CHAPTER 13

Section 13.1

1.

a. $\bar{x} = 15$ and $\sum (x_j - \bar{x})^2 = 250$, so s.d. of $Y_i - \hat{Y}_i$ is $10\sqrt{1 - \frac{1}{5} - \frac{(x_i - 15)^2}{250}} =$ 6.32, 8.37, 8.94, 8.37, and 6.32 for i = 1, 2, 3, 4, 5.

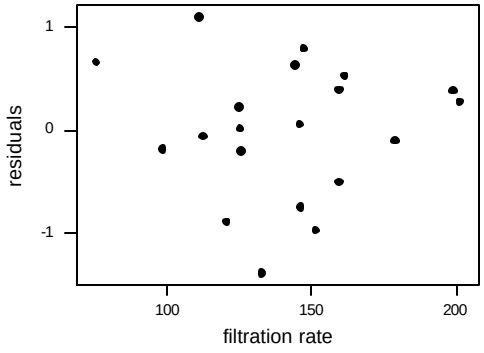
b. Now $\bar{x} = 20$ and $\sum (x_i - \bar{x})^2 = 1250$, giving standard deviations 7.87, 8.49, 8.83, 8.94, and 2.83 for i = 1, 2, 3, 4, 5.

c. The deviation from the estimated line is likely to be much smaller for the observation made in the experiment of **b** for x = 50 than for the experiment of **a** when x = 25. That is, the observation (50, Y) is more likely to fall close to the least squares line than is (25, Y).

2. The pattern gives no cause for questioning the appropriateness of the simple linear regression model, and no observation appears unusual.

3.

a. This plot indicates there are no outliers, the variance of ε is reasonably constant, and the ε are normally distributed. A straight-line regression function is a reasonable choice for a model.

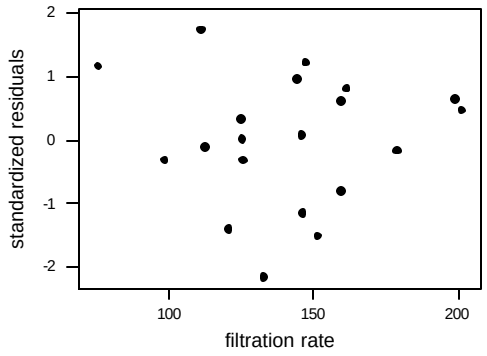


b. We need $S_{xx} = \sum (x_i - \bar{x})^2 = 415{,}914.85 - \frac{(2817.9)^2}{20} = 18{,}886.8295$. Then each e_i^* can be calculated as follows: $e_i^* = \frac{e_i}{.4427\sqrt{1 + \frac{1}{20} + \frac{(x_i - 140.895)^2}{18{,}886.8295}}}$. The table below shows the values:

standardized residuals	e / e_i^*	standardized residuals	e / 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 $e_i^* \sim e / s$, then $e / e_i^* \sim s$. All of the e / e_i^*'s range between .57 and .65, which are close to s.

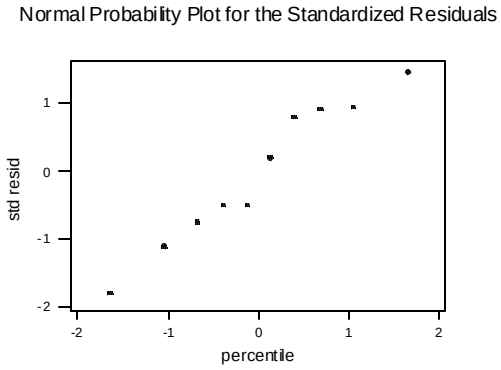
c. This plot looks very much the same as the one in part a.



4.

a. The (x, residual) pairs for the plot are (0, -.335), (7, -.508), (17. -.341), (114, .592), (133, .679), (142, .700), (190, .142), (218, 1.051), (237, -1.262), and (285, -.719). The plot shows substantial evidence of curvature.

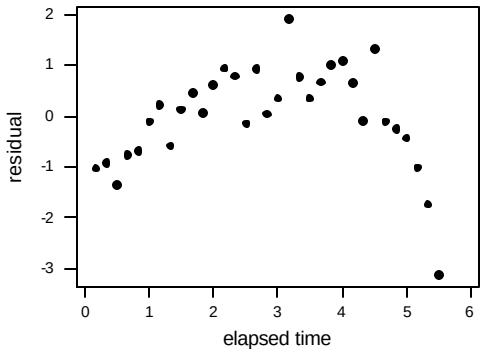
b. The standardized residuals (in order corresponding to increasing x) are -.50, -.75, -.50, .79, .90, .93, .19, 1.46, -1.80, and -1.12. A standardized residual plot shows the same pattern as the residual plot discussed in the previous exercise. The z percentiles for the normal probability plot are –1.645, -1.04, -.68, -.39, -.13, .13, .39, .68, 1.04, 1.645. The plot follows. The points follow a linear pattern, so the standardized residuals appear to have a normal distribution.



5.

a. 97.7% of the variation in ice thickness can be explained by the linear relationship between it and elapsed time. Based on this value, it appears that a linear model is reasonable.

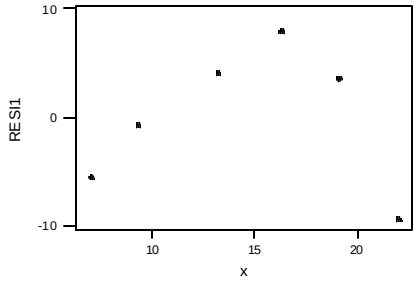
b. The residual plot shows a curve in the data, so perhaps a non-linear relationship exists. One observation (5.5, -3.14) is extreme.



6.

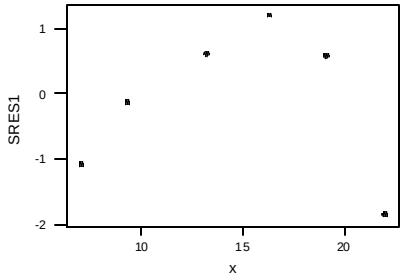
a. $H_o : b_1 = 0$ vs. $H_a : b_1 \neq 0$. The test statistic is $t = \dfrac{\hat{b}_1}{s_{\hat{b}_1}}$, and we will reject H_o if $t \geq t_{.025,4} = 2.776$ or if $t \leq -2.776$. $s_{\hat{b}_1} = \dfrac{s}{\sqrt{S_{xx}}} = \dfrac{7.265}{12.869} = .565$, and $t = \dfrac{6.19268}{.565} = 10.97$. Since $10.97 \geq 2.776$, we reject H_o and conclude that the model is useful.

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 may be inappropriate.

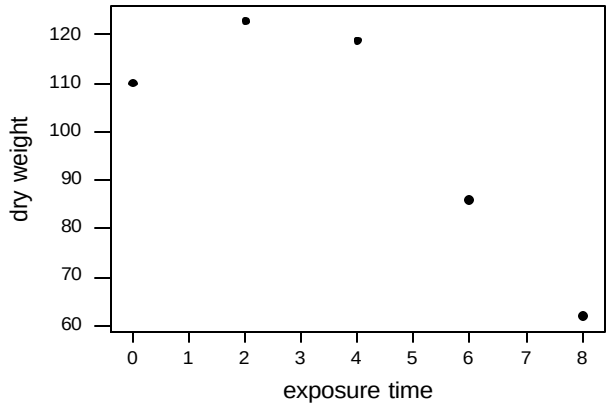
c. The standardized residuals are calculated as $e_1^* = \dfrac{-5.49}{7.265\sqrt{1 + \dfrac{1}{6} + \dfrac{(7.0 - 14.48)^2}{165.5983}}} = -1.074$, and similarly the others are -.123, .624, 1.208, .587, and –1.841. The plot of e* vs x follows :



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.

7.

a.

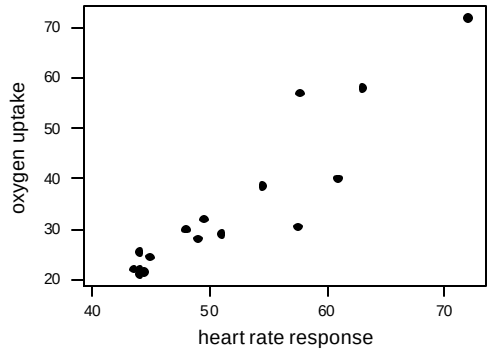


There is an obvious curved pattern in the scatter plot, which suggests that a simple linear model will not provide a good fit.

b. The $\hat{y}'s$, e's, and e*'s are given below:

x	y	$\hat{y}$	e	e*
0	110	126.6	-16.6	-1.55
2	123	113.3	9.7	.68
4	119	100.0	19.0	1.25
6	86	86.7	-.7	-.05
8	62	73.4	-11.4	-1.06

8. First, we will look at a scatter plot of the data, which is quite linear, so it seems reasonable to use linear regression.

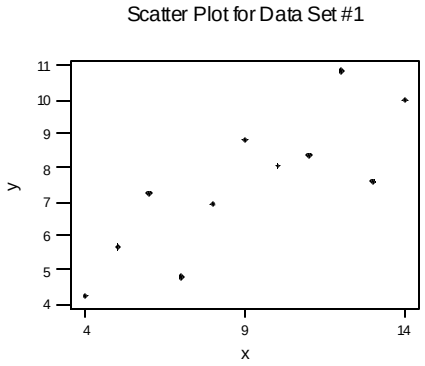


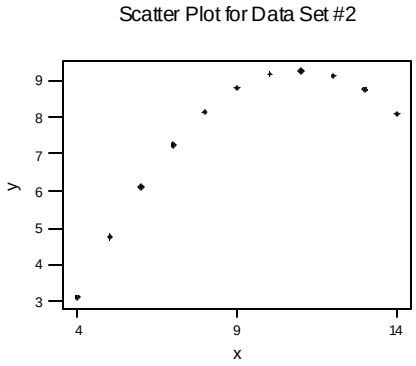
The linear regression output (Minitab) follows:

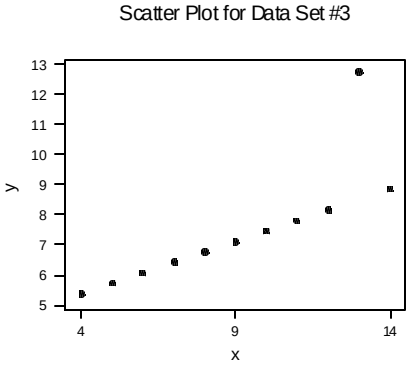
```
The regression equation is
y = - 51.4 + 1.66 x

Predictor        Coef       StDev          T        P
Constant      -51.355       9.795      -5.24    0.000
x              1.6580      0.1869       8.87    0.000

S = 6.119       R-Sq = 84.9%     R-Sq(adj) = 83.8%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1      2946.5      2946.5     78.69    0.000
Residual Error    14       524.2        37.4
Total             15      3470.7
```

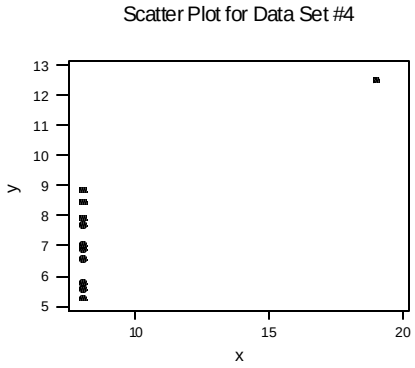
A quick look at the t and p values shows that the model is useful, and r^2 shows a strong relationship between the two variables.
The observation (72, 72) has large influence, since its x value is a distance from the others.
We could run the regression again, without this value, and get the line:
`oxygen uptake = - 44.8 + 1.52 heart rate response.`

9. 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. The signs of the residuals here (corresponding to increasing x) are + + + + - - - - - + -, and a residual plot would reflect this pattern and suggest a careful look at the chosen model. 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 (and its x value does lie far from the remaining ones).

