CHAPTER 12

Section 12.1

1.

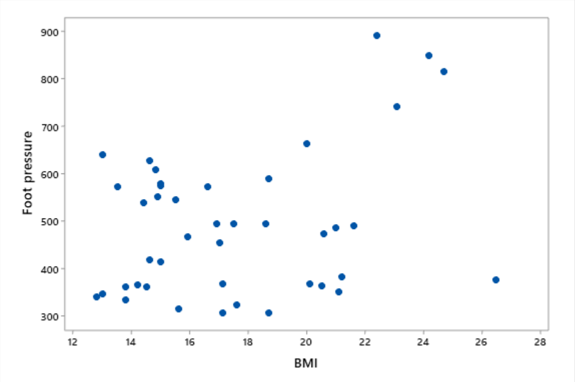
a. Both the BMI and peak foot pressure distributions appear positively skewed with some gaps and possible high outliers.

```
Stem-and-leaf of BMI

1    12  8
6    13  00588
13   14  2456689
19   15  000569
21   16  69
21   17  01156
16   18  677
13   19
13   20  0156
9    21  0126
5    22  4
4    23  1
3    24  27
1    25
1    26  5

  Leaf Unit = 0.1
```

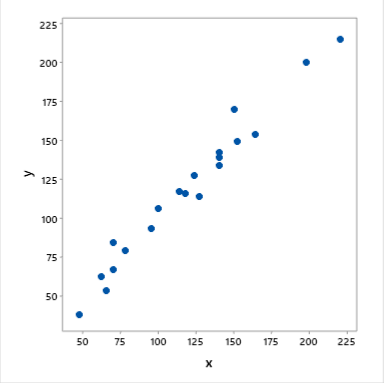
```
Stem-and-leaf of Foot pressure

7    3  0012344
16   3  566666678
18   4  11
(8)  4  56789999
16   5  34
14   5  577778
8    6  024
5    6  6
4    7  4
3    7
3    8  1
2    8  59

  Leaf Unit = 10
```

b. No, peak foot pressure cannot be uniquely determined by BMI. As a counterexample, the second and third children listed both have BMI = 13.0 but their peak foot pressures are very different.

c. The scatterplot suggests some positive association between BMI and peak foot pressure (the plot goes from lower-left to upper-right), so BMI may have some predictive power. But the relationship does not appear to be very strong, and there are many outliers from the overall pattern.

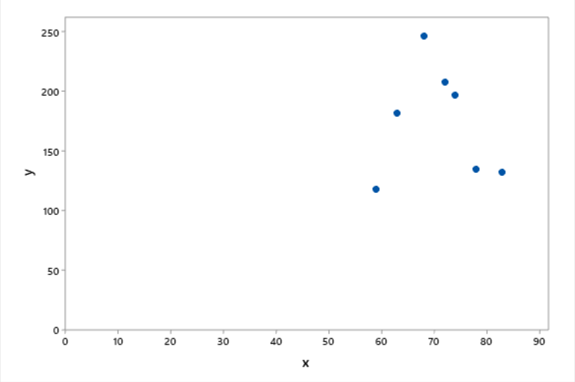


3. A scatter plot of the data appears below. The points fall very close to a straight line with an intercept of approximately 0 and a slope of about 1. This suggests that the two methods are producing substantially the same concentration measurements.

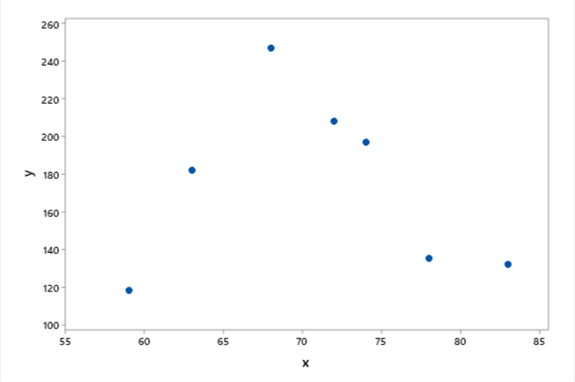


5.

a. The scatter plot with axes intersecting at (0,0) is shown below.



b. The scatter plot with axes intersecting at (55, 100) is shown below. This plot is certainly preferable, since the dots in the plot are not compressed into one corner (the plot in **a** leaves a lot of unused white space).



c. A parabola appears to provide a good fit to both graphs.

7.

a. Expected fuel efficiency when $x = 2500$ is $f(2500) = 70 - .0085(2500) = 48.75$ mpg.

b. If x = weight increases by 1 (lb), then $f(x)$ changes by $-.0085$. That is, for each 1-lb increase in car weight, expected fuel efficiency *decreases* by .0085 mpg.

c. Because the relationship is linear, the effect of a 500-lb increase is just 500 times the effect of a 1-lb increase. So, a 500-lb increase in car weight corresponds to a *decrease* in expected fuel efficiency equal to .0085(500) = 4.25 mpg.

d. Reversing part **c**, a 500-lb *decrease* in car weight corresponds to an *increase* of 4.25 mpg in expected fuel efficiency.

9.

a. β_1 = change in expected flow rate associated with a one-inch increase in pressure drop $= .095$.

b. We expect flow rate to decrease by $5\beta_1 = .475$.

c. $\mu_{Y\cdot 10} = -.12 + .095(10) = .83$, and $\mu_{Y\cdot 15} = -.12 + .095(15) = 1.305$.

d. $P(Y > .835) = P\left(Z > \dfrac{.835 - .830}{.025}\right) = P(Z > .20) = .4207$

$P(Y > .840) = P\left(Z > \dfrac{.840 - .830}{.025}\right) = P(Z > .40) = .3446$

e. Let Y_1 and Y_2 denote pressure drops for flow rates of 10 and 11, respectively. Then $\mu_{Y\cdot 11} = .925$, so $Y_1 - Y_2$ has expected value $.830 - .925 = -.095$ and sd $\sqrt{(.025)^2 + (.025)^2} = .035355$. Thus

$P(Y_1 > Y_2) = P(Y_1 - Y_2 > 0) = P\left(Z > \dfrac{0-(-.095)}{.035355}\right) = P(Z > 2.69) = .0036$.

11.

a. β_1 = expected change for a one-degree increase = -.01, and $10\beta_1 = -.1$ is the expected change for a 10-degree increase.

b. $\mu_{Y\cdot 200} = 5.00 - .01(200) = 3$, and $\mu_{Y\cdot 250} = 2.5$.

c. The probability that the first observation is between 2.4 and 2.6 is $P(2.4 \le Y \le 2.6) = P\left(\dfrac{2.4-2.5}{.075} \le Z \le \dfrac{2.6-2.5}{.075}\right) = P(-1.33 \le Z \le 1.33) = .8164$. The probability that any particular one of the other four observations is between 2.4 and 2.6 is also .8164, so the probability that all five are between 2.4 and 2.6 is $(.8164)^5 = .3627$.

d. Let Y_1 and Y_2 denote the times at the higher and lower temperatures, respectively. Then $Y_1 - Y_2$ has expected value $5.00-.01(x+1)-(5.00-.01x)=-.01$. The standard deviation of $Y_1 - Y_2$ is $\sqrt{(.075)^2+(.075)^2}=.10607$. Thus $P(Y_1-Y_2>0)=P\left(Z>\dfrac{0-(-.01)}{.10607}\right)=P(Z>.09)=.4641$.

Section 12.2

13.

a. $\hat{\beta}_1=\dfrac{S_{xy}}{S_{xx}}=\dfrac{13,048}{20,003}=.652$ and $\hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}=\dfrac{346}{14}-(.652)\dfrac{517}{14}=.626$, so the equation of the LSRL is $y=.626+.652x$.

b. $\hat{y}=.626+.652(35)=23.456$. The residual is $y-\hat{y}=21-23.456=-2.456$.

c. SSE = $S_{yy}-S_{xy}^2/S_{xx}$ = 8903 – $(13048)^2$/20003 = 392, so $\hat{\sigma}=\sqrt{\dfrac{\text{SSE}}{n-2}}=\sqrt{\dfrac{392}{14-2}}=5.7$.

d. $R^2=1-\dfrac{\text{SSE}}{\text{SST}}=1-\dfrac{392}{8903}=.956$.

e. Without the two upper extreme observations, the new summary values are $n=12,\Sigma x=272,\Sigma x^2=8322,\Sigma y=181,\Sigma y^2=3729,\Sigma xy=5320$. The new $S_{xx}=2156.667,S_{yy}=998.917,S_{xy}=1217.333$. New $\hat{\beta}_1=.56445$ and $\hat{\beta}_0=2.2891$, which yields the new equation $y=2.2891+.56445x$. Removing the two values changes the position of the line considerably, and the slope slightly. The new $R^2=1-\dfrac{311.79}{998.917}=.6879$, which is much worse than that of the original set of observations.

15.

a. With the aid of software, S_{xx} = 504.0, S_{yy} = 9.9153, S_{xy} = 45.8246, $\hat{\beta}_1$ = 4582.46/504 = .09092 and $\hat{\beta}_0=\overline{y}-\hat{\beta}_1\overline{x}=\dfrac{40.09}{24}-(.09092)\dfrac{4308}{24}$ = –14.6497. The equation of the LSRL is y = –14.6497 + .09092x.

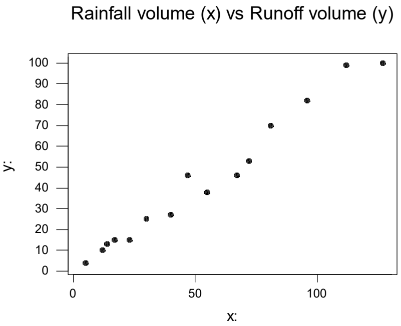
b. –14.6497 + .09092(182) = 1.8997.

c. The four observations for which temperature is 182 are: (182, .90), (182, 1.81), (182, 1.94), and (182, 2.68). Their corresponding residuals are: $.90-1.8997=-0.9977$, $1.81-1.8997=-0.0877$, $1.94-1.8997=0.0423$, $2.68-1.8997=0.7823$. These residuals do not all have the same sign because in the cases of the first two pairs of observations, the observed efficiency ratios were smaller than the predicted value of 1.8997. Whereas, in the cases of the last two pairs of observations, the observed efficiency ratios were larger than the predicted value.

d. SST = S_{yy} = 9.9153, SSE = 9.9153 – 45.8246^2/504.0 = 5.7489, R^2 = 1 – SSE/SST = 1 – 5.7489/9.9153 = .4202. 42.02% of the observed variation in efficiency ratio can be attributed to the approximate linear relationship between the efficiency ratio and the tank temperature.

17.

a. Yes, the scatterplot shows a strong linear relationship between rainfall volume and runoff volume, thus it supports the use of the simple linear regression model.



b. From software, $\overline{x} = 53.200$, $\overline{y} = 42.867$, $S_{xx} = 20{,}586.4$, $S_{yy} = 14{,}435.7$, and $S_{xy} = 17{,}024.4$.

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{17{,}024.4}{20{,}586.4} = .82697 \text{ and } \hat{\beta}_0 = 42.867 - (.82697)53.2 = -1.1278.$$

c. $\hat{\mu}_{Y|50} = -1.1278 + .82697(50) = 40.2207$.

d. SSE = 14435.7 – 17024.4^2/20586.4 = 357.07. $s_e = \hat{\sigma} = \sqrt{\dfrac{\text{SSE}}{n-2}} = \sqrt{\dfrac{357.07}{13}} = 5.24$.

e. $R^2 = 1 - \dfrac{\text{SSE}}{\text{SST}} = 1 - \dfrac{357.07}{14{,}435.7} = .9753$. So 97.53% of the observed variation in runoff volume can be attributed to the simple linear regression relationship between runoff and rainfall.

19.

