

Appendix—Answers to Selected Problems

Chapter 3

2. nominal, ordinal, interval, ratio
3. a. 0.8413 b. -0.842
4. a. 18.475 b. 0.699
5. a. -1.350 b. 0.05
6. a. 2.51 b. 0.025
7. a. 0 b. 0 c. 0
8. standard normal
9. a. 3.0 b. 3 c. 2.8 or 3.11
10. e
11. a. 5.0 b. (187.44, 192.56)
12. $t_{0.975,27} = 2.052$
13. (24.66, 35.33)
14. 3.6858
15. a. significant difference b. significant difference
16. nonsignificant difference
17. $t_{0.995,10} = 3.619$
18. b
19. a. Type I error b. correct decision c. correct decision d. Type II error
20. a, b
21. c
22. b
23. a. $1 - \alpha$ b. α c. β d. $1 - \beta$
24. Accept H_0 .
25. b

Chapter 5

1. a. Dry Weight (Y) does increase with increasing Age (X), but the relationship may not be linear. An exponential relationship between X and Y may better fit the data. Log Dry Weight (Z) increases linearly with increasing Age (X).
 - b. $Y = \beta_0 + \beta_1 X + E$ $Z = \beta'_0 + \beta'_1 X + E$
 - c. $\hat{Y} = -1.885 + 0.235X$ $\hat{Z} = -2.689 + 0.196X$
 - d. The regression line for Log_{10} Dry Weight regressed on Age has a better fit. It is more appropriate to run a linear regression of Z on X .
 - e. 95% confidence intervals: for β'_1 : (0.190, 0.202), for β'_0 : (-2.759, -2.620)
 - f. (-1.149, -1.096)
3. a. The relationship between Time (Y) and Inc (X) does not appear to be linear.
 - b. $\hat{\beta}_0 = 19.626$ $\hat{\beta}_1 = 0.0007$
 - c. $\hat{Y} = 19.626 + 0.0007X$. The regression line fits the data poorly.
 - d. The linearity assumption is not met.
 - e. $T = 2.023$, $P = 0.0582$ (from SAS output). We do not reject H_0 , since $P\text{-value} > 0.05$.
 - f. The scatter plot suggests that a parabola would better fit the data.
5. a. $\hat{Y} = 2.174 + 1.177X$. The line fits the data well.
 - b. No.
 - c. $T = 13.5$ $P < 0.0001$ (from SAS output).
Since $P\text{-value} < 0.05$, we reject H_0 .
 - d. $T = 0.954$ $0.15 < P < 0.25$.
Since $P\text{-value} > 0.05$, we do not reject H_0 .
 - e. (44.146, 47.276)
7. a. $\hat{Y}_1 = -122.345 + 6.227X$ $\hat{Y}_2 = -1.697 + 0.299X$
 - b. Y_2 regressed on X .
 - c. $T = -57.934$. Critical value: $t_{17} \sim 2.898$ under H_0 at $\alpha = 0.01$. We see that $|T| = 57.934 > 2.898$, so we reject H_0 at $\alpha = 0.01$.
 - d. (0.264, 0.334)
 - e. (11.18, 12.32)
9. a. $\hat{\beta}_0 = 2.936$ $\hat{\beta}_1 = -1.785$
 - b. $\hat{Y} = 2.936 - 1.785X$. The line fits the data well.
 - d. $\hat{Y} = 862.979\hat{X}^{-1.785}$
 - e. (6.266, 9.311), (321.366, 528.445)
 - f. One could plot the transformed data (X' , Y') and then draw the estimated regression line on the plot. Then one could compare the fit of the estimated line of (X , Y) versus that of (X' , Y'). If one did this comparison, the straight-line regression of (X , Y) gives a better fit.
11. a. $\hat{Y} = 3.707 - 0.012X$.
 - b. $T = -8.684$ $P = 0.001$ (from SAS output).
Since $P\text{-value} < 0.05$, we reject H_0 .
 - c. Including the data from the three experiments, rather than just using the average values, would provide more information and might improve the sensitivity of the analysis.
 - d. (1.243, 3.737)

- e. It is inappropriate, since the estimated model relies on data that do not include any information for average growth rate when exposed to a gas with a molecular weight of 200.
- f. The chosen X values are not uniformly distributed in the experiment. There are large gaps between the X values of 39.9, 83.8, and 131.3. This may result in a fitted line subject to inaccuracies for predicting Y based on X .
13. b. From the SAS output: $\hat{\beta}_0 = 0.116$ $\hat{\beta}_1 = 0.005$
- c. $T = 0.811$ $P = 0.433$ (from SAS output).
Since P -value > 0.10 , we do not reject H_0 .
- d. $T = 0.046$ $P = 0.964$ (from SAS output).
Since P -value > 0.10 , we do not reject H_0 .
- e. $\hat{Y} = 0.116 + 0.005X$
- f. The line does not differ from the line plotted in part (e). The evidence suggests that there is no significant linear relationship. Determining a well-fitting line is difficult, given the dispersion of the data.
15. a. As X increases, the dispersion of Y increases.
- b. $\hat{\beta}_0 = -2.546$ $\hat{\beta}_1 = 0.032$ $\hat{Y} = -2.546 + 0.032X$
- c. This chart implies a better linear relationship between X and Y .
- d. $\hat{\beta}_0 = -6.532$ $\hat{\beta}_1 = 1.430$ $\hat{Y} = -6.532 + 1.430X$
- e. The natural log transformation provides the best representation. The natural log plot better illustrates the linear relationship and the dispersion of the data is more similar at each level of toluene exposure. The first plot indicates that there may be a violation of homoscedasticity for the untransformed data.
17. a. Yes.
- b. $\hat{Y} = 1.643 + 1.057X$. The line appears to fit the data well.
- c. 95% CI for β_1 : (0.580, 1.534). The 95% CI does not include the null value of zero, indicating that there is a significant linear relationship at $\alpha = 0.05$.
- d. No, it is not appropriate, since the data used to estimate the regression line do not include a \$10 million advertising expenditure in its range.
19. a. There may be a slight negative linear relationship.
- b. $Y = \beta_0 + \beta_1X + E$ $\hat{\beta}_0 = 76.008$ $\hat{\beta}_1 = -0.015$.
The baseline OWNEROCC = 76%, and as OWNCOST increases by \$1,000, the percentage of OWNEROCC decreases by ~2%.
- c. $\hat{Y} = 76.008 - 0.015X$. The line fits the data well.
- d. $T = -2.607$ $P = 0.0155$ (from SAS output).
Since P -value < 0.05 , we reject H_0 .
- e. (-0.027, -0.003). We are 95% confident that the true slope is between -0.027 and -0.003. Since the interval does not contain zero, we conclude that the slope is not equal to zero (at $\alpha = 0.05$).

Chapter 6

1. a. (1) $r = 0.86$ (2) $r = 0.999$
- b. (1) (0.546, 0.964) (2) (0.996, 1.000)
- c. (1) $r^2 = 0.744$, so 74% of the variation in Y is explained with the help of X .
(2) $r^2 = 0.998$, so 99% of the variation in Z is explained with the help of X .
- d. The regression using Log_{10} Dry Weight appears to fit better. This agrees with Chapter 5, Problem 1(d).