10.

a. $e_i = y_i - \left(\hat{b}_0 - \hat{b}_1 x_i\right) = y_i - \bar{y} - \hat{b}_1\left(x_i - \bar{x}\right)$, so
$\Sigma e_i = \Sigma\left(y_i - \bar{y}\right) - \hat{b}_1\Sigma\left(x_i - \bar{x}\right) = 0 + \hat{b}_1 \cdot 0 = 0.$

b. Since $\Sigma e_i = 0$ always, the residuals cannot be independent. There is clearly a linear relationship between the residuals. If one e_I is large positive, then al least one other e_I would have to be negative to preserve $\Sigma e_i = 0$. This suggests a negative correlation between residuals (for fixed values of any n – 2, the other two obey a negative linear relationship).

c. $\Sigma x_i e_i = \Sigma x_i y_i - \Sigma x_i \bar{y} - \hat{b}_1 \Sigma x_i \left(x_i - \bar{x}\right) = \left[\Sigma x_i y_i - \dfrac{\left(\Sigma x_i\right)\left(\Sigma y_i\right)}{n}\right] - \hat{b}_1\left[\Sigma x_i^2 - \dfrac{\left(\Sigma x_i\right)^2}{n}\right]$

, but the first term in brackets is the numerator of $\hat{b}_1$, while the second term is the denominator of $\hat{b}_1$, so the difference becomes (numerator of $\hat{b}_1$) – (numerator of $\hat{b}_1$) = 0.

d. The five e_i^*'s from Exercise 7 above are –1.55, .68, 1.25, -.05, and –1.06, which sum to -.73. This sum differs too much from 0 to be explained by rounding. In general it is not true that $\Sigma e_i^* = 0$.

11.

a. $Y_i - \hat{Y}_i = Y_i - \overline{Y} - \hat{b}_1(x_i - \overline{x}) = Y_i - \frac{1}{n}\sum_j Y_j - \frac{(x_i - \overline{x})\Sigma_j(x_j - \overline{x})Y_j}{\Sigma_j(x_j - \overline{x})^2} = \sum_j c_j Y_j$,

where $c_j = 1 - \frac{1}{n} - \frac{(x_i - \overline{x})^2}{n\Sigma(x_j - \overline{x})^2}$ for j = i and $c_j = 1 - \frac{1}{n} - \frac{(x_i - \overline{x})(x_j - \overline{x})}{\Sigma(x_j - \overline{x})^2}$ for $j \neq i$. Thus $Var(Y_i - \hat{Y}_i) = \Sigma Var(c_j Y_j)$ (since the Y_j's are independent) = $s^2 \Sigma c_j^2$ which, after some algebra, gives equation (13.2).

b. $s^2 = Var(Y_i) = Var(\hat{Y}_i + (Y_i - \hat{Y}_i)) = Var(\hat{Y}_i) + Var(Y_i - \hat{Y}_i)$, so

$Var(Y_i - \hat{Y}_i) = s^2 - Var(\hat{Y}_i) = s^2 - s^2\left[\frac{1}{n} + \frac{(x_i - \overline{x})^2}{n\Sigma(x_j - \overline{x})^2}\right]$, which is exactly (13.2).

c. As x_i moves further from $\overline{x}$, $(x_i - \overline{x})^2$ grows larger, so $Var(\hat{Y}_i)$ increases (since $(x_i - \overline{x})^2$ has a positive sign in $Var(\hat{Y}_i)$), but $Var(Y_i - \hat{Y}_i)$ decreases (since $(x_i - \overline{x})^2$ has a negative sign).

12.

a. $\Sigma e_i = 34$, which is not = 0, so these cannot be the residuals.

b. Each $x_i e_i$ is positive (since x_i and e_i have the same sign) so $\Sigma x_i e_i > 0$, which contradicts the result of exercise 10**c**, so these cannot be the residuals for the given x values.

13. The distribution of any particular standardized residual is also a t distribution with n – 2 d.f., since e_i^* is obtained by taking standard normal variable $\frac{\left(Y_i - \hat{Y}_i\right)}{\left(s_{Y_i - \hat{Y}}\right)}$ and substituting the estimate of σ in the denominator (exactly as in the predicted value case). With E_i^* denoting the i[th] standardized residual as a random variable, when n = 25 E_i^* has a t distribution with 23 d.f. and $t_{.01,23} = 2.50$, so P(E_i^* outside (-2.50, 2.50)) = $P\left(E_i^* \geq 2.50\right) + P\left(E_i^* \leq -2.50\right) = .01 + .01 = .02$.

14. space

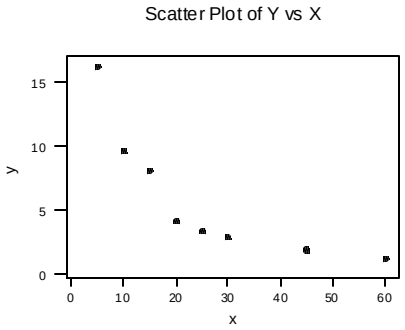
a. $n_1 = n_2 = 3$ (3 observations at 110 and 3 at 230), $n_3 = n_4 = 4$, $\bar{y}_{1.} = 202.0$, $\bar{y}_{2.} = 149.0$, $\bar{y}_{3.} = 110.5$, $\bar{y}_{4.} = 107.0$, $\Sigma\Sigma y_{ij}^2 = 288{,}013$, so $SSPE = 288{,}013 - \left[3(202.0)^2 + 3(149.0)^2 + 4(110.5)^2 + 4(107.0)^2\right] = 4361$. With $\Sigma x_i = 4480$, $\Sigma y_i = 1923$, $\Sigma x_i^2 = 1{,}733{,}500$, $\Sigma y_i^2 = 288{,}013$ (as above), and $\Sigma x_i\, y_i = 544{,}730$, SSE = 7241 so SSLF = 7241-4361=2880. With c – 2 = 2 and n – c = 10, $F_{.05,2,10} = 4.10$. $MSLF = \frac{2880}{2} = 1440$ and $SSPE = \frac{4361}{10} = 436.1$, so the computed value of F is $\frac{1440}{436.1} = 3.30$. Since 3.30 is not ≥ 4.10, we do not reject H_o. This formal test procedure does not suggest that a linear model is inappropriate.

b. The scatter plot clearly reveals a curved pattern which suggests that a nonlinear model would be more reasonable and provide a better fit than a linear model.

Section 13.2

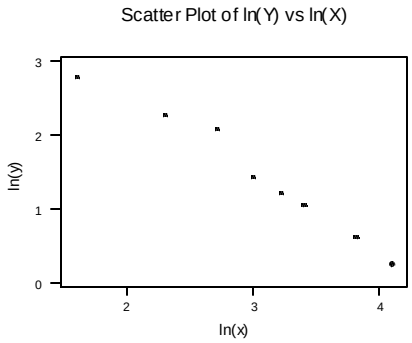
15.

a.



The points have a definite curved pattern. A linear model would not be appropriate.

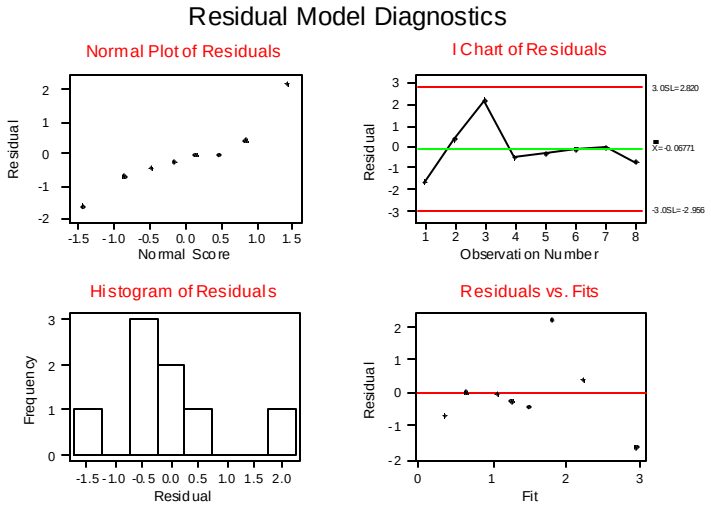
b. In this plot we have a strong linear pattern.



c. The linear pattern in **b** above would indicate that a transformed regression using the natural log of both x and y would be appropriate. The probabilistic model is then $y = ax^b \cdot e$. (The power function with an error term!)

d. A regression of ln(y) on ln(x) yields the equation $\ln(y) = 4.6384 - 1.04920\ln(x)$. Using Minitab we can get a P.I. for y when x = 20 by first transforming the x value: ln(20) = 2.996. The computer generated 95% P.I. for ln(y) when ln(x) = 2.996 is (1.1188,1.8712). We must now take the antilog to return to the original units of Y: $\left(e^{1.1188}, e^{1.8712}\right) = (3.06, 6.50)$.

e. A computer generated residual analysis:



Looking at the residual vs. fits (bottom right), one standardized residual, corresponding to the third observation, is a bit large. There are only two positive standardized residuals, but two others are essentially 0. The patterns in the residual plot and the normal probability plot (upper left) are marginally acceptable.

16.

a. $\Sigma x_i = 9.72$, $\Sigma y_i' = 313.10$, $\Sigma x_i^2 = 8.0976$, $\Sigma y_i'^2 = 288{,}013$, $\Sigma x_i\, y_i' = 255.11$, (all from computer printout, where $y_i' = \ln(L_{178})$), from which $\hat{b}_1 = 6.6667$ and $\hat{b}_0 = 20.6917$ (again from computer output). Thus $\hat{b} = \hat{b}_1 = 6.6667$ and $\hat{a} = e^{\hat{b}_0} = 968{,}927{,}163$.

b. We first predict y' using the linear model and then exponentiate:
$y' = 20.6917 + 6.6667(.75) = 25.6917$, so
$\hat{y} = \hat{L}_{178} = e^{25.6917} = 1.438051363 \times 10^{11}$.

c. We first compute a prediction interval for the transformed data and then exponentiate. With $t_{.025,10} = 2.228$, s = .5946, and $\sqrt{1 + \frac{1}{12} + \frac{(.95 - \bar{x})^2}{\Sigma x^2 - (\Sigma x)^2/12}} = 1.082$, the prediction interval for y' is
$27.0251 \pm (2.228)(.5496)(1.082) = 27.0251 \pm 1.4334 = (25.5917, 28.4585)$.
The P.I. for y is then $(e^{25.5917}, e^{28.4585})$.

17.

a. $\Sigma x_i' = 15.501$, $\Sigma y_i' = 13.352$, $\Sigma x_i'^2 = 20.228$, $\Sigma y_i'^2 = 16.572$, $\Sigma x_i' y_i' = 18.109$, from which $\hat{b}_1 = 1.254$ and $\hat{b}_0 = -.468$ so $\hat{b} = \hat{b}_1 = 1.254$ and $\hat{a} = e^{-.468} = .626$.

b. The plots give strong support to this choice of model; in addition, $r^2 = .960$ for the transformed data.

c. SSE = .11536 (computer printout), s = .1024, and the estimated sd of $\hat{b}_1$ is .0775, so $t = \dfrac{1.25 - 1.33}{.0775} = -1.07$. Since –1.07 is not $\le -t_{.05,11} = -1.796$, H_o cannot be rejected in favor of H_a.

d. The claim that $m_{Y \cdot 5} = 2m_{Y \cdot 2.5}$ is equivalent to $a \cdot 5^b = 2a(2.5)^b$, or that $b = 1$. Thus we wish test $H_o : b_1 = 1$ vs. $H_a : b_1 \neq 1$. With $t = \dfrac{1 - 1.33}{.0775} = -4.30$ and RR $-t_{.005,11} \le -3.106$, H_o is rejected at level .01 since $-4.30 \le -3.106$.

18. A scatter plot may point us in the direction of a power function, so we try $y = ax^b$. We transform $x' = \ln(x)$, so $y = a + b \ln(x)$. This transformation yields a linear regression equation $y = .0197 - .00128x'$ or $y = .0197 - .00128 \ln(x)$. Minitab output follows:

```
The regression equation is
y = 0.0197 - 0.00128 x

Predictor        Coef       StDev          T        P
Constant     0.019709    0.002633       7.49    0.000
x          -0.0012805   0.0003126      -4.10    0.001

S = 0.002668    R-Sq = 49.7%     R-Sq(adj) = 46.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1  0.00011943  0.00011943     16.78    0.001
Residual Error    17  0.00012103  0.00000712
Total             18  0.00024046
```

The model is useful, based on a t test, with a p value of .001. But $r^2 = 49.7$, so only 49.7% of the variation in y can be explained by its relationship with ln(x).