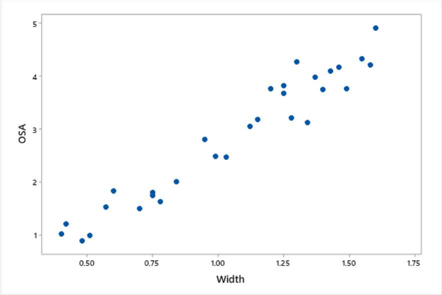
a. From statistical software, $\hat{\beta}_1 = -.20939$ and $\hat{\beta}_0 = 75.212$. So the equation of the LSRL is $y = 75.212 - .20939x$. For $x = 100$, we predict 75.212 – .20939(100) = 54.274.

b. From statistical software, SSE = 78.92 and SST = 377.17, so $R^2 = 1 - 78.92/377.17 = .791$. So, 79.1% of the variation in cetane number is explained by this linear model with predictor iodine. That is, the error sum of squares is reduced by 79.1% compared to predicting with just a constant.

c. $s_e = \hat{\sigma} = \sqrt{\dfrac{SSE}{n-2}} = \sqrt{\dfrac{78.92}{12}} = 2.56$, which is a typical deviation of an actual cetane number from the predicted cetane number calculated by the estimated regression line.

21.

a. The scatterplot shows a very strong, positive, linear relationship between palprebal fissure width and ocular surface area.



b. With the aid of statistical software, $\hat{\beta}_1$ = 3.080 and $\hat{\beta}_0$ = –0.398, so the LSRL is y = –0.398 + 3.080x.

c. A 1-cm increase in palprebal fissure width corresponds to an estimated 3.080 cm^2 increase in average/expected OSA.

d. –0.398 + 3.080(1.25) = 3.452 cm^2.

e. $\hat{\mu}_{Y|1.25}$ is also equal to 3.452 cm^2. (That is, the point prediction and point estimate at x = 1.25 cm are the same.)

23. With each $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma)$, their joint likelihood function is

$$f(y_1)\cdots f(y_n) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_1-[\beta_0+\beta_1 x_1])^2/2\sigma^2} \cdots \frac{1}{\sigma\sqrt{2\pi}} e^{-(y_n-[\beta_0+\beta_1 x_n])^2/2\sigma^2} = C\exp\left(-\frac{\sum(y_i-[\beta_0+\beta_1 x_i])^2}{2\sigma^2}\right)$$. The

mle's of β_0 and β_1 maximize this expression, but *maximizing* $Ce^{-w/2\sigma^2}$ is equivalent to *minimizing* the expression w. In the likelihood function that's $\sum(y_i-[\beta_0+\beta_1 x_i])^2$, which is exactly $g(\beta_0,\beta_1)$. Therefore, the least squares estimates — which, by definition, minimize $g(\beta_0,\beta_1)$ — are also the mle's.

25. The new slope and intercept will be $1.8\hat{\beta}_1$ and the new intercept will be $1.8\hat{\beta}_0 + 32$. To see why, notice that the x's are unchanged, so $\overline{x}$ and S_{xx} are unchanged. But with $y_i' = 1.8y_i + 32$, $\overline{y}' = 1.8\overline{y} + 32$ by linearity of means and $S_{xy'} = \sum(x_i-\overline{x})(y_i'-\overline{y}') = \sum(x_i-\overline{x})(1.8y_i+32-[1.8\overline{y}+32]) = \sum(x_i-\overline{x})(1.8y_i-1.8\overline{y})$ = $1.8\sum(x_i-\overline{x})(y_i-\overline{y}) = 1.8S_{xy}$. Therefore, the *new* slope is $\dfrac{S_{xy'}}{S_{xx}} = \dfrac{1.8S_{xy}}{S_{xx}} = 1.8\hat{\beta}_1$, and then the *new* intercept is $\overline{y}' - (1.8\hat{\beta}_1)\overline{x} = 1.8\overline{y} + 32 - 1.8\hat{\beta}_1\overline{x} = 1.8(\overline{y} - \hat{\beta}_1\overline{x}) + 32 = 1.8\hat{\beta}_0 + 32$.

27. The LSRL equation is $y = \hat{\beta}_0 + \hat{\beta}_1 x = \bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x$. Substitute $x = \bar{x}$ and you get $y = \bar{y}$, which shows the LSRL passes through $(\bar{x}, \bar{y})$.

29.

a. Subtracting $\bar{x}$ from each x_i shifts the plot in a rigid fashion $\bar{x}$ units to the left without otherwise altering its character. The last squares line for the new plot will thus have the same slope as the one for the old plot. Since the new line is $\bar{x}$ units to the left of the old one, the new y intercept (i.e., the height at $x = 0$) is the height of the old line at $x = \bar{x}$, which is $\hat{\beta}_0 + \hat{\beta}_1\bar{x} = \bar{y}$ (since from exercise 26, $(\bar{x}, \bar{y})$ is on the old line). Thus the new y intercept is $\bar{y}$.

b. We wish b_0 and b_1 to minimize $g(b_0, b_1) = \Sigma\left[y_i - \left(b_0 + b_1\left(x_i - \bar{x}\right)\right)\right]^2$. Equating $\dfrac{\partial g}{\partial b_0}$ and $\dfrac{\partial g}{\partial b_1}$ to 0 yields $nb_0 + b_1\Sigma\left(x_i - \bar{x}\right) = \Sigma y_i$, $b_0\Sigma\left(x_i - \bar{x}\right) + b_1\Sigma\left(x_i - \bar{x}\right)^2 = \Sigma\left(x_i - \bar{x}\right)^2 = \Sigma\left(x_i - \bar{x}\right)y_i$. Since $\Sigma\left(x_i - \bar{x}\right) = 0$, $b_0 = \bar{y}$. And since $\Sigma\left(x_i - \bar{x}\right)y_i = S_{xy}$ because $\Sigma\left(x_i - \bar{x}\right)\bar{y} = \bar{y}\Sigma\left(x_i - \bar{x}\right)$, $b_1 = \hat{\beta}_1$. Thus $\hat{\beta}_0^* = \bar{Y}$ and $\hat{\beta}_1^* = \hat{\beta}_1$.

Section 12.3

31.

a. With these x-values, $\bar{x}$ = 725 and $S_{xx} = \sum(x_i - 725)^2$ = 17,500. Thus, $V(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{xx}} = \dfrac{10^2}{17{,}500}$ and

$\sigma_{\hat{\beta}_1} = \dfrac{10}{\sqrt{17{,}500}}$ = 0.0756.

b. Under the model assumptions, the rv $\hat{\beta}_1$ has a normal distribution with mean β_1 = .25 and standard deviation 0.0756 from part **a**. Thus $P(.15 < \hat{\beta}_1 < .35) = P\left(\dfrac{.15 - .25}{.0756} < Z < \dfrac{.35 - .25}{.0756}\right)$ = $P(-1.32 < Z < 1.32) = \Phi(1.32) - \Phi(-1.32) = .813$.

c. With these n = 11 values, S_{xx} = 11,000, which is smaller than in **a**. Thus, even though we have a larger sample, the resulting standard deviation of $\hat{\beta}_1$ is larger. The n = 7 sample from **a** resulting in more precise estimation.

33. Let β_1 denote the true average change in runoff for each 1 m^3 increase in rainfall. To test the hypotheses $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq 0$, the calculated t statistic is $t = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}} = \dfrac{.82697}{.03652} = 22.64$ which (from the printout) has an associated P-value of P = 0.000. Therefore, since the P-value is so small, H_0 is rejected, and we conclude that there is a useful linear relationship between runoff and rainfall.

A confidence interval for β_1 is based on $n - 2 = 15 - 2 = 13$ degrees of freedom. $t_{.025,13} = 2.160$, so the interval estimate is $\hat{\beta}_1 \pm t_{.025,13} \cdot s_{\hat{\beta}_1} = .82697 \pm (2.160)(.03652) = (.748, .906)$. Therefore, we can be confident that the true change in average runoff, for each 1 m^3 increase in rainfall, is somewhere between .748 m^3 and .906 m^3.

35.

a. We want a 95% CI for β_1: $\hat{\beta}_1 \pm t_{.025,15} \cdot s_{\hat{\beta}_1}$. Using the given summary statistics, $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}} = \dfrac{238.112}{115.019} = 1.536$. Next, SSE = 783.88 – 238.11^2/155.02 = 418.2494, from which $s_e = \sqrt{\dfrac{418.2494}{15}} = 5.28$ and $s_{\hat{\beta}_1} = \dfrac{5.28}{\sqrt{155.02}} = .424$. With $t_{.025,15} = 2.131$, our CI is $1.536 \pm 2.131 \cdot (.424)$ = (.632, 2.440). With 95% confidence, we estimate that the change in reported nausea percentage for every one-unit change in motion sickness dose is between .632 and 2.440.

b. We test the hypotheses $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$, and the test statistic is $t = \dfrac{1.536}{.424} = 3.6226$. With df = 15, the two-tailed *P*-value = 2*P*(*T* > 3.6226) = 2(.001) = .002. With a *P*-value of .002, we would reject the null hypothesis at most reasonable significance levels. This suggests that there is a useful linear relationship between motion sickness dose and reported nausea.

c. No. A regression model is only useful for estimating values of nausea % when using dosages between 6.0 and 17.6 — the range of values sampled.

d. Removing the point (6.0, 2.50), the new values are (with the aid of software) $\hat{\beta}_1 = 1.561$, $\hat{\beta}_0 = -9.118$, SSE = 430.5264, $s_e = 5.55$, $s_{\hat{\beta}_1} = .551$, and the new CI is $1.561 \pm 2.145 \cdot (.551)$, or (.379, 2.743). The interval is a little wider. But removing the one observation did not change it that much. The observation does not seem to be exerting undue influence.

37.

a. From Exercise 19, SSE = 78.92, so $s_e = \sqrt{\dfrac{78.92}{14-2}} = 2.5645$ and $s_{\hat{\beta}_1} = \dfrac{2.5645}{82.479} = .03109$. Thus $t = \dfrac{-.20939}{.03109} = -6.73 < -4.318 = -t_{.0005,12}$ and *P*-value < .001. Because the *P*-value < .01, $H_0 : \beta_1 = 0$ is rejected at level .01 in favor of the conclusion that the model is useful $(\beta_1 \neq 0)$.

b. The CI for β_1 is $-.2094 \pm (2.179)(.03109) = -.2094 \pm .0677 = (-.277, -.142)$. Thus the CI for $10\beta_1$ is $(-2.77, -1.42)$.

39. Each Y_i has mean $\beta_0 + \beta_1 x_i$, so $E(\overline{Y}) = \frac{1}{n}\sum E(Y_i) = \frac{1}{n}\sum(\beta_0 + \beta_1 x_i) = \beta_0 + \beta_1\overline{x}$. Thus, using the fact that $E(\hat{\beta}_1) = \beta_1$, $E(\hat{\beta}_0) = E(\overline{Y} - \hat{\beta}_1\overline{x}) = E(\overline{Y}) - E(\hat{\beta}_1)\overline{x} = \beta_0 + \beta_1\overline{x} - \beta_1\overline{x} = \beta_0$.

41. Let $x' = cx$ and $y' = dy$. Then $S_{x'y'} = \sum(cx_i - c\overline{x})(dy_i - d\overline{y}) = cdS_{xy}$ and, similarly, $S_{x'x'} = c^2 S_{xx}$ and $S_{y'y'} = d^2 S_{yy}$. The new slope is $\hat{\beta}_1' = \frac{S_{x'y'}}{S_{x'x'}} = \frac{cdS_{xy}}{c^2 S_{xx}} = \frac{d}{c}\hat{\beta}_1$. Similarly, the new SSE is $\text{SSE}' = d^2\text{SSE}$, so $s_e' = \sqrt{\frac{\text{SSE}'}{n-2}} = \sqrt{\frac{d^2\text{SSE}}{n-2}} = ds_e$. Put it all together: $t' = \frac{\hat{\beta}_1'}{s_e'/\sqrt{S_{x'x'}}} = \frac{(d/c)\hat{\beta}_1}{ds_e/\sqrt{c^2 S_{xx}}} = \frac{\hat{\beta}_1}{s_e/\sqrt{S_{xx}}} = t$, as claimed.