3. a. $r = 0.980$
 b. Substitute $r(S_Y/S_X)$ and $\frac{(n-1)}{(n-2)}(S_Y^2 - \hat{\beta}_1^2 S_X^2)$ for $\hat{\beta}_1$ and $S_{Y|X}^2$ in formula (5.9).
 c. $T' = 13.929$; $T \sim t_8$ under $H_0: \rho = 0$. The P -value for this test is less than 0.001; therefore, we reject H_0 .
 d. The graph of Y vs X does not illustrate a linear relationship.
5. a. $r^2 = 0.601$, $r = 0.775$. 60% of the variation in SBP (Y) is explained by AGE (X).
 b. (0.504, 0.907). Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = 0.01$.
7. a. $r^2 = 0.1853$, $r = 0.430$. 19% of the variation in SBP (Y) is explained by AGE (X).
 b. $T = 2.02$; Critical value: $t_{18,0.975} = 2.101$. At $\alpha = 0.05$, since $|T| <$ critical value, we would not reject H_0 .
 c. (-0.015, 0.733). Since $\rho = 0$ is included in the interval, we do not reject $H_0: \rho = 0$ at $\alpha = 0.05$.
9. a. $r^2 = 0.9101$, $r = 0.954$. 91% of the variation in Y is explained by X .
 b. $T = 13.49$; Critical value: $t_{18,0.975} = 2.101$. At $\alpha = 0.05$, since $|T| >$ critical value, we would reject H_0 .
 c. (0.885, 0.982). Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = 0.05$.
11. a. $r^2 = 0.9728$, $r = 0.986$, so 98% of the variation in Y_2 is explained by X .
 b. $T = 24.640$; Critical value: $t_{17,0.975} = 2.110$. At $\alpha = 0.05$, since $|T| >$ critical value, we would reject H_0 .
 c. (0.963, 0.995). Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = 0.05$.
13. a. $r = 0.971$
 b. (0.813, 0.996). Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = 0.05$.
15. $Z = 3.62$; Critical value = 1.96. Since $|Z|$ is $>$ critical value, we reject the null hypothesis.
17. a. $r^2 = 0.9558$, $r = 0.978$. 96% of the variation in Y is explained by X . *includes 16th obs.
 b. $T = 17.424$; Critical value: $t_{14,0.975} = 2.145$. At $\alpha = 0.05$, we would reject H_0 .
 c. (0.936, 0.993). Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = 0.05$.
19. a. $r^2 = 0.0310$, $r = -0.176$. 3% of the variation in Y is explained by X .
 b. $T = 0.759$; Critical value: $t_{18,0.975} = 2.101$. At $\alpha = 0.05$, we would not reject H_0 .
 c. (-0.574, 0.289). Since $\rho = 0$ is included in the interval, we do not reject $H_0: \rho = 0$ at $\alpha = 0.05$.
21. a. $r^2 = 0.9813$, $r = 0.991$. 98% of the variation in Y is explained by X .
 b. $T = 55.088$; Critical value: $t_{38,0.975} = 2.0$. At $\alpha = 0.05$, we would reject H_0 .
 c. (0.985, 0.995). Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = 0.05$.
23. a. $r^2 = 0.9044$, $r = 0.951$. 90% of the variation in Y is explained by X .
 b. $T = 6.155$; Critical value: $t_{4,0.975} = 2.776$. At $\alpha = 0.05$, we would reject H_0 .
 c. (0.611, 0.995). Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = 0.05$.

25. a. $r^2 = 0.2207, r = 0.470$. 22% of the variation in Y is explained by X .
 b. $T = 2.608$; Critical value: $t_{24,0.975} = 2.064$. At $\alpha = 0.05$, we would reject H_0 .
 c. $(0.101, 0.725)$. Since $\rho = 0$ is not included in the interval, we reject $H_0: \rho = 0$ at $\alpha = 0.05$.

Chapter 7

1. a. (1)

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	6.0785	6.0785	26.18
Residual	9	2.0896	0.2322	

a. (2)

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	4.2211	4.2211	5355.59
Residual	9	0.0071	0.0008	
	10	4.2282		

b. (1) $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; Critical value: $F_{1,9,0.95} = 5.12$. Since $26.18 > 5.12$, we reject H_0 at $\alpha = 0.05$.

b. (2) $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; Critical value: $F_{1,9,0.95} = 5.12$. Since $5355.59 > 5.12$, we reject H_0 at $\alpha = 0.05$.

5. a.

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	450.8673	450.8673	4.09
Residual	18	1982.9141	110.1619	
	19	2433.7814		

b. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; Critical value: $F_{1,18,0.95} = 4.41$. $4.09 < 4.41$, so we do not reject H_0 at $\alpha = 0.05$.

c. $T^2 = (2.023)^2 = 4.09$. The values are the same.

d. The hypotheses for each test are equivalent. As mentioned in the text, the F statistic and T statistic are equivalent after squaring T . Using the information that the tests of hypotheses are equivalent (as are the test statistics), one may infer that the resulting P -values for each test should also be equivalent.

9. a.

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	76858486.06	76858486.06	60.76
Residual	28	35419546.54	1264983.00	
	29	112278032.60		

b. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; Critical value: $F_{1,28,0.95} = 4.20$. Since $60.76 > 4.20$, we reject H_0 at $\alpha = 0.05$.

13. a.

	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	12846354	12846354	302.99
Residual	14	593585	42399	
	15	13439939		

b. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; Critical value: $F_{1,14,0.95} = 4.60$

Since $302.99 > 4.60$, we would reject H_0 and conclude that there is a significant linear relationship of Y on X at $\alpha = 0.05$.

17. a.	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	177.5297	177.5297	3045.56
Residual	58	3.3809	0.0583	
	59	180.9106		

b. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; Critical value: $F_{1,58,0.95} = 4.00$. Since $3045.56 > 4.00$, we reject H_0 at $\alpha = 0.05$.

21. a.	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	1	132.6203	132.6203	6.80
Residual	24	468.3412	19.5142	
	25	600.9615		

b. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$; Critical value: $F_{1,24,0.95} = 4.26$. Since $6.80 > 4.26$, we reject H_0 at $\alpha = 0.05$.

Chapter 8

1. a. i $\hat{Y} = 145.771$ ii $\hat{Y} = 135.825$ iii $\hat{Y} = 141.475$

As QUET increases from 3.0 to 3.5, average SBP increases by an estimated 4.296 points, from 141.475 to 145.771.

b. SBP on AGE: $R^2 = 0.601$; SBP on AGE and SMK: $R^2 = 0.730$; SBP on AGE, SMK, and QUET: $R^2 = 0.761$.

The model using AGE and SMK to predict SBP appears to be the best choice.

5. a.	d.f.	Sum of Squares	Mean Sum of Squares	F
Regression	3	25974.40	8658.00	80.87
Residual	21	2248.23	107.06	
Total	24	28222.63		

b. $R^2 = 0.920$. There is a strong positive linear relationship between education resources and student performance.

9. a. As temperature increases from 20 to 25, the average oxygen consumption increases by an estimated 0.197 units.

b. As weight increases from 0.25 to 0.5, the average oxygen consumption increases by an estimated 0.148 units.

c. i $R^2 = 0.019$ ii $R^2 = 0.814$ iii $R^2 = 0.943$

13. a. $\hat{Y} = 6.874 - 0.004X_2 - 0.234X_3$ b. $\hat{Y} = 3.734$

c. $R^2 = 0.553$. The model explains about half of the variation in Yield (Y). The model has a limited ability to predict the yield for a company using 1989 ranking and P-E ratio as predictors.