To estimate y_{5000}, we need $x' = \ln(5000) = 8.51718$. A point estimate for y when x =5000 is y = .009906. A 95 % prediction interval for y_{5000} is $(.002257, 017555)$.

19.

a. No, there is definite curvature in the plot.

b. $Y' = b_0 + b_1(x') + e$ where $x' = \dfrac{1}{temp}$ and $y' = \ln(lifetime)$. Plotting y' vs. x' gives a plot which has a pronounced linear appearance (and in fact $r^2 = .954$ for the straight line fit).

c. $\Sigma x_i' = .082273$, $\Sigma y_i' = 123.64$, $\Sigma x_i'^2 = .00037813$, $\Sigma y_i'^2 = 879.88$, $\Sigma x_i' y_i' = .57295$, from which $\hat{b}_1 = 3735.4485$ and $\hat{b}_0 = -10.2045$ (values read from computer output). With x = 220, $x' = .00445$ so $\hat{y}' = -10.2045 + 3735.4485(.00445) = 6.7748$ and thus $\hat{y} = e^{\hat{y}'} = 875.50$.

d. For the transformed data, SSE = 1.39857, and $n_1 = n_2 = n_3 = 6$, $\bar{y}_{1.}' = 8.44695$, $\bar{y}_{2.}' = 6.83157$, $\bar{y}_{3.}' = 5.32891$, from which SSPE = 1.36594, SSLF = .02993, $f = \dfrac{.02993/1}{1.36594/15} = .33$. Comparing this to $F_{.01,1,15} = 8.68$, it is clear that H_o cannot be rejected.

20. After examining a scatter plot and a residual plot for each of the five suggested models as well as for y vs. x, I felt that the power model $Y = ax^b \cdot e$ $(y' = \ln(y)$ vs. $x' = \ln(x))$ provided the bet fit. The transformation seemed to remove most of the curvature from the scatter plot, the residual plot appeared quite random, $|e_i'^*| < 1.65$ for every i, there was no indication of any influential observations, and $r^2 = .785$ for the transformed data.

21.

a. The suggested model is $Y = b_0 + b_1(x') + e$ where $x' = \dfrac{10^4}{x}$. The summary quantities are $\Sigma x_i' = 159.01$, $\Sigma y_i = 121.50$, $\Sigma x_i'^2 = 4058.8$, $\Sigma y_i^2 = 1865.2$, $\Sigma x_i' y_i = 2281.6$, from which $\hat{b}_1 = -.1485$ and $\hat{b}_0 = 18.1391$, and the estimated regression function is $y = 18.1391 - \dfrac{1485}{x}$.

b. $x = 500 \Rightarrow \hat{y} = 18.1391 - \dfrac{1485}{500} = 15.17$.

22.

a. $\frac{1}{y} = \alpha + \beta x$, so with $y' = \frac{1}{y}$, $y' = \alpha + \beta x$. The corresponding probabilistic model is $\frac{1}{y} = \alpha + \beta x + \varepsilon$.

b. $\frac{1}{y} - 1 = e^{\alpha + \beta x}$, so $\ln\left(\frac{1}{y} - 1\right) = \alpha + \beta x$. Thus with $y' = \ln\left(\frac{1}{y} - 1\right)$, $y' = \alpha + \beta x$. The corresponding probabilistic model is $Y' = \alpha + \beta x + \varepsilon'$, or equivalently $Y = \frac{1}{1 + e^{\alpha + \beta x} \cdot \varepsilon}$ where $\varepsilon = e^{\varepsilon'}$.

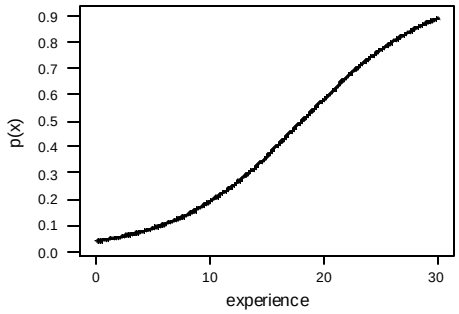
c. $\ln(y) = e^{\alpha + \beta x} = \ln(\ln(y)) = \alpha + \beta x$. Thus with $y' = \ln(\ln(y))$, $y' = \alpha + \beta x$. The probabilistic model is $Y' = \alpha + \beta x + \varepsilon'$, or equivalently, $Y = e^{e^{\alpha + \beta x}} \cdot \varepsilon$ where $\varepsilon = e^{\varepsilon'}$.

d. This function cannot be linearized.

23. $Var(Y) = Var(\alpha e^{\beta x} \cdot \varepsilon) = \left[\alpha e^{\beta x}\right]^2 \cdot Var(\varepsilon) = \alpha^2 e^{2\beta x} \cdot \tau^2$ where we have set $Var(\varepsilon) = \tau^2$. If $\beta > 0$, this is an increasing function of x so we expect more spread in y for large x than for small x, while the situation is reversed if $\beta < 0$. It is important to realize that a scatter plot of data generated from this model will not spread out uniformly about the exponential regression function throughout the range of x values; the spread will only be uniform on the transformed scale. Similar results hold for the multiplicative power model.

24. $H_0 : \beta_1 = 0$ vs $H_a : \beta_1 \neq 0$. The value of the test statistic is z = .73, with a corresponding p-value of .463. Since the p-value is greater than any sensible choice of alpha we do not reject H_o. There is insufficient evidence to claim that age has a significant impact on the presence of kyphosis.

25. The point estimate of b_1 is $\hat{b}_1 = .17772$, so the estimate of the odds ratio is $e^{\hat{b}_1} = e^{.17772} \approx 1.194$. That is , when the amount of experience increases by one year (i.e. a one unit increase in x), we estimate that the odds ratio increase by about 1.194. The z value of 2.70 and its corresponding p-value of .007 imply that the null hypothesis $H_0 : b_1 = 0$ can be rejected at any of the usual significance levels (e.g., .10, .05, .025, .01). Therefore, there is clear evidence that b_1 is not zero, which means that experience does appear to affect the likelihood of successfully performing the task. This is consistent with the confidence interval (1.05, 1.36) for the odds ratio given in the printout, since this interval does not contain the value 1. A graph of $\hat{p}$ appears below.



Section 13.3

26.

a. There is a slight curve to this scatter plot. It could be consistent with a quadratic regression.

b. We desire R^2, which we find in the output: $R^2 = 93.8\%$

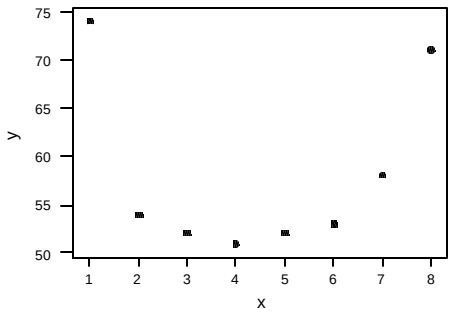
c. $H_0 : b_1 = b_2 = 0$ vs H_a : at least one $b_i \neq 0$. The test statistic is $f = \dfrac{MSR}{MSE} = 22.51$, and the corresponding p-value is .016. Since the p-value < .05, we reject H_o and conclude that the model is useful.

d. We want a 99% confidence interval, but the output gives us a 95% confidence interval of (452.71, 529.48), which can be rewritten as 491.10 ± 38.38; $t_{.025,3} = 3.182$, so $s_{\hat{y}\cdot 14} = \dfrac{38.38}{3.182} = 12.06$; Now, $t_{.005,3} = 5.841$, so the 99% C.I. is $491.10 \pm 5.841(12.06) = 491.10 \pm 70.45 = (420.65, 561.55)$.

e. $H_0 : b_2 = 0$ vs $H_a : b_2 \neq 0$. The test statistic is t = -3.81, with a corresponding p-value of .032, which is < .05, so we reject H_o. the quadratic term appears to be useful in this model.

27.

a. A scatter plot of the data indicated a quadratic regression model might be appropriate.



b. $\hat{y} = 84.482 - 15.875(6) + 1.7679(6)^2 = 52.88$; residual = $y_6 - \hat{y}_6 = 53 - 52.88 = .12;$

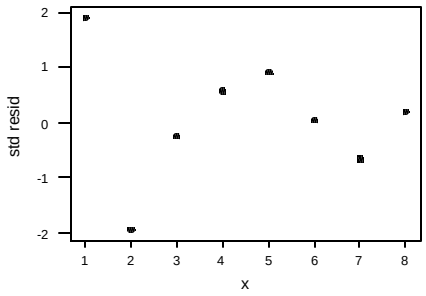
c. $SST = \Sigma y_i^2 - \dfrac{(\Sigma y_i)^2}{n} = 586.88$, so $R^2 = 1 - \dfrac{61.77}{586.88} = .895$.

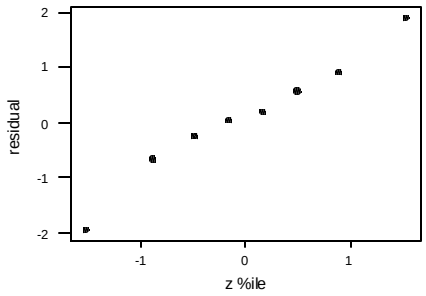
d. The first two residuals are the largest, but they are both within the interval (-2, 2). Otherwise, the standardized residual plot does not exhibit any troublesome features. For the Normal Probability Plot:

Residual	Zth percentile
-1.95	-1.53
-.66	-.89
-.25	-.49
.04	-.16
.20	.16
.58	.49
.90	.89
1.91	1.53

(continued)

The normal probability plot does not exhibit any troublesome features.





e. $\hat{m}_{Y\cdot 6} = 52.88$ (from **b**) and $t_{.025,n-3} = t_{.025,5} = 2.571$, so the C.I. is $52.88 \pm (2.571)(1.69) = 52.88 \pm 4.34 = (48.54, 57.22)$.

f. SSE = 61.77 so $s^2 = \dfrac{61.77}{5} = 12.35$ and $\sqrt{12.35 + (1.69)^2} = 3.90$. The P.I. is $52.88 \pm (2.571)(3.90) = 52.88 \pm 10.03 = (42.85, 62.91)$.

28.

a. $\hat{m}_{Y\cdot 75} = \hat{b}_0 + \hat{b}_1(75) + \hat{b}_2(75)^2 = -113.0937 + 3.36684(75) - .01780(75)^2 = 39.41$

b. $\hat{y} = \hat{b}_0 + \hat{b}_1(60) + \hat{b}_2(60)^2 = 24.93.$

c. $SSE = \Sigma y_i^2 - \hat{b}_0 \Sigma y_i - \hat{b}_1 \Sigma x_i y_i - \hat{b}_2 \Sigma x_i^2 y_i = 8386.43 - (-113.0937)(210.70)$
$- (3.3684)(17{,}002) - (-.0178)(1{,}419{,}780) = 217.82,$
$s^2 = \dfrac{SSE}{n-3} = \dfrac{217.82}{3} = 72.61$, s = 8.52

d. $R^2 = 1 - \dfrac{217.82}{987.35} = .779$

e. H_o will be rejected in favor of H_a if either $t \geq t_{.005,3} = 5.841$ or if $t \leq -5.841$. The computed value of t is $t = \dfrac{-.01780}{.00226} = -7.88$, and since $-7.88 \leq -5.841$, we reject H_o.

