Least-Squares Line

by Sophia

WHAT'S COVERED

In this tutorial, you're going to learn about how to find a line of best fit using the method of least squares. Our discussion breaks down as follows:

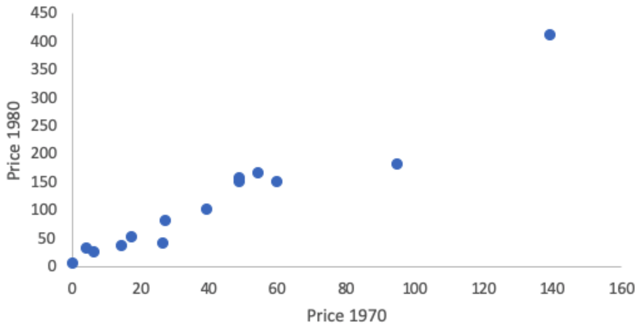
1. Least-Squares Line

When you look at data on a scatterplot, there are lots of lines that provide good fits for the data. You can usually eyeball them. In fact, there are many criteria for which you can create what's called a best-fit line.

The **least-squares lines** is one of the most common types of best-fit line and focuses on the residuals. Recall that residuals are the distance from the predicted values and the actual values on the scatterplot.

You can use Excel or other statistical software to create a least-squares line for a set of data. The least-squares line is calculated by minimizing the sum of the squares of the vertical differences from the line of best fit to each point. We will cover how to calculate this by hand in a later tutorial.

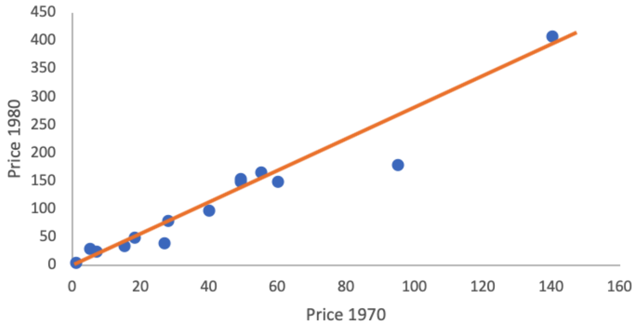
EXAMPLE This is the price of seafood, for different types of seafood. :Unsurprisingly, the most expensive ones in 1970 were still the most expensive in 1980. The trend is linear.



Draw a regression line and test to see how great of a fit it is to the rest of the data.

Good Fit:

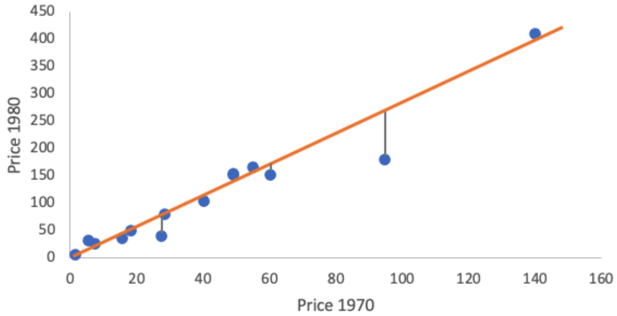
Suppose we draw the following regression line.



The equation of this line is the predicted price of 1980 equals three times the price of 1970's price. Note that there is a hat symbol of the Price 1980, which indicates what is being predicted.

$$\widehat{Price\ 1980} = 3(Price\ 1970)$$

If you take a look at that line, it seems fair. There are a couple of points that are noticeably lower than the line, but regardless of the line we could draw, we will end up with residuals. So what makes it a good fit?

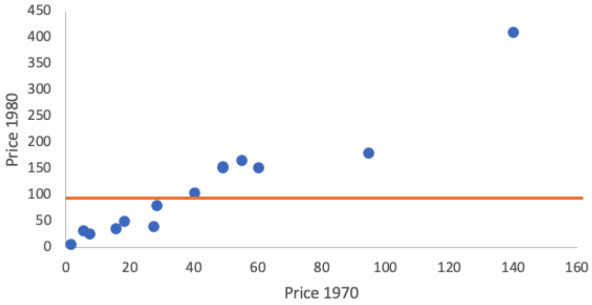


Every point has a residual. If it lies exactly on the line, its residual is zero. There are three points not on this line that have fairly large negative residuals. So, if you look at all the residuals put together, the sum of the residuals is negative 189.

You want the sum of the residuals to be low because that would mean that the points are close to the line. Let's create a line that we know for sure is a worse fit and compare the values of those residuals to the value of negative 189 to see if this first example is really a good fit.