Section 12.4

43.

a. The mean of the *x* values is $\overline{x}$ = 613.5. Since *x* = 600 is closer to 613.5 than is *x* = 750, the quantity $(600-\overline{x})^2$ must be smaller than $(750-\overline{x})^2$. Therefore, since these quantities are the only ones that are different in the two $s_{\hat{Y}}$ values, the $s_{\hat{Y}}$ value for *x* = 600 must necessarily be smaller than the $s_{\hat{Y}}$ for *x* = 750. Said briefly, the closer *x* is to $\overline{x}$, the smaller the value of $s_{\hat{Y}}$.

b. Error degrees of freedom = *n* – 2 = 6. $t_{.025,6} = 2.447$, so the interval estimate when *x* = 600 is $2.723 \pm (2.447)(.190)$ = (2.258,3.188).

c. The 95% prediction interval is $\hat{y} \pm t_{.025,6}\sqrt{s_e^2 + s_{\hat{Y}}^2} = 2.723 \pm (2.447)\sqrt{(.534)^2 + (.190)^2}$ = (1.336,4.110). Note that the prediction interval is much wider than the CI.

d. For two 95% intervals, the simultaneous confidence level is at least 100(1 – 2(.05)) = 90%.

45. The accompanying Minitab output will be used throughout.

a. From software, the least squares regression line is $\hat{y}$ = –1.5846 + 2.58494*x*. The coefficient of determination is R^2 = 83.73% or .8373.

b. From software, a 95% CI for β_1 is roughly (2.16, 3.01). We are 95% confident that a one-unit increase in tannin concentration is associated with an increase in expected perceived astringency between 2.16 units and 3.01 units. (Since a 1-unit increase is unrealistically large, it would make more sense to say a 0.1-unit increase in *x* is associated with an increase between .216 and .301 in the expected value of *y*.)

c. From software, a 95% CI for $\mu_{Y|.6}$, the mean perceived astringency when $x = x^* = .6$, is roughly (–0.125, 0.058).

d. From software, a 95% PI for $Y|.6$, a single astringency value when $x = x^* = .6$, is roughly (–0.559, 0.491). Notice the PI is much wider than the corresponding CI, since we are making a prediction for a single future value rather than an estimate for a mean.

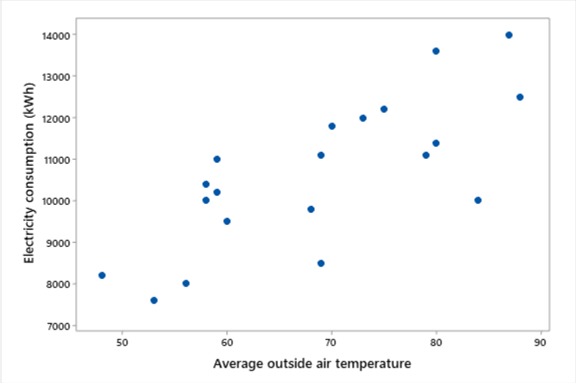
e. The hypotheses are H_0: $\mu_{Y|.7} = 0$ versus H_a: $\mu_{Y|.7} \neq 0$, where $\mu_{Y|.7}$ is the true mean astringency when $x = x^* = .7$. Since this is a two-sided test, the simplest approach is to use the 95% CI for $\mu_{Y|.7}$ provided by software. That CI, as seen in the output is roughly (0.125, 0.325). In particular, since this interval does not include 0, we reject H_0. There is evidence at the .05 level that the true mean astringency when tannin concentration equals .7 is something other than 0.

```
Regression Equation

y  =  -1.5846 + 2.58494 x

Coefficients

Term           Coef   SE Coef          T      P          95% CI
Constant  -1.58460  0.133860  -11.8377  0.000  (-1.85798, -1.31122)
x          2.58494  0.208042   12.4251  0.000  ( 2.16007,  3.00982)

Summary of Model

S = 0.253259     R-Sq = 83.73%        R-Sq(adj) = 83.19%

Predicted Values for New Observations

New Obs        Fit     SE Fit          95% CI                  95% PI
      1  -0.033635  0.0447899  (-0.125108, 0.057838)  (-0.558885, 0.491615)
      2   0.224859  0.0488238  ( 0.125148, 0.324571)  (-0.301888, 0.751606)


Values of Predictors for New Observations

New Obs    x
      1  0.6
      2  0.7
```

47. The midpoint of the CI is the point estimate: $\hat{y} = \dfrac{92.1+117.7}{2}$ = 104.9. The margin of error is 117.7 – 104.9 = 12.8 and also = $t_{.025,10-2}s_{\hat{Y}}$, so $s_{\hat{Y}} = \dfrac{12.8}{t_{.025,8}} = \dfrac{12.8}{2.306}$ = 5.55. From these, the 99% CI for $\mu_{Y|5}$ is $\hat{y} \pm t_{.005,10-2}s_{\hat{Y}}$ = 104.9 ± 3.355(5.55) = 104.9 ± 18.6 = (86.3, 123.5).

49. We will need SSE = $S_{yy} - S_{xy}^2 / S_{xx}$ = 60,089,500 – 303,515^2/2692.55 = 25,876,076 and $s_e = \sqrt{\frac{\text{SSE}}{n-2}}$ = $\sqrt{\frac{25{,}876{,}076}{20-2}}$ = 1198.98.

a. The scatterplot certainly suggests a useful relationship, but let's formally test H_0: $\beta_1 = 0$ vs H_a: $\beta_1 \neq 0$. The test statistic is $t = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \frac{112.7 - 0}{1198.98/\sqrt{2692.55}}$ = 4.88, and the P-value at 18 df is ≈ 0. So, we clearly reject H_0 at any reasonable significance level and conclude that a useful relationship exists.



b. A 95% CI for β_1 is $\hat{\beta}_1 \pm t_{.025,20-2} s_{\hat{\beta}_1} = 112.7 \pm 2.101 \cdot \frac{1198.98}{\sqrt{2692.55}}$ = (64.2, 161.3). At the 95% confidence level, a 1°F increase in average outside air temperature is associated with a increase in expected electricity consumption between 64.2 and 161.3 kWh.

c. The predicted value at x^* = 70 is 2906 + 112.7(70) = 10,795. With the aid of software, the standard error of $\hat{Y}$ when x^* = 70 is $s_{\hat{Y}} = 269.9$. The CI is 10,795 ± 2.101(269.9) = (10228, 11362).

d. 10,795 ± $2.101\sqrt{1198.98^2 + 269.9^2}$ = (8215, 13379).

e. Wider, because 85 is farther from the mean x-value of 68.65 than is 70.

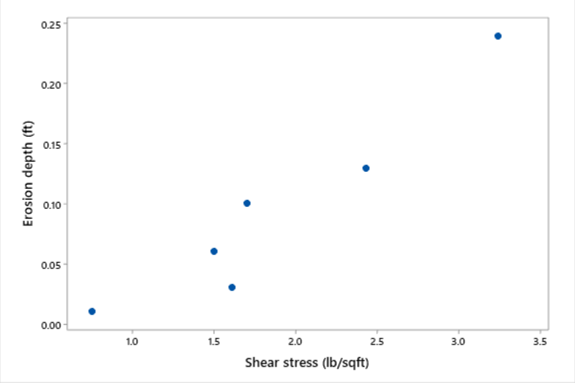
f. No: Looking at the scatterplot, x = 95 is well outside the scope of the observed data. This suggests that estimates or predictions made at x = 95 are not necessarily trustworthy, since we don't know that the apparently linear trend will continue.

g. To achieve a simultaneous confidence level of at least 97% (α = .03) for three intervals, we need to use individual confidence level 100(1 – .03/3)% = 100(1 – .01)% = 99%. That is, we'll construct three 99% CI's. The t critical value is $t_{.005,18}$ = 2.878.

x^*	$\hat{y}$	$s_{\hat{Y}}$	99% CI for $\mu_{Y\|x^*}$
60	9670	334.4	(8707, 10633)
70	10,795	269.9	(10020, 11574)
80	11,924	375.0	(10845, 13004)

51.

a. Yes, the overall trend is a strong, positive, linear association between shear stress and erosion depth.



b. Test H_0: $\beta_1 = 0$ vs H_a: $\beta_1 \neq 0$. With the aid of software, $\hat{\beta}_1 = .0931$ and $s_{\hat{\beta}_1} = .0144$, so the test statistic is $t = \dfrac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \dfrac{.0931 - 0}{.0144} = 6.45$, and the P-value at 6 – 2 = 4 df is ≈ .003. So, we reject H_0 at even the .005 significance level and conclude that a useful relationship exists.

c. At $x^* = 1.75$, $\hat{y} = .08367$ and $s_{\hat{y}} = .01144$ from software. So, a 95% CI for $\mu_{Y|1.75}$ is $\hat{y} \pm t_{.025,4} s_{\hat{Y}}$ = $.08367 \pm 2.776(.01144) = (.05191, .11544)$.

d. From software, $s_e = .02769$, so the 95% PI for $Y|1.75$ is $.08367 \pm 2.776\sqrt{.02769^2 + .01144^2}$ = (.00048, .16687).

Section 12.5

53. Most people acquire a license as soon as they become eligible. If, for example, the minimum age for obtaining a license is 16, then the time since acquiring a license, y, is usually related to age by the equation $y \approx x - 16$, which is the equation of a straight line. In other words, the majority of people in a sample will have y values that closely follow the line $y = x - 16$.

55.

a. We are testing $H_0 : \rho = 0$ vs $H_a : \rho > 0$. $r = \dfrac{7377.704}{\sqrt{36.9839}\sqrt{2,628,930.359}} = .7482$, and

$t = \dfrac{.7482\sqrt{12}}{\sqrt{1 - .7482^2}} = 3.9066$. We reject H_0 since $t = 3.9066 \geq t_{.05,12} = 1.782$. There is evidence that a positive correlation exists between maximum lactate level and muscular endurance.

b. We are looking for R^2, the coefficient of determination. $R^2 = r^2 = (.7482)^2 = .5598$. It is the same no matter which variable is the predictor.

57.

a. $H_0 : \rho = 0$ vs $H_\text{a} : \rho \neq 0$; reject H_0 at level .05 if $|t| \geq t_{.025,12} = 2.179$. $t = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \dfrac{(.449)\sqrt{12}}{\sqrt{1-(.449)^2}} = 1.74$.

Hence we fail to reject H_0; the data does not suggest that the population correlation coefficient differs significantly from 0.

b. $(.449)^2 = .20$, so 20 percent of the observed variation in gas porosity can be accounted for by its linear relationship to hydrogen content.

59.

a. Perform the log-transform specified in the section: $v = \frac{1}{2}\ln\left(\frac{1+.878}{1-.878}\right)$ = 1.367. A 90% CI for $\mu_V = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ is $v \pm \frac{z_{.05}}{\sqrt{83-2}} = 1.367 \pm \frac{1.645}{\sqrt{81}}$ = (1.184, 1.550). Inverting the transformation, a 90% CI for ρ is $\left(\frac{e^{2(1.184)}-1}{e^{2(1.184)}+1}, \frac{e^{2(1.550)}-1}{e^{2(1.550)}+1}\right)$ = (.829, .914).

b. Using the same log-transform, we have $\mu_V = \frac{1}{2}\ln\left(\frac{1+.8}{1-.8}\right)$ = 1.099 when H_0 is true. Thus, the test statistic value is $z = \frac{1.367-1.099}{1/\sqrt{83-2}}$ = 2.412 and P-value = $P(Z \geq 2.412) = 1 - \Phi(2.412) \approx .008$. Because .008 < .05, we reject H_0 in favor of H_a and conclude that the population correlation coefficient between digital caliper and laser arm measurements exceeds .8. *Note:* Since the lower bound of the CI in part (a) is also a 95% *lower* confidence bound for ρ, we could have rejected H_0 because $\rho > .829 > .8$.

c. $R^2 = r^2 = .878^2 = .771$, or 77.1%.

d. When x and y are reversed, neither r nor R^2 change, so the answer is still 77.1%.

61.

a. Because P-value = .00032 < α = .001, H_0 should be rejected at this significance level.

b. Not necessarily. For this n, the test statistic t has approximately a standard normal distribution when $H_0 : \rho = 0$ is true, and a P-value of .00032 corresponds to $z = \pm 3.60$. Solving $3.60 = \dfrac{r\sqrt{498}}{\sqrt{1-r^2}}$ for r yields $r = .159$. This r suggests only a weak linear relationship between x and y, one that would typically have little practical importance.

c. $t = \dfrac{.022\sqrt{9998}}{\sqrt{1-.022^2}} = 2.20 \geq t_{.025,9998} = 1.96$, so H_0 is rejected in favor of H_a. The value $t = 2.20$ is statistically significant — i.e., it cannot be attributed just to sampling variability in the case $\rho = 0$. But with this n, $r = .022$ implies $\rho \approx .022$, which in turn shows an extremely weak linear relationship.