15. a. $\hat{Y} = 130.468 + 0.089 \text{ zooplankton} - 0.025 \text{ phytoplankton}$

b. $R^2 = 0.1332$. The model fit is poor.

Chapter 9

1. a. i $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + E$)
 $F(X_1)_{1,30} = 45.18, P = 0.0001$. At $\alpha = 0.05$, we reject H_0 .
- ii $H_0: \beta_1 = \beta_2 = 0$ $H_A: \text{at least one } \beta_i \neq 0 (i = 1, 2)$
 (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_1, X_2)_{2,29} = 39.16, P < 0.0001$. At $\alpha = 0.05$, we reject H_0 .
- iii $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ $H_A: \text{at least one } \beta_i \neq 0 (i = 1, 2, 3)$
 (Full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$).
 $F(X_1, X_2, X_3)_{3,28} = 29.71, P < 0.0001$. At $\alpha = 0.05$, we reject H_0 .
- b. One would choose the parsimonious model tested in (a(i)).
5. a. i $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + E$)
 $F(X_1)_{1,40} = 0.40, P > 0.25$. At $\alpha = 0.05$, we do not reject H_0 .
- ii $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ (Full model: $Y = \beta_0 + \beta_2 X_2 + E$)
 $F(X_2)_{1,40} = 0.19, P > 0.25$. At $\alpha = 0.05$, we do not reject H_0 .
- iii $H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$ (Full model: $Y = \beta_0 + \beta_3 X_3 + E$)
 $F(X_3)_{1,40} = 7.58, P = 0.009$. At $\alpha = 0.05$, we reject H_0 .
- b. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ $H_A: \text{at least one } \beta_i \neq 0 (i = 1, 2, 3)$
 (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$)
 $F(X_1, X_2, X_3)_{3,38} = 2.74, 0.05 < P < 0.10$. At $\alpha = 0.05$, we do not reject H_0 .
- c. $H_0: \beta_4 = \beta_5 = 0$ $H_A: \text{at least one } \beta_i \neq 0 (i = 4, 5)$
 (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3 + \beta_5 X_2 X_3 + E$)
 $F(X_1 X_3, X_2 X_3 | X_1, X_2, X_3)_{2,36} = 0.362, P > 0.25$. At $\alpha = 0.05$, we do not reject H_0 .
- d. $H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$ (Full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E$).
 $F(X_3 | X_1, X_2)_{1,38} = 6.85, 0.01 < P < 0.025$. At $\alpha = 0.05$, we reject H_0 .
- e. X_3 is associated with Y , but the other two independent variables are not.
9. a. $H_0: \beta_1 = \beta_2 = 0$ $H_A: \text{at least one } \beta_i \neq 0 (i = 1, 2)$
 (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_1, X_2)_{2,4} = 20.03, P = 0.0082$. At $\alpha = 0.05$, we reject H_0 .
- b. i $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ in the model $Y = \beta_0 + \beta_1 X_1 + E$.
 $F(X_1)_{1,5} = 40.06, P = 0.0032$. At $\alpha = 0.05$, we reject H_0 . Note: We use the residual MS from the largest model. See Section 9.5.2.
- ii $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ in the model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$.
 $F(X_2 | X_1)_{1,4} = 0.01, P = 0.9344$. At $\alpha = 0.05$, we do not reject H_0 .
- c. i $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ in the model $Y = \beta_0 + \beta_2 X_2 + E$.
 $F(X_2)_{1,5} = 37.83, 0.001 < P < 0.005$. At $\alpha = 0.05$, we reject H_0 .
- ii $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_1 | X_2)_{1,4} = 9.80, P = 0.0352$. At $\alpha = 0.05$, we reject H_0 .
- d. Source | d.f. | SS | MS | F | R^2
 $X_1 | X_2$ | 1 | 1402.315 | 1402.315 | 9.8 | 0.91
 $X_2 | X_1$ | 1 | 1.098 | 1.098 | 0.01 |
 Residual | 4 | 572.393 | 143.098 |
- e. X_1 is the only necessary predictor.

13. a. $H_0: \beta_1 = \beta_2 = 0$ H_A : at least one $\beta_i \neq 0$ ($i = 1, 2$)
 ($X_1 = \text{OWNCOST}, X_2 = \text{URBAN}$)
 (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_1, X_2)_{2,23} = 8.52, P = 0.017$. At $\alpha = 0.05$, we reject H_0 .
- b. i $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + E$)
 $F(X_1)_{1,24} = 8.84, P = 0.0068$. At $\alpha = 0.05$, we reject H_0 .
- ii $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_2 | X_1)_{1,23} = 8.21, P = 0.0088$. At $\alpha = 0.05$, we reject H_0 .
- c. i $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ (Full model: $Y = \beta_0 + \beta_2 X_2 + E$)
 $F(X_1)_{1,24} = 13.17, 0.001 < P < 0.005$. At $\alpha = 0.05$, we reject H_0 .
- ii $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_1 | X_2)_{1,23} = 4.42, P = 0.0467$. At $\alpha = 0.05$, we reject H_0 .

d. Source	d.f.	SS	MS	F	R ²
$X_1 X_2$	1	66.334	66.334	4.42	0.43
$X_3 X_2$	1	123.158	123.158	8.21	
Residual	23	345.183	15.008		

e. Both predictors are necessary.

Chapter 10

1. a. Age with an r value of 0.7752.
- b. i $r_{\text{SBP,SMK}|\text{AGE}} = 0.568$ ii $r_{\text{SBP,QUET}|\text{AGE}} = 0.318$
- c. $H_0: \rho_{\text{SBP,SMK}|\text{AGE}} = 0$ $H_A: \rho_{\text{SBP,SMK}|\text{AGE}} \neq 0$
 $F(\text{SMK} | \text{AGE})_{1,29} = 13.83, P < 0.001$. At $\alpha = 0.05$, we reject H_0 .
- d. $H_0: \rho_{\text{SBP,QUET}|\text{AGE,SMK}} = 0$ $H_A: \rho_{\text{SBP,QUET}|\text{AGE,SMK}} \neq 0$
 $T_{28} = 1.91, P = 0.066$. At $\alpha = 0.05$, we do not reject H_0 .
- e. Based on the results for a–d, we find that the following variables (ranked in order of their significance) helped explain the variation in SBP: (1) AGE, (2) SMK, (3) QUET.
- f. $r_{\text{SBP}(\text{QUET,SMK})|\text{AGE}}^2 = 0.401$
 $H_0: \rho_{\text{SBP}(\text{QUET,SMK})|\text{AGE}} = 0$ $H_A: \rho_{\text{SBP}(\text{QUET,SMK})|\text{AGE}} \neq 0$
 $F(\text{QUET, SMK} | \text{AGE})_{2,28} = 9.371, P < 0.001$.
 The highly significant P -value suggests that both SMK and QUET are important variables, but there is room for debate since the increase in r^2 going from model 1, with only AGE (0.601), to model 3, with all 3 variables (0.761), is small (0.160).
5. a. i $r_{YX_1|X_3}^2 = 0.076$. ii $r_{YX_2|X_3}^2 = 0.214$.
 iii $r_{YX_2|X_3}^2 = 0.215$. The computations are nearly equivalent with the difference being due to round-off error.
- b. X_2 should be considered next for entry into the model because X_2 has a higher partial correlation than does X_1 .
- c. $H_0: \rho_{YX_2|X_3} = 0$ $H_A: \rho_{YX_2|X_3} \neq 0$
 $T_{17} = 2.154, 0.02 < P < 0.05$. At $\alpha = 0.05$, we reject H_0 .
- d. $r_{YX_1|X_2,X_3}^2 = 0.082$
 $H_0: \rho_{YX_1|X_2,X_3} = 0$ $H_A: \rho_{YX_1|X_2,X_3} \neq 0$
 $T_{16} = 1.199, P = 0.248$. At $\alpha = 0.05$, we do not reject H_0 .

