Appendix F
Answers for selected problems

This appendix contains *brief* answers for even-numbered problems—these are intended to help you review material. The odd-numbered problems tend to be more challenging and normally require more extensive solutions than those given here.

Chapter 1

1.2 (a) $\Pr(Z < 1.282) = 0.90 \Leftrightarrow \Pr((Y - 70)/10 < 1.282) = 0.90$
$\Leftrightarrow \Pr(Y < 70 + 1.282(10)) = 0.90 \Leftrightarrow \Pr(Y < 82.8) = 0.90.$

(b) $\Pr(Z < 2.326) = 0.99 \Leftrightarrow \Pr((Y - 70)/10 < 2.326) = 0.99$
$\Leftrightarrow \Pr(Y < 70 + 2.326(10)) = 0.99 \Leftrightarrow \Pr(Y < 93.3) = 0.99.$

(c) $\Pr(Z < -1.645) = 0.05 \Leftrightarrow \Pr((Y - 70)/10 < -1.645) = 0.05$
$\Leftrightarrow \Pr(Y < 70 - 1.645(10)) = 0.05 \Leftrightarrow \Pr(Y < 53.6) = 0.05.$

(d) $\Pr(Z < 1.282) = 0.90 \Leftrightarrow \Pr((M_Y - 70)/(10/\sqrt{25}) < 1.282) = 0.90$
$\Leftrightarrow \Pr(M_Y < 70 + 1.282(10/\sqrt{25})) = 0.90 \Leftrightarrow \Pr(M_Y < 72.6) = 0.90.$

(e) $\Pr(Z < 2.326) = 0.99 \Leftrightarrow \Pr((M_Y - 70)/(10/\sqrt{25}) < 2.326) = 0.99$
$\Leftrightarrow \Pr(M_Y < 70 + 2.326(10/\sqrt{25})) = 0.99 \Leftrightarrow \Pr(M_Y < 74.7) = 0.99.$

(f) $\Pr(Z < -1.645) = 0.05 \Leftrightarrow \Pr((M_Y - 70)/(10/\sqrt{25}) < -1.645) = 0.05$
$\Leftrightarrow \Pr(M_Y < 70 - 1.645(10/\sqrt{25})) = 0.05 \Leftrightarrow \Pr(M_Y < 66.7) = 0.05.$

(g) If the bottom 5% of the class fail, the cut-off percentage to pass the class is 53.6%, the 5th percentile of the distribution of Y.

(h) Since the 99th percentile of the sampling distribution of M_Y is 74.7%, the requirement seem feasible.

1.4 (a) $\Pr(Y > 34.1) = \Pr((Y-3)/10 > (34.1-3)/10)$
$= \Pr(Z > (34.1-3)/10) = \Pr(Z > 3.11) \approx 0.001$ (from last row of Table B.1).

(b) $\Pr(Y > 15.7) = \Pr((Y-3)/10 > (15.7-3)/10)$
$= \Pr(Z > (15.7-3)/10) = \Pr(Z > 1.27) \approx 0.1$ (from last row of Table B.1).

(c) $\Pr(Y < -13.3) = \Pr((Y-3)/10 < (-13.3-3)/10)$
$= \Pr(Z < (-13.3-3)/10) = \Pr(Z < -1.63) \approx 0.05$ (from last row of Table B.1).

(d) $\Pr(M_Y > 7.4) = \Pr((M_Y-3)/(10/\sqrt{50}) > (7.4-3)/(10/\sqrt{50}))$
$= \Pr(Z > (7.4-3)/(10/\sqrt{50})) = \Pr(Z > 3.11) \approx 0.001$ (from last row of Table B.1).

(e) $\Pr(M_Y > 4.8) = \Pr((M_Y-3)/(10/\sqrt{50}) > (4.8-3)/(10/\sqrt{50}))$
$= \Pr(Z > (4.8-3)/(10/\sqrt{50})) = \Pr(Z > 1.27) \approx 0.1$ (from last row of Table B.1).

(f) $\Pr(M_Y < 0.7) = \Pr((M_Y-3)/(10/\sqrt{50}) < (0.7-3)/(10/\sqrt{50}))$
$= \Pr(Z < (0.7-3)/(10/\sqrt{50})) = \Pr(Z < -1.63) \approx 0.05$ (from last row of Table B.1).

1.6 (a) Mean: 68.480; standard deviation: 14.6367.

(b) $m_Y \pm 95\text{th percentile}\,(s_Y/\sqrt{n}) = 68.480 \pm 1.660(14.6367/\sqrt{100})$
$= 68.480 \pm 2.430 = (66.050, 70.910)$.

1.8 (a) Since the t-statistic of $(68.480-66)/(14.6367/\sqrt{100}) = 1.69$ is more than 1.660 (the 95th percentile of the t-distribution with $n-1=99$ degrees of freedom), reject NH: $E(Y)=66$ in favor of AH: $E(Y)>66$.

The p-value is between 0.025 (critical value 1.984) and 0.05 (critical value 1.660).

(b) Since the t-statistic of $(68.480 - 73)/(14.6367/\sqrt{100}) = -3.09$ is less than -1.660 (the 5th percentile of the t-distribution with $n-1=99$ degrees of freedom), reject NH: $\mathrm{E}(Y) = 66$ in favor of AH: $\mathrm{E}(Y) < 73$. The p-value is between 0.001 (critical value -3.174) and 0.005 (critical value -2.626).

