

Stat 218 - Day 33 Least Squares Regression

Recall that we have been studying relationships between two *quantitative* variables.

- A **scatterplot** is a graphical display of the relationship between two quantitative variables.
- We examine a scatterplot for evidence of **association** between the variables.
 - We look for the **form** (linear, other), **direction** (positive, negative), and **strength** (strong, moderate, weak) of an association.
- The **correlation coefficient** provides a measure of the linear association between the two variables.

Today we will study using a straight line to summarize the relationship between two quantitative variables.

Example: 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 “Java applets” and then on “Least squares regression” from our course web page.

(a) 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.

(b) Did the students across from you obtain the same equation? Did everyone in the class?

(c) Does your line provide a better fit than other students’? Suggest a criterion for deciding which line “best” summarizes the relationship.

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 (SAE) under your equation. Record this SAE value for your line. What is the best SAE in the class?

(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. What is the best SSE in the class?

(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. Now what is the best SSE in the class?

(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. Did everyone obtain the same equation? How does it compare to your line? (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.

(j) 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.

(k) Uncheck the “Your line” box to remove it from the display. Click on one student’s point in the scatterplot and drag the point up and down (changing the height, without changing the foot length). Does the regression line change much as you change this student’s height?

(l) Repeat the previous question, using a student with a low x (foot size) value and then a point with an x value near the middle and then a point with a large x value. Which of these seem(s) to have the most *influence* on the least squares regression line? Explain.

- An observation or set of observations is considered *influential* if removing the observation from the data set substantially changes the values of the correlation coefficient and/or the least squares regression equation. Typically, observations that have extreme explanatory variable outcomes (far below or far above \bar{x}) are potentially influential.

(m) “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.

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

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

- 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 , 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.

(o) Verify that your answer to (n) equals the square of the correlation coefficient.

Example: 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.
- (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.
- (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?
- The **residual standard deviation**, calculated as: $s_{y|x} = \sqrt{\frac{SS(resid)}{n-2}}$, is a measure of how far above or below the regression line the data points tend to fall.
- (h) Use Minitab to determine `SS(resid)` by selecting (`Stat> Regression> Regression`). Look under `SS(residual error)` in the resulting ANOVA table. Then use the above expression to calculate the residual standard deviation for the airfare data.