- e. $r_{Y(X_1, X_2)|X_3}^2 = 0.279$
 $H_0: \rho_{Y(X_1, X_2)|X_3} = 0 \quad H_A: \rho_{Y(X_1, X_2)|X_3} \neq 0$
 $F(X_1, X_2|X_3)_{2,16} = 3.098, 0.05 < P < 0.10$. At $\alpha = 0.05$, we do not reject H_0 .
- f. Based on the above results, only X_3 should be included in the model at $\alpha = 0.05$.
9. a. i $H_0: \rho_{YX_1} = 0 \quad H_A: \rho_{YX_1} \neq 0$
 $F(X_1)_{1,45} = 0.89, P = 0.35$. At $\alpha = 0.05$, we do not reject H_0 .
 ii $H_0: \rho_{YX_2} = 0 \quad H_A: \rho_{YX_2} \neq 0$
 $F(X_2)_{1,45} = 197.58, P < 0.0001$. At $\alpha = 0.05$, we reject H_0 .
- b. i $H_0: \rho_{YX_1|X_2} = 0 \quad H_A: \rho_{YX_1|X_2} \neq 0$
 $F(X_1, X_2)_{1,44} = 98.605, P < 0.001$. At $\alpha = 0.05$, we reject H_0 .
 ii $H_0: \rho_{YX_2|X_1} = 0 \quad H_A: \rho_{YX_2|X_1} \neq 0$
 $F(X_2|X_1)_{1,44} = 709.641, P < 0.001$. At $\alpha = 0.05$, we reject H_0 .
- c. Both X_1 and X_2 should be included in the model with X_2 being more important than X_1 .
13. a. $r_{Y|X_1, X_2}^2 = 0.909$ b. $r_{YX_2|X_1} = -\sqrt{0.002} = -0.045$
 c. $r_{YX_1|X_2} = \sqrt{0.710} = 0.843$
 d. $H_0: \rho_{YX_2|X_1} = 0 \quad H_A: \rho_{YX_2|X_1} \neq 0$
 $T_4 = -0.09, P > 0.90$. At $\alpha = 0.05$, we do not reject H_0 .
 e. $H_0: \rho_{YX_1|X_2} = 0 \quad H_A: \rho_{YX_1|X_2} \neq 0$
 $T_4 = 3.131, 0.02 < P < 0.05$. At $\alpha = 0.05$, we reject H_0 .
 f. X_1 should be included in the model, while X_2 should not be included.
17. a. $r_{Y|X_1, X_2}^2 = 0.426$ b. $r_{YX_2|X_1} = -\sqrt{0.263} = -0.513$
 c. $r_{YX_1|X_2} = \sqrt{0.161} = 0.401$
 d. $H_0: \rho_{YX_2|X_1} = 0 \quad H_A: \rho_{YX_2|X_1} \neq 0$
 $T_{23} = -2.865, 0.001 < P < 0.01$. At $\alpha = 0.05$, we reject H_0 .
 e. $H_0: \rho_{YX_1|X_2} = 0 \quad H_A: \rho_{YX_1|X_2} \neq 0$
 $T_{23} = -2.1, 0.02 < P < 0.05$. At $\alpha = 0.05$, we reject H_0 .
 f. Both variables should be included in the model with X_2 being the more important predictor of Y .

Chapter 11

1. a. $WGT = \beta_0 + \beta_1HGT + \beta_2AGE + \beta_3AGE^2 + E$
 b. $\hat{\beta}_1$ does not change when either AGE or AGE² are removed from the model. However, $\hat{\beta}_1$ changes "significantly" when both AGE and AGE² are removed from the model. Thus, there is confounding due to AGE and AGE².
 c. AGE² can be dropped from the model because $\hat{\beta}_1$ does not change significantly.
 d. AGE² should not be retained in the model, because the 95% CI for β_1 is narrower when AGE² is absent from the model.
 e. Considering the change in $\hat{\beta}_1$ and the width of the 95% CI, the final model should be $WGT = \beta_0 + \beta_1HGT + \beta_2AGE + E$.
 f. Revise the initial model as $WGT = \beta_0 + \beta_1HGT + \beta_2AGE + \beta_3AGE^2 + \beta_4HGT*AGE + \beta_5HGT*AGE^2 + E$.
 g. We would test for interaction by performing a multiple-partial F test for $H_0: \beta_4 = \beta_5 = 0$. If this test proved significant, we would perform separate partial F tests to assess $H_0: \beta_4 = 0$ and $H_0: \beta_5 = 0$.

n order of their
 QUET.

mportant vari-
 el 1, with only

rence

er partial cor-

3. a. There is no confounding due to X_2 , because $\hat{\beta}_1$ does not change when X_2 is removed from the model.
- b. $r_{YX_1} = 0.265$ $r_{YX_1|X_2} = \sqrt{0.5} = 0.707$
 Since the two correlation coefficients are significantly different, we conclude that confounding exists.
- c. The conclusions for confounding depend on the definition of confounding.
- d. Since $H_0: \beta_2 = 0$ is rejected ($P = 0.0005$), we conclude that confounding exists, which is contradictory to part (a).
5. a. i $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ $F(X_1)_{1,40} = 0.4, P = 0.528$. At $\alpha = 0.05$, we do not reject H_0 .
- ii $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ $F(X_2)_{1,40} = 0.19, P = 0.669$. At $\alpha = 0.05$, we do not reject H_0 .
- iii $H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$ $F(X_3)_{1,40} = 7.58, P = 0.009$. At $\alpha = 0.05$, we reject H_0 .
- b. $H_0: \beta_1 = \beta_2 = \beta_3 = 0$ $H_A: \text{At least one } \beta_i \neq 0 (i = 1, 2, 3)$
 $F(X_1, X_2, X_3)_{3,38} = 2.737, 0.05 < P < 0.1$. At $\alpha = 0.05$, we do not reject H_0 .
- c. $H_0: \rho_{Y(X_1, X_2, X_3)|X_1, X_2, X_3} = 0$ $H_A: \rho_{Y(X_1, X_2, X_3)|X_1, X_2, X_3} \neq 0$
 $F(X_1, X_2, X_3)_{2,36} = 0.362, P > 0.25$. At $\alpha = 0.05$, we do not reject H_0 .
 The two regression lines are parallel.
- d. $H_0: \rho_{YX_3|X_1, X_2} = 0$ $H_A: \rho_{YX_3|X_1, X_2} \neq 0$
 $F(X_3|X_1, X_2)_{1,38} = 6.855, P = 0.014$. At $\alpha = 0.05$, we do not reject H_0 .
- e. First fit the full model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + E \quad (1)$$
 Next, fit the reduced models