(c) Since the t-statistic of $(68.480 - 66)/(14.6367/\sqrt{100}) = 1.69$ is between -1.984 and 1.984 (the 2.5th and 97.5th percentiles of the t-distribution with $n-1=99$ degrees of freedom), fail to reject NH: $\mathrm{E}(Y)=66$ in favor of AH: $\mathrm{E}(Y)\neq 66$. The p-value/2 is between 0.025 (critical value 1.984) and 0.05 (critical value 1.660), so the p-value is between 0.05 and 0.10.

1.10 (a) $\Pr(Price < 215) = \Pr((Price - 280)/50 < (215 - 280)/50) = \Pr(Z < -1.30)$ = slightly less than 10% (since $\Pr(Z < -1.282) = 10\%$).

(b) $m_Y \pm 95\text{th percentile}\,(s_Y/\sqrt{n}) = 278.6033 \pm 1.699(53.8656/\sqrt{30})$
$= 278.6033 \pm 16.7088 = (262, 295)$.

(c) (i) Since the t-statistic of $(278.6033 - 265)/(53.8656/\sqrt{30}) = 1.38$ is less than 1.699 (the 95th percentile of the t-distribution with $n-1=29$ degrees of freedom), fail to reject NH in favor of AH;

(ii) Since the t-statistic of $(278.6033 - 300)/(53.8656/\sqrt{30}) = -2.18$ is less than -1.699 (the 5th percentile of the t-distribution with $n-1=29$ degrees of freedom), reject NH in favor of AH;

(iii) Since the t-statistic of $(278.6033 - 290)/(53.8656/\sqrt{30}) = -1.16$ is more than -1.699 (the 5th percentile of the t-distribution with $n-1=29$ degrees of freedom), fail to reject NH in favor of AH;

(iv) Since the t-statistic of $(278.6033 - 265)/(53.8656/\sqrt{30}) = 1.38$ is less than 2.045 (the 97.5th percentile of the t-distribution with $n-1=29$ degrees of freedom), fail to reject NH in favor of AH.

(d) $m_Y \pm$95th percentile $(s_Y\sqrt{1+1/n}) = 278.6033 \pm 1.699(53.8656\sqrt{1+1/30})$ $= 278.6033 \pm 93.0304 = (186, 372)$.

1.12 (a) $m_Y \pm$ 95th percentile $(s_Y/\sqrt{n}) = 2.9554 \pm 2.037(1.48104/\sqrt{33})$ $= 2.9554 \pm 0.5252 = (2.43, 3.48)$.

(b) $m_Y \pm$ 95th percentile $(s_Y\sqrt{1+1/n}) = 2.9554 \pm 2.037(1.48104\sqrt{1+1/33})$ $= 2.9554 \pm 3.0622 = (-0.107, 6.018)$.

Since the data range is 1.642 to 7.787, this does not seem very reasonable.

(c) $0.3956 \pm 2.037(0.12764/\sqrt{33}) = 0.3956 \pm 0.0453 = (0.3503, 0.4409)$.

(d) $(1/0.4409, 1/0.3503) = (2.27, 2.85)$.

(e) $0.3956 \pm 2.037(0.12764\sqrt{1+1/33}) = 0.3956 \pm 0.2639 = (0.1317, 0.6595)$.

In original units this corresponds to $(1.52, 7.59)$, a much more reasonable interval based on the range of Y-values in the data. Although not as obvious, the confidence interval in part (d) is also more reasonable than the confidence interval in part (a)—if you look at a histogram of the Y-values, it seems very unlikely that the "center" could be as high as 3.5, as suggested by the confidence interval in part (a). The data in original units were far from normally distributed, whereas taking the reciprocal transformation made the sample values look more normal—confidence intervals and (particularly) prediction intervals tend to be more effective the closer to normal the data look.

2.2 (a) The slope should be positive since a higher batting average should result in more wins, all other things being equal.

(b) The points in the scatterplot (not shown) with *Win* on the vertical axis and *Bat* on the horizontal axis have a general upward trend from left to right, so this does agree with the answer in part (a).

(c) $\widehat{Win} = -127 + 813Bat$.

(d) The line seems to represent the linear trend in the data reasonably well and is not overly influenced by any isolated points in the scatterplot.

(e) There is some kind of linear association, but it would probably not be described as "strong." The variation of the data points about the least squares line is somewhat less than the overall variation in the number of wins, but there still remains quite a lot of unexplained variation—winning baseball games depends on more than just a team's batting average.

(f) The estimated intercept of -127 has little practical interpretation since it corresponds to the expected number of wins for a team that has a zero batting average—a nonsensical value in this context. The estimated slope of 813 corresponds to the expected change in the number of wins as a team's batting average increases by 1 unit. It might be clearer to say that we expect the number of wins to increase by 8.13 games on average when a team's batting average increases by 0.01 unit (e.g., from 0.250 to 0.260).

(g) The linear association between the number of wins and the batting average is not as strong for the American League teams as it is for the National League teams.

2.4 (a) $\widehat{Arm} = 9.599 + 0.611 Foot$.

(b) $\widehat{Arm} = 9.599 + 0.611(26) = 25.485$ cm.

(c) We expect right forearm length to increase by 0.611 cm for a 1 cm increase in right foot length.