63. Re-write both statistics in terms of the original sums of squares. The test statistic from Section 12.3 is $\frac{b_1 - 0}{S_e / \sqrt{S_{xx}}} = \frac{S_{xy} / S_{xx}}{\sqrt{\text{SSE} / n-2} / \sqrt{S_{xx}}} = \frac{S_{xy}\sqrt{n-2}}{\sqrt{\text{SSE}(S_{xx})}}$. Meanwhile, since SST and S_{yy} are the same thing, $T = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{(S_{xy} / \sqrt{S_{xx}S_{yy}})\sqrt{n-2}}{\sqrt{\text{SSE/SST}}} = \frac{S_{xy}\sqrt{n-2}}{\sqrt{\text{SSE}(S_{xx})}}$.

65.

a. We used software to calculate the r_i's: $r_1 = 0.184$, $r_2 = –0.238$, and $r_3 = –0.426$.

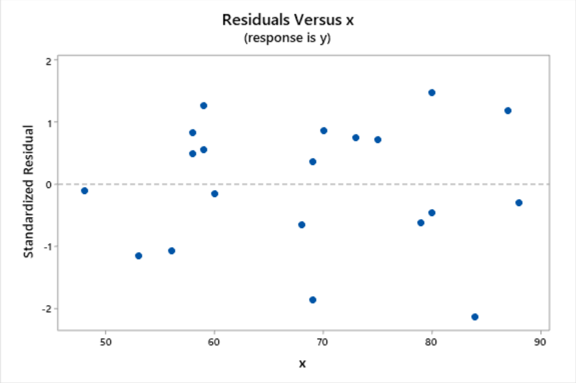
b. The only difference between lag autocorrelation coefficients and regular correlation is the number of terms in the numerator summand: the sum only runs from 1 to $n – 1$, but $\bar{x}$ is based upon all n observations. In regular correlation, $\bar{x}$ would be replaced in each part of the numerator by the mean of just the relevant $n – 1$ values (1 through $n – 1$ in the first parentheses, 2 through n in the second). As n gets larger, the difference between $\bar{x}$ and these "truncated" means becomes negligible. A similar comment applies to lag 2.

c. $\frac{2}{\sqrt{100}} = .2$. We reject H_0 if $|r_i| \geq .2$. For all lags, r_i does not fall in the rejection region, so we cannot reject H_0. There is not evidence of theoretical autocorrelation at the first 3 lags.

d. If we want an approximate .05 significance level for the simultaneous hypotheses, we would have to use smaller individual significance level. If the individual confidence levels were .95, then the simultaneous confidence levels would be approximately (.95)(.95)(.95) = .857.

Section 12.6

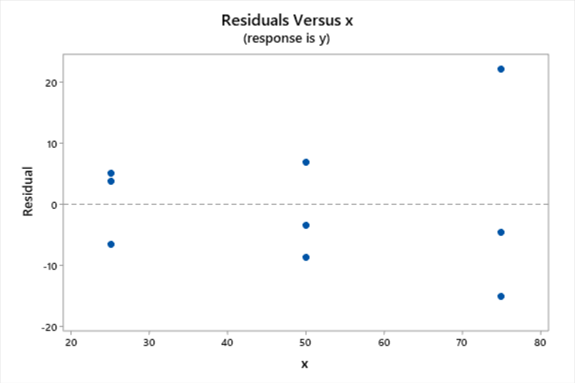
67. The accompanying graph is a plot of e^* versus x. As desired, the plot exhibits neither curvature nor a pattern of increasing/decreasing vertical (i.e., residual) spread. These suggest that the regression model assumptions of linearity/model adequacy and constant error variance are both plausible.



69.

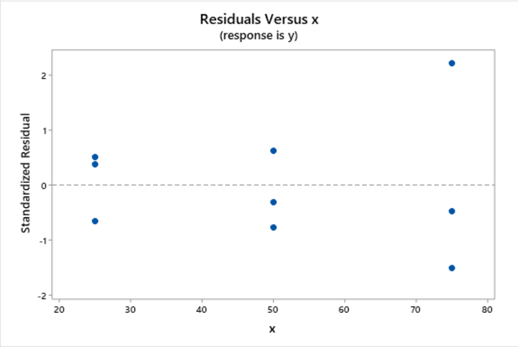
a. For each observation, $e_i = y_i - \hat{y}_i = y_i - (182.7 + 3.29x_i)$. The accompanying plot of *e*'s versus *x*'s does not show curvature (that's good), but it shows greater variability at *x* = 75 than at other *x* values. The latter suggests that the equal variance assumption of the simple linear regression model is *not* satisfied.

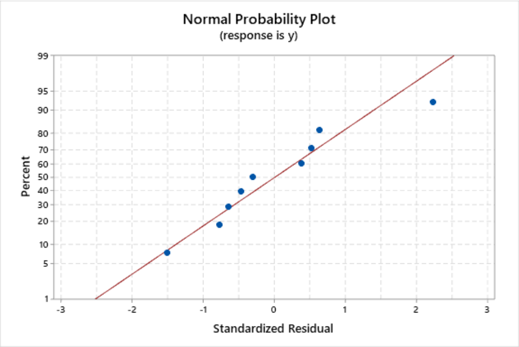
x_i	25	25	25	50	50	50	75	75	75
e_i	–6.50	3.82	5.20	–8.65	–3.37	7.01	–15.12	–4.68	22.32



b. The table below shows the standardized residuals. The standardized residual plot shows the same issue as the previous graph: a lack of constant variance across all *x*-values. The normal probability plot of the standardized residuals is at least roughly linear (no huge deviations from the reference line), so normality of the true errors is at least plausible.

x_i	25	25	25	50	50	50	75	75	75
e_i^*	–0.65	0.38	0.52	–0.78	–0.30	0.63	–1.51	–0.47	2.23



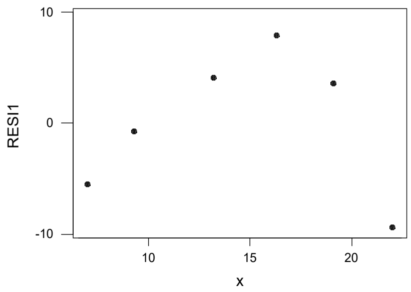


71.

a. H_0: $\beta_1 = 0$ vs H_a: $\beta_1 \neq 0$. The test statistic is $t = \dfrac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$, and we will reject H_0 if $|t| \geq t_{.025,4} = 2.776$.

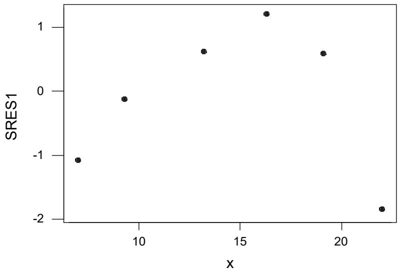
$s_{\hat{\beta}_1} = \dfrac{s_e}{\sqrt{S_{xx}}} = \dfrac{7.265}{12.869} = .565$, and $t = \dfrac{6.19268}{.565} = 10.97$. Since $10.97 \geq 2.776$, we reject H_0. We are *tempted* to conclude that the linear model is useful. However, this test assumes that a true linear relationship exists between x and y, which is contradicted by the residual plots below.

b. $\hat{y}_{(7.0)} = 1008.14 + 6.19268(7.0) = 1051.49$, from which the residual is $y - \hat{y}_{(7.0)} = 1046 - 1051.49 = -5.49$. Similarly, the other residuals are –.73, 4.11, 7.91, 3.58, and –9.38. The plot of the residuals vs x follows:



Because a curved pattern appears, a linear regression function is inadequate.

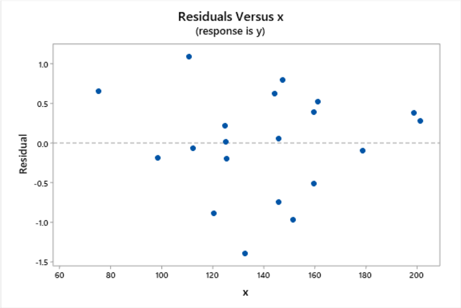
c. The standardized residuals are calculated as $e_1{}^* = \dfrac{-5.49}{7.265\sqrt{1 - \frac{1}{6} - \frac{(7.0 - 14.48)^2}{165.5983}}} = -1.074$, and similarly the others are $-.123$, .624, 1.208, .587, and –1.841. The plot of e^* vs x follows :



This plot gives the same information as the previous plot. No values are exceptionally large, but the e^* of –1.841 is close to 2 std deviations away from the expected value of 0.

73.

a. This plot indicates there are no outliers, but there appears to be higher variance for middle values of filtration rate.



b. We need $S_{xx} = \sum \left(x_i - \bar{x}\right)^2 = 18{,}886.8295$. Then each e_i^* can be calculated as follows:

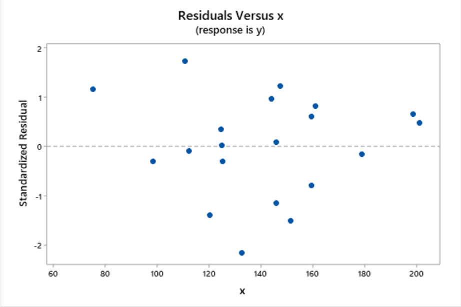
$$e_i^* = \frac{e_i}{.665\sqrt{1 - \frac{1}{20} - \frac{(x_i - 140.895)^2}{18{,}886.8295}}}$$

. The table below shows the values:

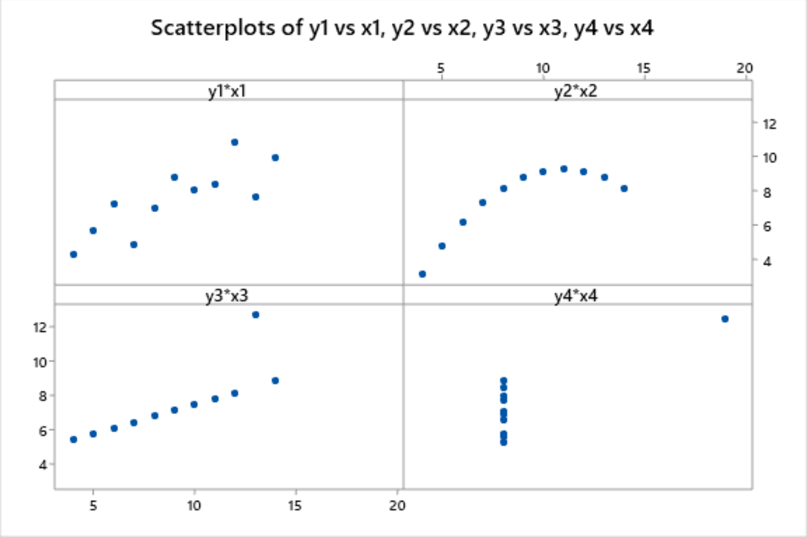
standardized residuals	e_i / e_i^*	**standardized residuals**	e_i / e_i^*
-0.31064	0.644053	0.6175	0.64218
-0.30593	0.614697	0.09062	0.64802
0.4791	0.578669	1.16776	0.565003
1.2307	0.647714	-1.50205	0.646461
-1.15021	0.648002	0.96313	0.648257
0.34881	0.643706	0.019	0.643881
-0.09872	0.633428	0.65644	0.584858
-1.39034	0.640683	-2.1562	0.647182
0.82185	0.640975	-0.79038	0.642113
-0.15998	0.621857	1.73943	0.631795

Notice that if $e_i^* \approx e_i / s_e$, then $e_i / e_i^* \approx s_e$. All of the e_i / e_i^* 's range between .57 and .65, which are close to s_e.

c. This plot looks very much the same as the one in part a.



75. Both a scatter plot and residual plot (based on the simple linear regression model) for the first data set suggest that a simple linear regression model is reasonable, with no pattern or influential data points which would indicate that the model should be modified. However, scatter plots for the other three data sets reveal difficulties. For data set #2, a quadratic function would clearly provide a much better fit. For data set #3, the relationship is perfectly linear except one outlier, which has obviously greatly influenced the fit even though its x value is not unusually large or small. One might investigate this observation to see whether it was mistyped and/or it merits deletion. For data set #4 it is clear that the slope of the least squares line has been determined entirely by the outlier, so this point is extremely influential. A linear model is completely inappropriate for data set #4. And all of this is true despite the fact that the summary statistics for all four data sets are practically identical! *The lesson:* Always graph your data!



