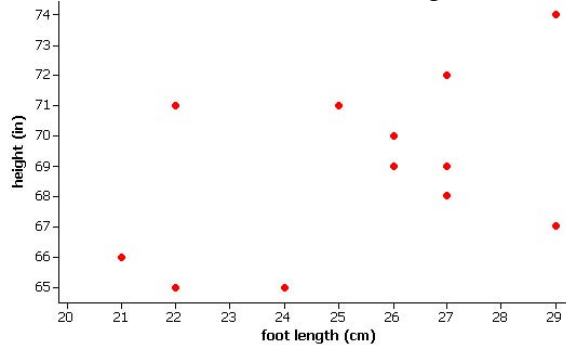


## Stat 150 – Day 26 Least Squares Regression

### Example 1: Predicting Height from Foot Length

Criminal investigators sometimes predict someone's height from his/her footprint. The data collected from 14 individuals in a class reveals the following relationship



- Identify the explanatory variable (also called a *predictor* variable) and the response variable.
- Sketch a line on the scatterplot that appears to summarize the overall relationship as well as a line can.
- Suppose that we find a 28cm footprint. What height would your line predict for this person?
- Do you believe you have predicted this individual's height exactly? Explain.

A **residual** is the vertical distance between the observed value of the response and the predicted value of the response for an observed value of the explanatory variable (residual = observed – predicted).

- An individual in the class reported a 27cm foot and height 68 inches. What is the residual of your prediction for this individual?
- On the graph, mark the points with positive residual values. Also, mark the negative residual values.
- Did everyone in your class make the same prediction in (c)?

h) Does your line provide a better fit than other students' lines? Suggest a criterion (rule) for deciding which line "best" summarizes the relationship.

Open the "Least Squares Regression" applet (follow the link to "Java applets" from Dr. Rossman's webpage). You will see a scatterplot of the 14 students' height and foot measurements. Click the "Your line" box, which adds a blue line to the scatterplot.

i) Record the equation of this blue line, and comment on the predictions made by this line. Also make a conjecture for how this line was selected.

j) Note three things about how statisticians like to record the equation of a line, which are probably different from what you've seen in most math classes:

k) Change the line in the applet by clicking on the "Move line" button. Place your mouse over one of the ends and drag to change the *slope* of the line. Move the green dot up and down vertically to change the *intercept* of the line. Move the line around until you believe that you have found the line that "best" summarizes the relationship between height and foot length for these data. Record the equation for this line, using good statistical notation.

l) Click the “Show residuals” box to visually represent the residuals for your line on the scatterplot. The applet reports the sum of the absolute residuals (called SAE, for sum of absolute errors) under your equation. Record this SAE value for your line. What is the best SAE in the class? (Does “best” correspond to the smallest or the largest value of SAE?)

m) A more common criterion for determining the “best” line is to instead minimize the sum of *squared* residuals (also called sum of squared errors, SSE). Click the “Show squared residuals” box to visually represent them and to determine SSE for your line. Record your SSE value. What is the best SSE in the class?

n) 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 smallest SSE value in the class?

o) Now check the “Regression line” box to determine and display the equation for the line that actually does minimize the sum of the squared residuals. Record its equation and SSE value. Now did everyone obtain the same equation? How does it compare to your line?

The **least squares regression line**  $\hat{y} = b_0 + b_1x$  is determined by finding the values of the intercept  $b_0$  and slope  $b_1$  that minimize the sum of the squared residuals.

p) Use the least-squares regression line to predict the height of a person with a 28 cm foot length. Then repeat for a person with a 29 cm foot length. Calculate the difference in these two height predictions. Does this value look familiar? Explain.

q) Provide an interpretation of the slope coefficient ( $b_1$ ) in this context. Be sure to include a probabilistic component to your interpretation.

r) Provide an interpretation of the intercept coefficient ( $b_0$ ). Does this interpretation make sense in this context? Explain.

s) Suppose that the person with a 28cm foot and 60 inch height had incorrectly reported his or her height to be 48 inches. Would changing just this one value affect the least squares line substantially? Click on this point in the graph and hold the mouse button down and drag the point down to a lower height. Is there an appreciable change in the least-squares regression line?

t) Click the Reset button to return the point to its original position and add the regression line again. Now change the height for the individual with a 22cm foot length from 63 inches to 47 inches. Does the regression line change as much as it did with the previous person? Suggest an explanation for the difference in behavior of the regression line in these two situations.

An observation is considered **influential** if removing the observation from the dataset substantially changes the least squares regression equation. Typically, observations that have extreme explanatory variable values (far below or far above the sample mean  $\bar{x}$ ) have the potential to be influential.

u) Now 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 height ( $y$ ) values. Click the “Show squared residuals” box to determine the SSE if we were to use  $\bar{y}$  as our predicted value for every value of foot length ( $x$ ), as if we knew nothing about the foot lengths. Record this SSE value.

v) Recall the SSE value for the least squares regression line. Determine the *percentage change* in the SSE between the  $\bar{y}$  line and the least squares line. *Hint*: Calculate  $100\% \times [\text{SSE}(\bar{y}) - \text{SSE}(\text{least-squares})] / \text{SSE}(\bar{y})$ .

This percentage change in SSE 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**, interpreted as the percentage of the variability in the response variable that is explained by the least squares regression line with the explanatory variable. The coefficient of determination is equal to the square of the correlation coefficient, so it is often denoted by  $r^2$  or  $R^2$ .

### Example 2: House Prices

Open the Minitab worksheet `HousePrices150.mtw`, available from the “Datasets” link on our course webpage. This file contains data on prices and sizes (in square feet) for a random sample of houses that sold in the year 2006 in Arroyo Grande, California

a) Produce a scatterplot of the data with the least squares regression line superimposed (`Stat > Regression > Fitted Line Plot...`). Record the equation of the line, using good statistical notation. Does the line appear to summarize the relationship between house price and size fairly well?

b) Use this least squares regression line to predict the price of the house at 845 Pearl Drive, which has a size of 1242 square feet. (When you finish the calculation, look at the scatterplot and line above to make sure that your prediction is reasonable.)

c) The actual (observed) price for the house at 845 Pearl Drive was \$459,000. Was your prediction in (b) too high or too low? By how much? Calculate the residual value for this house.

d) Based on the scatterplot (with the least squares regression line drawn on it) alone, which house has the largest negative residual?

e) Write a sentence interpreting the value of the *slope* coefficient in this context. Again be sure to include a probabilistic component to your interpretation.

f) How much does the line predict the price of a house to increase for each additional 100 square feet of size?

g) Write a sentence interpreting the value of the *intercept* coefficient. Does this make sense in this context? Explain.

h) What proportion of the variability in house prices is explained by the least squares line with size? *Hint*: This value appears in the Minitab graph..

i) Use the line to predict the price of a 1500 square foot house in Arroyo Grande in 2006.

j) Would it be reasonable to use this least squares line to predict the price of a 3500 square foot house? Explain.

k) Would it be reasonable to use this least squares line to predict the price of a 1500 square foot house in Canton, New York (a small town in the far northern part of New York state)? Explain.

l) Would it be reasonable to use this least squares line to predict the price of a 1500 square foot house in Arroyo Grande in the year 2010? Explain.