

How well does the linear mathematical model fit the data?

This activity explores the goodness of fit of a linear regression model to data using an interactive Excel spreadsheet. The questions are grouped into categories that use one of the four different worksheets as labeled by the tabs at the bottom of the computer screen. For any of the worksheets, adjusting the scroll bar parameters will generate a new set of data, which will re-plot and the regression will respond accordingly.

Scatter in data - Scatter tab

A common measure of the goodness of fit is the r^2 value. Vary the scatter by moving the scroll bar or clicking on the arrows on the worksheet. How does r^2 vary with the scatter of the data about the regression line?

What is the value of r^2 for a perfect fit?

The use of r^2 as a measure of the goodness of fit is realistic because it describes the fraction (or if multiplied by 100, the percentage) of the y-variable that is explained by the variation of the x-variable. A perfect fit of a regression equation to the data generates an r^2 value of one.

For 10 data points, an $r^2 > 0.795$ is needed at the 95% confidence level to have statistical significance. Can you generate data with enough noise or scatter so it is not a significant relationship?

How would two data sets appear for a graph where the regression line was $y = 0.79x + 0.31$ with an $r^2 = 0.98$ compared to the same regression line with an $r^2 = 0.84$?

An outlier in the data at the extremes - Outlier tab

This worksheet has a datum point at the high end of the range that can be adjusted. Initially it is on the regression line as are all the other points. What is the value of r^2 for a perfect fit?

Can one datum point influence the results of a linear regression?

Move the last point to add scatter above the line. What happens to the slope and intercept? How about the value of r^2 ?

Now repeat this with the scatter below the line. What happens?

If you were using a regression line to predict the y-variable at an x-variable where $x = 12$, how would the scatter of the last point influence the prediction results for the following situations.

Situation	Prediction ($y = ?$)
no scatter	
scattered above the line	
scattered below the line	

Would an outlier at the low end of the range influence the regression line the same way as one at the high end of the range? Explain.

An outlier in the data at the mid-range - Outlier II tab

This worksheet has a datum point at the mid-range that can be adjusted. Click on the arrows of the scroll to add a small amount of scatter. How does the scatter influence the regression line?

Add considerable scatter to the mid-range datum point, what happens to the regression line now?

For a small amount of scatter, how does this influence the regression line if the outlier is at the extreme as compared to the mid-range?

When life throws you a curve - Scatter-Curvature-Residuals tab

This worksheet allows you to explore scatter as well as introduce curvature to the data. We also introduce a new way to examine the goodness of fit through the use of residuals. The residual is defined as the difference in the actual y-datum minus the y-value calculated from the regression equation.

Change the scatter by clicking on the arrows of the scroll and examine the value of r^2 and the behavior of the residuals plot. What did you observe?

Now set the scatter to low and introduce some curvature. If a small amount of curvature occurs, would the value of r^2 alert you to it?

How did the residuals plot change with the introduction of curvature?

Residuals should be randomly distributed for a linear relationship, if a pattern develops, a non-linear fit is suggested. Using only r^2 , you will not discover the presence of curvature.

Now, with a contribution of curvature, as seen in the residuals plot, increase the noise/scatter. What happens?

Noise or scatter in data can disguise the non-linear character of data. Careful collection of data is always required and making multiple measurements when possible as a check.