29.

a. From computer output:

$\hat{y}$:	111.89	120.66	114.71	94.06	58.69
$y-\hat{y}$:	-1.89	2.34	4.29	-8.06	3.31

Thus $SSE=(-1.89)^2+...+(3.31)^2=103.37$, $s^2=\dfrac{103.37}{2}=51.69$, $s=7.19$.

b. $SST=\Sigma y_i^2-\dfrac{(\Sigma y_i)^2}{n}=2630$, so $R^2=1-\dfrac{103.37}{2630}=.961$.

c. $H_0: b_2=0$ will be rejected in favor of $H_a: b_2\neq 0$ if either $t\geq t_{.025,2}=4.303$ or if $t\leq -4.303$. With $t=\dfrac{-1.84}{.480}=-3.83$, H_o cannot be rejected; the data does not argue strongly for the inclusion of the quadratic term.

d. To obtain joint confidence of at least 95%, we compute a 98% C.I. for each coefficient using $t_{.01,2}=6.965$. For b_1 the C.I. is $8.06\pm(6.965)(4.01)=(-19.87,35.99)$ (an extremely wide interval), and for b_2 the C.I. is $-1.84\pm(6.965)(.480)$ $=(-5.18,1.50)$.

e. $t_{.05,2}=2.920$ and $\hat{b}_0+4\hat{b}_1+16\hat{b}_2=114.71$, so the C.I. is $114.71\pm(2.920)(5.01)$ $=114.71\pm 14.63=(100.08,129.34)$.

f. If we knew $\hat{b}_0,\hat{b}_1,\hat{b}_2$, the value of x which maximizes $\hat{b}_0+\hat{b}_1x+\hat{b}_2x^2$ would be obtained by setting the derivative of this to 0 and solving:

$b_1+2b_2x=0\Rightarrow x=-\dfrac{b_1}{2b_2}$. The estimate of this is $x=-\dfrac{\hat{b}_1}{2\hat{b}_2}=2.19$.

30.

a. $R^2 = 0.853$. This means 85.3% of the variation in wheat yield is accounted for by the model.

b. $-135.44 \pm (2.201)(41.97) = (-227.82, -43.06)$

c. $H_0 : m_{y \cdot 2.5} = 1500;\ H_a : m_{y \cdot 2.5} < 1500;\ RR : t \le -t_{.01,11} = -2.718$

When x = 2.5, $\hat{y} = 1402.15$

$$t = \frac{1,402.15 - 1500}{53.5} = -1.83$$

Fail to reject H_o. The data does not indicate $m_{y \cdot 2.5}$ is less than 1500.

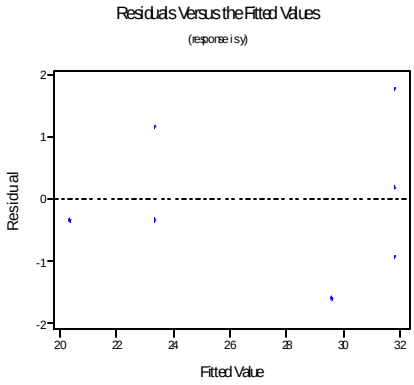
d. $1402.15 \pm (2.201)\sqrt{(136.5)^2 + (53.5)^2} = (1081.3, 1725.0)$

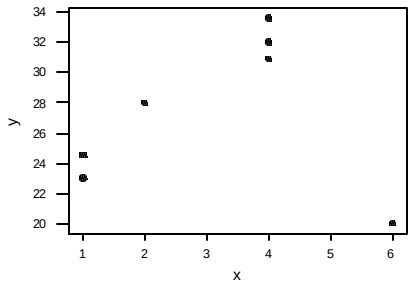
31.

a. Using Minitab, the regression equation is `y = 13.6 + 11.4x - 1.72x²`.

b. Again, using Minitab, the predicted and residual values are:

$\hat{y}$:	23.327	23.327	29.587	31.814	31.814	31.814	20.317
$y - \hat{y}$:	-.327	1.173	1.587	.914	.186	1.786	-.317





The residual plot is consistent with a quadratic model (no pattern which would suggest modification), but it is clear from the scatter plot that the point (6, 20) has had a great influence on the fit – it is the point which forced the fitted quadratic to have a maximum between 3 and 4 rather than, for example, continuing to curve slowly upward to a maximum someplace to the right of x = 6.

c. From Minitab output, $s^2 = MSE = 2.040$, and $R^2 = 94.7\%$. The quadratic model thus explains 94.7% of the variation in the observed y's , which suggests that the model fits the data quite well.

d. $s^2 = Var(\hat{Y}_i) + Var\left(Y_i - \hat{Y}_i\right)$ suggests that we can estimate $Var\left(Y_i - \hat{Y}_i\right)$ by $s^2 - s_{\hat{y}}^2$ and then take the square root to obtain the estimated standard deviation of each residual. This gives $\sqrt{2.040 - (.955)^2} = 1.059$, (and similarly for all points) 10.59, 1.236, 1.196, 1.196, 1.196, and .233 as the estimated std dev's of the residuals. The standardized residuals are then computed as $\frac{-.327}{1.059} = -.31$, (and similarly) 1.10, -1.28, -.76, .16, 1.49, and –1.28, none of which are unusually large. (Note: Minitab regression output can produce these values.) The resulting residual plot is virtually identical to the plot of **b**. $\frac{y - \hat{y}}{s} = \frac{-.327}{1.426} = -.229 \neq -.31$, so standardizing using just s would not yield the correct standardized residuals.

e. $Var(Y_f) + Var(\hat{Y}_f)$ is estimated by $2.040 + (.777)^2 = 2.638$, so $s_{y_f + \hat{y}_f} = \sqrt{2.638} = 1.624$. With $\hat{y} = 31.81$ and $t_{.05,4} = 2.132$, the desired P.I. is $31.81 \pm (2.132)(1.624) = (28.35, 35.27)$.

32.

a. $.3463 - 1.2933(x - \bar{x}) + 2.3964(x - \bar{x})^2 - 2.3968(x - \bar{x})^3$.

b. From **a**, the coefficient of x^3 is -2.3968, so $\hat{b}_3 = -2.3968$. There sill be a contribution to x^2 both from $2.3964(x - 4.3456)^2$ and from $-2.3968(x - 4.3456)^3$. Expanding these and adding yields 33.6430 as the coefficient of x^2, so $\hat{b}_2 = 33.6430$.

c. $x = 4.5 \Rightarrow x' = x - \bar{x} = .1544$; substituting into **a** yields $\hat{y} = .1949$.

d. $t = \frac{-2.3968}{2.4590} = -.97$, which is not significant ($H_0 : b_3 = 0$ cannot be rejected), so the inclusion of the cubic term is not justified.

33.

a. $\bar{x} = 20$ and $s_x = 10.8012$ so $x' = \dfrac{x - 20}{10.8012}$. For x = 20, $x' = 0$, and $\hat{y} = \hat{b}_0^* = .9671$. For x = 25, $x' = .4629$, so $\hat{y} = .9671 - .0502(.4629) - .0176(.4629)^2 + .0062(.4629)^3 = .9407$.

b. $\hat{y} = .9671 - .0502\left(\dfrac{x-20}{10.8012}\right) - .0176\left(\dfrac{x-20}{10.8012}\right)^2 + .0062\left(\dfrac{x-20}{10.8012}\right)^3$
$.00000492x^3 - .000446058x^2 + .007290688x + .96034944$.

c. $t = \dfrac{.0062}{.0031} = 2.00$. We reject H_o if either $t \geq t_{.025,n-4} = t_{.025,3} = 3.182$ or if $t \leq -3.182$. Since 2.00 is neither ≥ 3.182 nor ≤ -3.182, we cannot reject H_o; the cubic term should be deleted.

d. $SSE = \Sigma(y_i - \hat{y}_i)$ and the $\hat{y}_i's$ are the same from the standardized as from the unstandardized model, so SSE, SST, and R^2 will be identical for the two models.

e. $\Sigma y_i^2 = 6.355538$, $\Sigma y_i = 6.664$, so SST = .011410. For the quadratic model R^2 = .987 and for the cubic mo del, R^2 = .994; The two R^2 values are very close, suggesting intuitively that the cubic term is relatively unimportant.

34.

a. $\bar{x} = 49.9231$ and $s_x = 41.3652$ so for x = 50, $x' = \dfrac{x - 49.9231}{41.3652} = .001859$ and $\hat{m}_{Y \cdot 50} = .8733 - .3255(.001859) + .0448(.001859)^2 = .873$.

b. SST = 1.456923 and SSE = .117521, so R^2 = .919.

c. $.8733 - .3255\left(\dfrac{x - 49.9231}{41.3652}\right) + .0448\left(\dfrac{x - 49.9231}{41.3652}\right)^2$
$1.200887 - .01048314x + .00002618x^2$.

d. $\hat{b}_2 = \dfrac{\hat{b}_2^*}{s_x^2}$ so the estimated sd of $\hat{b}_2$ is the estimated sd of $\hat{b}_2^*$ multiplied by $\dfrac{1}{s_x}$:
$s_{\hat{b}_2} = (.0319)\left(\dfrac{1}{41.3652}\right) = .00077118$.

e. $t = \dfrac{.0448}{.0319} = 1.40$ which is not significant (compared to $\pm t_{.025,9}$ at level .05), so the quadratic term should not be retained.

35. $Y' = \ln(Y) = \ln a + bx + gx^2 + \ln(e) = b_0 + b_1 x + b_2 x^2 + e'$ where $e' = \ln(e)$, $b_0 = \ln(a)$, $b_1 = b$, and $b_2 = g$. That is, we should fit a quadratic to $(x, \ln(y))$. The resulting estimated quadratic (from computer output) is $2.00397 + .1799x - .0022x^2$, so $\hat{b} = .1799$, $\hat{g} = -.0022$, and $\hat{a} = e^{2.0397} = 7.6883$. (The ln(y)'s are 3.6136, 4.2499, 4.6977, 5.1773, and 5.4189, and the summary quantities can then be computed as before.)

Section 13.4

36.

a. Holding age, time, and heart rate constant, maximum oxygen uptake will increase by .01 L/min for each 1 kg increase in weight. Similarly, holding weight, age, and heart rate constant, the maximum oxygen uptake decreases by .13 L/min with every 1 minute increase in the time necessary to walk 1 mile.

b. $\hat{y}_{76,20,12,140} = 5.0 + .01(76) - .05(20) - .13(12) - .01(140) = 1.8$ L/min.

c. $\hat{y} = 1.8$ from **b**, and $s = .4$, so, assuming y follows a normal distribution,

$$P(1.00 < Y < 2.60) = P\left(\frac{1.00 - 1.8}{.4} < Z < \frac{2.6 - 1.8}{.4}\right) = P(-2.0 < Z < 2.0) = .9544$$

37.

a. The mean value of y when $x_1 = 50$ and $x_2 = 3$ is
$m_{y \cdot 50,3} = -.800 + .060(50) + .900(3) = 4.9$ hours.

b. When the number of deliveries (x_2) is held fixed, then average change in travel time associated with a one-mile (i.e. one unit) increase in distance traveled (x_1) is .060 hours. Similarly, when distance traveled (x_1) is held fixed, then the average change in travel time associated with on extra delivery (i.e., a one unit increase in x_2) is .900 hours.

c. Under the assumption that y follows a normal distribution, the mean and standard deviation of this distribution are 4.9 (because $x_1 = 50$ and $x_2 = 3$) and $s = .5$ (since the standard deviation is assumed to be constant regardless of the values of x_1 and x_2).

Therefore $P(y \le 6) = P\left(z \le \frac{6 - 4.9}{.5}\right) = P(z \le 2.20) = .9861$. That is, in the long run, about 98.6% of all days will result in a travel time of at most 6 hours.

38.

a. mean life $= 125 + 7.75(40) + .0950(1100) - .009(40)(1100) = 143.50$

b. First, the mean life when $x_1 = 30$ is equal to $125 + 7.75(30) + .0950x_2 - .009(30)x_2 = 357.50 - .175x_2$. So when the load increases by 1, the mean life decreases by .175. Second, the mean life when $x_1 = 40$ is equal to $125 + 7.75(40) + .0950x_2 - .009(40)x_2 = 435 - .265x_2$. So when the load increases by 1, the mean life decreases by .265.