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + E \quad (2)$$

$$Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + E \quad (3)$$

$$Y = \beta_0 + \beta_3 X_3 + E \quad (4)$$
 We assess confounding by noting how $\hat{\beta}_3$ changes for the different models. In particular, if $\hat{\beta}_3$ from model (2), (3), or (4) differs from $\hat{\beta}_3$ from model (1), then X_1 , X_2 , or X_1 and X_2 , respectively, are confounders. To assess precision, we note how the $100(1 - \alpha)\%$ CI's for $\hat{\beta}_3$ change. We only eliminate potential confounders from the model if the width of the CI for $\hat{\beta}_3$ does not widen significantly.
- f. From the information provided, we can assess the confounding effects of X_1 or X_2 alone with respect to X_3 but not for X_1 and X_2 taken together.
7. a. No, there is no meaningful change in the estimate for $\hat{\beta}_1$ when X_2 is added to the model.
- b. No, the confidence interval for $\hat{\beta}_1$ is narrower when only X_1 is in the model.
- c. No, there is not evidence suggesting that including X_2 to the model improves the precision and/or the validity of the estimated relationship between X_1 and Y .
9. a. Let $\text{OWNCOST} = X_1$ and $\text{INCOME} = X_2$
 $H_0: \beta_1 = \beta_2 = 0$ $H_A: \text{at least one } \beta_i \text{ does not equal } 0$.
 (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_1, X_2)_{2,23} = 6.38, P = 0.006$. At $\alpha = 0.05$, we reject H_0 .
- b. $H_0: \beta_1 = 0$ $H_A: \beta_1 \neq 0$ (Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_1|X_2)_{1,23} = 11.47, P = 0.003$. At $\alpha = 0.05$, we reject H_0 .

- c. $H_0: \beta_2 = 0 \quad H_A: \beta_2 \neq 0$ (Full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + E$)
 $F(X_2|X_1)_{1,23} = 4.87, P = 0.038$. At $\alpha = 0.05$, we reject H_0 .
- d. Including X_2 does meaningfully change $\hat{\beta}_1$, and it should therefore be included in the model as a confounder, assuming there is no interaction between X_1 and X_2 .

Chapter 12

- 1. a. For smokers: $\hat{Y} = 79.225 + 20.118X$. For nonsmokers: $\hat{Y} = 49.312 + 26.303X$.
- b. $H_0: \beta_{1SMK} = \beta_{1NSMK} \quad H_A: \beta_{1SMK} < \beta_{1NSMK}$
 $T_{28} = 0.892, 0.15 < P < 0.25$. At $\alpha = 0.05$, we do not reject H_0 that the slopes for smokers and nonsmokers are the same.
- c. $H_0: \beta_{0SMK} = \beta_{0NSMK} \quad H_A: \beta_{0SMK} \neq \beta_{0NSMK}$
 $T_{28} = -1.24, 0.20 < P < 0.30$. At $\alpha = 0.05$, we do not reject H_0 , and we conclude that the two intercepts are equal.
- d. The straight lines for smokers and nonsmokers are coincident since both tests failed to reject H_0 .
- 5. a. For NY: $\hat{Y} = 2.174 + 1.177X$. For CA: $\hat{Y} = 8.030 + 1.036X$.
- b. $H_0: \beta_{1NY} = \beta_{1CA} \quad H_A: \beta_{1NY} > \beta_{1CA}$
 $T_{33} = 1.115, 0.10 < P < 0.15$. At $\alpha = 0.05$, we do not reject H_0 , and we conclude that the slopes are the same for NY and CA.
- c. $H_0: \beta_{0NY} = \beta_{0CA} \quad H_A: \beta_{0NY} > \beta_{0CA}$
 $T_{33} = 1.219, 0.85 < P < 0.90$. At $\alpha = 0.05$, we do not reject H_0 that the two intercepts are equal for NY and CA.
- d. Since the tests for equal slopes and equal intercepts did not lead to rejection, we can conclude that the lines are coincident.
- e. $H_0: \rho_{NY} = \rho_{CA} \quad H_A: \rho_{NY} \neq \rho_{CA}$
 $Z = 0.252, P > 0.80$. At $\alpha = 0.05$, we do not reject H_0 , and we conclude that the correlation coefficients for each straight line regression are not significantly different.
- 9. a. $SBP = \beta_0 + \beta_1 AGE + \beta_2 QUET + \beta_3 SMK + \beta_4 AGE*SMK + \beta_5 QUET*SMK + E$
 Smokers: $SBP = (\beta_0 + \beta_3) + (\beta_1 + \beta_4)AGE + (\beta_2 + \beta_5)QUET + E$
 Nonsmokers: $SBP = \beta_0 + \beta_1 AGE + \beta_2 QUET + E$
- b. Smokers: $\widehat{SBP} = 48.076 + 1.466(AGE) + 6.744(QUET)$
 Nonsmokers: $\widehat{SBP} = 48.613 + 1.029(AGE) + 10.451(QUET)$
- c. $H_0: \beta_4 = \beta_5 = 0 \quad H_A: \text{at least one } \beta_i \neq 0$ (Full Model given in part (a).)
 $F(QUET*SMK, AGE*SMK|AGE, QUET, SMK)_{2,26} = 0.222, P > 0.25$
 At $\alpha = 0.05$, we do not reject H_0 and conclude the two lines are coincident.
- d. $H_0: \beta_3 = \beta_4 = \beta_5 = 0 \quad H_A: \text{At least one } \beta_i \neq 0$ (Full Model given in part (a).)
 $F(SMK, QUET*SMK, AGE*SMK|AGE, QUET)_{3,26} = 4.562, 0.01 < P < 0.025$
 At $\alpha = 0.05$, we reject H_0 and conclude the two lines are not coincident.
- 13. a. $R = 1$ and $TD = 1: \widehat{SBPL} = (\hat{\beta}_0 + \hat{\beta}_2 + \hat{\beta}_4 + \hat{\beta}_9) + \hat{\beta}_1(SBP1) + (\hat{\beta}_6 + \hat{\beta}_{11})RW$
 $R = 0$ and $TD = 1: \widehat{SBPL} = (\hat{\beta}_0 + \hat{\beta}_3 + \hat{\beta}_4 + \hat{\beta}_7) + \hat{\beta}_1(SBP1) + (\hat{\beta}_6 + \hat{\beta}_{13})RW$
- b. $H_0: \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0 \quad H_A: \text{At least one } \beta_i \neq 0$.
 $F(X_{11}, X_{12}, X_{13}, X_{14}|X_1, \dots, X_{10})_{4,89} = 1.410, 0.1 < P < 0.25$. At $\alpha = 0.05$, we do not reject H_0 , and we conclude that the lines are parallel.
- c. H_0 : The three regression lines corresponding to rural, town, and urban background are parallel (i.e., $H_0: \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0$).

H_A : At least one $\beta_i \neq 0$.

$F(X_7, \dots, X_{14}|X_1, \dots, X_6)_{8,89} = 1.103$ with $P > 0.25$. At $\alpha = 0.05$, we do not reject H_0 , and we conclude that the three regression lines are parallel.

d. $H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = 0$

H_A : At least one $\beta_i \neq 0$.

