Your Turn

1. You have five machines. The following data gives the age of the machines in years and the annual maintenance cost in thousands of dollars:

   ```
   > machine = data.frame(age=c(2,5,9,3,8), cost=c(6,13,23,5,22))
   ```

   1. (a) Draw a scatterplot and describe the relationship.
   2. (b) Find the correlation between age and cost. What does it tell you?
   3. (c) Find the equation predicting cost from age. Interpret the slope and intercept.
   4. (d) Superimpose the regression line on your scatterplot.
   5. (e) What would you expect the annual maintenance to be for a machine that is 7 years old?
   6. (f) What is a typical size for the prediction errors?
   7. (g) What fraction of the variation in cost is explained by knowing the age of the machines?
   8. (h) Is the relationship between age and cost statistically significant?
   9. (i) A colleague suggested to use $20,000 per year per machine for planning purposes. Perform a test at the 5% level to see if this is reasonable.

2. The amounts of a chemical compound y, which was dissolved in 100 grams of water at various temperatures, x° C, were recorded:

   ```
   dat = data.frame(
   x=c(0,0,0, 15,15,15, 30,30,30, 45,45,45, 60,60,60, 75,75,75), y=c(8,6,8, 12,10,14,
   25,21,24, 31,33,28, 44,39,42, 48,51,44))
   ```

   (a) Find the equation of the regression line. (b) Graph the line on a scatterplot.

   (c) Compute and interpret the standard error of the estimate. (d) Compute and interpret the coefficient of determination.

   (e) Find a 99% CI for β_0 (f) Find a 99% CI for β_1

   (g) Test at the 1% level if the slope differs from 0.
   (h) Estimate the mean amount that will dissolve in 100 grams of water at 50° C.

   (i) Find a 99% CI for the mean amount that will dissolve at 50° C. (j) Find a 99% PI for the mean amount that will dissolve at 50° C.

3. Consider the circulation and the open line rate (price per line for an ad placed just once) for selected large newspapers shown below

```
dat=data.frame(
circ = c(2081995,1374858,1284613,1057536,970051,963069,828236,779259,768288,
691771,663693,657015,645623,533384,528777,514702,492002,486426, 443592,349182),
linerate=c(37.65,18.48,14.50,14.61,16.47,16.07,13.82,13.05,13.78,12.25,10.53,14.18,12.83,7.81,5
.17,11.08,6.58,8.77,6.03,6.77),
row.names=c("WSJ","NY Daily News","USA Today","LA Times","NYT", "NY Post",
"Philadelphia","Chi Tribune","Wash Post","SF Chronicle","Chi Sun Times", "Detroit
News","Detroit Free Press","Long Island Newsday","KC Times", "Miami
Herald","Cleveland","Milwaukee","Houston","Baltimore"))
```

1. (a) Create a scatterplot of the open line rate against circulation. Comment.
2. (b) Find and interpret the correlation of open line rate with circulation. Is the correlation reasonable from a business perspective?
3. (c) Find the regression equation to predict open line rate from circulation. Superimpose the regression line on your scatterplot.
4. (d) Test if the association between the open-line rate and circulation is significant ($\alpha = .05$).
5. (e) Find the predicted and residual value for the New York Times. Interpret these values. In particular, is the open line rate higher or lower than what you would expect for a newspaper with its circulation?
6. (f) There is an outlier visible in the scatterplot. Let's test to see if it could reasonably be from the same population as the others by treating it as a new observation. Remove The Wall Street Journal (WSJ) from the data set and find the regression equation to predict the open line rate from circulation for the other newspapers.
7. (g) Find the two-sided 95% prediction interval (PI) for a new observation, with X_0 being the circulation of WSJ.
8. (h) Test whether or not WSJ is an outlier by seeing if its open line rate is in the PI.
9. (i) The milline rate is defined as the open line rate divided by the circulation, in millions. Thus, it is the cost per line oof advertising per million circulation. This adjustment should take care of some of the differences in advertising rates due to circulation. That is, one explanation of the open line rate is that it is proportional to circulation. If it is just proportional, there should be nothing left in the milline rate to be explained by circulation. On the other hand, if there is an additional advantage or penalty to being big, the circulation should help explain the variation in milline rates. Let's use regression analysis to see if three is anything left in the milline rate to be explained by circulation. Draw a scatterplot of milline rate against circulation.
10. (j) Find and interpret the correlation between circulation and milline rate.
11. (k) What percentage of the variation in milline rate is explained by circulation?
12. (l) Test to see if there is a significant relationship between circulation and milline rate.

13. (m) Write a paragraph explaining and interpreting your results.

4. The data below gives mailing-list size (thousands of names) and sales (thousands of dollars) for a group of catalogs.

```
dat = data.frame(
size=c(168, 21, 94, 39, 249, 43, 589, 41),
sales = c(5178, 2370,3591, 2056, 7325, 2449, 15708, 2469))
```

1. (a) How strong is the association between these two variables? Find the appropriate summary measure and interpret it.
2. (b) Find the equation to predict sales from the size of the mailing list.
3. (c) What level of sales would you expect for a a catalog mailed to 5,000 people?
4. (d) What percent of the variation in the list size can be explained by the fact that some generated more sales than others?
5. (e) Is there a significant relationship between list size and sales? How do you know?

5. JWHT problem 8a,b on pages 121–2 (Hint: see §2.3.4 on page 48–49.) Click here for data. If you use the data from the author's website you will need to read about the na.strings option. Note: omit part c for now. Answer these questions about the output. Type the following to read it in:

```
auto = read.table("Downloads/auto.csv", header=T, na.strings="?") auto$origin =
factor(auto$origin, 1:3, c("US", "Europe", "Japan"))
```

1. (a) What is the estimated regression equation?
2. (b) What does the slope tell you?
3. (c) How much uncertainty is associated with the slope estimate?
4. (d) What does the residual standard error tell you?
5. (e) Using this model, is there a significant relationship between mpg and horsepower?
6. (f) What fraction of the variation in mpg is explained by using this linear function of horsepower?
7. (g) What is the predicted mpg associated with a horsepower of 98?
8. (h) What is the 95% prediction interval for the predicted mpg associated with a horsepower of 98?
9. (i) What is the 99% confidence interval for the mean prediction of mpg when horsepower is 98?
10. (j) What is a 90% confidence interval for the slope?
11. (k) In looking at the scatterplot and fitted model, note any violations of the model assumptions.