(d) The estimated standard deviation of the random errors is 1.555 cm, which tells us approximately how far, on average, the observed right forearm lengths are from their predicted values.

(e) 54.4% of the variation in right forearm lengths (about their mean) is "explained" by a linear association between right forearm length and right foot length.

2.6 (a) Scatterplot of the data.

(b) $\widehat{Height} = 120.046 + 0.734\, Weight$.

(c) Add least squares line to the scatterplot.

(d) Conduct this upper-tail test: NH: $b_1 = 0$ versus AH: $b_1 > 0$. Since the t-statistic of $0.734/0.185 = 3.964$ is greater than 1.771 (the 95th percentile of the t-distribution with $n-2 = 13$ degrees of freedom), reject NH in favor of AH. (Or, since the p-value of $0.0016/2 = 0.0008$ is less than 0.05, reject NH in favor of AH.) Thus, the sample data suggest that the population slope is greater than 0 (at a 5% significance level). In other words, this supports the claim that there is a positive linear association between *Height* and *Weight*.

(e) $\widehat{Height} = 120.046 + 0.734(75) = 175.1$ cm.

2.8 (a) $\hat{b}_0 = 9.14 - 1.30(1) = 7.84$.

(b) $\widehat{pH}_1 = 7.84 + 1.30(0) = 7.84$.

(c) $pH_1 = 7.84 - 0.54 = 7.3$.

(d) $\widehat{pH}_4 = 7.84 + 1.30(3) = 11.74$, $\hat{e}_4 = 11.4 - 11.74 = -0.34$,

$\widehat{pH}_5 = 7.84 + 1.30(4) = 13.04,\ \hat{e}_5 = 12.8 - 13.04 = -0.24.$

(e) $\hat{e}_3 = -(-0.54 + 0.26 - 0.34 - 0.24) = 0.86.$

(f) $\widehat{pH}_3 = 7.84 + 1.30(2) = 10.44,\ pH_3 = 10.44 + 0.86 = 11.3.$

(g) $\text{RSS} = (-0.54)^2 + 0.26^2 + 0.86^2 + (-0.34)^2 + (-0.24)^2 = 1.272.$

(h) $s = \sqrt{\frac{1.272}{5-2}} = 0.6512.$

(i) $m_{pH} = 10.44.\ pH - m_{pH} = -3.14, -1.04, 0.86, 0.96, 2.36.$

(j) $\text{TSS} = (-3.14)^2 + (-1.04)^2 + 0.86^2 + 0.96^2 + 2.36^2 = 18.172.$

(k) $R^2 = 1 - \frac{1.272}{18.172} = 93.0\%.$

2.10 (a) $\hat{b}_1 \pm 95\text{th percentile}\,(s_{\hat{b}_1}) = 40.800 \pm 2.353 \times 5.684 = 40.800 \pm 13.374 =$ (27.4, 54.2).

(b) Conduct this upper-tail test: NH: $b_1 = 20$ versus AH: $b_1 > 20$. Since the t-statistic of $(40.800 - 20)/5.684 = 3.66$ is greater than 2.353 (the 95th percentile of the t-distribution with $n-2=3$ degrees of freedom), reject NH in favor of AH. Thus, the sample data suggest that the population slope is greater than 20 (at a 5% significance level). In other words, putting a 500-square foot addition onto a house could be expected to increase its sale price by \$10,000 or more.

2.12 (a) NH: $b_1 = 0$ vs. AH : $b_1 > 0$.

(b) Yes, since the p-value of the test (0.0005) is less than the significance level (0.05).

(c) For every additional customer, we estimate costs to increase by \$9.87.

(d) For a day with no customers (presumably a day the restaurant is closed), we estimate costs (presumably fixed costs) to be \$1,192.

(e) We expect about 95% of the observed cost values to lie within \$448 of their least squares predicted values.

(f) About 90.5% of the total variation in the sample cost values (about their mean) can be explained by (or attributed to) the linear association between cost and number of customers.

2.14 (a) The zero mean assumption appears to fail for the upper-right plot since it has a U-shaped pattern. The constant variance assumptions appears to fail for the lower-left plot since it has a fan/megaphone pattern. The apparent gap in the middle of the X-values for the lower-right plot is nothing to worry about as this has nothing to do with the assumptions, which relate to the residuals not the X-values.

(b) The normality assumption appears to fail for the upper-left, upper-right, and lower-left plots since the points do not lie reasonably close to the diagonal line.

2.16 (a) Our best point estimate for E(*Height*) at *Weight* = 75 is

$\widehat{Height} = 120.046 + 0.734 \times 75 = 175.1.$

Then, $\widehat{Height} \pm$ 95th percentile $(s_{\hat{Y}}) = 175.1 \pm 2.160 \times 2.908$

$= 175.1 \pm 6.3 = (168.8, 181.4).$

(b) Our best point estimate for $Height^*$ at *Weight* = 75 is

$\widehat{Height} = 120.046 + 0.734 \times 75 = 175.1.$

Then, $\widehat{Height}^* \pm$ 95th percentile $(s_{\hat{Y}^*}) = 175.1 \pm 2.160 \times 10.951$

$= 175.1 \pm 23.7 = (151.4, 198.8).$

(c) The interval in part (b) is appropriate for quantifying the uncertainty around predicting the height for a student who weighs 75 kilograms.