39.

a. For $x_1 = 2$, $x_2 = 8$ (remember the units of x_2 are in 1000,s) and $x_3 = 1$ (since the outlet has a drive-up window) the average sales are $\hat{y} = 10.00 - 1.2(2) + 6.8(8) + 15.3(1) = 77.3$ (i.e., \$77,300).

b. For $x_1 = 3$, $x_2 = 5$, and $x_3 = 0$ the average sales are $\hat{y} = 10.00 - 1.2(3) + 6.8(5) + 15.3(0) = 40.4$ (i.e., \$40,400).

c. When the number of competing outlets (x_1) and the number of people within a 1-mile radius (x_2) remain fixed, the sales will increase by \$15,300 when an outlet has a drive-up window.

40.

a. $\hat{m}_{Y \cdot 10,.5,50,100} = 1.52 + .02(10) - 1.40(.5) + .02(50) - .0006(100) = 1.96$

b. $\hat{m}_{Y \cdot 20,.5,50,30} = 1.52 + .02(20) - 1.40(.5) + .02(50) - .0006(30) = 1.40$

c. $\hat{b}_4 = -.0006;\ 100\hat{b}_4 = -.06$.

d. There are no interaction predictors – e.g., $x_5 = x_1x_4$ -- in the model. There would be dependence if interaction predictors involving x_4 had been included.

e. $R^2 = 1 - \frac{20.0}{39.2} = .490$. For testing $H_0 : b_1 = b_2 = b_3 = b_4 = 0$ vs. H_a: at least one among $b_1, \ldots, b_4$ is not zero, the test statistic is $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$. H_o will be rejected if $f \geq F_{.05,4,25} = 2.76$. $f = \frac{.490/4}{.510/25} = 6.0$. Because $6.0 \geq 2.76$, H_o is rejected and the model is judged useful (this even though the value of R^2 is not all that impressive).

41. $H_0 : b_1 = b_2 = \ldots = b_6 = 0$ vs. H_a: at least one among $b_1, \ldots, b_6$ is not zero. The test statistic is $F = \dfrac{R^2/k}{(1-R^2)/(n-k-1)}$. H_o will be rejected if $f \geq F_{.05,6,30} = 2.42$.

$f = \dfrac{.83/6}{(1-.83)/30} = 24.41$. Because $24.41 \geq 2.42$, H_o is rejected and the model is judged useful.

42.

a. To test $H_0 : b_1 = b_2 = 0$ vs. H_a : at least one $b_i \neq 0$, the test statistic is $f = \dfrac{MSR}{MSE} = 319.31$ (from output). The associated p-value is 0, so at any reasonable level of significance, H_o should be rejected. There does appear to be a useful linear relationship between temperature difference and at leas one of the two predictors.

b. The degrees of freedom for SSE = n – (k + 1) = 9 – (2 – 1) = 6 (which you could simply read in the DF column of the printout), and $t_{.025,6} = 2.447$, so the desired confidence interval is $3.000 \pm (2.447)(.4321) = 3.000 \pm 1.0573$, or about $(1.943, 4.057)$. Holding furnace temperature fixed, we estimate that the average change in temperature difference on the die surface will be somewhere between 1.943 and 4.057.

c. When $x_1 = 1300$ and $x_2 = 7$, the estimated average temperature difference is $\hat{y} = -199.56 + .2100x_1 + 3.000x_2 = -199.56 + .2100(1300) + 3.000(7) = 94.44$. The desired confidence interval is then $94.44 \pm (2.447)(.353) = 94.44 \pm .864$, or $(93.58, 95.30)$.

d. From the printout, s = 1.058, so the prediction interval is $94.44 \pm (2.447)\sqrt{(1.058)^2 + (.353)^2} = 94.44 \pm 2.729 = (91.71, 97.17)$.

43.

a. $x_1 = 2.6$, $x_2 = 250$, and $x_1x_2 = (2.6)(250) = 650$, so
$\hat{y} = 185.49 - 45.97(2.6) - 0.3015(250) + 0.0888(650) = 48.313$

b. No, it is not legitimate to interpret b_1 in this way. It is not possible to increase by 1 unit the cobalt content, x_1, while keeping the interaction predictor, x_3, fixed. When x_1 changes, so does x_3, since $x_3 = x_1x_2$.

c. Yes, there appears to be a useful linear relationship between y and the predictors. We determine this by observing that the p-value corresponding to the model utility test is < .0001 (F test statistic = 18.924).

d. We wish to test $H_0 : b_3 = 0$ vs. $H_a : b_3 \neq 0$. The test statistic is t=3.496, with a corresponding p-value of .0030. Since the p-value is < alpha = .01, we reject H_o and conclude that the interaction predictor does provide useful information about y.

e. A 95% C.I. for the mean value of surface area under the stated circumstances requires the following quantities:
$\hat{y} = 185.49 - 45.97(2) - 0.3015(500) + 0.0888(2)(500) = 31.598$. Next,
$t_{.025,16} = 2.120$, so the 95% confidence interval is
$31.598 \pm (2.120)(4.69) = 31.598 \pm 9.9428 = (21.6552, 41.5408)$

44.

a. Holding starch damage constant, for every 1% increase in flour protein, the absorption rate will increase by 1.44%. Similarly, holding flour protein percentage constant, the absorption rate will increase by .336% for every 1-unit increase in starch damage.

b. $R^2 = .96447$, so 96.447% of the observed variation in absorption can be explained by the model relationship.

c. To answer the question, we test $H_0 : b_1 = b_2 = 0$ vs H_a : at least one $b_i \neq 0$. The test statistic is $f = 339.31092$, and has a corresponding p-value of zero, so at any significance level we will reject H_o. There is a useful relationship between absorption and at least one of the two predictor variables.

d. We would be testing $H_a : b_2 \neq 0$. We could calculate the test statistic $t = \dfrac{b_2}{s_{b_2}}$, or we could look at the 95% C.I. given in the output. Since the interval (.29828, 37298) does not contain the value 0, we can reject H_o and conclude that 'starch damage' should not be removed from the model.

e. The 95% C.I. is $42.253 \pm (2.060)(.350) = 42.253 \pm 0.721 = (41.532, 42.974)$.
The 95% P.I. is
$42.253 \pm (2.060)\left(\sqrt{1.09412^2 + .350^2}\right) = 42.253 \pm 2.366 = (39.887, 44.619)$.

f. We test $H_a : b_3 \neq 0$, with $t = \dfrac{-.04304}{.01773} = -2.428$. The p-value is approximately 2(.012) = .024. At significance level .01 we do not reject H_o. The interaction term should not be retained.

45.

a. The appropriate hypotheses are $H_0 : b_1 = b_2 = b_3 = b_4 = 0$ vs. H_a : at least one $b_i \neq 0$. The test statistic is $f = \dfrac{R^2/k}{(1-R^2)/(n-k-1)} = \dfrac{.946/4}{(1-.946)/20} = 87.6 \geq 7.10 = F_{.001,4,20}$ (the smallest available significance level from Table A.9), so we can reject H_o 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 $1 - \dfrac{n-1}{n-(k+1)}\left(\dfrac{SSE}{SST}\right) = 1 - \dfrac{n-1}{n-(k+1)}\left(1-R^2\right)$

$= 1 - \dfrac{24}{20}(1-.946) = .935$, 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

$\hat{y} = 6.121 - .082x + .113x + .256x - .219x$

$\hat{y} = 6.121 - .082(16.5) + .113(50) + .256(3) - .219(5) = 10.091$. For a 99% C.I., $t_{.005,20} = 2.845$, 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.

46.

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:

$H_0 : b_1 = b_2 = 0$ vs. H_a : at least one $b_i \neq 0$. We reject H_o if $f \geq F_{.05,2,9} = 4.26$.

The calculated statistic is $f = \dfrac{SSR/k}{SSE/(n-k-1)} = \dfrac{MSR}{MSE} = \dfrac{10.63/2}{(20.9)/9} = \dfrac{5.315}{.232} = 22.91$. Since $22.91 \geq 4.26$, we reject H_o and conclude that at least one of the two predictor variables is useful.

b. We will reject $H_0 : b_2 = 0$ in favor of $H_a : b_2 \neq 0$ if $|t| \geq t_{.005,9} = 3.25$. The test statistic is $t = \dfrac{1.250}{.312} = 4.01$ which is ≥ 3.25, so we reject H_o 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 b_3 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. $\hat{y} = .950 + .400(6) + 1.250(1) = 4.6$, $s^2 = MSE = .23222$, and $t_{.005,9} = 3.25$, so the 99% P.I. 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.

47.

a. For a 1% increase in the percentage plastics, we would expect a 28.9 kcal/kg increase in energy content. Also, for a 1% increase in the moisture, we would expect a 37.4 kcal/kg decrease in energy content.

b. The appropriate hypotheses are $H_0 : b_1 = b_2 = b_3 = b_4 = 0$ vs. H_a : at least one $b_i \neq 0$. The value of the F test statistic is 167.71, with a corresponding p-value that is extremely small. So, we reject H_o and conclude that at least one of the four predictors is useful in predicting energy content, using a linear model.

c. $H_0 : b_3 = 0$ vs. $H_a : b_3 \neq 0$. The value of the t test statistic is t = 2.24, with a corresponding p-value of .034, which is less than the significance level of .05. So we can reject H_o and conclude that percentage garbage provides useful information about energy consumption, given that the other three predictors remain in the model.

d. $\hat{y} = 2244.9 + 28.925(20) + 7.644(25) + 4.297(40) - 37.354(45) = 1505.5$, and $t_{.025,25} = 2.060$. (Note an error in the text: $s_{\hat{y}} = 12.47$, not 7.46). So a 95% C.I for the true average energy content under these circumstances is $1505.5 \pm (2.060)(12.47) = 1505.5 \pm 25.69 = (1479.8, 1531.1)$. Because the interval is reasonably narrow, we would conclude that the mean energy content has been precisely estimated.

e. A 95% prediction interval for the energy content of a waste sample having the specified characteristics is $1505.5 \pm (2.060)\sqrt{(31.48)^2 + (12.47)^2}$
$= 1505.5 \pm 69.75 = (1435.7, 1575.2)$.

48.

a. $H_0 : b_1 = b_2 = \ldots = b_9 = 0$

H_a : at least one $b_i \neq 0$

RR: $f \geq F_{.01,9,5} = 10.16$

$$f = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{.938/9}{(1-.938)/(5)} = 8.41$$

Fail to reject H_o . The model does not appear to specify a useful relationship.

b. $\hat{m}_y = 21.967$, $t_{a/2,n-(k+1)} = t_{.025,5} = 2.571$, so the C.I. is

$21.967 \pm (2.571)(1.248) = (18.76, 25.18)$.

c. $s^2 = \dfrac{SSE}{n-(k+1)} = \dfrac{23.379}{5} = 4.6758$, and the C.I. is

$21.967 \pm (2.571)\sqrt{4.6758 + (1.248)^2} = (15.55, 28.39)$.

d. $SSE_k = 23.379,\ SSE_l = 203.82,$

$H_0 : b_4 = b_5 = \ldots = b_9 = 0$

H_a : at least one of the above $b_i \neq 0$

RR: $f \geq F_{a,k-l,n-(k+1)} = F_{.05,6,5} = 4.95$

$$f = \frac{(203.82-23.379)/(9-3)}{(23.379)/(5)} = 6.43.$$

Reject H_o. At least one of the second order predictors appears useful.

49.

a. $\hat{m}_{y \cdot 18.9,43} = 96.8303$; Residual = 91 – 96.8303 = -5.8303.

b. $H_0 : b_1 = b_2 = 0$; H_a : at least one $b_i \neq 0$

RR: $f \geq F_{.05,2,9} = 8.02$

$$f = \frac{R^2/k}{(1-R^2)/(n-k-1)} = \frac{.768/2}{(1-.768)/9} = 14.90$$. Reject H_o. The model appears useful.

c. $96.8303 \pm (2.262)(8.20) = (78.28, 115.38)$

d. $96.8303 \pm (2.262)\sqrt{24.45^2 + 8.20^2} = (38.50, 155.16)$

e. We find the center of the given 95% interval, 93.875, and half of the width, 57.845. This latter value is equal to $t_{.025,9}(s_{\hat{y}}) = 2.262(s_{\hat{y}})$, so $s_{\hat{y}} = 25.5725$. Then the 90% interval is $93.785 \pm (1.833)(25.5725) = (46.911, 140.659)$

f. With the p-value for $H_a : b_1 \neq 0$ being 0.208 (from given output), we would fail to reject H_o. This factor is not significant given x_2 is in the model.

g. With $R_k^2 = .768$ (full model) and $R_l^2 = .721$ (reduced model), we can use an alternative f statistic (compare formulas 13.19 and 13.20). $F = \dfrac{R_k^2 - R_l^2 \,/\, k-l}{(1-R_k^2) \,/\, n-(k+1)}$. With n=12, k=2 and l=1, we have $F = \dfrac{.768 - .721}{(1-.768)/9} = \dfrac{.047}{.0257} = 1.83$.

