Today we continue our study of situations involving two quantitative variables, changing our focus to mathematical models for predicting the values of one variable based on the other.

**Example 19-1: Predicting heights from footprints**
Can a footprint taken at the scene of a crime help to determine the height of the criminal? In other words, is there an association between height and foot length? A sample of 20 students measured their height (in inches) and their foot length (in centimeters). To see a scatterplot of the data, click on “applets” and then on “Least squares regression” from our course web page.

a) Describe the association between the variables (direction, strength, form).

b) Make a guess for the value of the correlation coefficient between height and foot length.

c) Click the “Your line” box to add a blue line to the scatterplot. This line is movable. Click the Move Line button. If you now place your mouse over one of the ends and drag, you can change the slope of the line. You can also use the mouse to move the green dot up and down vertically to change the intercept of the line. Move the line until you believe your line “best” summarizes the relationship between height and foot length for these data. Write down the resulting equation for your line.

One way to measure the fit of your line is to calculate the residuals for all of the observational units. A **residual** is the difference between the observed $y$ value and the $y$ value predicted by your line for a particular $x$ value: $\text{residual}_i = y_i - \hat{y}_i$.

d) Click the “Show residuals” box to visually represent these residuals for your line on the scatterplot. The applet also reports the sum of the absolute residuals/errors (SAE) under your equation. Record this SAE value for your line.

e) A more common criterion for determining the “best” line is to instead look at the sum of the **squared residuals** (SSE). Click the “Show squared residuals” to visually represent them and to determine SSE for your line. Record this value.
f) Now continue to adjust your line until you think you have minimized the sum of the squared residuals. Report your new equation and new SSE value.

g) Now click on “Regression line” to determine and display the equation for the line that actually does minimize (as shown using some calculus) the sum of the squared residuals. Record its equation. (You can also display the residuals and the squared residuals for this line.)

h) Use the regression line to predict the height of someone whose foot length is 28 cm. Does this prediction seem reasonable, based on the scatterplot?

i) Use the regression line to predict the height of someone whose foot length is 29 cm.

i) By how much do these predictions differ? Does this number look familiar? Explain.

- The **slope coefficient** of a least squares regression model is interpreted as the predicted change in the response (y-) variable for a one-unit change in the explanatory (x-) variable.

j) “Reload” the applet and click the “Your line” box to redisplay the blue line. Notice that this line is flat at the mean of the y (height) values. Click the “Show squared residuals” box to determine the SSE if we were to use \( \bar{y} \) as our predicted value for every x (foot size). Record this value.

k) Recall the SSE value for the regression line. Determine the percentage change in the SSE between the \( \bar{y} \) line and the regression line:

\[
100\% \times \frac{SS_{resid(\bar{y})} - SS_{resid(least-squares)}}{SS_{resid(\bar{y})}} =
\]

- This expression indicates the reduction in the prediction errors from using the least squares line instead of the \( \bar{y} \) line. This is referred to as the **coefficient of determination**, denoted by \( r^2 \) or \( R^2 \), and is interpreted as the percentage of the variability in the response variable that is explained by the least-squares regression on the explanatory variable. This provides us with a measure of how accurate our predictions will be and is most useful for comparing different models (e.g., different choices of explanatory variable). The coefficient of determination is equal to the square of the correlation coefficient.
l) Verify that your answer to (k) equals the square of the correlation coefficient.

The slope and intercept coefficients of the least squares line can be calculated from summary statistics. Let the equation of a generic least squares line be: \( \hat{y} = b_0 + b_1 x \), so \( b_0 \) is the intercept coefficient and \( b_1 \) is the slope coefficient. (The “hat” on the \( y \)-variable indicates that the line produces an estimate, or predicted value, for the response.)

- The value of the slope coefficient can be calculated as: \( b_1 = r \frac{s_y}{s_x} \).
- The value of the intercept coefficient can be calculated as: \( b_0 = \bar{y} - b_1 \bar{x} \).