(d) The prediction interval in part (b) has a width of 47.3, which is wider than the confidence interval in part (a) with a width of 12.6.

2.18 (a) Our best point estimate for E(*Price*) at *Floor* = 2 is

$\widehat{Price} = 190.318 + 40.800 \times 2 = 271.918$.

Then, $\widehat{Price} \pm$ 95th percentile $(s_{\hat{Y}}) = 271.918 \pm 2.353 \times 1.313$

$= 271.918 \pm 3.089 = (268.8, 275.0)$.

(b) Our best point estimate for $Price^*$ at $Floor = 2$ is

$\widehat{Price} = 190.318 + 40.800 \times 2 = 271.918$.

Then, $\widehat{Price}^* \pm$ 95th percentile $(s_{\hat{Y}^*}) = 271.918 \pm 2.353 \times 3.080$

$= 271.918 \pm 7.247 = (264.7, 279.2)$.

Chapter 3

3.2 (a) $\hat{b}_1 \pm$ 95th percentile $(s_{\hat{b}_1}) = 6.074 \pm 1.753 \times 2.662 = 6.074 \pm 4.666k =$ (1.41, 10.74).

(b) $\hat{b}_1 \pm$ 95th percentile $(s_{\hat{b}_1}) = 5.001 \pm 1.740 \times 2.261 = 5.001 \pm 3.934 =$ (1.07, 8.94). This interval is narrower (more precise) than the one in (a) because the two-predictor model is more accurate than the four-predictor model (which contains two unimportant predictors).

3.4 (a) $n = 50$.

(b) $k = 3$.

(c) $k + 1 = 4$.

(d) Global F-statistic = 126.459/3 = 42.153.

(e) p-value = 0.000.

(f) Individual t-statistic = 0.406/0.145 = 2.799.

(g) p-value = 0.007.

(h) MSE = 97.961/46 = 2.130.

(i) $s = \sqrt{2.130} = 1.459$.

(j) 1.459.

(k) $R^2 = 1 - 97.961/224.420 = 0.5635$.

(l) 56.35%.

(m) $\widehat{Disease} = -0.679 + 0.406Risk1 + 0.379Risk2 + 0.112Risk3$.

(n) When *Risk1* and *Risk3* are held constant, we expect *Disease* to increase by 0.379 units when *Risk2* increases by 1 unit.

(o) $\widehat{Disease} = -0.679 + 0.406(7) + 0.379(5) + 0.112(6) = 4.73$.

3.6 (a) Our best point estimate for E(*Lab*) at *Tws*=6 and *Asw*=20 is $\widehat{Lab} = 110.431 + 5.001 \times 6 - 2.012 \times 20 = 100.2$.

Then, $\widehat{Lab}$ ± 95th percentile $(s_{\hat{Y}}) = 100.2 \pm 1.740 \times 2.293 = 100.2 \pm 3.990 = (96.2, 104.2)$.

(b) Our best point estimate for Lab^* at *Tws*=6 and *Asw*=20 is $\widehat{Lab} = 110.431 + 5.001 \times 6 - 2.012 \times 20 = 100.2$.

Then, $\widehat{Lab}^*$ ± 95th percentile $(s_{\hat{Y}^*}) = 100.2 \pm 1.740 \times 9.109 = 100.2 \pm 15.850 = (84.4, 116.1)$.

3.8 (a) The least squares equation is $\widehat{Mort} = 1{,}006.244 - 15.346Edu + 4.214Nwt - 2.150Jant + 1.624Rain + 18.548Nox + 0.537Hum - 0.345Inc$.

(b) NH: $b_6 = b_7 = 0$ versus AH: at least one of b_6 or b_7 is not equal to zero.

$$\begin{aligned} \text{Nested F-statistic} &= \frac{(\text{RSS}_\text{R} - \text{RSS}_\text{C})/(k-r)}{\text{RSS}_\text{C}/(n-k-1)} \\ &= \frac{(60{,}948 - 60{,}417)/(7-5)}{60{,}417/(56-7-1)} \\ &= 0.211. \end{aligned}$$

Since this is less than the 95th percentile of the F-distribution with 2 numerator degrees of freedom and 48 denominator degrees of freedom (3.19), do not reject the null hypothesis in favor of the alternative. In other words, *Hum* and *Inc* do not provide significant

information about the response, *Mort*, beyond the information provided by the other predictor variables, and the reduced model is preferable.

(c) NH: $b_p = 0$ versus AH: $b_p \neq 0$. Individual t-statistics (from statistical software output) are −2.418, 6.334, −3.347, 2.942, and 3.479. Since the absolute values of these statistics are all greater than the 97.5th percentile of the t-distribution with 50 degrees of freedom (2.01), we reject the null hypothesis (in each case) in favor of the alternative. In other words, each of the variables has a significant linear association with *Mort*, controlling for the effects of all the others.

(d) Looking at residual plots and a histogram and QQ-plot of the residuals (not shown), the four assumptions seem reasonable.

(e) $\widehat{Mort} = 1{,}028.232 - 15.589Edu + 4.181Nwt - 2.131Jant + 1.633Rain + 18.413Nox$. This model shows positive associations between mortality and each of percentage nonwhite, rainfall, and nitrous oxide, and negative associations between mortality and each of education and temperature. All of these associations might have been expected.