Section 12.7

77.

a. Since $E(\varepsilon) = 0$, the expected sales when there are $x_1 = 2$ competing outlets and $x_2 = 8$ thousand people in a one-mile radius is 10000 – 1400(2) + 2100(8) = $24,000.

b. Similarly, 10000 – 1400(3) + 2100(5) = $16,300.

c. $\beta_1 = -1400$: Adjusting for the size of the nearby population, an increase of one competing outlet corresponds to a $1400 *decrease* in expected weekly sales.
$\beta_2 = 2100$: Adjusting for the number of competing outlets, an increase of 1 thousand people within a one-mile radius corresponds to a $2100 increase in expected weekly sales.

d. $\beta_0 = 10000$: In an area with 0 competing outlets and 0 people living within a one-mile radius, expected weekly sales are $10,000. This *might* make sense for a highway/roadside fast-food outlet in the middle of nowhere (so no competition but also no surrounding population).

79.

a. Adjusting for fit, arch support, and stability, a one-point increase in a shoe's cushioning rating from any particular person is associated with a .34 increase in its estimated overall preference score from that person.

b. $\hat{y}$ = -.66 + .35(9.0) + .34(8.7) + .09(8.9) + .32(9.2) = 9.193 points (out of 15 max). It would be more informative to provide a *confidence interval* for the mean overall preference score at these settings.

c. Test $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ vs H_a: not all β's are 0. With k = 4 predictors and n = 100 runners/observations, the test statistic value is $f = \dfrac{R^2}{1-R^2}\dfrac{n-(k+1)}{k} = \dfrac{.777}{1-.777}\dfrac{100-5}{4}$ = 82.75. This is an extremely large F statistic; in particular, 82.75 > $F_{.01,4,95} = 3.52$, so H_0 is resoundingly rejected. This indicates that *at least one* of the four predictor variables has a significant relationship with overall score, but not necessarily that *all* of them do.

d. To achieve a family-wise .01 significance level requires testing each of the 4 null hypotheses at α = .01/4 = .0025. With error df = 95, the critical value for each of the four tests H_0: β_j = 0 vs H_a: $\beta_j \neq 0$ is $t_{\alpha/2,n-(k+1)} = t_{.00125,95}$ = 3.106. With the t-values provided, all null hypotheses are rejected except j = 3. Thus, variables x_1 and x_2 and x_4 are deemed useful, but after adjusting for those variables, x_3 is not deemed a statistically significant predictor.

81.

a. Software provides $\hat{\beta}_0 = -77$, $\hat{\beta}_1$ = 4.397, and $\hat{\beta}_2$ = 165. Therefore, with y = price and x_1 = size and x_2 = L/B ratio, the estimated regression equation is $y = -77 + 4.397x_1 + 165x_2$.

b. Interpreting the intercept doesn't make sense here. $\hat{\beta}_1$ = 4.397 means that after adjusting for the effects of land-to-building ratio, a 1 thousand square foot increase in size is associated with an estimated increase in expected price of 4.397 thousand dollars ($4,397). $\hat{\beta}_2$ = 165 means that after adjusting for the effects of size, an increase of 1 in the L/B ratio (e.g., from 2:1 to 3:1) corresponds to an estimated increase of 165 thousand dollars ($165,000) in expected price.

c. $\hat{y}$ = –77 + 4.397(500) + 165(4.0) = 2781.5 = $2,781,500.

83.

a. With the aid of software, SST = $\sum(y_i - \bar{y})^2$ = 17.024 and SSE = $\sum e_i^2$ = 11.226. From these, $R^2 = 1 - \dfrac{11.226}{17.024}$ = .3405, or 34.05%. That is, 34.05% of the observed variation in electrical conductivity can be explained by a regression model with CNT weight, CN$_x$ height, and water volume as predictors. Also, $s_e = \sqrt{\dfrac{11.226}{16-(3+1)}}$ = 0.967. The typical difference between the actual and predicted electrical conductivity of a CNT specimen is ± 0.967 S/cm.

b. The hypotheses of the model utility test are $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ vs H_a: not all β's are 0. With k = 3 predictors and n = 16 specimens, the test statistic value is $f = \dfrac{R^2}{1-R^2}\dfrac{n-(k+1)}{k} = \dfrac{.3405}{1-.3405}\dfrac{16-4}{3} =$

2.065. Since 2.065 < $F_{.05,3,12}$ = 3.49, we *do not* reject H_0 at the .05 level. That is, the data do *not* provide convincing evidence that at least one of the three explanatory variables is useful for predicting *y*.

c. Yes, in the sense that the test in part **b** failed to detect a statistically significant relationship between any of the *x*'s and electrical conductivity. (Of course, that is not to say we *proved* H_0 is true.)

85.

a. With the aid of software, the estimated regression equation is $y = 148 - 133x_1 + 128.5x_2 + 0.0351x_3$.

b. Information for the three variable utility tests appear below. The *P*-values suggest that, at any reasonable significance level, only x_2 is a statistically significant predictor of *y*.

$\hat{\beta}_j$	$s_{\hat{\beta}_j}$	*t*-statistic	*P*-value (df = 19)
–133	511	–0.26	.798
128.5	13.6	9.43	< .0001
0.0351	0.0247	1.42	.171

c. From software, SSE = 385,801 and SST = 2,822,482, from which $R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$ = .8633 or 86.33%, while $R_a^2 = 1 - \frac{\text{MSE}}{\text{MST}} = 1 - \frac{385{,}801/(23-(3+1))}{2{,}822{,}482/(23-1)}$ = .8417 or 84.17%.

d. For this two variable-regression, SSE = 387,170, while SST remains at 2,822,482. From the updated SSE, R^2 = .8628 or 86.28% and $R_a^2 = 1 - \frac{387{,}170/(23-(2+1))}{2{,}822{,}482/(23-1)}$ = .8491 or 84.91%. The R^2 value is (slightly) larger under the full model (k = 3) than under the reduced model (k = 2). This must always be true: SSE cannot increase when more predictors are included, so R^2 can never be smaller with a larger set of predictors. However, adjusted R^2 is (slightly) *larger* for the reduced model: 84.91% vs 84.17%. This suggests that if we adjust for the number of predictors in the model, the reduced (k = 2) model does a better job than does the full (k = 3) model.

87.

a. The hypotheses are H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ vs. H_a: at least one $\beta_i \neq 0$. The test statistic is f = $\frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{.946/4}{(1-.946)/20}$ = 87.6 ≥ $F_{.001,4,20}$ = 7.10 (the smallest available significance level from Table A.8), so we can reject H_0 at any significance level. We conclude that at least one of the four predictor variables appears to provide useful information about tenacity.

b. The adjusted R^2 value is $1 - \frac{n-1}{n-(k+1)}\left(\frac{\text{SSE}}{\text{SST}}\right) = 1 - \frac{n-1}{n-(k+1)}(1-R^2) = 1 - \frac{24}{20}(1-.946) = .935$, which does not differ much from R^2 = .946.

c. The estimated average tenacity when x_1 = 16.5, x_2 = 50, x_3 = 3, and x_4 = 5 is $\hat{y} = 6.121 - .082(16.5) + .113(50) + .256(3) - .219(5) = 10.091$. For a 99% CI, $t_{.005,20} = 2.845$, so the interval is $10.091 \pm 2.845(.350) = (9.095, 11.087)$. Therefore, when the four predictors are as specified in this problem, the true average tenacity is estimated to be between 9.095 and 11.087.

Section 12.8

89.

a. The vertex of the least-squares parabola occurs at $x = -\frac{b}{2a} = -\frac{-.0191}{-1.92\times 10^{-5}} = -994.8^{\circ}\text{C}$. Since the quadratic coefficient is negative, the equation suggests that elastic modulus increases with temperatures up to –994.8°C and decreases after that. But since temperatures that low do not exist (that's below absolute zero), we conclude that elastic modulus decreases with temperature through all physically possible temperatures.

b. $\hat{y} = -1.92\times 10^{-5}(800)^2 - .0191(800) + 89.0 = 61.432$ GPa.

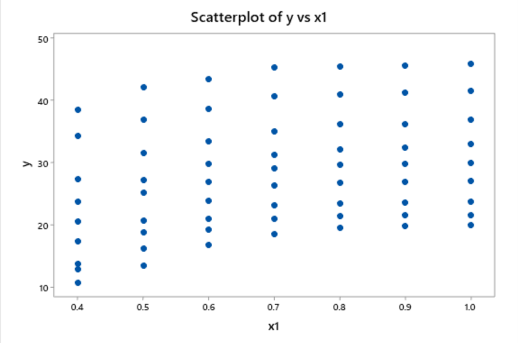
c. We test $H_0: \beta_1 = \beta_2 = 0$ versus H_a: not all β's are 0. With $k = 2$ terms in the model, $n = 28$, and $R^2 =$ .948, the model utility test statistic is $f = \frac{R^2}{1-R^2}\frac{n-(k+1)}{k} = \frac{.948}{1-.948}\frac{28-(2+1)}{2} = 227.88$. This is an extremely large F-value; in particular, 227.88 > $F_{.01,2,25} = 5.568$. Hence, H_0 is rejected at the .01 level, and we conclude that at least one of the two terms in the quadratic model is useful for predicting y.

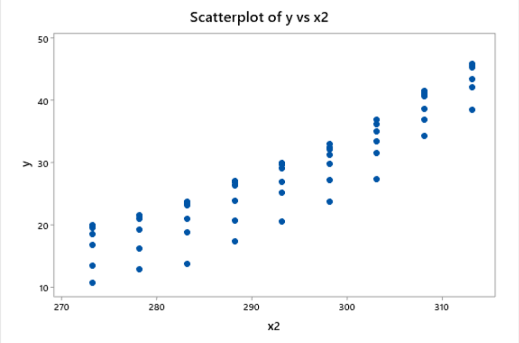
d. From part **b**, $\hat{y} = 61.432$. Since $t_{.025,25} = 2.060$, a 95% CI for $\mu_{Y|800}$ is 61.432 ± 2.060(2.9) = (55.458, 67.406).

e. The 95% PI is $\hat{y} \pm 2.060\sqrt{s_e^2 + s_{\hat{Y}}^2} = 61.432 \pm 2.060\sqrt{2.37^2 + 2.9^2} = 61.432 \pm 7.715 = (53.717, 69.147)$. With 95% confidence, the elastic modulus of a single ceria specimen at 800°C will be between 53.717 and 69.147 GPa.

91.

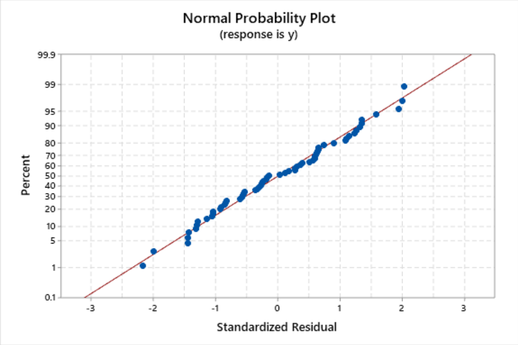
a. Both scatterplots exhibit some curvature, suggesting that quadratic terms may be appropriate.