**Example 19-2: Car data (cont.)**
Reconsider the car data, and consider predicting a car’s highway MPG rating from its weight.

a) Examine and comment on a scatterplot of these data (cars99.mtw). Remember to put the response variable on the vertical axis.

b) Use Minitab to calculate the following descriptive statistics:

<table>
<thead>
<tr>
<th>Mean hwy MPG</th>
<th>SD hwy MPG</th>
<th>Mean weight</th>
<th>SD weight</th>
<th>Correlation</th>
</tr>
</thead>
</table>

b) Use Minitab to calculate the following descriptive statistics:

<table>
<thead>
<tr>
<th>Mean hwy MPG</th>
<th>SD hwy MPG</th>
<th>Mean weight</th>
<th>SD weight</th>
<th>Correlation</th>
</tr>
</thead>
</table>

c) Use these statistics to determine the least squares line for predicting a car’s highway MPG rating from its weight. [Hint: Be sure to write this as an equation, and get in the habit of using variable names rather than generic \( y \) and \( x \) symbols.]

d) Use Minitab to confirm these calculations and to superimpose the regression line on the scatterplot (Stat> Regression> Fitted Line Plot).

e) Interpret the value of the slope coefficient of this line.

f) Does the value of the intercept coefficient make sense in this context? Explain.
g) What highway MPG rating would the least squares line predict for a car weighing 3600 pounds?

h) What proportion of the variability in highway MPG ratings is explained by the least squares line with weight?

Now we consider how to make inferences about a population regression line based on a sample regression line.

- Population regression line: \( y = \beta_0 + \beta_1 x \)
- Estimate \( \beta_0 \) by \( b_0 \), \( \beta_1 \) by \( b_1 \).
- We can conduct hypothesis tests and form confidence intervals for the population slope coefficient \( \beta_1 \).
  - Typical null hypothesis is that \( \beta_1 = 0 \) (no linear relationship)
  - Standard error of \( b_1 \): \( SE(b_1) = \frac{s_e}{\sqrt{\sum (x_i - \bar{x})^2}} \), where \( s_e = \sqrt{\frac{SS(resid)}{n-2}} \)
  - Compare to \( t \)-distribution with \((n-2)\) degrees of freedom
  - Minitab calculates standard error, test statistic, two-sided \( p \)-value

**Example 19-3: House prices (cont.)**
Recall the data on sales price and size of 20 houses in Arroyo Grande in February 2007 (HousePricesAG.mtw).

a) Produce regression output for predicting price from size (Stat> Regression> Regression> Fit Regression Model). Report the regression equation, along with the sample slope coefficient and its standard error.

b) State the null and alternative hypotheses for testing whether there is a positive slope coefficient in the population. Also indicate what the population of interest is for these data.

c) Determine (by hand) the test statistic and \( p \)-value.
d) Summarize the conclusion that you would draw from this test.

e) Produce and interpret a 95% confidence interval for the population slope coefficient.

Now we consider confidence intervals for two additional quantities:
- the mean value of the response variable at a particular value of the explanatory variable
- a single observation of the response variable at a particular value of the explanatory variable

We will let Minitab do the calculations for us, but we will examine the relevant formulas and make some general observations based on them.

Example 19-4: House prices (cont.)
Recall yet again the data on sales price and size of 20 houses in Arroyo Grande in February 2007 (HousePricesAG.mtw).

a) Reproduce the scatterplot with the regression line superimposed. Also report the equation of that line.

b) Use the regression line to predict the price of a 1200 square foot house.

c) What value would you use to estimate the mean/average price among all 1200 square foot houses in Arroyo Grande? Explain.

d) Do you believe that either of these (point) estimates is likely to be exactly right?
e) If you were to use an interval to predict the price of an specific, individual 1200 square foot house and also use an interval to estimate the mean/average price among all 1200 square foot houses, which would you expect to be wider? Explain. [Hint: which varies more – averages or individual values?]

We can determine interval estimates (confidence intervals) for both of these quantities:

- **Confidence interval for the mean value of y at a particular value of x (denoted by $x_p$):**
  \[
  \hat{y} \pm t^* s_{\bar{y}} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}
  \]

- **Prediction interval for an individual value of y at a particular value of x (denoted by $x_p$):**
  \[
  \hat{y} \pm t^* s_y \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}}
  \]

f) Identify how these formulas are similar and how they differ. Comment on how these two intervals will compare, based on that difference.

g) Use Minitab to produce both of these 95% confidence interval estimates (Stat > Regression > Regression > Predict, select “Options” and enter 1200 for the prediction interval). Record both intervals. Also interpret what each interval means.

h) Which interval is wider? Is this consistent with your previous answers?

i) How do the midpoints of the intervals compare? Explain why this makes sense.

j) Based on the formulas, at what value of the explanatory variable will these intervals be most narrow? Explain.
• A prediction interval for an individual value is always wider than the corresponding confidence interval for a mean value.
• Both types of intervals are narrowest at the sample mean of the explanatory variable (x) values.

k) For what size house will a prediction interval, or a confidence interval for the mean price, be most narrow?

l) Use Minitab to produce confidence and prediction bands for all square foot values within the data (Stat> Regression> Fitted line plot, select “Options,” and check off to display the confidence and prediction intervals.) Are these bands consistent with your earlier findings? Explain.