, and $\hat{\beta}_3 = 36.05$. However, the standard errors are extremely large, resulting in miniscule Z-test values. One finds that X_2 and X_3 may be deleted based on the chi-square test discussed in section 12.6; the test statistics for X_2 and X_3 are essentially equal to zero. In fact, all 63 of the remaining observations may be correctly classified merely according to whether $X_1 > 8.0$. So, yes, there has been a substantial change in the results.

12.2
Index plots of the Pearson residuals and the deviance residuals (as well as their plots against the fitted probabilities) suggest that observation 9 is unusual. Especially with respect to X_1, observation 9 more nearly resembles the solvent firms than the other bankrupt firms.

12.3
(a) The fitted logits are given by $15.04 - 0.232Temp$. Interpretation: The relative odds of Pr(Failure)/Pr(No Failure) are multiplied by $\exp(-0.232) = 0.793$ for each degree increase in temperature.
(b) The fitted logits are given by $23.40 - 0.361Temp$. Interpretation: The relative odds of Pr(Failure)/Pr(No Failure) are multiplied by $\exp(-0.361) = 0.697$ for each degree increase in temperature.
(c) The fitted logit for a temperature of 31 degrees is 12.21, yielding a fitted probability of almost 1.
(d) Knowing the result of part (c), no.

12.4
(a) The fitted logits for the NFL are given by $2.49 - 0.0131X - 0.00151X^2$; the fitted logits for the AFL are given by $4.89 - 0.197X + 0.00160X^2$.
(b) The fitted logits are given by $3.52 - 0.0959X - 0.00011X^2 + 0.104Z$.
(c) The value of the test statistic described in section 12.6 is 0.0218, suggesting that a quadratic term is not needed in the model of part (b).
(d) The test statistic for whether Z is needed is 0.373 if we have retained X^2 and 0.372 if we have not. In either case, it does not appear that Z is needed.