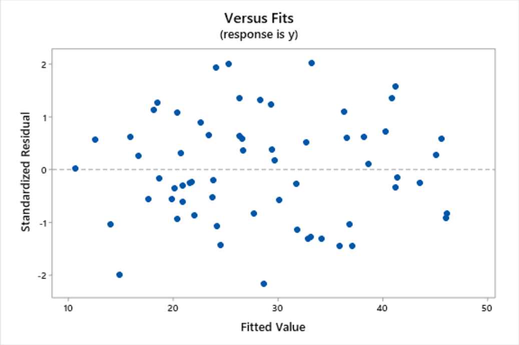




b. No: Interaction is a property of the simultaneous relationship between x_1, x_2, and y. A scatter plot of (x_1, x_2) pairs could not indicate whether the effect of x_1 on y depends on x_2 and vice versa.

c. With the aid of software, a multiple regression was performed with response variable y and predictors $x_1, x_2, x_1^2, x_2^2, x_1x_2$. A normal probability plot of the resulting standardized residuals appears below, as does a plot of the standardized residuals against the fitted values (i.e., e_i^* versus $\hat{y}_i$). The linearity of the normal plot suggests that it's very plausible that errors are normally distributed. The residual-v-fit plot shows neither curvature nor varying vertical spread, so the model adequacy and constant variance assumptions are both reasonable as well.





d. We test separately the null hypotheses $H_0 : \beta_3 = 0$, $H_0 : \beta_4 = 0$, and $H_0 : \beta_5 = 0$. From software, the test statistic values are $t_3 = \dfrac{-36.19}{2.18} = -16.58$, $t_4 = \dfrac{.008850}{.000517} = 17.12$, and $t_5 = \dfrac{-.0870}{.0293} = -2.97$. The two-sided *P*-values at 57 df are ≈ .000, ≈ .000, and .004, respectively. At most reasonable significance levels, all three of the aforementioned null hypotheses are rejected, suggesting all three second-order terms should be retained in the model. (With higher-order terms retained, we don’t bother to test the first-order terms and they, too, persist in our model.)

93.

a. Yes, there does appear to be a useful linear relationship between repair time and the two model predictors. We determine this by conducting a model utility test $H_0 : \beta_1 = \beta_2 = 0$ vs. H_a: not all β’s are 0. We reject H_0 if $f \ge F_{.05,2,9} = 4.26$. The calculated statistic is $f = \dfrac{\text{SSR}/k}{\text{SSE}/(n-k-1)} = \dfrac{10.63/2}{20.9/9} = \dfrac{5.315}{.232} = 22.91$. Since $22.91 \ge 4.26$, we reject H_0 and conclude that at least one of the two predictor variables is useful.

b. We will reject $H_0 : \beta_2 = 0$ in favor of H_a: $\beta_2 \ne 0$ if $|t| \ge t_{.005,9} = 3.25$. The test statistic is $t = \dfrac{1.250}{.312} = 4.01$ which is ≥ 3.25, so we reject H_0 and conclude that the “type of repair” variable does provide useful information about repair time, given that the “elapsed time since the last service” variable remains in the model.

c. A 95% confidence interval for β_2 is: $1.250 \pm (2.262)(.312) = (.5443, 1.9557)$. We estimate, with a high degree of confidence, that when an electrical repair is required the repair time will be between .54 and 1.96 hours longer than when a mechanical repair is required, while the “elapsed time” predictor remains fixed.

d. $\hat{y} = .950 + .400(6) + 1.250(1) = 4.6$, $s_e^2 = \text{MSE} = .23222$, and $t_{.005,9} = 3.25$, so the 99% PI is $4.6 \pm (3.25)\sqrt{(.23222) + (.192)^2} = 4.6 \pm 1.69 = (2.91, 6.29)$ The prediction interval is quite wide, suggesting a variable estimate for repair time under these conditions.

95.

a. The complete second-order model obviously provides a much better fit, so there is a need to account for quadratic and interaction effects from these three predictors.

b. A complete second-order model based on three predictors has $3+3+\binom{3}{2}$ = 3 + 3 + 3 = 9 terms, so degrees of freedom = n – (k + 1) = 20 – (9 + 1) = 10. A 95% PI for Y|(30,30,10) is $\hat{y} \pm t_{.025,10}\sqrt{s_e^2 + s_{\hat{Y}}^2}$ = $.66573 \pm 2.228\sqrt{.044^2 + .01785^2}$ = (.560, .771).

97.

a. From Minitab, here are the correlations and corresponding P-values:

```
            IBU     ABV
ABV       0.843
          0.000

Rating    0.843   0.621
          0.000   0.001
```

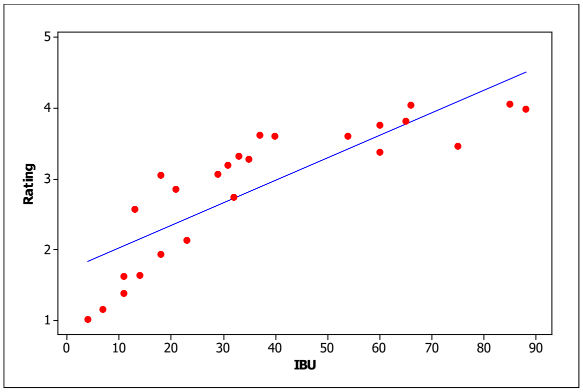
The correlations are all strongly significant, including the correlation between the two predictors.

b. Here is some of the Minitab regression output:

```
The regression equation is
Rating = 2.24 + 0.0419 IBU - 0.166 ABV

Predictor       Coef   SE Coef       T      P
Constant      2.2383    0.3961    5.65  0.000
IBU         0.041940  0.007688    5.46  0.000
ABV          -0.1661    0.1078   -1.54  0.138

S = 0.507612   R-Sq = 73.9%   R-Sq(adj) = 71.5%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       2  16.0266  8.0133  31.10  0.000
Residual Error  22   5.6687  0.2577
Total           24  21.6953
```

Although ABU has a strongly significant, ABV does not. This means that, with ABU in the model, ABV is not needed. Even though ABV has a strongly significant relationship with Rating, ABV is redundant when ABU is included. The idea is that ABV is strongly correlated with ABU, so when ABU is already in the model, ABV has very little new to add to the model.

c. Here is the plot of Rating against IBU. Notice that the plot is not linear; the slope decreases as we move to the right. This suggests including a quadratic term.



d. Here is some of the Minitab output with the quadratic term included:

```
The regression equation is
Rating = 0.214 + 0.0953 IBU + 0.131 ABV - 0.000801 IBUsq

Predictor        Coef     SE Coef       T      P
Constant       0.2142      0.5181    0.41  0.683
IBU           0.09533     0.01269    7.51  0.000
ABV            0.1311      0.1001    1.31  0.205
IBUsq      -0.0008014   0.0001716   -4.67  0.000

S = 0.363873   R-Sq = 87.2%   R-Sq(adj) = 85.4%

Analysis of Variance

Source          DF       SS      MS      F      P
Regression       3  18.9148  6.3049  47.62  0.000
Residual Error  21   2.7805  0.1324
Total           24  21.6953
```

Plots of Rating and the residuals against IBU no longer show curvature, the normal plot is reasonably straight, and there is no reason to doubt constant variance.

e. Notice that the quadratic term is highly significant, but the ABV term is still not needed. The R-Squared, adjusted R-Squared, and *s* are substantially improved. Notice that the quadratic coefficient is negative, in accord with the decreasing slope.

f. The model now does a good job of fitting the relationship of Rating to IBU. ABV is redundant when IBU is included.

Section 12.9

99.

a. The data and response matrices are $\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix}$ and $\mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 4 \end{bmatrix}$. The normal equations are $\mathbf{X'Xb} = \mathbf{X'y}$, which here become $\begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}\mathbf{b} = \begin{bmatrix} 6 \\ 2 \\ 4 \end{bmatrix}$.

b. Since $\mathbf{X'X} = 4\mathbf{I}$, $(\mathbf{X'X})^{-1} = .25\mathbf{I}$, and $\mathbf{b} = .25\mathbf{X'y} = \begin{bmatrix} 1.5 \\ 0.5 \\ 1.0 \end{bmatrix}$.

c. $\hat{\mathbf{y}} = \mathbf{Xb} = \begin{bmatrix} 0 \\ 2 \\ 1 \\ 3 \end{bmatrix}$, from which SSE = $\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = (1)^2 + (-1)^2 + (-1)^2 + (1)^2 = 4$, and MSE = SSE/[4-(2+1)] = SSE/1 = 4.

d. $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X'X})^{-1} \approx \text{MSE}(\mathbf{X'X})^{-1} = 4 \cdot .25\mathbf{I} = \mathbf{I}$. So, in particular, $s^2_{\hat{\beta}_1}$ = (2,2) entry of the estimated covariance matrix = (2,2) entry of **I** = 1. Thus a 95% CI for β_1 is $\hat{\beta}_1 \pm t_{.025,4-3} s_{\hat{\beta}_1} = 0.5 \pm 12.706(1) =$ (–12.206,13.206). The CI is so large because we only have 1 df (4 observations, 3 parameters).

e. The *t* statistic here is $t = \dfrac{0.5 - 0}{1} = 0.5$, which at 1 df has a 2-sided *P*-value of 2(.352) = .704. We certainly fail to reject the hypothesis that $\beta_1 = 0$. This is consistent with our 95% CI from part (d).

f. $\bar{y}$ = 6/4 = 1.5, so SSR = $\|\hat{\mathbf{y}} - \bar{\mathbf{y}}\|^2 = (-1.5)^2 + (.5)^2 + (-.5)^2 + (1.5)^2 = 5$. The rest of the ANOVA table below follows. In particular, the *F* test statistic is *f* = 0.63 with a corresponding P-value of .667, so we definitely fail to reject H_0. Both slopes could plausibly be zero, and so it appears neither x_1 nor x_2 is a useful predictor for *y*. Finally, R^2 = SSR/SST = 5/9 = 55.56%; that is, ~56% of the variability in *y* can be explained by the linear regression model that involves predictors x_1 and x_2.

```
Source          DF     SS      MS      F      P
Regression       2  5.000   2.500   0.63  0.667
Residual Error   1  4.000   4.000
Total            3  9.000
```

101.

a. First,

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}$$

Then, using the matrix inverse formula provided,

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2} \begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix}$$

b. $\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$, so from part (a)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2} \begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix} \begin{bmatrix} \Sigma y_i \\ \Sigma x_i y_i \end{bmatrix}$$

$$= \frac{1}{n\Sigma x_i^2 - (\Sigma x_i)^2} \begin{bmatrix} (\Sigma x_i^2)(\Sigma y_i) - (\Sigma x_i)(\Sigma x_i y_i) \\ n\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i) \end{bmatrix}$$

To make these resemble the earlier formulas, make the following substitutions: $\Sigma x_i = n\bar{x}$, $\Sigma y_i = n\bar{y}$, $\Sigma x_i y_i = S_{xy} + n\bar{x}\,\bar{y}$ (based on the hint) and, similarly, $\Sigma x_i^2 = S_{xx} + n\bar{x}^2$. The fraction outside the matrix simplifies to nS_{xx}, and

$$\hat{\boldsymbol{\beta}} = \frac{1}{nS_{xx}} \begin{bmatrix} (S_{xx} + n\bar{x}^2)(n\bar{y}) - (n\bar{x})(S_{xy} + n\bar{x}\,\bar{y}) \\ n(S_{xy} + n\bar{x}\,\bar{y}) - n^2\bar{x}\,\bar{y} \end{bmatrix} = \frac{1}{nS_{xx}} \begin{bmatrix} n\bar{y}S_{xx} - n\bar{x}S_{xy} \\ nS_{xy} \end{bmatrix} = \begin{bmatrix} \bar{y} - (S_{xy}/S_{xx})\bar{x} \\ S_{xy}/S_{xx} \end{bmatrix}$$

These match our previous formulas: $\hat{\beta}_1 = S_{xy}/S_{xx}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$.

c. From part (b), $(\mathbf{X}'\mathbf{X})^{-1}$ may be rewritten as $(\mathbf{X}'\mathbf{X})^{-1} = \dfrac{1}{nS_{xx}} \begin{bmatrix} S_{xx} + n\bar{x}^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix}$. Since $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, we have $V(\hat{\beta}_1) = \sigma^2 \cdot n/(nS_{xx}) = \sigma^2/S_{xx}$ and $V(\hat{\beta}_0) = \sigma^2 \cdot (S_{xx} + n\bar{x}^2)/(nS_{xx}) = \sigma^2 \cdot (1/n + \bar{x}^2/S_{xx})$.

The first formula matches $V(\hat{\beta}_1)$ given in Section 12.3. As for the second, substitute $x^* = 0$ into the LSRL; according to Section 12.4, the resulting variable $\hat{Y} = \hat{\beta}_0$ has variance $V(\hat{Y}) = \sigma^2\left(1/n + (0-\bar{x})^2/S_{xx}\right) = \sigma^2\left(1/n + \bar{x}^2/S_{xx}\right)$, which matches $V(\hat{\beta}_0)$ above.