$t^2 = (-1.36)^2 = 1.85$. The discrepancy can be attributed to rounding error.

50.

a. Here k = 5, n – (k+1) = 6, so H_o will be rejected in favor of H_a at level .05 if either $t \geq t_{.025,6} = 2.447$ or $t \leq -2.447$. The computed value of t is $t = \dfrac{.557}{.94} = .59$, so H_o cannot be rejected and inclusion of x_1x_2 as a carrier in the model is not justified.

b. No, in the presence of the other four carriers, any particular carrier is relatively unimportant, but this is not equivalent to the statement that all carriers are unimportant.

c. $SSE_k = SST(1 - R^2) = 3224.65$, so $f = \dfrac{(5384.18 - 3224.65)/3}{(3224.65)/6} = 1.34$, and since 1.34 is not $\geq F_{.05,3,6} = 4.76$, H_o cannot be rejected; the data does not argue for the inclusion of any second order terms.

51.

a. No, there is no pattern in the plots which would indicate that a transformation or the inclusion of other terms in the model would produce a substantially better fit.

b. k = 5, n – (k+1) = 8, so $H_0 : b_1 = \ldots = b_5 = 0$ will be rejected if $f \geq F_{.05,5,8} = 3.69$; $f = \dfrac{(.759)/5}{(.241)/8} = 5.04 \geq 3.69$, so we reject H_o. At least one of the coefficients is not equal to zero.

c. When $x_1 = 8.0$ and $x_2 = 33.1$ the residual is e = 2.71 and the standardized residual is e* = .44; since e* = e/(sd of the residual), sd of residual = e/e* = 6.16. Thus the estimated variance of $\hat{Y}$ is $(6.99)^2 - (6.16)^2 = 10.915$, so the estimated sd is 3.304. Since $\hat{y} = 24.29$ and $t_{.025,8} = 2.306$, the desired C.I. is $24.29 \pm 2.306(3.304) = (16.67, 31.91)$.

d. $F_{.05,3,8} = 4.07$, so $H_0 : b_3 = b_4 = b_5 = 0$ will be rejected if $f \geq 4.07$. With $SSE_k = 8, s^2 = 390.88,$ and $f = \dfrac{(894.95 - 390.88)/3}{(390.88)/8} = 3.44$, and since 3.44 is not ≥ 4.07, H_o cannot be rejected and the quadratic terms should all be deleted. (n.b.: this is not a modification which would be suggested by a residual plot.

52.

a. The complete 2nd order model obviously provides a better fit, so there is a need to account for interaction between the three predictors.

b. A 95% CI for y when $x_1 = x_2 = 30$ and $x_3 = 10$ is $.66573 \pm 2.120(.01785) = (.6279, .7036)$

53. Some possible questions might be:

Is this model useful in predicting deposition of poly-aromatic hydrocarbons? A test of model utility gives us an F = 84.39, with a p-value of 0.000. Thus, the model is useful.

Is x_1 a significant predictor of y while holding x_2 constant? A test of $H_0 : b_1 = 0$ vs the two-tailed alternative gives us a t = 6.98 with a p-value of 0.000., so this predictor is significant.

A similar question, and solution for testing x_2 as a predictor yields a similar conclusion: With a p-value of 0.046, we would accept this predictor as significant if our significance level were anything larger than 0.046.

54.

a. For $x_1 = x_2 = x_3 = x_4 = +1$, $\hat{y} = 84.67 + .650 - .258 + \ldots + .050 = 85.390$. The single y corresponding to these x_i values is 85.4, so $y - \hat{y} = 85.4 - 85.390 = .010$.

b. Letting $x_1', \ldots, x_4'$ denote the uncoded variables, $x_1' = .1x_1 + .3$, $x_2' = .1x_2 + .3$, $x_3' = x_3 + 2.5$, and $x_4' = 15x_4 + 160$; Substitution of $x_1 = 10x_1' - 3$, $x_2 = 10x_2' - 3$, $x_3 = x_3' - 2.5$, and $x_4 = \dfrac{x_4' + 160}{15}$ yields the uncoded function.

c. For the full model k = 14 and for the reduced model l – 4, while n – (k + 1) = 16. Thus $H_0 : b_5 = \ldots = b_{14} = 0$ will be rejected if $f \geq F_{.05,10,16} = 2.49$. $SSE = (1 - R^2)SST$ so $SSE_k = 1.9845$ and $SSE_l = 4.8146$, giving $f = \dfrac{(4.8146 - 1.9845)/10}{(1.9845)/16} = 2.28$. Since 2.28 is not ≥ 2.49, H_o cannot be rejected, so all higher order terms should be deleted.

d. $H_0 : m_{Y \cdot 0,0,0,0} = 85.0$ will be rejected in favor of $H_a : m_{Y \cdot 0,0,0,0} < 85.0$ if $t \leq -t_{.05,26} = -1.706$. With $\hat{m} = \hat{b}_0 = 85.5548$, $t = \dfrac{85.5548 - 85}{.0772} = 7.19$, which is certainly not ≤ -1.706, so H_o is not rejected and prior belief is not contradicted by the data.

Section 13.5

55.

a. $\ln(Q) = Y = \ln(a) + b\ln(a) + g\ln(b) + \ln(e) = b_0 + b_1 x_1 + b_2 x_2 + e'$ where $x_1 = \ln(a), x_2 = \ln(b), b_0 = \ln(a), b_1 = b, b_2 = g$ and $e' = \ln(e)$. Thus we transform to $(y, x_1, x_2) = (\ln(Q), \ln(a), \ln(b))$ (take the natural log of the values of each variable) and do a multiple linear regression. A computer analysis gave $\hat{b}_0 = 1.5652$, $\hat{b}_1 = .9450$, and $\hat{b}_2 = .1815$. For a = 10 and b = .01, $x_1 = \ln(10) = 2.3026$ and $x_2 = \ln(.01) = -4.6052$, from which $\hat{y} = 2.9053$ and $\hat{Q} = e^{2.9053} = 18.27$.

b. Again taking the natural log, $Y = \ln(Q) = \ln(a) + ba + gb + \ln(e)$, so to fit this model it is necessary to take the natural log of each Q value (and not transform a or b) before using multiple regression analysis.

c. We simply exponentiate each endpoint: $(e^{.217}, e^{1.755}) = (1.24, 5.78)$.

56.

a. $n = 20, k = 5, n-(k+1) = 14$, so $H_0 : b_1 = ... = b_5 = 0$ will be rejected in favor of H_a : at least one among $b_1, ..., b_5 \neq 0$, if $f \geq F_{.01,5,14} = 4.69$. With $f = \frac{(.769)/5}{(.231)/14} = 9.32 \geq 4.69$, so H_o is rejected. Wood specific gravity appears to be linearly related to at lest one of the five carriers.

b. For the full model, adjusted $R^2 = \frac{(19)(.769)-5}{14} = .687$, while for the reduced model, the adjusted $R^2 = \frac{(19)(.769)-4}{15} = .707$.

c. From **a**, $SSE_k = (.231)(.0196610) = .004542$, and $SSE_l = (.346)(.0196610) = .006803$, so $f = \frac{(.002261)/3}{(.004542)/14} = 2.32$. Since $F_{.05,3,14} = 3.34$ and 2.32 is not ≥ 3.34, we conclude that $b_1 = b_2 = b_4 = 0$.

d. $x_3' = \frac{x_3 - 52.540}{5.4447} = -.4665$ and $x_5' = \frac{x_5 - 89.195}{3.6660} = .2196$, so $\hat{y} = .5255 - (.0236)(-.4665) + (.0097)(.2196) = .5386$.

e. $t_{.025,17} = 2.110$ (error df = n – (k+1) = 20 – (2+1) = 17 for the two carrier model), so the desired C.I. is $-.0236 \pm 2.110(.0046) = (-.0333, -.0139)$.

f. $y = .5255 - .0236\left(\frac{x_3 - 52.540}{5.4447}\right) + .0097\left(\frac{x_5 - 89.195}{3.6660}\right)$, so $\hat{b}_3$ for the unstandardized model $= \frac{-.0236}{5.447} = -.004334$. The estimated sd of the unstandardized $\hat{b}_3$ is $= \frac{.0046}{5.447} = -.000845$.

g. $\hat{y} = .532$ and $\sqrt{s^2 + s_{\hat{b}_0 + \hat{b}_3 x_3' + \hat{b}_5 x_5'}} = .02058$, so the P.I. is $.532 \pm (2.110)(.02058) = .532 \pm .043 = (.489, .575)$.

57.

k	R^2	Adj. R^2	$C_k = \frac{SSE_k}{s^2} + 2(k+1) - n$
1	.676	.647	138.2
2	.979	.975	2.7
3	.9819	.976	3.2
4	.9824		4

Where $s^2 = 5.9825$

a. Clearly the model with k = 2 is recommended on all counts.

b. No. Forward selection would let x_4 enter first and would not delete it at the next stage.

58. At step #1 (in which the model with all 4 predictors was fit), t = .83 was the t ratio smallest in absolute magnitude. The corresponding predictor x_3 was then dropped from the model, and a model with predictors x_1, x_2, and x_4 was fit. The t ratio for x_4 , -1.53, was the smallest in absolute magnitude and $1.53 < 2.00$, so the predictor x_4 was deleted. When the model with predictors x_1 and x_2 only was fit, both t ratios considerably exceeded 2 in absolute value, so no further deletion is necessary.

59. The choice of a "best" model seems reasonably clear–cut. The model with 4 variables including all but the summerwood fiber variable would seem bests. R^2 is as large as any of the models, including the 5 variable model. R^2 adjusted is at its maximum and CP is at its minimum . As a second choice, one might consider the model with k = 3 which excludes the summerwood fiber and springwood % variables.

60. Backwards Stepping:

Step 1: A model with all 5 variables is fit; the smallest t-ratio is t = .12, associated with variable x_2 (summerwood fiber %). Since $t = .12 < 2$, the variable x_2 was eliminated.

Step 2: A model with all variables except x_2 was fit. Variable x_4 (springwood light absorption) has the smallest t-ratio (t = -1.76), whose magnitude is smaller than 2. Therefore, x_4 is the next variable to be eliminated.

Step 3: A model with variables x_3 and x_5 is fit. Both t-ratios have magnitudes that exceed 2, so both variables are kept and the backwards stepping procedure stops at this step. The final model identified by the backwards stepping method is the one containing x_3 and x_5.

(continued)

Forward Stepping:

Step 1: After fitting all 5 one-variable models, the model with x_3 had the t-ratio with the largest magnitude (t = -4.82). Because the absolute value of this t-ratio exceeds 2, x_3 was the first variable to enter the model.

Step 2: All 4 two-variable models that include x_3 were fit. That is, the models $\{x_3, x_1\}$, $\{x_3, x_2\}$, $\{x_3, x_4\}$, $\{x_3, x_5\}$ were all fit. Of all 4 models, the t-ratio 2.12 (for variable x_5) was largest in absolute value. Because this t-ratio exceeds 2, x_5 is the next variable to enter the model.

Step 3: (not printed): All possible tree-variable models involving x_3 and x_5 and another predictor, None of the t-ratios for the added variables have absolute values that exceed 2, so no more variables are added. There is no need to print anything in this case, so the results of these tests are not shown.

