STAT 312 – Handout 7
Least Squares Regression

We have begun to study relationships between two quantitative variables.

- **Graphical display:**
  - Scatterplot
- **Association:**
  - Direction: Positive, negative, none
  - Strength: Weak, moderate, strong
  - Form: Linear, non-linear
- **Correlation coefficient:**
  - \(-1 \leq r \leq 1\)
  - Sign indicates direction; magnitude indicates strength
  - Measures only linear association; not resistant to outliers

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 7-1: 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 least squares 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 least squares 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:

$$100\% \times \frac{\text{SS}_\text{resid}(\bar{y}) - \text{SS}_\text{resid(least-squares)}}{\text{SS}_\text{resid}(\bar{y})} =$$
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.

**Example 7-2: Airfares**

We will predict the cheapest airfare to a destination based on the distance to that destination. The data consist of distances (in miles, from Baltimore) to various cities and the cheapest airfare (as reported by the Sunday newspaper) to those cities (airfare.mtw).

a) Which is the explanatory and which is the response variable?

b) Use Minitab to produce a scatterplot, with the response variable on the vertical axis. Comment on the form, direction, and strength of the relationship as revealed by the scatterplot. Also make a guess for the value of the correlation coefficient.

c) Use Minitab to superimpose the least squares regression line on the scatterplot (Stat> Regression> Fitted Line Plot). Report the equation of this line. Comment on whether it seems to summarize the relationship in the data well.

d) Use the regression line to predict the airfare to a destination that is 750 miles away.

e) Use the regression line to predict the airfare to a destination that is 5000 miles away. Explain why it is not advisable to use the line for this prediction.

- **Extrapolation** refers to making predictions far beyond the scope of the data. Extrapolation is generally ill-advised.

f) By how much does the line predict the airfare to increase for each additional mile of travel? For each additional 100 miles?
g) What percentage of the variation in airfares is explained by the least squares line with distance?

One important issue that we have yet to consider the question of how to calculate the slope and intercept coefficients of the least squares line. 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 \textit{slope} coefficient can be calculated as: 
  \[ b_1 = r \frac{s_y}{s_x}. \]
- The value of the \textit{intercept} coefficient can be calculated as: 
  \[ b_0 = \bar{y} - b_1 \bar{x}. \]

\textbf{Example 7-3: Car data (cont.)}

Reconsider the car data, and consider predicting a car’s highway MPG rating from its weight.

\[ \text{a) Examine and comment on a scatterplot of these data (cars99.mtw). Remember to put the response variable on the vertical axis.} \]

\[ \text{b) Use Minitab to calculate the following descriptive statistics:} \]

<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>

\[ \text{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.]} \]

\[ \text{d) Use Minitab to confirm these calculations and to superimpose the regression line on the scatterplot (Stat> Regression> Fitted Line Plot).} \]

\[ \text{e) Interpret the value of the slope coefficient of this line.} \]

\[ \text{f) What highway MPG rating would the least squares line predict for a car weighing 3600 pounds?} \]

\[ \text{g) What proportion of the variability in highway MPG ratings is explained by the least squares line with weight?} \]