(f) Using statistical software, the 95% confidence interval is (946, 979).

(g) Using statistical software, the 95% prediction interval is (891, 1,035).

Chapter 4

4.2 (a) Conduct this two-tail test: NH: $b_1 = 0$ versus AH: $b_1 \neq 0$. Since the t-statistic of $0.0296/0.0142 = 2.086$ is greater than 2.052 (the 97.5th percentile of the t-distribution with $n - 3 = 27$ degrees of freedom), reject NH in favor of AH. (Or, since the p-value of 0.047 is less than 0.05, reject NH in favor of AH.) Thus, the sample data suggest that

$b_1 \neq 0$ (at a 5% significance level), adjusting for predictor variable *Age*. In other words, the Age^2 term is statistically significant in the quadratic model.

(b) $\widehat{Muscle} = 249.2038 - 4.5253(60) + 0.0296(60^2) = 84.2$.

4.4 95% prediction intervals for model 1 at \$100k, \$150k, and \$200k are (\$466, \$1,032), (\$808, \$1,377), and (\$1,146, \$1,727), while the equivalent intervals for model 2 are (\$534, \$1,018), (\$794, \$1,520), and (\$1,049, \$2,026). The model 2 intervals seem to be more appropriate than the model 1 intervals based on visual inspection of a scatterplot of *Tax* versus *Price*.

4.6 When the prevailing interest rate is 3%, we expect to increase sales by $1.836 - 0.126(3) = \$1.46$m for each additional \$1m we spend on advertising.

4.8 (a) Conduct this two-tail test: NH: $b_3 = 0$ versus AH: $b_3 \neq 0$. Since the t-statistic of $-0.086/0.040 = -2.135$ is less than -2.028 (the 2.5th percentile of the t-distribution with $n-4=36$ degrees of freedom), reject NH in favor of AH. (Or, since the p-value of 0.040 is less than 0.05, reject NH in favor of AH.) Thus, the sample data suggest that $b_3 \neq 0$ (at a 5% significance level), adjusting for the other predictor variables in the model. In other words, the linear association between *Calories* and *Carbs* depends on *Fats*.

(b) Yes.

(c) $\widehat{Calories} = 53.666 + 6.584 Carbs + 9.385 Fats - 0.086 Carbs Fats$.

(d) $\widehat{Calories} = 53.666 + 6.584(30) + 9.385 Fats - 0.086(30) Fats = 251.186 + 6.805 Fats$.

(e) $\widehat{Calories} = 53.666 + 6.584(40) + 9.385 Fats - 0.086(40) Fats = 317.026 +$

5.945*Fats*.

(f) The "slope" for *Fats* decreases from 6.805 when *Carbs* = 30 to 5.945 when *Carbs* = 40.

(g) Predicted calories for a meal with *Carbs* = 30 and *Fats* = 30 is 251.186 + 6.805(30) = 455.3. Predicted calories for a meal with *Carbs* = 30 and *Fats* = 20 is 251.186 + 6.805(20) = 387.3. Difference: 455.3 − 387.3 = 68.0.

(h) Predicted calories for a meal with *Carbs* = 40 and *Fats* = 30 is 317.026 + 5.945(30) = 495.4. Predicted calories for a meal with *Carbs* = 40 and *Fats* = 20 is 317.026 + 5.945(20) = 435.9. Difference: 495.4 − 435.9 = 59.5.

(i) The difference in predicted valories when *Fats* decreases from 30 to 20 is less when *Carbs* = 40 than when *Carbs* = 30.

4.10 (a) NH: $b_1 = b_2 = \cdots = b_5 = 0$ versus AH: at least one of b_1, b_2, ..., b_5 is not equal to zero.

(b) The test statistic calculation is as follows:

$$\begin{aligned}\text{global F-statistic} &= \frac{(\text{TSS}-\text{RSS})/k}{\text{RSS}/(n-k-1)} = \frac{(733.520-97.194)/5}{97.194/(200-5-1)} \\ &= 254.022.\end{aligned}$$

Since this is greater than the 95th percentile of the F-distribution with 5 numerator degrees of freedom and 194 denominator degrees of freedom (2.26), reject the null hypothesis in favor of the alternative. In other words, at least one of the predictor terms—*Freq*, *Amt*, *FreqAmt*, $Freq^2$, and Amt^2—is linearly associated with *Score*.

(c) NH: $b_3 = b_4 = b_5 = 0$ versus AH: at least one of b_3, b_4, or b_5 is not

equal to zero.

(d) The test statistic calculation is as follows:

$$\begin{aligned}\text{nested F-statistic} &= \frac{(\text{RSS}_\text{R}-\text{RSS}_\text{C})/(k-r)}{\text{RSS}_\text{C}/(n-k-1)} \\ &= \frac{(124.483-97.194)/(5-2)}{97.194/(200-5-1)} \\ &= 18.156.\end{aligned}$$

Since this is greater than the 95th percentile of the F-distribution with 3 numerator degrees of freedom and 194 denominator degrees of freedom (2.65), reject the null hypothesis in favor of the alternative. In other words, $FreqAmt$, $Freq^2$, and Amt^2 provide useful information about *Score* beyond the information provided by *Freq* and *Amt* alone.

(e) Use the complete model to predict *Score* (assuming that it passes the regression assumption checks) since it provides significantly more predictive power than the reduced model.