Note; Both the forwards and backwards stepping methods arrived at the same final model, $\{x_3, x_5\}$, in this problem. This often happens, but not always. There are cases when the different stepwise methods will arrive at slightly different collections of predictor variables.

61. If multicollinearity were present, at least one of the four R^2 values would be very close to 1, which is not the case. Therefore, we conclude that multicollinearity is not a problem in this data.

62. Looking at the h_{ii} column and using $\frac{2(k+1)}{n} = \frac{8}{19} = .421$ as the criteria, three observations appear to have large influence. With h_{ii} values of .712933, .516298, and .513214, observations 14, 15, 16, correspond to response (y) values 22.8, 41.8, and 48.6.

63. We would need to investigate further the impact these two observations have on the equation. Removing observation #7 is reasonable, but removing #67 should be considered as well, before regressing again.

64.

a. $\frac{2(k+1)}{n} = \frac{6}{10} = .6;$ since $h_{44} > .6$, data point #4 would appear to have large influence.
(Note: Formulas involving matrix algebra appear in the first edition.)

b. For data point #2, $x'_{(2)} = (1 \quad 3.453 \quad -4.920)$, so $\hat{b} - \hat{b}_{(2)} =$

$$\frac{-.766}{1-.302}(X'X)^{-1}\begin{pmatrix} 1 \\ 3.453 \\ -4.920 \end{pmatrix} = -1.0974\begin{pmatrix} .3032 \\ .1644 \\ .1156 \end{pmatrix} = \begin{pmatrix} -.333 \\ -.180 \\ -.127 \end{pmatrix}$$ and similar

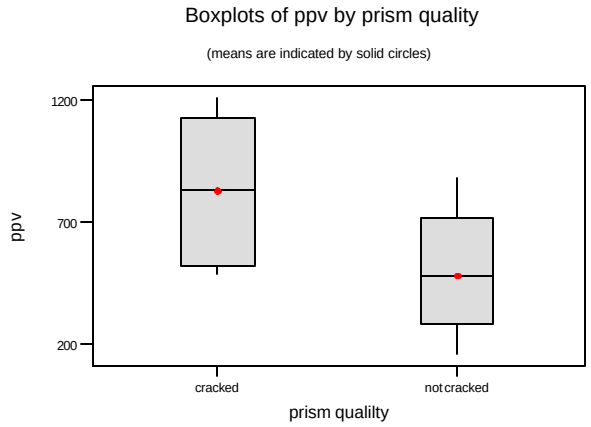
calculations yield $\hat{b} - \hat{b}_{(4)} = \begin{pmatrix} .106 \\ -.040 \\ .030 \end{pmatrix}$.

c. Comparing the changes in the $\hat{b}_i$'s to the $s_{\hat{b}_i}$'s, none of the changes is all that substantial (the largest is 1.2sd's for the change in $\hat{b}_1$ when point #2 is deleted). Thus although h_{44} is large, indicating a potential high influence of point #4 on the fit, the actual influence does not appear to be great.

Supplementary Exercises

65.

a.



A two-sample t confidence interval, generated by Minitab:

```
Two sample T for ppv

prism qu     N      Mean     StDev   SE Mean
cracked     12       827       295        85
not cracke  18       483       234        55

95% CI for mu (cracked   ) - mu (not cracke): ( 132,  557)
```

b. The simple linear regression results in a significant model, r^2 is .577, but we have an extreme observation, with std resid = -4.11. Minitab output is below. Also run, but not included here was a model with an indicator for cracked/ not cracked, and for a model with the indicator and an interaction term. Neither improved the fit significantly.

```
The regression equation is
ratio = 1.00 -0.000018 ppv

Predictor        Coef       StDev          T        P
Constant      1.00161     0.00204     491.18    0.000
ppv       -0.00001827  0.00000295      -6.19    0.000

S = 0.004892    R-Sq = 57.7%     R-Sq(adj) = 56.2%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1  0.00091571  0.00091571     38.26    0.000
Residual Error    28  0.00067016  0.00002393
Total             29  0.00158587

Unusual Observations
Obs        ppv      ratio         Fit   StDev Fit    Residual    St Resid
 29       1144   0.962000    0.980704    0.001786   -0.018704       -4.11R

R denotes an observation with a large standardized residual
```

66.

a. For every 1 cm^{-1} increase in inverse foil thickness (x), we estimate that we would expect steady-state permeation flux to increase by $.26042 mA/cm^2$. Also, 98% of the observed variation in steady-state permeation flux can be explained by its relationship to inverse foil thickness.

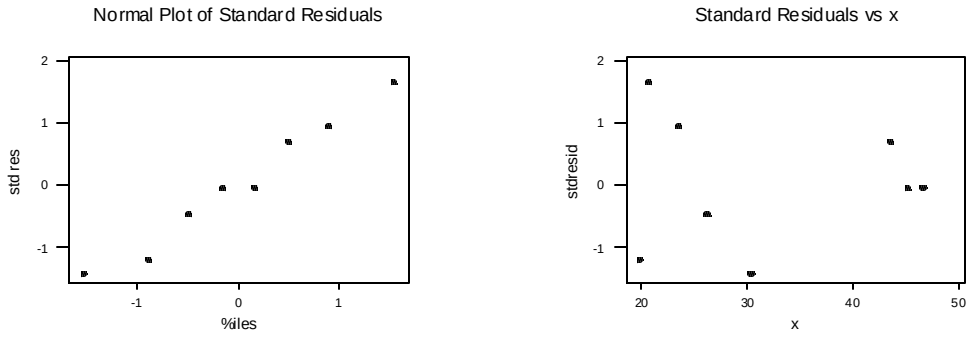
b. A point estimate of flux when inverse foil thickness is 23.5 can be found in the Observation 3 row of the Minitab output: $\hat{y} = 5.722 mA/cm^2$.

c. To test model usefulness, we test the hypotheses $H_0 : b_1 = 0$ vs. $H_a : b_1 \neq 0$. The test statistic is t = 17034, with associated p-value of .000, which is less than any significance level, so we reject H_o and conclude that the model is useful.

d. With $t_{.025,6} = 2.447$, a 95% Prediction interval for $Y_{(45)}$ is

$11.321 \pm 2.447\sqrt{.203 + (.253)^2} = 11.321 \pm 1.264 = (10.057, 12.585)$. That is, we are confident that when inverse foil thickness is 45 cm^{-1}, a predicted value of steady-state flux will be between 10.057 and 12.585 mA/cm^2.

e.



The normal plot gives no indication to question the normality assumption, and the residual plots against both x and y (only vs x shown) show no detectable pattern, so we judge the model adequate.

67.

a. For a one-minute increase in the 1-mile walk time, we would expect the VO_2max to decrease by .0996, while keeping the other predictor variables fixed.

b. We would expect male to have an increase of .6566 in VO_2max over females, while keeping the other predictor variables fixed.

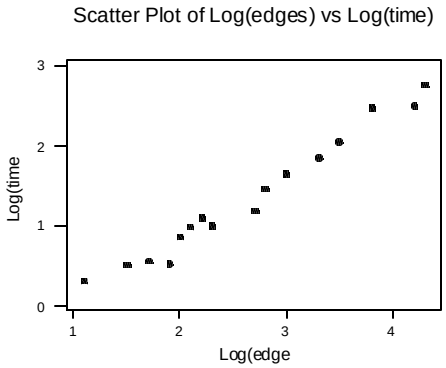
c. $\hat{y} = 3.5959 + .6566(1) + .0096(170) - .0996(11) - .0880(140) = 3.67$. The residual is $\hat{y} = (3.15 - 3.67) = -.52$.

d. $R^2 = 1 - \dfrac{SSE}{SST} = 1 - \dfrac{30.1033}{102.3922} = .706,$ or 70.6% of the observed variations in VO_2max can be attributed to the model relationship.

e. $H_0 : b_1 = b_2 = b_3 = b_4 = 0$ will be rejected in favor of H_a : at least one among $b_1, ..., b_4 \neq 0$, if $f \geq F_{.05,4,15} = 8.25$. With $f = \dfrac{(.706)/4}{(1-.706)/15} = 9.005 \geq 8.25$, so H_o is rejected. It appears that the model specifies a useful relationship between VO_2max and at least one of the other predictors.

68.

a.



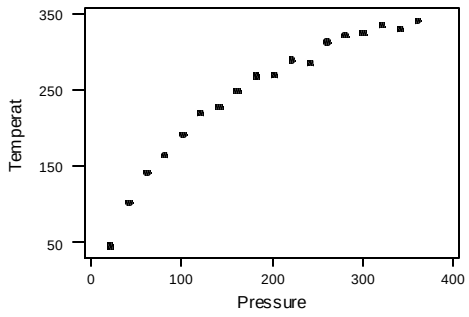
Yes, the scatter plot of the two transformed variables appears quite linear, and thus suggests a linear relationship between the two.

b. Letting y denote the variable 'time', the regression model for the variables y' and x' is $\log_{10}(y) = y' = a + bx' + e'$. Exponentiating (taking the antilogs of) both sides gives $y = 10^{a+b\log(x)+e'} = (10^a)(x^b)10^{e'} = g_0 x^{g_1} \cdot e$; i.e., the model is $y = g_0 x^{g_1} \cdot e$ where $g_0 = a$ and $g_1 = b$. This model is often called a "power function" regression model.

c. Using the transformed variables y' and x', the necessary sums of squares are $S_{x'y'} = 68.640 - \frac{(42.4)(21.69)}{16} = 11.1615$ and $S_{x'x'} = 126.34 - \frac{(42.4)^2}{16} = 13.98$. Therefore $\hat{b}_1 = \frac{S_{x'y'}}{S_{x'x'}} = \frac{11.1615}{13.98} = .79839$ and $\hat{b}_0 = \frac{21.69}{16} - (.79839)\left(\frac{42.4}{16}\right) = -.76011$. The estimate of g_1 is $\hat{g}_1 = .7984$ and $g_0 = 10^a = 10^{-.76011} = .1737$. The estimated power function model is then $y = .1737x^{.7984}$. For x = 300, the predicted value of y is $\hat{y} = .1737(300)^{.7984} 16.502$, or about 16.5 seconds.

69.

a. Based on a scatter plot (below), a simple linear regression model would not be appropriate. Because of the slight, but obvious curvature, a quadratic model would probably be more appropriate.



b. Using a quadratic model, a Minitab generated regression equation is $\hat{y} = 35.423 + 1.7191x - .0024753x^2$, and a point estimate of temperature when pressure is 200 is $\hat{y} = 280.23$. Minitab will also generate a 95% prediction interval of (256.25, 304.22). That is, we are confident that when pressure is 200 psi, a single value of temperature will be between 256.25 and 304.22 ^{o}F.

70.

a. For the model excluding the interaction term, $R^2 = 1 - \frac{5.18}{8.55} = .394$, or 39.4% of the observed variation in lift/drag ratio can be explained by the model without the interaction accounted for. However, including the interaction term increases the amount of variation in lift/drag ratio that can be explained by the model to $R^2 = 1 - \frac{3.07}{8.55} = .641$, or 64.1%.

b. Without interaction, we are testing $H_0 : b_1 = b_2 = 0$ vs. H_a : either b_1 or $b_2 \neq 0$. The test statistic is $f = \frac{R^2/k}{(1-R^2)/(n-k-1)}$, The rejection region is $f \geq F_{.05,2,6} = 5.14$, and the calculated statistic is $f = \frac{.394/2}{(1-.394)/6} = 1.95$, which does not fall in the rejection region, so we fail to reject H_o. This model is not useful. With the interaction term, we are testing $H_0 : b_1 = b_2 = b_3 = 0$ vs. H_a : at least one of the $b_i's \neq 0$. With rejection region $f \geq F_{.05,3,5} = 5.41$ and calculated statistic $f = \frac{.641/3}{(1-.641)/5} = 2.98$, we still fail to reject the null hypothesis. Even with the interaction term, there is not enough of a significant relationship between lift/drag ratio and the two predictor variables to make the model useful (a bit of a surprise!)