103. The design matrix is now just the column vector $\mathbf{X} = [1,\ldots,1]'$, so $\mathbf{X}'\mathbf{X} = n$, $\mathbf{X}'\mathbf{y} = \Sigma y_i$ and $[\hat{\beta}_0] = \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = [1/n][\Sigma y_i] = [\bar{y}]$. Next, $V(\hat{\beta}_0) = \mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2/n$, so $SD(\hat{\beta}_0) = \sigma/\sqrt{n}$ and $s_{\hat{\beta}_0} = s_e/\sqrt{n}$. Finally, for every *i* we have $\hat{y}_i = \hat{\beta}_0 = \bar{y}$, so $e_i = y_i - \hat{y}_i = y_i - \bar{y}$ and $s_e^2 = \dfrac{\Sigma e_i^2}{n-(k+1)} = \dfrac{\Sigma(y_i - \bar{y})^2}{n-1} = s_y^2$. Therefore, a CI for $\beta_0 = E(Y)$ is given by

$\hat{\beta}_0 \pm t_{\alpha/2,n-(k+1)} \cdot s_{\hat{\beta}_0} = \bar{y} \pm t_{\alpha/2,n-1} \cdot s_e/\sqrt{n} = \bar{y} \pm t_{\alpha/2,n-1} \cdot s_y/\sqrt{n}$. This is exactly the one-sample *t* CI from Chapter 8. In other words, the one-sample *t* procedures are the $k = 0$ special case of the regression model!

105.

a. The proposed design matrix and response vector are $\mathbf{X}' = \begin{bmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ .5 & \cdots & .5 & -.5 & \cdots & -.5 \end{bmatrix}$ and $\mathbf{y} = [y_{11} \quad \cdots \quad y_{m1} \quad y_{(m+1)1} \quad \cdots \quad y_{(m+n)1}]$. Assuming $m = n$, the sum of the second column of **X** is 0, and so

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} (m+n) & 0 \\ 0 & .25(m+n) \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{i=1}^{m+n} y_i \\ .5\Sigma_{i=1}^{m} y_i - .5\Sigma_{i=m+1}^{m+n} y_i \end{bmatrix}$$

$$= \begin{bmatrix} 1/(m+n) & 0 \\ 0 & 4/(m+n) \end{bmatrix} \begin{bmatrix} m\bar{y}_1 + n\bar{y}_2 \\ .5m\bar{y}_1 - .5n\bar{y}_2 \end{bmatrix}$$

$$= \begin{bmatrix} 1/(2n) & 0 \\ 0 & 2/n \end{bmatrix} \begin{bmatrix} n(\bar{y}_1 + \bar{y}_2) \\ n(\bar{y}_1 - \bar{y}_2)/2 \end{bmatrix} = \begin{bmatrix} (\bar{y}_1 + \bar{y}_2)/2 \\ \bar{y}_1 - \bar{y}_2 \end{bmatrix} \Rightarrow \hat{\beta}_0 = \frac{\bar{y}_1 + \bar{y}_2}{2}, \hat{\beta}_1 = \bar{y}_1 - \bar{y}_2$$

b. $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1 & \cdots & 1 & 1 & \cdots & 1 \\ .5 & \cdots & .5 & -.5 & \cdots & -.5 \end{bmatrix}' \begin{bmatrix} (\bar{y}_1 + \bar{y}_2)/2 \\ \bar{y}_1 - \bar{y}_2 \end{bmatrix} = \begin{bmatrix} \bar{y}_1 & \cdots & \bar{y}_1 & \bar{y}_2 & \cdots & \bar{y}_2 \end{bmatrix}'$.

$\text{SSE} = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \Sigma(y_i - \hat{y}_i)^2 = \Sigma_{i=1}^{m}(y_i - \bar{y}_1)^2 + \Sigma_{i=m+1}^{n}(y_i - \bar{y}_2)^2$.

$s_e^2 = \dfrac{\text{SSE}}{(m+n)-(k+1)} = \dfrac{\text{SSE}}{m+n-2}$.

$\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \sigma^2/(2n) & 0 \\ 0 & 2\sigma^2/n \end{bmatrix} \Rightarrow V(\hat{\beta}_1) = 2\sigma^2/n \Rightarrow s_{\hat{\beta}_1} = \sqrt{2s_e^2/n} = s_e\sqrt{2/n}$. Note: In the general case where m and n may differ, $2/n$ becomes $1/m + 1/n$.

c.

$$\hat{\beta}_1 \pm t_{\alpha/2,m+n-2} \cdot s_{\hat{\beta}_1} = (\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2,m+n-2} \cdot s_e\sqrt{2/n}$$

$$= (\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2,m+n-2} \cdot \sqrt{\frac{\Sigma_{i=1}^{m}(y_i - \hat{y}_1)^2 + \Sigma_{i=m+1}^{n}(y_i - \bar{y}_2)^2}{m+n-2}}\sqrt{\frac{1}{m} + \frac{1}{n}}$$

d. With $\mathbf{X}' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ .5 & .5 & .5 & -.5 & -.5 & -.5 \end{bmatrix}$ and **y′** = [117 119 127 129 138 139], we get the following: **b′** = [128.166, –14.333]; $\hat{\mathbf{y}}'$ = [121 121 121 135.33 135.33 135.33]; SSE = … = 116.666, s_e = 5.4. Finally, the 95% CI for β_1 is –14.333 ± 2.776(5.4)$\sqrt{2/3}$ = (–26.58, –2.09).

107.

a. $\mathbf{H}^2 = \mathbf{HH} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{XI}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{H}$.

b. Write $\mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$. Using the covariance matrix proposition and properties of the matrices **I** and **H** and the random vector **Y**,

$$\begin{aligned}\text{Cov}(\mathbf{Y}-\hat{\mathbf{Y}}) &= (\mathbf{I}-\mathbf{H})'\text{Cov}(\mathbf{Y})(\mathbf{I}-\mathbf{H}) = (\mathbf{I}'-\mathbf{H}')\sigma^2\mathbf{I}(\mathbf{I}-\mathbf{H}) = \sigma^2(\mathbf{I}'-\mathbf{H}')(\mathbf{I}-\mathbf{H}) \\ &= \sigma^2(\mathbf{I}'\mathbf{I}-\mathbf{H}'\mathbf{I}-\mathbf{I}'\mathbf{H}+\mathbf{H}'\mathbf{H}) = \sigma^2(\mathbf{I}-\mathbf{H}'-\mathbf{H}+\mathbf{H}'\mathbf{H}) \\ &= \sigma^2(\mathbf{I}-\mathbf{H}-\mathbf{H}+\mathbf{H}^2) \quad \text{because } \mathbf{H} \text{ is symmetric} \\ &= \sigma^2(\mathbf{I}-\mathbf{H}-\mathbf{H}+\mathbf{H}) \quad \text{part (a)} \\ &= \sigma^2(\mathbf{I}-\mathbf{H})\end{aligned}$$

109.

a. Using the results in this section, the vector of coefficients is $\mathbf{b}' = [35.0 \;\; 3.18 \;\; -.006]$. The model utility test gives $f = 12.04$ (P-value < .001); the t test for β_1 (foot) is $t = 2.96$ (P-value = .021); the t test for β_2 (height) is $t = -0.02$ (P-value = .981). That is, the overall model is useful for predicting wingspan, and foot size is a useful predictor. However, in the presence of foot size, height is a basically useless addition to the model.

b. The diagonal entries of **H**, in order, are: .55, .31, .13, .11, .88, .17, .31, .15, .18, .20. Observation #5 has the highest leverage by far, by grace of the fact that the height (54”) is much lower than any other observed height. 54” is 4’6”, suggesting that student #5 mis-recorded his own height (perhaps it should be 64”). It’s also hard to believe that a 4’6” person would wear a size 9 shoe.

c. Students #1 and #7 (h = .55, .31) are very tall and have very big feet. Student #2 has rather small feet, both for his height and for the group overall.

d. Student #2 has a very large, negative residual. It seems that a 56” wing span for a 66” person is rather short.

e. If numbers were clearly mis-recorded, these observations should be corrected or deleted. In general, though, we do not delete a “correct” observation simply because it doesn’t follow the pattern suggested by the other observations.

Section 12.10

111. The logistic regression model specifies $p(x) = \dfrac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} = \dfrac{e^{-3.75+0.1x}}{1+e^{-3.75+0.1x}}$.

a. $p(10) = \dfrac{e^{-3.75+0.1(10)}}{1+e^{-3.75+0.1(10)}} = \dfrac{e^{-2.75}}{1+e^{-2.75}} \approx .060$, while $p(50) = \dfrac{e^{-3.75+0.1(50)}}{1+e^{-3.75+0.1(50)}} = \dfrac{e^{1.25}}{1+e^{1.25}} \approx .777$. According to the model, there’s a 6% chance someone will redeem a \$10 discount coupon, while there’s a 77.7% chance that someone will redeem a \$50 discount coupon.

b. $odds(10) = \dfrac{p(10)}{1-p(10)} = e^{-2.75} = .0639$, while $odds(50) = \dfrac{p(50)}{1-p(50)} = e^{1.25} = 3.49$. The odds of a \$10 coupon being redeemed are .0639:1 (quite unlikely), while the odds of a \$50 coupon being redeemed are 3.49:1 (3.49 times more likely than not).

c. $\beta_1 = 0.1$: For each \$1 increase in the value of the emailed coupon, the log-odds of the coupon being redeemed increase by 0.1. Equivalently, since $e^{\beta_1} = e^{0.1} = 1.105$, for each \$1 increase in the value of the emailed coupon, the odds of the coupon being redeemed increase by a *multiplicative* factor of 1.105 (i.e., the odds increase by 10.5%).

d. $p(x) = .5 \rightarrow odds(x) = 1 \rightarrow \beta_0 + \beta_1 x = \text{log-odds} = \ln(1) = 0 \rightarrow x = -\dfrac{\beta_0}{\beta_1} = -\dfrac{-3.75}{0.1} = \$37.50$.

113.

a. We test H_0: $\beta_1 = 0$ vs H_0: $\beta_1 \neq 0$. Using the large-sample z test, we will reject H_0 if $|z| \geq z_{.025} = 1.96$. The test statistic value is $z = \frac{\hat{\beta}_1 - 0}{s_{\hat{\beta}_1}} = \frac{-.1998 - 0}{.0986} = -2.026$. Since $|-2.026| \geq 1.96$, H_0 is rejected at the .05 significance level. We conclude that sleep indeed has an effect on the likelihood of driving drunk among American teenagers.

b. A 95% CI for β_1 is $\hat{\beta}_1 \pm z_{.025}s_{\hat{\beta}_1} = -.1998 \pm 1.96(.0986) = (-.3931, -.0065)$. Thus a 95% CI for e^{β_1} has endpoints $(e^{-.3931}, e^{-.0065}) = (.675, .993)$.

c. Multiplication by .675 = (1 – .325) is equivalent to a 32.5% decrease, while multiplying by .993 is the same as a 0.7% decrease because .993 = 1 – .007. Thus, with 95% confidence, a 1-hour increase in a teenager's typical number of sleep hours per night is associated with a 0.7% to 32.5% decrease in the odds of driving drunk.

115.

a. From software, $\hat{\beta}_0 = -.0573$ and $\hat{\beta}_1 = .00430$. So, the estimated logistic regression function is $\hat{p}(x) = \frac{e^{-.0573+.00430x}}{1+e^{-.0573+.00430x}}$.

b. $e^{\hat{\beta}_1} = 1.0043$, so a 1-month increase in age is associated with an estimated 0.43% increase in the odds of having kyphosis.

c. We test H_0: $\beta_1 = 0$ vs H_0: $\beta_1 \neq 0$. Using the large-sample z test, we will reject H_0 if $|z| \geq z_{.025} = 1.96$. From software, $s_{\hat{\beta}_1} = .00585$, so the test statistic value is $z = \frac{.00430 - 0}{.00585} = 0.74$. Since $|0.74| < 1.96$, we do not reject H_0 at the .05 level (or at any reasonable significance level). The data do not provide convincing evidence that age has an impact on the presence of kyphosis.