4.12 (a) Bachelors: $\text{E}(Salary) = b_0 + b_1(0) + b_2(0) + b_3\,YrsExp = b_0 + b_3\,YrsExp$.

Masters: $\text{E}(Salary) = b_0 + b_1(1) + b_2(0) + b_3\,YrsExp = (b_0 + b_1) + b_3\,YrsExp$.

Doctoral: $\text{E}(Salary) = b_0 + b_1(0) + b_2(1) + b_3\,YrsExp = (b_0 + b_2) + b_3\,YrsExp$.

(b) (i) b_1 represents the difference in $\text{E}(Salary)$ between Masters and Bachelors for a fixed *YrsExp*.

(ii) b_2 represents the difference in $\text{E}(Salary)$ between Doctoral and Bachelors for a fixed *YrsExp*.

(iii) b_3 represents the difference in $\text{E}(Salary)$ for a 1-year increase

in *YrsExp* for a fixed type of degree.

(c) If we reject the null hypothesis in favor of the alternative at a selected significance level, then there is a significant difference in E(*Salary*) between Masters and Bachelors or between Doctoral and Bachelors (or both) for a fixed *YrsExp*.

(d) The complete model is $\mathrm{E}(Salary) = b_0 + b_1 D_M + b_2 D_D + b_3\, YrsExp$. The reduced model is $\mathrm{E}(Salary) = b_0 + b_3\, YrsExp$. The nested F-statistic $= \frac{(\mathrm{RSS_R} - \mathrm{RSS_C})/(3-1)}{\mathrm{RSS_C}/(75-3-1)}$. Numerator degrees of freedom is $3 - 1 = 2$ and denominator degrees of frredom is $75 - 3 - 1 = 71$.

(e) $\mathrm{E}(Salary) = b_0 + b_1 D_M + b_2 D_D + b_3\, YrsExp + b_4 D_M\, YrsExp + b_5 D_D\, YrsExp$.

(f) NH: $b_4 = b_5 = 0$.

4.14 (a) $\mathrm{E}(Y) = b_0 + b_1 X$.

(b) $\mathrm{E}(Y) = b_0 + b_1 X + b_2 D_1 + b_3 D_2$, where D_1 and D_2 are indicator variables for two of the levels relative to the third (reference) level.

(c) $\mathrm{E}(Y) = b_0 + b_1 X + b_2 D_1 + b_3 D_2 + b_4 D_1 X + b_5 D_2 X$.

(d) When $b_4 = b_5 = 0$.

(e) When $b_2 = b_3 = b_4 = b_5 = 0$.

Chapter 5

5.2 (a) $\widehat{Price} = 44.21 + 47.42 D_{Ha} + 114.41 Floor$.

(b) Observation 40 has a studentized residual of 3.17 because $Price = 435$ is much higher than $\widehat{Price} = 302.0$.

(c) Observation 8 has a leverage of 0.237, which is greater than $3(k+1)/n = 3(2+1)/40 = 0.225$. The leverage is so high because this house has the smallest value of *Floor* by some distance.

(d) Observation 40 has the highest Cook's distance of 0.362, but this is

not greater than 0.5.

(e) $\widehat{Price} = 123.01 + 53.95D_{Ha} + 71.49Floor$.

(f) Observation 40 is influential enough that it should probably be omitted from the analysis of this dataset since the estimated regression equations are quite different with and without observation 40. This is despite the Cook's distance not exceeding 0.5.

5.4 (a) It would be inappropriate to fit a simple linear regression model with *Sales* as the response variable and *Time* as the predictor variable because this model cannot account for the strong seasonal pattern in sales that is evident in the scatterplot (not shown).

(b) A multiple linear regression model with *Sales* as the response variable and (D_1, D_2, D_3, *Time*) as the predictor variables is also inappropriate because there is a strong autocorrelation pattern evident in the residual plot (not shown).

(c) Compared with the residual plot from part (b), the residual plot for the model including *LagSales* shows no strong autocorrelation patterns, so including *LagSales* does appear to correct the autocorrelation problem.

(d) The prediction errors for the four quarters in 1999 are (−717, 165, −88, −549) for model (a), (−383, −160, −245, −380) for model (b), and (6, 110, −139, −211) for model (c), indicating that model (c) provides the best predictions overall.

5.6 (a) The scatterplot matrix indicates a high correlation between *Risk1* and *Risk2*, which could cause data-based multicollinearity problems in a multiple linear regression model with both *Risk1* and *Risk2* as predictor variables, including reduced precision of the estimated

regression parameters and increased difficulty interpreting the parameters.

(b) $\widehat{Dis} = -4.759 + 1.273Risk1 + 0.520Risk2$.

(c) The VIFs for *Risk1* and *Risk2* are both 9.43, so greater than 4. This indicates a potential multicollinearity problem. To mitigate data-based multicollinearity like this we could try removing one or more of the violating predictors from the regression model or try to collect additional data under different experimental or observational conditions.

(d) The scatterplot matrix indicates less correlation between the two predictor variables than the scatterplot matrix in part (a), which could mitigate the data-based multicollinearity problem evident in the model in part (b).

(e) $\widehat{DisFull} = -4.648 + 1.346Risk1Full + 0.403Risk2Full$.