$F(X_2, X_3, X_4, X_5, X_7, \dots, X_{14}|X_1, X_6)_{12,89}$

$$= \frac{\text{SS Regression (full model)} - \text{SS Regression } (X_1, X_6)}{12}$$

$$= \frac{\quad}{\text{MS Residual (full model)}}$$

17. a. $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E$ where $Z = 1$ if cool, 0 if warm.
 b. For cool: $\hat{Y} = 104.003 + 2.465X$. For warm: $\hat{Y} = 96.830 + 3.485X$.
 c. $H_0: \beta_2 = \beta_3 = 0$ H_A : At least one $\beta_i \neq 0$ (Full model given in part (a).)
 $F(Z, XZ|X)_{2,14} = 41.875$, $P = 0.0076$. At $\alpha = 0.05$, we reject H_0 and conclude that the lines do not coincide.
 d. $H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$ (Full model given in part (a).)
 $F(XZ|X, Z)_{1,14} = 9.699$, $P < 0.01$. At $\alpha = 0.05$, we reject H_0 and conclude that the lines are not parallel.
 e. Baseline sales are higher during the warm season than in the cool season. Advertising expenditures are higher in the cool season than in the warm season. By spending more money in advertising during the cool season, retailers are able to surpass the sales revenue of the warm season.
21. a. $Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + E$
 b. $H_0: \beta_2 = \beta_3 = 0$ H_A : At least one $\beta_i \neq 0$ (Full model given in part (a).)
 $F(Z, X_1 Z|X_1)_{2,50} = 2.704$, $0.05 < P < 0.10$. At $\alpha = 0.05$, we do not reject H_0 , and we conclude that the lines coincide.
 d. $H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$ (Full model given in part (a).)
 $F(X_1 Z|X_1, Z)_{1,50} = 3.587$, $P = 0.064$. At $\alpha = 0.05$, we do not reject H_0 , and we conclude that the lines are parallel.
 e. The change in refraction-baseline refractive relationship is the same for males and females.

Chapter 13

1. a. $\text{SBP} = \beta_0 + \beta_1(\text{QUET}) + \beta_2\text{SMK} + E$
 b. For smokers: $\text{SBP} = (\beta_0 + \beta_2) + \beta_1(\text{QUET}) + E$; $\overline{\text{SBP}}_{(\text{adj})} = 148.548$
 For nonsmokers: $\text{SBP} = \beta_0 + \beta_1(\text{QUET}) + E$; $\overline{\text{SBP}}_{(\text{adj})} = 139.977$
 c. $H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$ in the model $\text{SBP} = \beta_0 + \beta_1(\text{QUET}) + \beta_2\text{SMK} + E$.
 $T_{29} = 2.707$, $P = 0.011$. At $\alpha = 0.05$, we reject H_0 and conclude that mean SBP differs for smokers and for nonsmokers, after adjusting for QUET.
 d. Finding the 95% confidence interval for the true difference in adjusted mean SBP is equivalent to finding the 95% confidence interval for β_2 . The 95% confidence interval for β_2 is (2.094, 15.048).
5. a. $\text{VIAD} = \beta_0 + \beta_1\text{IQM} + \beta_2\text{IQF} + \beta_3 Z + E$ where $Z_1 = 1$ if female, 0 if male.
 b. For males: $\text{VIAD (adj)} = -3.307$ vs -3.00 unadjusted.
 For females: $\text{VIAD (adj)} = 1.889$ vs 1.60 unadjusted.

c. $H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$ in the model $VIAD = \beta_0 + \beta_1IQM + \beta_2IQF + \beta_3Z + E$
 $T_{16} = 1.659, P = 0.117$. At $\alpha = 0.05$, we do not reject H_0 , and we conclude that the mean scores do not significantly differ by gender, after adjusting for IQM and IQF.

d. 95% CI: $(-1.446, 11.838)$

9. a. $LN_BRNTL = \beta_0 + \beta_1WGT + \beta_2Z_1 + \beta_3Z_2 + \beta_4Z_3 + E$, where $Z_1 = 1$ if 100 ppm, 0 otherwise; $Z_2 = 1$ if 500 ppm, 0 otherwise; $Z_3 = 1$ if 1000 ppm, 0 otherwise.

b. $\hat{\beta}_0 = -0.764$ $\hat{\beta}_1 = 0.0006$ $\hat{\beta}_2 = 0.828$ $\hat{\beta}_3 = 3.571$ $\hat{\beta}_4 = 4.214$

PPM_TOLU	Adjusted Means	Unadjusted Means
50	-0.537	-0.548
100	0.291	0.282
500	3.034	3.019
1000	3.677	3.668

d. $H_0: \beta_2 = \beta_3 = \beta_4 = 0$ $H_A: \text{At least one } \beta_i \neq 0 (i = 2, 3, 4)$
 (Full model: $LN_BRNTL = \beta_0 + \beta_1WGT + \beta_2Z_1 + \beta_3Z_2 + \beta_4Z_3 + E$)
 $F(Z_1, Z_2, Z_3 | WGT)_{3,55} = 1662.526, P < 0.001$. At $\alpha = 0.05$, we reject H_0 and conclude that the adjusted means significantly differ.

11. a. Using PPM_TOLU as the reference group, we define the cross-product terms as

$XZ_1 = WGT$ if PPM_TOLU = 100, 0 otherwise

$XZ_2 = WGT$ if PPM_TOLU = 500, 0 otherwise

$XZ_3 = WGT$ if PPM_TOLU = 1000, 0 otherwise

in which X stands for WGT.

b. The appropriate regression model is

$LN_BRNTL = \beta_0 + \beta_1X + \beta_2Z_1 + \beta_3Z_2 + \beta_4Z_3 + \beta_5XZ_1 + \beta_6XZ_2 + \beta_7XZ_3 + E$

in which X stands for WGT.

c. The hypothesis of no interaction is tested in terms of model parameters as $H_0: \beta_5 = \beta_6 = \beta_7 = 0$.

13. a. $Y = \beta_0 + \beta_1AGE + \beta_2Z + E$

b. $Y = \beta_0 + \beta_1AGE + \beta_2Z + \beta_3AGE*Z + E$

$H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$ (Full model: $Y = \beta_0 + \beta_1AGE + \beta_2Z + \beta_3AGE*Z + E$)

$T_{26} = -1.677, P = 0.106$. At $\alpha = 0.05$, we do not reject H_0 , and we conclude that the ANCOVA model in part (a) is appropriate.

Location	Adjusted Means	Unadjusted Means
Intown/inner	81.526	82.187
Outer	85.46	84.807

$H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$

$T_{27} = 1.212, P = 0.236$. At $\alpha = 0.05$, we do not reject H_0 , and we conclude that the adjusted means do not significantly differ.

15. a. Let X denote OWNCOST and Y denote OWNEROCC.

$Y = \beta_0 + \beta_1X + \beta_2Z + E$

b. $Y = \beta_0 + \beta_1X + \beta_2Z + \beta_3XZ + E$

$H_0: \beta_3 = 0$ $H_A: \beta_3 \neq 0$

$F(XZ|X, Z)_{1,22} = \frac{0.696}{17.580} = 0.04$, with $P \gg 0.25$.

