Exercise 3.10

a. There is evidence of a linear trend: winning times appear to decrease with year. A straight-line model for predicting winning times based on year is:

\[ \text{Winning time} = \beta_0 + \beta_1 \text{ year} + \text{error}, \]

where errors are assumed to be normally distributed with mean zero and some unknown variance \( \sigma^2 \). The slope of the line will be negative because time decreases as year increases.

b. Similar response.

c. In absolute value, the slope associated to women will be larger, because times appear to decrease more rapidly for women than for men. For example, winning times between 1980 and 2000 appear to have decreased approximately 10 minutes for men but over 30 minutes for women.

d. It might be reasonable to use the model for men, but for women the predicted times 20 years out would appear to be too low. It is quite possible that winning times for women will decrease more slowly in the future, and thus the straight line model would probably under-predict actual winning times in 2020.

Exercise 3.13

a. The least squares estimators of intercept and slope are (from JMP output):
\[
\begin{align*}
    b_0 &= 6.25 \\
    b_1 &= -0.0023
\end{align*}
\]

b. Results indicate the following:

- The sweetness index of orange juice is negatively associated to the water-soluble pectin content in the juice. When pectin increases by 1 ppm, the sweetness index is expected to decrease by 0.0023 units. The scale of the data indicate that pectin content ranges from about 200 to about 400 ppm in orange juice. Thus a clearer interpretation of results is that, for example, every 10 ppm increase in pectin is expected to result in a 0.023 decrease in sweetness index units.
- When pectin is not present (0 ppm in the juice), sweetness can be expected to be as high as 6.25 units.
c. In juice with 300 ppm pectin content, expected sweetness is $6.25 - 0.0023 \times 300 = 5.56$ units.

Exercise 3.14

The SAS program that was used to answer the questions in this problem is given below. The path in the infile statement needs to be the appropriate path for your system.

data birds;
  infile 'C:\Stat 328\Text Data\Data Files\Chap3\Condor2.dat';
  input sample nkill nestoccup;
run;

symbol1 v = dot i = none h = 2;
axis1 order = (0 to 5 by 1) label = (f = triplex a = 90 h = 2 'No. Flycatchers killed')
  minor = none length = 60 pct;
axis2 order = (20 to 70 by 5) label = (f = triplex h = 2 '% next box tit occupancy')
  minor = none length = 75 pct;

proc gplot data = birds;
  plot nkill * nestoccup / vaxis = axis1 haxis = axis2 frame;
  title f = triplex h = 3 'Birds data';
run;

proc reg data = birds;
  model nkill = nestoccup;
run;

a. The scatter plot of the data obtained from SAS is below. The frequency of flycatchers killed, while larger when the proportion of nests is occupied by tits, seems to be only modestly linearly associated to it. A curve (rather than a straight line) might be a better model for these data. A straight line might be used as a first rough model.
b. See SAS output below. The values of the estimated intercept and slope are, respectively -3.05 and 0.108. Those values are interpreted as follows:

- The number of flycatchers killed can be expected to increase by about 0.108 when next box tit occupancy increases by 1%.
- When next box tit occupancy is 0%, the number of flycatchers killed can be expected to be -3.05. This of course is nonsensical from a subject-matter point of view and suggests that a curve rather than a straight-line might be a better model.
Exercise 3.16

a. With $n = 9$, we have $n-2 = 7$ degrees of freedom for error. Therefore, $S^2 = \frac{SSE}{n-2} = \frac{0.219}{7} = 0.0313$.

b. $S = 0.177$ (the square root of $S^2$). The value $S$ is a measure of the variability of the observations around the fitted regression line. The tighter the observations cluster around the line, the smaller the value of $S$. We expect about 95% of the observations to lie within +/- 2$S$ of their predicted value. In this example, we expect that most observations will lie within +/- 0.39 of their predicted value.

Exercise 3.20

See JMP output below:

a. The least squares line is:

Predicted heat transfer = 0.21 + 2.43 x unflooded area.

b. See plot above, at the top.

c. SSE is 4.53 and $S^2$ is 0.206.

d. $S$ is 0.454. We expect about 95% of the observations to lie within +/- 2$S$ of their predicted value. Thus, most of our observations would be expected to lie within +/- 0.91 of their predicted value.
Response H-transf

Whole Model

Regression Plot

Actual by Predicted Plot

Leverage Plot

Summary of Fit

RSquare 0.837041
RSquare Adj 0.829634
Root Mean Square Error 0.453826
Mean of Response 4.775
Observations (or Sum Wgts) 24

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>23.273921</td>
<td>23.2739</td>
<td>23.2739</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>22</td>
<td>0.2060</td>
<td>0.0931</td>
<td>0.0931</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>C. Total</td>
<td>23</td>
<td>23.273921</td>
<td>23.2739</td>
<td>23.2739</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Lack Of Fit

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack Of Fit</td>
<td>19</td>
<td>3.5210791</td>
<td>0.185320</td>
<td>0.185320</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pure Error</td>
<td>3</td>
<td>1.0100000</td>
<td>0.336667</td>
<td>0.336667</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Total Error</td>
<td>22</td>
<td>4.5310791</td>
<td>0.2060</td>
<td>0.2060</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Max RSq 0.9637

Parameter Estimates

| Term  | Estimate | Std Error | t Ratio  | Prob>|z| |
|-------|----------|-----------|----------|-----|---|
| intercept | 0.2133892 | 0.439 | 0.49 | 0.6317 |<.0001 |
| Area   | 2.4628687 | 0.228252 | 10.63 |<.0001 |

Effect Tests

<table>
<thead>
<tr>
<th>Source</th>
<th>Nparm</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>F Ratio</th>
<th>Prob &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>1</td>
<td>1</td>
<td>23.273921</td>
<td>113.0032</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Residual by Predicted Plot
Exercise 3.22

There is sufficient evidence to conclude that sale price increases with assessed values for homes in Tampa. The slope is estimated to be 0.8845 and the test of hypothesis that slope is equal to zero versus the alternative that the slope is different from zero leads to concluding $H_a$ at the 0.01 level (in fact, the p-value for the test is below 0.001).

The 95% confidence interval for the true slope is 0.84 to 0.93. In English this can be interpreted as follows: with high probability, the interval 0.84 to 0.93 covers the true value of the slope. Thus, we are 95% confident that the true slope is no smaller than 0.84 or larger than 0.93. The interval does not include the value zero and thus we are quite confident that the true slope is not equal to zero.

A narrower confidence interval can be obtained in one of two ways:

- By decreasing the coverage level from 95% to perhaps 90%, but that would decrease our confidence in our conclusions.
- By increasing the sample size to include data on more homes. If $n$ increases, $S$ is likely to decrease and thus the 95% CI for the true slope can be expected to be narrower.

Exercise 3.23

A 90% CI for the true slope is given by

$$b_1 +/- \frac{t_{\alpha/2, n-2}}{S_{b_1}} S_{b_1}.$$ 

In orange juice example, $b_1 = -0.0023$ and $S_{b_1} = 0.000905$. For $\alpha = 0.10$, and $n-2 = 22$, we get:

$$-0.0023 +/- 1.717 \times 0.000905 = (-0.0038 \text{ to } -0.00074).$$

The 90% confidence interval does not cover 0. Thus we conclude that we have evidence to say that pectin in orange juice is linearly associated to sweetness. The higher the pectin content the lower the sweetness as measured by the index.