(f) The VIFs for *Risk1Full* and *Risk2Full* are both 1.37, so less than 4, which indicates that the data-based multicollinearity problem has been mitigated.

(g) $\widehat{DisFull} = -1.520 + 0.025Risk1Full + 0.135Risk1Full^2 + 0.400Risk2Full$.

(h) The VIFs for *Risk1Full* and $Risk1Full^2$ are 78.41 and 78.39, respectively, so greater than 10. To mitigate structural multicollinearity like this we could try centering the *Risk1Full* predictor.

(i) $\widehat{DisFull} = 1.999 + 1.381Risk1FullC + 0.135Risk1FullC^2 + 0.400Risk2Full$.

(j) The VIFs for *Risk1FullC* and $Risk1FullC^2$ are 1.42 and 1.05, respectively, so less than 4, which indicates that the structural multicollinearity problem has been mitigated.

(k) $\widehat{DisFull} = -1.520 + 0.025(4) + 0.135(4^2) + 0.400(6) = 3.14$.

$\widehat{DisFull} = 1.999 + 1.381(-1.01) + 0.135((-1.01)^2) + 0.400(6) = 3.14.$

The predictions are the same because the fitted values are unchanged after centering a predictor.

5.8 (a) $s = 0.7656$, $R^2 = 0.8835$, and adjusted $R^2 = 0.8573$.

(b) Nested model F-test p-value is 0.7995.

(c) $s = 0.7448$, $R^2 = 0.8788$, and adjusted $R^2 = 0.8650$.

(d) RMSE under the first model is 5.46, while RMSE under the second model is 5.19.

5.10 (a) AIC values are (X_1): 122.8; (X_4): 129.0; (X_1, X_2): 123.2; (X_3, X_4): 122.2; (X_1, X_2, X_3): 106.1; (X_1, X_3, X_4): 105.7; (X_1, X_2, X_3, X_4): 106.8. The model with X_1, X_3, and X_4 is "best" under this criterion.

(b) Forward selection would not find the model with X_1, X_3, and X_4 since it would start by adding X_1, then X_2, then X_3.

5.12 (a) The nested F-statistic = 42.294 and p-value = 0.000 suggest that model A is relatively poor and model B is relatively good. Model A is poor because it doesn't include the D_1X_2 interaction, which is very significant in model B (individual p-value = 0.000).

(b) The nested F-statistic = 0.9473 and p-value = 0.3916 suggest that model D is relatively poor and model C is relatively good. Model D is poor because it includes X_4 and X_5 in addition to X_6, and these three predictors are collinear (individual p-values for X_4, X_5, and X_6 in model D are 0.174, 0.183, and 0.374, respectively, but the individual p-value for X_6 in model C is 0.000)—see also part (h).

(c) The nested F-statistic = 26.435 and p-value = 0.000 suggest that model E is relatively poor and model F is relatively good. Model E is poor because it doesn't include the X_3^2 transformation, which is

very significant in model F (individual p-value = 0.000).

(d) A: X_2 effect = $8.19 - 0.05X_2$ (for $D_1 = 0$), $5.50 - 0.05X_2$ (for $D_1 = 1$).

B: X_2 effect = $5.98 + 0.79X_2$ (for $D_1 = 0$), $8.18 - 1.09X_2$ (for $D_1 = 1$).

C: X_2 effect = $5.98 + 0.78X_2$ (for $D_1 = 0$), $8.20 - 1.10X_2$ (for $D_1 = 1$).

F: X_2 effect = $5.94 + 0.80X_2$ (for $D_1 = 0$), $8.13 - 1.08X_2$ (for $D_1 = 1$).

The X_2 predictor effects are very different for model A.

(e) B: X_3 effect = $6.76+2.78X_3-1.28X_3^2$ ($D_1 = 0$), $4.02+2.78X_3-1.28X_3^2$ ($D_1 = 1$).

C: X_3 effect = $6.76+2.76X_3-1.26X_3^2$ ($D_1 = 0$), $4.03+2.76X_3-1.26X_3^2$ ($D_1 = 1$).

E: X_3 effect = $9.44 - 1.30X_3$ (for $D_1 = 0$), $6.87 - 1.30X_3$ (for $D_1 = 1$).

F: X_3 effect = $6.77+2.74X_3-1.26X_3^2$ ($D_1 = 0$), $4.03+2.74X_3-1.26X_3^2$ ($D_1 = 1$).

The X_3 predictor effects are very different for model E.

(f) A: X_4 effect = $6.80 + 0.43X_4$ (for $D_1 = 0$), $4.11 + 0.43X_4$ (for $D_1 = 1$).

B: X_4 effect = $6.58 + 0.50X_4$ (for $D_1 = 0$), $3.84 + 0.50X_4$ (for $D_1 = 1$).

D: X_4 effect = $3.77 + 1.46X_4$ (for $D_1 = 0$), $1.03 + 1.46X_4$ (for $D_1 = 1$).

F: X_4 effect = $6.54 + 0.51X_4$ (for $D_1 = 0$), $3.80 + 0.51X_4$ (for $D_1 = 1$).

The X_4 predictor effects are very different for model D.

(g) A: X_5 effect = $6.95 + 0.44X_5$ (for $D_1 = 0$), $4.26 + 0.44X_5$ (for $D_1 = 1$).

B: X_5 effect = $6.87 + 0.47X_5$ (for $D_1 = 0$), $4.13 + 0.47X_5$ (for $D_1 = 1$).