At $\alpha = 0.05$, we do not reject H_0 and conclude the ANACOVA model in part (a) is appropriate.

c. Location	Adjusted Means	Unadjusted Means
Urban $\geq 75\%$	56.15%	64.5%
Urban $\leq 75\%$	60.06%	69.25%

$H_0: \beta_2 = 0$ $H_A: \beta_2 \neq 0$

$T_{23} = -2.191, P = 0.039.$

At $\alpha = 0.05$, we reject H_0 and conclude the adjusted means significantly differ.

Chapter 14

1. a. The plot of jackknife residuals versus the predicted values shows a distinct pattern, indicating that an assumption has been violated. In this case, the obvious curvilinear pattern indicates a violation of the linearity assumption (much as the simple plot of Y vs X did in Chapter 5, Problem 1). A linearizing transformation should be applied, e.g., $\log_{10}(\text{dry weight})$ can be used as the dependent variable.
 - b. The skewness statistic ($=1.66$) and kurtosis statistic ($=3.21$) suggest that at least a moderate violation of the normality assumption has occurred. The normal probability plot also looks fairly non-linear. With such few data values, it is difficult to conclude that a gross violation of the linearity assumption has occurred. Since a violation of the linearity assumption has occurred (see part (a)), no attempt should be made to correct for the possible violation of normality until the linearity issue is addressed.
 - c. Observation 11 has a Cook's distance value greater than 1. No observations have leverage values greater than $2(k + 1)/n = 0.36$. The data for observation 11 should be double-checked and corrected, if necessary; if the recorded value is correct, it should be judged as to plausibility. If judged to be implausible, the observation can be removed from the analysis. If plausible, no corrective action is taken; instead, the analysis can be run with and without the observation included and the regression results compared to judge the impact of the outlier.
3. a. The plot of jackknife residuals versus the predicted values shows a distinct pattern, indicating that an assumption has been violated. In this case, the obvious curvilinear pattern indicates a violation of the linearity assumption (much as the simple plot of Y vs X did in Chapter 5, Problem 3). In this problem, it may help to add an X^2 term to the model.
 - b. The skewness and kurtosis statistics are both less than 1 in magnitude; the normal probability plot is not grossly non-linear. This suggests that there is no gross violation of the normality assumption.
 - c. No Cook's distance values are greater than 1. Four observations have leverage values greater than $2(k + 1)/n = 0.20$. The data for these observations should be double-checked and corrected, if necessary; if the recorded values are correct, they should be judged as to plausibility. If judged to be implausible, an observation can be removed from the analysis. If plausible, no corrective action is taken; instead, the analysis can be run with and without the observation included and the regression results compared to judge the impact of the outlier.
19. a. The plot of jackknife residuals vs. predicted values looks like a random scatter of points; since no pattern is evident, no assumptions appear to be violated based on this plot.
 - b. The skewness and kurtosis statistics are both less than 1 in magnitude; the normal probability plot is fairly linear. This suggests that there is no gross violation of the normality assumption.

- c. No Cook's distance values are greater than 1. The leverage value for observation 18 is greater than $2(k + 1)/n = 0.42$. The data for this observation should be double-checked and corrected, if necessary.
 - d. None of the variance inflation factors are larger than 10 and none of the condition indexes is greater than 30. Therefore, no collinearity problem exists.
25. a. $Y = \beta_0 + \beta_1 (\text{ADVERTISING}) + E$, where Y denotes sales.
- b. The largest studentized residual (absolute value) = 1.527, which is no cause for alarm.
 - c. The plot of the jackknife residuals versus the predictor does not suggest any troublesome problems.

Chapter 15

1. b. From the computer output, we find:
- (1) Degree 1: $\hat{Y} = -1.932 + 0.246X$
 - (2) Degree 2: $\hat{Y} = 3.172 - 0.781X + 0.047X^2$
 - (3) $\ln Y$ on X : $\ln \hat{Y} = -6.21 + 0.451X$
 - (4) The above fitted equations are plotted on the graphs presented for 1(a).

c.

Source	d.f.	SS	MS	F
Regression	1	12.705	12.705	43.69
Lack of fit	4	4.419	1.105	57.03
Residual	16	4.651	0.2908	
Pure error	12	0.232	0.0194	
Total	17	17.357		

d.

Source	d.f.	SS	MS	F
Degree 1(X)	1	12.705	12.705	43.69
Regression	2	16.61	8.305	
Degree 2 ($X^2 X$)	1	3.905	3.905	78.46
Lack of fit	3	0.514	0.171	8.85
Residual	15	0.746	0.0497	
Pure error	12	0.232	0.0194	
Total	17	17.357		

e. $r^2_{XY} = 0.732$; $r^2(\text{quadratic}) = 0.957$

f. Test for significance of straight-line regression of Y on X

H_0 : The straight-line regression is not significant.

$F_{1,16} = 43.69$, $P < 0.001$. At $\alpha = 0.05$, we reject H_0 and conclude that the straight-line regression is significant.

Test for adequacy of straight-line model

H_0 : The straight-line model is adequate.

$F_{4,12} = 56.959$, $P < 0.001$. At $\alpha = 0.05$, we reject H_0 and conclude that the straight-line model is not adequate.

g. Test for significance of quadratic regression

H_0 : The quadratic regression is not significant.

$F_{2,15} = 166.77$, $P = 0.0001$. At $\alpha = 0.05$, we reject H_0 and conclude that the quadratic regression is significant.

Test for addition of X^2 term

H_0 : The addition of X^2 to a model already containing X is not significant.

Partial $F(X^2|X)_{1,15} = 78.41$, $P = 0.0001$. At $\alpha = 0.05$, we reject H_0 and conclude that the addition of X^2 is significant.

Test for adequacy of quadratic model

H_0 : The quadratic model is adequate.

$F_{3,12} = 8.84$, $P = 0.002$. At $\alpha = 0.05$, we reject H_0 and conclude that the quadratic model is not adequate.

h. Test for significance of straight-line regression of $\ln Y$ on X

H_0 : The straight-line regression is not significant.

$F_{1,16} = 4277.167$, $P = 0.0001$. At $\alpha = 0.05$, we reject H_0 and conclude that the straight-line regression is significant.

Test for adequacy of straight-line model of $\ln Y$ on X

H_0 : The straight-line model is adequate.

$F_{4,12} = 0.896$, $P = 0.471$. At $\alpha = 0.05$, we do not reject H_0 and conclude that the straight-line model is adequate.

i. R^2 (straight-line regression of $\ln Y$ on X) = 0.9965

R^2 (quadratic regression of Y on X) = 0.957

A comparison of the above two R^2 shows that the straight-line fit of $\ln Y$ on X provides a better fit.

j. (1) Homoscedasticity assumption appears to be much more reasonable when using $\ln Y$ on X than when using Y on X .

(2) The straight-line regression of $\ln Y$ on X is preferred.

k. The independence assumption is violated.**5. b. Test for significance of straight-line regression**

H_0 : The straight-line regression is not significant.

$F_{1,24} = 978.04$, $P < 0.001$. At $\alpha = 0.05$, we reject H_0 and conclude that the straight-line regression is significant.

Test for adequacy of straight-line model

H_0 : The straight-line model is adequate.

$F_{16,8} = 1.57$, $P > 0.25$. At $\alpha = 0.05$, we do not reject H_0 and conclude that the straight-line model is adequate.

c. Test for addition of X^2 to the model

H_0 : The addition of X^2 is not significant.