71.

a. Using Minitab to generate the first order regression model, we test the model utility (to see if any of the predictors are useful), and with $f = 21.03$ and a p-value of .000, we determine that at least one of the predictors is useful in predicting palladium content. Looking at the individual predictors, the p-value associated with the pH predictor has value .169, which would indicate that this predictor is unimportant in the presence of the others.

b. Testing $H_0 : b_1 = \ldots = b_{20} = 0$ vs. H_a : at least one of the $b_i's \neq 0$. With calculated statistic $f = 6.29$, and p-value .002, this model is also useful at any reasonable significance level.

c. Testing $H_0 : b_6 = \ldots = b_{20} = 0$ vs. H_a : at least one of the listed $b_i's \neq 0$, the test statistic is $f = \frac{(SSE_l - SSE_k)/k-l}{(SSE_k)/n-k-1} = \frac{(716.10-290.27)/(20-5)}{290.27/(32-20-1)} = 1.07$. Using significance level .05, the rejection region would be $f \geq F_{.05,15,11} = 2.72$. Since 1.07 < 2.72, we fail to reject H_o and conclude that all the quadratic and interaction terms should not be included in the model. They do not add enough information to make this model significantly better than the simple first order model.

d. Partial output from Minitab follows, which shows all predictors as significant at level .05:

```
The regression equation is
pdconc = - 305 + 0.405 niconc + 69.3 pH - 0.161 temp + 0.993 currdens
           + 0.355 pallcont - 4.14 pHsq

Predictor        Coef       StDev          T        P
Constant      -304.85       93.98      -3.24    0.003
niconc        0.40484     0.09432       4.29    0.000
pH              69.27       21.96       3.15    0.004
temp          -0.16134     0.07055      -2.29    0.031
currdens        0.9929      0.3570       2.78    0.010
pallcont       0.35460     0.03381      10.49    0.000
pHsq            -4.138       1.293      -3.20    0.004
```

72.

a. $R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{.80017}{16.18555} = .9506$, or 95.06% of the observed variation in weld strength can be attributed to the given model.

b. The complete second order model consists of nine predictors and nine corresponding coefficients. The hypotheses are $H_0 : b_1 = \ldots = b_9 = 0$ vs. H_a : at least one of the $b_i's \neq 0$. The test statistic is $f = \frac{R^2/k}{(1-R^2)/(n-k-1)}$, where k = 9, and n = 37.The rejection region is $f \geq F_{.05,9,27} = 2.25$. The calculated statistic is $f = \frac{.9506/9}{(1-.9506)/27} = 57.68$ which is ≥ 2.25, so we reject the null hypothesis. The complete second order model is useful.

c. To test $H_0 : b_7 = 0$ vs $H_a : b_7 \neq 0$ (the coefficient corresponding to the wc*wt predictor), $t = \sqrt{f} = \sqrt{2.32} = 1.52$. With df = 27, the p-value $\approx 2(.073) = .146$ (from Table A.8). With such a large p-value, this predictor is not useful in the presence of all the others, so it can be eliminated.

d. The point estimate is $\hat{y} = 3.352 + .098(10) + .222(12) + .297(6) - .0102(10^2) - .037(6^2) + .0128(10)(12) = 7.962$. With $t_{.025,27} = 2.052$, the 95% P.I. would be $7.962 \pm 2.052(.0750) = 7.962 \pm .154 = (7.808, 8.116)$. Because of the narrowness of the interval, it appears that the value of strength can be accurately predicted.

73.

a. We wish to test $H_0 : b_1 = b_2 = 0$ vs. H_a : either b_1 or $b_2 \neq 0$. The test statistic is $f = \frac{R^2/k}{(1-R^2)/(n-k-1)}$, where k = 2 for the quadratic model. The rejection region is $f \geq F_{a,k,n-k-1} = F_{.01,2,5} = 13.27$. $R^2 = 1 - \frac{.29}{202.88} = .9986$, giving f = 1783. No doubt about it, folks – the quadratic model is useful!

b. The relevant hypotheses are $H_0 : b_2 = 0$ vs. $H_a : b_2 \neq 0$. The test statistic value is $t = \frac{\hat{b}_2}{s_{\hat{b}_2}}$, and H_o will be rejected at level .001 if either $t \geq 6.869$ or $t \leq -6.869$ (df = n – 3 = 5). Since $t = \frac{-.00163141}{.00003391} = -48.1 \leq -6.869$, H_o is rejected. The quadratic predictor should be retained.

c. No. R^2 is extremely high for the quadratic model, so the marginal benefit of including the cubic predictor would be essentially nil – and a scatter plot doesn't show the type of curvature associated with a cubic model.

d. $t_{.025,5} = 2.571$, and $\hat{b}_0 + \hat{b}_1(100) + \hat{b}_2(100)^2 = 21.36$, so the C.I. is
$21.36 \pm (2.571)(.1141) = 21.36 \pm .69 = (20.67, 22.05)$

e. First, we need to figure out s^2 based on the information we have been given.
$s^2 = MSE = \frac{SSE}{df} = \frac{.29}{5} = .058$. Then, the 95% P.I. is
$21.36 \pm 2.571\left(\sqrt{.058 + .1141}\right) = 21.36 \pm 1.067 = (20.293, 22.427)$

74. A scatter plot of $y' = \log_{10}(y)$ vs. x shows a substantial linear pattern, suggesting the model $Y = a \cdot (10)^{bx} \cdot e$, i.e. $Y' = \log(a) + bx + \log(e) = b_0 + b_1 x + e'$. The necessary summary quantities are
$\Sigma x_i = 397,\ \Sigma x_i^2 = 14{,}263,\ \Sigma y_i' = -74.3,\ \Sigma y_i'^2 = 47{,}081,$ and $\Sigma x_i y_i' = -2358.1$,
giving $\hat{b}_1 = \dfrac{12(-2358.1) - (397)(-74.3)}{12(14{,}263) - (397)^2} = .08857312$ and $\hat{b}_0 = -9.12196058$.
Thus $\hat{b} = .08857312$ and $a = 10^{-9.12196058}$. The predicted value of y' when x = 35 is
$-9.12196058 + .08857312(35) = -6.0219$, so $\hat{y} = 10^{-6.0219}$.

75.

a. $H_0 : b_1 = b_2 = 0$ will be rejected in favor of H_a: either b_1 or $b_2 \neq 0$ if
$f = \dfrac{R^2/k}{(1-R^2)/(n-k-1)} \geq F_{a,k,n-k-1} = F_{.01,2,7} = 9.55$. $SST = \Sigma y^2 - \dfrac{(\Sigma y)}{n} = 264.5$, so
$R^2 = 1 - \dfrac{26.98}{264.5} = .898$, and $f = \dfrac{.898/2}{(.102)/7} = 30.8$. Because $30.8 \geq 9.55$ H_o is rejected at significance level .01 and the quadratic model is judged useful.

b. The hypotheses are $H_0 : b_2 = 0$ vs. $H_a : b_2 \neq 0$. The test statistic value is
$t = \dfrac{\hat{b}_2}{s_{\hat{b}_2}} = \dfrac{-2.3621}{.3073} = -7.69$, and $t_{.0005,7} = 5.408$, so H_o is rejected at level .001 and p-value < .001. The quadratic predictor should not be eliminated.

c. x = 1 here, and $\hat{m}_{Y\cdot 1} = \hat{b}_0 + \hat{b}_1(1) + \hat{b}_2(1)^2 = 45.96$. $t_{.025,7} = 1.895$, giving the C.I.
$45.96 \pm (1.895)(1.031) = (44.01, 47.91)$.

76.

a. 80.79

b. Yes, p-value = .007 which is less than .01.

c. No, p-value = .043 which is less than .05.

d. $.14167 \pm (2.447)(.03301) = (.0609, .2224)$

e. $\hat{m}_{y \cdot 9,66} = 6.3067$, using $a = .05$, the interval is
$6.3067 \pm (2.447)\sqrt{(.4851)^2 + (.162)^2} = (5.06, 7.56)$

77.

a. Estimate = $\hat{b}_0 + \hat{b}_1(15) + \hat{b}_2(3.5)^2 = 180 + (1)(15) + (10.5)(3.5) = 231.75$

b. $R^2 = 1 - \dfrac{117.4}{1210.30} = .903$

c. $H_0 : b_1 = b_2 = 0$ vs. H_a : either b_1 or $b_2 \neq 0$ (or both). $f = \dfrac{.903/2}{.097/9} = 41.9$, which greatly exceeds $F_{.01,2,9}$ so there appears to be a useful linear relationship.

d. $s^2 = \dfrac{117.40}{12-3} = 13.044$, $\sqrt{s^2 + (est.st.dev)^2} = 3.806$, $t_{.025,9} = 2.262$. The P.I. is
$229.5 \pm (2.262)(3.806) = (220.9, 238.1)$

78. The second order model has predictors $x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, x_2x_3$ with corresponding coefficients $b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8, b_9$. We wish to test $H_0 : b_4 = b_5 = b_6 = b_7 = b_8 = b_9 = 0$ vs. the alternative that at least one of these six b_i's is not zero. The test statistic value is $f = \dfrac{(821.5 - 5027.1)/(9-3)}{(5027.1)/(20-10)} = \dfrac{530.9}{502.71} = 1.1$. Since $1.1 < F_{.05,6,10} = 3.22$, H_o cannot be rejected. It doesn't appear as though any of the quadratic or interaction carriers should be included in the model.

79. There are obviously several reasonable choices in each case.

a. The model with 6 carriers is a defensible choice on all three grounds, as are those with 7 and 8 carriers.

b. The models with 7, 8, or 9 carriers here merit serious consideration. These models merit consideration because R_k^2, MSE_k, and CK meet the variable selection criteria given in Section 13.5.

80.

a. $f = \dfrac{(.90)/(15)}{(.10)/(4)} = 2.4$. Because 2.4 < 5.86, $H_0 : b_1 = \ldots = b_{15} = 0$ cannot be rejected. There does not appear to be a useful linear relationship.

b. The high R^2 value resulted from saturating the model with predictors. In general, one would be suspicious of a model yielding a high R^2 value when K is large relative to n.

c. $\dfrac{(R^2)/(15)}{(1-R^2)/(4)} \geq 5.86$ iff $\dfrac{R^2}{1-R^2} \geq 21.975$ iff $R^2 \geq \dfrac{21.975}{22.975} = .9565$

81.

a. The relevant hypotheses are $H_0 : b_1 = \ldots = b_5 = 0$ vs. H_a: at least one among $b_1, \ldots, b_5$ is not 0. $F_{.05,5,111} = 2.29$ and $f = \dfrac{(.827)/(5)}{(.173)/(111)} = 106.1$. Because $106.1 \geq 2.29$, H_o is rejected in favor of the conclusion that there is a useful linear relationship between Y and at least one of the predictors.

b. $t_{.05,111} = 1.66$, so the C.I. is $.041 \pm (1.66)(.016) = .041 \pm .027 = (.014, .068)$. b_1 is the expected change in mortality rate associated with a one-unit increase in the particle reading when the other four predictors are held fixed; we cab be 90% confident that .014 < b_1 < .068.

c. $H_0 : b_4 = 0$ will be rejected in favor of $H_a : b_4 \neq 0$ if $t = \dfrac{\hat{b}_4}{s_{\hat{b}_4}}$ is either ≥ 2.62 or ≤ -2.62. $t = \dfrac{.014}{.007} = 5.9 \geq 2.62$, so H_o is rejected and this predictor is judged important.

d. $\hat{y} = 19.607 + .041(166) + .071(60) + .001(788) + .041(68) + .687(.95) = 99.514$ and the corresponding residual is 103 – 99.514 = 3.486.

82.

a. The set $x_1, x_3, x_4, x_5, x_6, x_8$ includes both x_1, x_4, x_5, x_8 and x_1, x_3, x_5, x_6, so $R^2_{1,3,4,5,6,8} \geq \max\left(R^2_{1,4,5,8}, R^2_{1,3,5,6}\right) = .723$.

b. $R^2_{1,4} \leq R^2_{1,4,5,8} = .723$, but it is not necessarily $\leq .689$ since x_1, x_4 is not a subset of x_1, x_3, x_5, x_6.