D: X_5 effect = $4.53 + 1.39X_5$ (for $D_1 = 0$), $1.79 + 1.39X_5$ (for $D_1 = 1$).

F: X_5 effect = $6.86 + 0.47X_5$ (for $D_1 = 0$), $4.11 + 0.47X_5$ (for $D_1 = 1$).

The X_5 predictor effects are very different for model D.

(h) C: X_6 effect = $6.46 + 0.49X_6$ (for $D_1 = 0$), $3.72 + 0.49X_6$ (for $D_1 = 1$).

D: X_6 effect = $11.09 - 0.94X_6$ (for $D_1 = 0$), $8.35 - 0.94X_6$ (for $D_1 = 1$).

The slope part of the predictor effect for X_6 in model C (0.49) is approximately the same as the average of the slope parts of the predictor effects for X_4 and X_5 in model F $((0.51+0.47)/2 = 0.49)$. The X_6 predictor effect is different in models C and D.

Chapter 7

7.2 (a) $\log\left(\frac{\Pr(Offer=1)}{1-\Pr(Offer=1)}\right) = -41.31161 + 0.25713 Exam + 5.97117 Gpa$.

(b) $\exp(\hat{b}_1) = \exp(0.25713) = 1.293$, which means that we estimate the odds of an offer increases by 1.293 times when *Exam* increases by 1 and *Gpa* is fixed.

(c) $\frac{\Pr(Offer=1)}{1-\Pr(Offer=1)} = \exp(-41.31161 + 0.25713(80) + 5.97117(3.6))$
$= \exp(0.755002) = 2.128$.

(d) $\frac{\Pr(Offer=1)}{1-\Pr(Offer=1)} = \exp(-41.31161 + 0.25713(81) + 5.97117(3.6))$
$= \exp(1.012132) = 2.751$.

(e) $2.751/2.128 = 1.293$.

(f) $\Pr(Offer = 1) = \exp(-41.31161 + 0.25713(80) + 5.97117(3.6))/(1 + \exp(-41.31161 + 0.25713(80) + 5.97117(3.6))) = 2.128/(1 + 2.128) = 0.680$.

(g) Non-offers: 39 correctly predicted, 5 predicted to be offers.
Offers: 21 correctly predicted, 5 predicted to be non-offers.

Appendix A

A.2 (a) Histogram of *Vol* (not shown).

(b) The histogram shows interior passenger and cargo volumes ranging between 50 and 180 cubic feet, with many values tending to cluster around 90–130 cubic feet rather than at the extremes. The

distribution seems relatively symmetric.

(c) Mean =1.10 (110 cubic feet), median =1.11 (111 cubic feet).

(d) For symmetric data such as these, the mean has nice technical properties that make it a more appropriate measure of the center. By contrast, for highly skewed data, the median is a better summary than the mean of the central tendency of the data. For example, when a dataset has a few very large values, this can cause the mean to be relatively large in comparison to the median (which is not affected to the same extent by these large values). Since the mean and median are quite close together in this case, any skewness in these data is practically nonexistent.

(e) 25th percentile (first quartile) =1.00 (100 cubic feet);
50th percentile (second quartile, or median) =1.11 (111 cubic feet);
75th percentile (third quartile) =1.20 (120 cubic feet).

(f) The middle 50% of the volumes fall between 1.00 (100 cubic feet) and 1.20 (120 cubic feet).

A.4 (a) Cross-tabulation (not shown).

(b) There are 36 front-wheel drive midsize cars.

(c) 43% of midsize cars are front-wheel drive.

(d) 28% of front-wheel drive vehicles are midsize cars.

(e) The most common vehicle class is midsize car (83 vehicles), mostly front-wheel drive (43%) or all-wheel drive (30%). The next most common classes are compact, large, and subcompact cars, with compact cars mostly front-wheel drive (54%), and large and subcompact cars mostly rear-wheel drive (50% and 51%, respectively). Station wagons and minicompacts/two-seater cars are the next most fre-

quent, with station wagons mostly front-wheel drive (50%), and minicompacts/two-seater cars mostly rear-wheel drive (57%). Finally, there are just 20 sport utility vehicles, mostly all-wheel drive (40%) and four-wheel drive (30%). The most common front-wheel drive vehicles are compact and midsize cars (28% each), while the most common rear-wheel drive vehicles are subcompact and large cars (25% each). The most common all-wheel drive vehicles are midsize cars (30%), while the most common four-wheel drive vehicles are minicompacts/two-seater cars and sport utility vehicles (30% each).

A.6 (a) Scatterplot matrix (not shown).

(b) There are fairly strong negative associations between *Cmpg* and each of *Eng* and *Cyl*, but the associations appear curved—steeper for low values of *Eng* and *Cyl* and becoming shallower as *Eng* and *Cyl* each increase. There doesn't appear to be much of an association between *Cmpg* and *Vol* other than a slight tendency for the few vehicles with very large volumes to have lower fuel efficiencies. There is a positive, reasonably linear association between *Eng* and *Cyl*, but no clear association between *Eng* and *Vol* or between *Cyl* and *Vol*. A few vehicles "stick out" from the dominant patterns in the plots. For example, vehicle 29 (Bentley Mulsanne 6.8L) with 8 cylinders has a particularly high value of *Eng*.