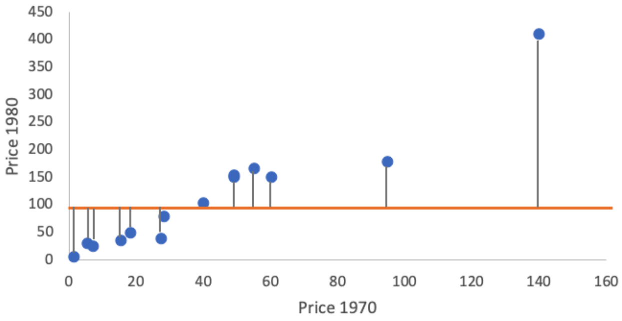
Bad Fit:

Suppose we now draw this regression line.



This regression line shows that all 1980 prices are going to be predicted to be 109.8. Regardless of it was really cheap or really expensive in 1970, just say that it will be predicted to be 109.8 cents per pound for everything in 1980.

Now, that's a bad idea. This is a poor fit for a line. You can see a lot of points are above and below the line. It's not a good fit; it doesn't go through the pack.

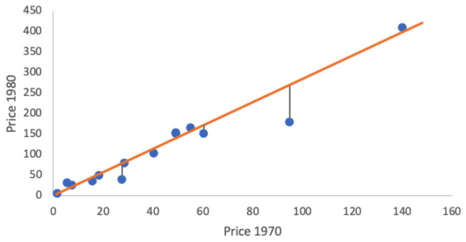
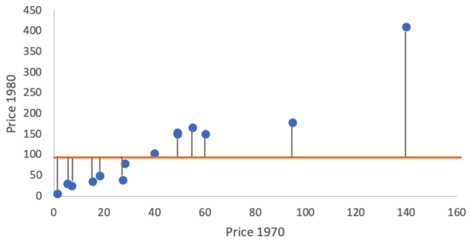


You can see visually there are some very large residuals here. But the problem is, when you add up all the residuals, it actually equals zero. What happened here?

Well, there are some large positive residuals that are canceled out by adding together several of these fairly large negative residuals. They end up canceling each other out so that even though the blue line is a poor fit for the data, the sum of the residuals is equal to zero. However, it’s agreed that this first model was, in fact, a better fit than the second model. So how can that be reconciled?

Instead of minimizing the sum of the residuals, you will use the method of least squares, which involves minimizing the sum of the *squares* of the residuals. What that means is the negative residuals, when you square them, become positive. The positive ones, when you square them, also become positive so that this negates the effect of having positive and negative residuals that might cancel each other out.

Now check which line is a better fit, using the method of least squares.

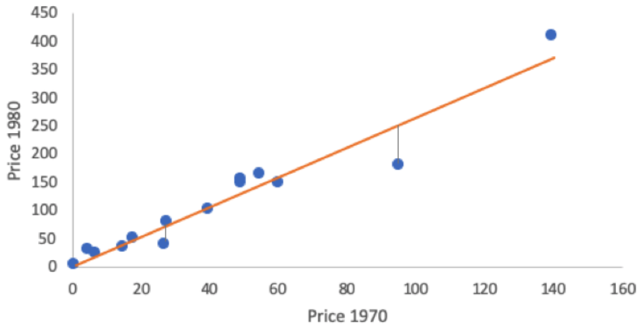
Type of Fit	Regression Line	Sum of Squares of Residuals
Good Fit		13,519
Bad Fit		143,838

Sure enough, 13,519 is a lot smaller than 143,838, indicating that the first line is a better fit for the data than the second line below.

Best Fit:

The first line is not even the best fit for the line. Actually, when calculated correctly, the best-fit line is the predicted 1980 price equal to 2.7 times the 1970 price, minus 1.2 cents per pound.

$$\widehat{Price\ 1980} = 2.7(Price\ 1970) - 1.2$$



In this case, with this line being the model, the sum of the squares of residuals is 9,326, which is even better than the 13,519. 9,326 is the smallest that the sum of squares can be, which makes this line the best-fit line.

TERM TO KNOW

Least-Squares Line

A best-fit line that is found through a process of minimizing the sum of the squared residuals.

SUMMARY

The method of least squares requires that you minimize the sum of the squares of the residuals. The line that does this is the best-fit line. It is called the least-squares line or the least-squares regression line.

Good luck!

Source: THIS TUTORIAL WAS AUTHORED BY JONATHAN OSTERS FOR SOPHIA LEARNING. PLEASE SEE OUR TERMS OF USE.

TERMS TO KNOW

Least-squares Line

The regression line where the sum of the squares of the residuals are the smallest.