117.

a. The log-odds are .8247 + .0073(35) + .0041(65) + .9910(1) + .0224(0) = 2.3377, so the estimated probability is $\hat{p}(35, 65, 1, 0) = \frac{e^{2.3377}}{1+e^{2.3377}} = .912$.

b. Now the log-odds are .8247 + .0073(35) + .0041(65) + .9910(0) + .0224(0) = 1.3467 and the estimated probability is $\hat{p}(35, 65, 0, 0) = \frac{e^{1.3467}}{1+e^{1.3467}} = .794$.

c. Adjusting for a customer's income, sex, and child status (has some or not), a 1-year increase in age corresponds to an estimated $e^{.0073} = 1.007$ multiplicative increase (aka 0.7% increase) in the odds a customer wants GM chicken products labeled.

d. Adjusting for a customer's age, income, and child status, the odds that a female customer wants GM chicken products labeled is $e^{.9910} = 2.69$ times higher than for a male customer.

119.

a. Statistical software confirms these estimated coefficients. :)

b. The negative coefficient in front of x_1 signifies that the chance of seeing a whale decreases as the number of days after final salmon release increases. (That's logical — there's incrementally less food for the whales over time.) The positive coefficient on x_2 means that the chance of seeing a whale increases the longer you visit the site (duh).

c. log-odds = –5.68 – .096(7) + .210(30) = –0.052, so $\hat{p}(7,30) = \dfrac{e^{-0.052}}{1+e^{-0.052}} = .484$.

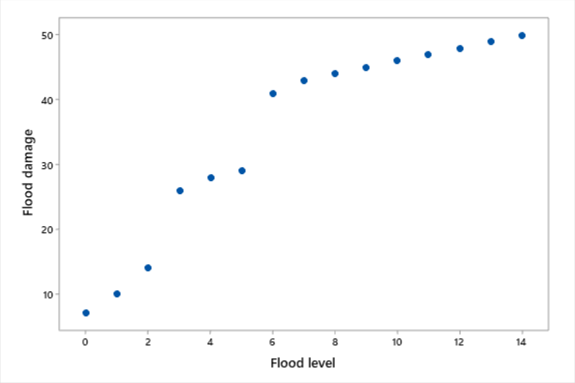
d. The test statistic values are $z_1 = \dfrac{-.096-0}{.253} = -0.38$ and $z_2 = \dfrac{.210-0}{.120} = 1.75$. At the .1 level, we reject H_0 if $|z| \geq z_{.05} = 1.645$. So, based on the z-values, we *do not* reject H_0: $\beta_1 = 0$ but we *do* reject H_0: $\beta_2 = 0$ in favor of H_a: $\beta_2 \neq 0$.

e. $e^{-.096} = .9084$ means that a 1-day increase in the time since the final salmon release corresponds to a (1 – .9084) = .0916 = 9.16% estimated decrease in the odds of seeing a whale. $e^{.210} = 1.2337$ means that a 1-minute increase in the duration of your visit corresponds to a 23.37% estimated increase in the odds of seeing a whale.

Supplementary Exercises

121.

a. As flood level increases, so does flood damage, not surprisingly. But there are two "jumps" in the pattern: the amount of flood damage increases suddenly from x = 2ft to 3ft and at x = 5ft to 6ft.



b. No: A single straight line would not accurately describe the relationship in the scatterplot. If anything, three lines are required for three ranges (perhaps 0-2.5ft, 2.5-5.5ft, and 5.5ft+).

123.

a. $R^2 = .5073$ or 50.73%.

b. $r = +\sqrt{R^2} = \sqrt{.5073} = .7122$ (positive because $\hat{\beta}_1$ is positive.)

c. We test $H_0 : \beta_1 = 0$ v. $H_a : \beta_1 \neq 0$. The test statistic $t = 3.93$ gives *P*-value = .0013, which is < .01, the given level of significance, therefore we reject H_0 and conclude that the model is useful.

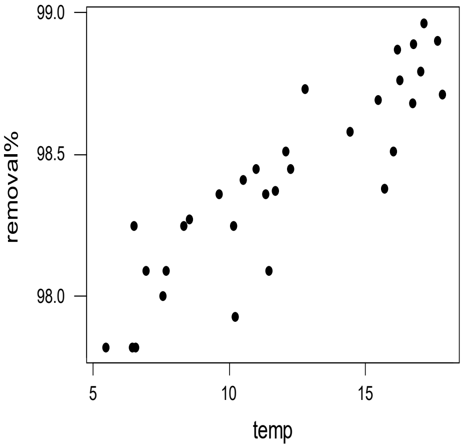
d. We use a 95% CI for $\mu_{Y\cdot 50}$. $\hat{y}_{(50)} = .787218 + .007570(50) = 1.165718$, $t_{.025,15} = 2.131$, s_e = "Root MSE" = .020308, so $s_{\hat{y}_{(50)}} = .20308\sqrt{\frac{1}{17} + \frac{17(50-42.33)^2}{17(41{,}575)-(719.60)^2}} = .051422$. The interval is, then, $1.165718 \pm 2.131(.051422) = 1.165718 \pm .109581 = (1.056137, 1.275299)$.

e. $\hat{y}_{(30)} = .787218 + .007570(30) = 1.0143.$ The residual is $y - \hat{y} = .80 - 1.0143 = -.2143$.

125. The value of the sample correlation coefficient using the squared *y* values would not necessarily be approximately 1. If the *y* values are greater than 1, then the squared *y* values would differ from each other by more than the *y* values differ from one another. Hence, the relationship between *x* and y^2 would be less like a straight line, and the resulting value of the correlation coefficient would decrease.

127.

a. A scatterplot suggests the linear model is appropriate.



b. Minitab Output:

```
The regression equation is
removal% = 97.5 + 0.0757 temp

Predictor        Coef       StDev          T        P
Constant      97.4986      0.0889    1096.17    0.000
temp         0.075691    0.007046      10.74    0.000

S = 0.1552      R-Sq = 79.4%     R-Sq(adj) = 78.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      2.7786      2.7786    115.40    0.000
Residual Error    30      0.7224      0.0241
Total             31      3.5010
```

Minitab will output all the residual information if the option is chosen, from which you can find the point prediction value $\hat{y}_{10.5} = 98.2933$, the observed value y = 98.41, so the residual = .0294.

c. Roughly s_e = .1552.

d. R^2 = 79.4.

e. A 95% CI for β_1, using $t_{.025,30} = 2.042$: $.075691 \pm 2.042(.007046) = (.061303, .090079)$.

f. The slope of the regression line is steeper. The value of s_e is almost doubled (to 0.291), and the value of R^2 drops to 61.6%.

129.

a. Using the techniques from a previous chapter, we can perform a t test for the difference of two means based on paired data. A paired t test for equality of means gives t = 3.54, with a P-value of .002, which suggests that the average bf% reading for the two methods is not the same.

b. A scatterplot (not shown) indicates that using linear regression to predict HW from BOD POD seems reasonable. The least squares linear regression equation, as well as the test statistic and P-value for a model utility test, can be found in the output below. We see that we do have significance, and the coefficient of determination shows that about 75% of the variation in HW can be explained by the variation in BOD.

```
The regression equation is
HW = 4.79 + 0.743 BOD

Predictor        Coef       StDev          T        P
Constant        4.788       1.215       3.94    0.001
BOD            0.7432      0.1003       7.41    0.000

S = 2.146       R-Sq = 75.3%     R-Sq(adj) = 73.9%
```

131. Use what we already know about MLE's of normal random samples. In the unrestricted case, $\hat{\sigma}^2 = \frac{1}{n}\Sigma(y_i - \hat{\mu})^2 = \frac{1}{n}\Sigma(y_i - b_0 - b_1x_i)^2 = \frac{1}{n}\Sigma(y_i - \hat{y}_i)^2$. Under H$_0$: $\beta_1 = 0$, $\mu = \beta_0$, so $\hat{\beta}_0 = \bar{y}$ and $\hat{\sigma}_0^2 = \frac{1}{n}\Sigma(y_i - \hat{\mu})^2 = \frac{1}{n}\Sigma(y_i - \hat{\beta}_0 - 0x_i)^2 = \frac{1}{n}\Sigma(y_i - \bar{y})^2$. Lastly, the exponential terms in the likelihood simplify to exp(–n/2) in both cases, for a likelihood ratio equal to $\frac{(2\pi\hat{\sigma}_0^2)^{-n/2}\exp(-n/2)}{(2\pi\hat{\sigma}^2)^{-n/2}\exp(-n/2)} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} = \left(\frac{SSR}{SST}\right)^{n/2}$. We reject H_0 when this ratio is small, which (by the ANOVA equation) is equivalent to SSR/SSE being large, or F = MSR/MSE being large.

133.

a. Take logs of both sides of the model to get ln(Y) = ln(α) + βx + ln(ε), or $Y' = \beta_0 + \beta_1 x + \varepsilon'$. If ε is lognormal, then (by definition) ε' is normal, and we have our usual regression model.

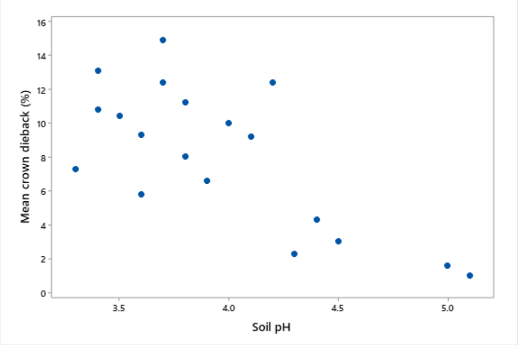
b. Software gives the estimated regression equation $y = –4.6 + 0.4057x$. However, residual plots show a strong curved patter among the residuals, and that the residuals are non-normal. The former indicates the simple linear model is <u>not</u> adequate.

c. A scatter plot (y vs. x) does show a pattern consistent with an exponential model. And a scatter plot of ln(y) versus x is quite linear. If we regress ln(y) on x, software gives the estimated regression equation ln(y) = 3.1564 + 0.004811x. Software also gives R^2 = 96.75%, a good sign of fit, and residual plots are at least somewhat better, although equal variance concerns persist. The estimates of the original parameters are $a = \exp(b_0) = 23.486$ and $b = b_1 = 0.004811$.

d. From software, a 95% PI for ln(Y) when x = 250 is (3.99671, 4.72167). Thus, a 95% PI for Y when x = 250 is ($e^{3.99671}$, $e^{4.72167}$) = (54.42, 112.36).

135.

a. The scatterplot suggests a linear relationship between pH and the mean response is plausible.



b. Software provides the estimated regression equation $y = 31.04 – 5.79x$. The estimated standard error of the slope is 1.36. So, the model utility test statistic is $t = \dfrac{-5.79}{1.36} = -4.25$. Comparing this to a t distribution with $n – 2 = 17$ df, the associated P-value is roughly 0. Hence, we reject H_0: $\beta_1 = 0$ and conclude that soil pH is a statistically significant predictor of mean crown dieback.

c. From software, a 95% PI for a new Y when x = 4.0 is (1.41657, 14.3251) while a 95% CI for $\mu_{Y|4.0}$ is (6.42391, 9.31772). The PI is considerably wider than the CI, consistent with what we’ve learned about simple linear regression (and about CI’s versus PI’s in general).

d. The PI and CI at x = 3.4 are (4.69265, 17.9996) and (9.17703, 13.5152), respectively. These are somewhat wider than the matching intervals in part **c**. That makes sense, because x = 4.0 is closer to the average x-value in the data set than is x = 3.4.

137.

a. With the aid of software, a first-order model yields $y = 84.82 + .1643x_1 - 79.67x_2$ and $R_a^2 = .654$.

b. The adjusted R^2 value jumps to .831 when the interaction term is added. For the full second-order model, $R_a^2 = .7207$. Looking at the R_a^2 values, it appears that the model with an interaction term but without quadratic terms is preferred.

c. The interaction model is $y = 6.22 + 5.779x_1 + 51.33x_2 - 9.357x_1x_2$. Substituting, the predicted compressive strength is $\hat{y} = 6.22 + 5.779(14) + 51.33(.60) - 9.357(14)(.60) = 39.32$ MPa.

d. First-order: $R_a^2 = 66.22\%$; with interaction, $R_a^2 = 68.27\%$; full second-order: $R_a^2 = 70.42\%$. These suggest that the full second-order model is "best" for predicting adsorbability.