Chapter 3.2 Day #3 R^2

How Well the Line Fits the Data: The Role of s and r^2 in Regression

Standard deviation for the residuals is the average distance we are off from the predicted value.

$$s = \sqrt{\frac{\sum \text{residuals}^2}{n-2}}$$

DEFINITION: Standard deviation of the residuals (s)

If we use a least-squares line to predict the values of a response variable y from an explanatory variable x, the **standard deviation of the residuals (s)** is given by

$$s = \sqrt{\frac{\sum \text{residuals}^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}$$

This value gives the approximate size of a "typical" prediction error (residual).

Sep 22-11:38 AM

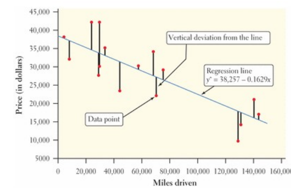
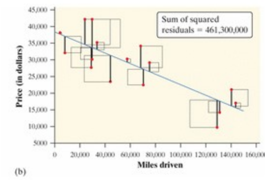


Figure 3.9 Scatterplot of the Ford F-150 data with a regression line added. A good regression line should make the prediction errors (shown as bold vertical segments) as small as possible.



$$s = \sqrt{\frac{461{,}300{,}000}{14}} = 5740 \text{ dollars}$$

If we add up the residuals, the positive and negative numbers cancel each other out. This is similar to finding the standard deviation, so we had to take the squares of each difference.

Sep 21-3:29 PM

If you don't know the mileage on the truck, what do you think we could use as the best predictor for the price?

$\bar{y}$ = mean price

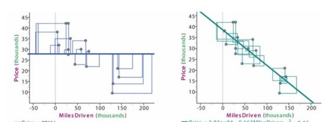


Figure 3.15 (a) The sum of squared residuals is 1,374,000,000 if we use the mean price as our prediction for all 16 trucks. (b) The sum of squares from the least-squares regression line is 461,300,000.

The ratio of these two quantities tells us what proportion of the total variation in y still remains after using the regression line to predict the values of the response variable. In this case,

$$\frac{461{,}300{,}000}{1{,}374{,}000{,}000} = 0.336$$

This means that 33.6% of the variation in price is *unaccounted for* by the least-squares regression line using x = miles driven. This unaccounted-for variation is likely due to other factors, including the age of the truck or its condition. Taking this one step further, the proportion of the total variation in y that *is accounted for* by the regression line is

$$1 - 0.336 = 0.664$$

* We interpret this by saying that "66.4% *of the variation in price is accounted for by the linear model relating price to miles driven.*"

Sep 22-11:40 AM

DEFINITION: The coefficient of determination: r2

The **coefficient of determination** r^2 is the fraction of the variation in the values of y that is accounted for by the least-squares regression line of y on x. We can calculate r^2 using the following formula:

$$r^2 = 1 - \frac{\sum \text{residuals}^2}{\sum (y_i - \bar{y})^2}$$

*** It is the correlation squared***

If given r^2 and asked to find r, you must determine if it's positive or negative.

AP® EXAM TIP Students often have a hard time interpreting the value of r^2 on AP® exam questions. They frequently leave out key words in the definition. Our advice: Treat this as a fill-in-the-blank exercise. Write "____% of the variation in [response variable name] is accounted for by the linear model relating [response variable name] to [explanatory variable name]."

Sep 22-11:55 AM

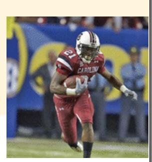
In **Section 3.1**, we looked at the relationship between the average number of points scored per game x and the number of wins y for the 12 college football teams in the Southeastern Conference. A scatterplot with the least-squares regression line and a residual plot are shown. The equation of the least-squares regression line is $\hat{y}$ = −3.75 + 0.437x. Also, s = 1.24 and r^2 = 0.88.

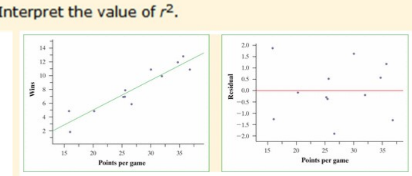
(a) Calculate and interpret the residual for South Carolina, which scored 30.1 points per game and had 11 wins.

(b) Is a linear model appropriate for these data? Explain.

(c) Interpret the value of s.

(d) Interpret the value of r^2.





SOLUTION:

Sep 22-12:00 PM

SOLUTION:

(a) The predicted amount of wins for South Carolina is

$$\hat{y} = -3.75 + 0.437(30.1) = 9.40 \text{ wins}$$

The residual for South Carolina is

$$\text{residual} = y - \hat{y} = 11 - 9.40 = 1.60 \text{ wins}$$

South Carolina won 1.60 more games than expected, based on the number of points they scored per game.

(b) Because there is no obvious pattern left over in the residual plot, the linear model is appropriate.

(c) When using the least-squares regression line with x = points per game to predict y = the number of wins, we will typically be off by about 1.24 wins.

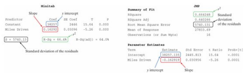
(d) About 88% of the variation in wins is accounted for by the linear model relating wins to points per game.

Sep 22-12:05 PM

3.2.6 Interpreting Computer Regression output

Printed Page 181

Figure 3.16 displays the basic regression output for the Ford F-150 data from two statistical software packages: Minitab and JMP. Other software produces very similar output. Each output records the slope and *y* intercept of the least-squares line. The software also provides information that we don't yet need (or understand!), although we will use much of it later. Be sure that you can locate the slope, the *y* intercept, and the values of *s* and r^2 on both computer outputs. *Once you understand the statistical ideas, you can read and work with almost any software output.*

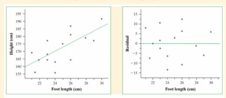


Example 14 Using Feet to Predict Height

Interpreting regression output

A random sample of 15 high school students was selected from the U.S. CensusAtSchool database. The foot length (in centimeters) and height (in centimeters) of each student in the sample were recorded. Least-squares regression was performed on the data. A scatterplot with the regression line added, a residual plot, and some computer output from the regression are shown below.

```
Predictor       Coef     SE Coef    T      P
Constant        103.41   19.50      5.30   0.000
Foot length     2.7469   0.7833     3.51   0.004
S = 7.95126     R-Sq = 48.6%    R-Sq(adj) = 44.7%
```



PROBLEM:

(a) What is the equation of the least-squares regression line that describes the relationship between foot length and height? Define any variables that you use.

(b) Interpret the slope of the regression line in context.

(c) Find the correlation.

(d) Is a line an appropriate model to use for these data? Explain how you know.

Sep 22-12:06 PM

SOLUTION:

(a) The equation is $\hat{y}$ = 103.41 + 2.7469 *x*, where $\hat{y}$ = predicted height (in centimeters) and *x* is foot length (in centimeters). We could also write

predicted height = 103.41 + 2.7469 (foot length)

(b) For each additional centimeter of foot length, the least-squares regression line predicts an increase of 2.7469 cm in height.

(c) To find the correlation, we take the square root of r^2:
$r = \pm\sqrt{0.486} = \pm 0.697$. Because the scatterplot shows a positive association, *r* = 0.697.

(d) Because the scatterplot shows a linear association and the residual plot has no obvious leftover patterns, a line is an appropriate model to use for these data.

Sep 22-12:09 PM