Partial $F(X^2|X)_{1,23} = 0.55$, $P > 0.25$. At $\alpha = 0.05$, we do not reject H_0 .

d. The straight-line model is most appropriate.**9. a. Fit the model $VOC_SIZE = \beta_0 + \beta_1(AGE^*) + \beta_2(AGE^*)^2 + \beta_3(AGE^*)^3 + E$, where $AGE^* = AGE - 2.867$.**

$$\hat{\beta}_0 = 741.84 \quad \hat{\beta}_1 = 645.60 \quad \hat{\beta}_2 = 70.43 \quad \hat{\beta}_3 = -31.18$$

b. Using variables-added-in-order tests, the best model includes AGE^* , $(AGE^*)^2$, and $(AGE^*)^3$.**c. Using variables-added-last tests, the best model includes AGE^* , $(AGE^*)^2$, and $(AGE^*)^3$.****d. The only large predictor correlation is between AGE^* and $(AGE^*)^3$. The largest condition index ($CI_3 = 6.05$) suggests that the centered data do not have any serious collinearity problems.**

- f. The estimated regression coefficients differ from those in Problem 8, but the best model includes the linear, quadratic, and cubic terms as in Problem 8. Also, the sums of squares for the variables-added-in-order test are the same as in Problem 8. The centering of AGE greatly reduced the previous collinearity problems. Centering does not affect the residual diagnostics.
13. a. The estimated equation is $\hat{Y} = 10.64 + 94.42(\text{LIN_TOL}) + 14.59(\text{QUAD_TOL}) - 0.73(\text{CUB_TOL})$, where LN_TOL, QUAD_TOL, and COL_TOL denote the centered PPM_TOLU orthogonal polynomials.
- b. Using the variable-added-in-order tests, the best model includes LIN_TOL and QUAD_TOL.
- c. Using the variable-added-last tests, the best model includes LIN_TOL and QUAD_TOL, which is the same model as in part (b).
- d. The orthogonal polynomials are uncorrelated with each other, which implies that any collinearities are eliminated as shown by the condition indices.
- e. The residual plots suggest that the variances increase as the predicted values increase.

Chapter 16

1. a. The final model by forward selection is $\widehat{\text{WGT}} = 37.600 + 0.053(\text{AGE}*\text{HGT})$
- b. The final model by backward elimination is $\widehat{\text{WGT}} = 37.600 + 0.053(\text{AGE}*\text{HGT})$
- c. The best model resulting from this approach is:
 $\widehat{\text{WGT}} = 6.553 + 0.722(\text{HGT}) + 2.050(\text{AGE})$
- d. The first two approaches result in the model including only the AGE*HGT interaction. It is difficult to interpret results for such a model, since the variables that make up the interaction term are not included. The third approach results in a model that is easier to interpret. It is possible to conduct modified forward and backward stepwise strategies in which interaction terms are only considered for inclusion if the terms that make up the interaction term are already in the model.
3. For smokers: the final estimated model is $\widehat{\text{SBP}} = 102.200 + 0.252(\text{AGE}*\text{QUET})$
 For nonsmokers: the final estimated model is $\widehat{\text{SBP}} = 93.073 + 0.250(\text{AGE}*\text{QUET})$
 These two models are different from those obtained from 2(d) by putting SMK = 1 (for smokers) and SMK = 0 (for nonsmokers).
5. The best regression models using the sequential procedure of adding AGE first for Females and Males are
 Females: $\widehat{\text{DEP}} = 190.012 - 1.099\text{AGE} - 1.217\text{MC}$
 Males: $\widehat{\text{DEP}} = 270.056 - 2.514\text{AGE} - 1.065\text{MC}$
 For females, the above model was selected as best because of its high r^2 (0.401), satisfactory $C(P)$ (2.238), and low MSE (3274.364). The next best model was the full model with $r^2 = 0.413$, $C(P) = 4.0$, MSE = 3478.318. Emphasizing parsimony, we find that the above model is more favorable than the full model.
 The reasoning is similar for selecting the model shown above for males.
7. a. Using the $C(P)$ criterion exclusively, we find that the three models with the most favorable $C(P)$ values contain AGE alone (1.23), AGE and WGT (2.63), or all three variables (4). None of these models is particularly impressive, and since it would be hard to argue that either of the multiple-variable models is better than the model with AGE alone, we could take AGE alone as the best model. Upon further investigation, the difficulty is seen to be partly due to none of these models having a significant overall F test.

- b. Since there seems to be no rationale for grouping the variables (age, weight, and height) in any way, chunks are taken to be the three variable specific pairs of linear/quadratic terms (i.e., AGE_C and AGE_CSQ; HGT_C and HGT_CSQ; WGT_C and WGT_CSQ, where the _C terms are the centered variables, and the _CSQ terms are the squared centered terms). A plausible forward chunkwise strategy is to treat each chunk/pair as a distinct entity that cannot be split and then proceed in the usual forward manner (use $\alpha = 0.10$):
- Step 1: WGT_C, WGT_CSQ added to the model. F test = 2.78
 - Step 2: AGE_C, AGE_CSQ added to the model. F test = 3.49
 - Step 3: Stop HGT_C, HGT_CSQ not significant. F test = 0.16
- c. The all possible regressions method yields the following “best” models for each of the model sizes:

Number in Model	Variables	$C(P)$	R^2	MSE
1	WGT_CSQ	8.026	0.078	0.771
2	WGT_C, WGT_CSQ	5.213	0.300	0.631
3	WGT_C, WGT_CSQ, AGE_CSQ	4.337	0.432	0.554
4	WGT_C, WGT_CSQ, AGE_C, AGE_CSQ	3.312	0.571	0.456
5	WGT_C, WGT_CSQ, AGE_C, AGE_CSQ, HGT_CSQ	5.225	0.576	0.497
6	Full Model	7.00	0.586	0.539

The best model is most likely the four-variable model including WGT_C, WGT_CSQ, AGE_C, and AGE_CSQ. The R^2 , $C(P)$, and MSE for this model are better than for any of the smaller models; and they are similar to, if not better than, the statistics for the larger models, which of course are less parsimonious.

- d. Any model containing only first-order terms (part (a)) is seriously deficient. The model in parts (b) and (c) is the best one.
9. a. A model containing X_1 , X_3 , and X_4 is the best model by this method. The model has a relatively high R^2 , a satisfactory $C(P)$, and one of the lowest MSEs. The model also has the benefit of parsimony compared to the larger models.
- b. The selected model contains X_1 , X_3 , and X_4 .
 - c. The selected model contains X_1 , X_3 , and X_4 .
 - d. All three methods selected the same model, and it appears to be the best model for the reasons cited in (a).
11. a. The model containing X_1 and X_3 would be the recommended model. Its R^2 , $C(P)$, and MSE are clearly superior to the best one-variable model statistics, and they are similar to the statistics for the less parsimonious full model.
- b. The results of the stepwise regression show that the model containing X_1 and X_3 is the best model.
 - c. The model: $Y = \beta_0 + \beta_1X_1 + \beta_3X_3 + E$ appears to be best. This model is chosen, given the results in parts (a) and (b).

Chapter 17

1. a.

Treatment	Mean	S
1	7.5	1.643
2	5	1.265
3	4.333	1.033
4	5.167	1.472
5	6.167	2.041