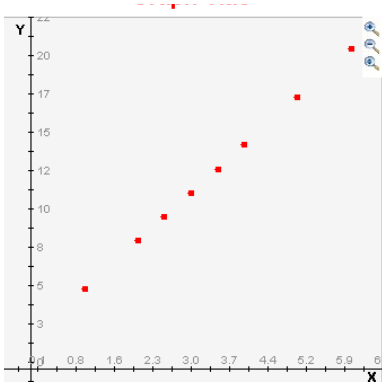
Notes on Regression Analysis

What is a function

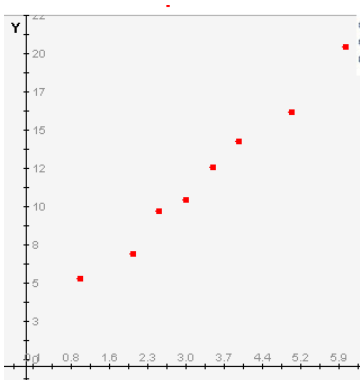
A function is a relationship between and independent variable, x, and a dependent variable, y. We have studied various forms of this relationship (linear, quadratic, polynomial, exponential, logarithmic…) and we have looked at relationships that were perfect. If we knew the value of x we could predict the value of y exactly.

Imperfect Function Relationships

In using a function to model a real world phenomena, the relationship is typically not perfect. Below you have 2 scatterplots of data. Plot A shows a perfect linear relationship between x and y and plot B shows an imperfect relationship between x and y, but it could still be modeled with a linear function.



Plot A



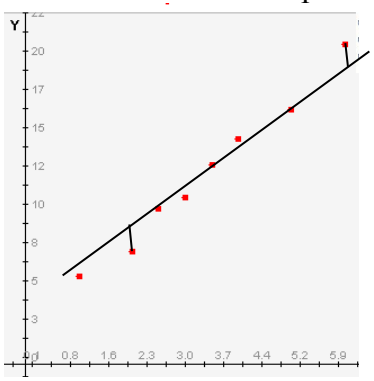
Plot B

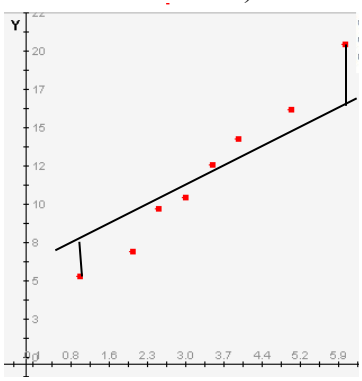
Finding the equation - Perfect Functions

The linear function for plot A is y=3x+2. The best line of fit through plot B is y=2.9x+2.1 For plot A, I got the equation of my line by calculating the slopes between each pair of points and seeing they were all the same and equal to 3. I then used a point slope formula with one of the data points and rearranged into slope intercept form.

Finding the Equation – Imperfect Functions

How did I decide which line was the "best" fit for the points in plot 2? Not all of the pairs of points have the same slope here. In this case I wanted a line that was as close as possible to all of the data points while remaining a straight line. To test the goodness of any line I measure the vertical distance of each point from the line, square this distance and then add up the squared distances to get a Sum of squared distances of the points from the line. (Two of the distances from the pictured line are illustrated below on left)

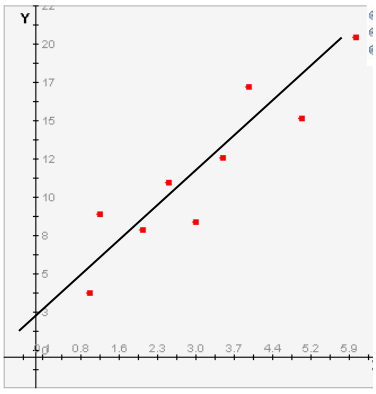


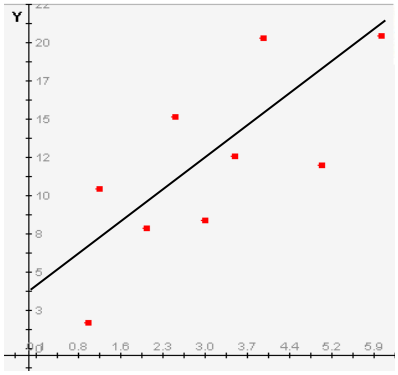


The plot on the right shows 2 distances from a different line. Clearly the sum of the distances from the right line would be greater than the sum of the distances from the line on the left. The line on the left is a better fit. Your calculator tries all possible lines and reports the one that fits best.

Goodness of Fit

So how good is the fit? Your calculator computes a value called R^2 that is a measure of how good the fit is. For the function in plot A if we asked the calculator to calculate an R^2 the value would be $R^2=1$, a perfect fit. For the best fit function to plot B. the $R^2$is .986, a pretty good fit of points to line, but not perfect. Below on the left is another scatter plot with the best fit line drawn in. It is easy to see that the points are more spread from the line than in plot B. The $R^2$for this line and these points is $R^2=.84$





For the best fit line of the points in the scatter plot on the right the R^2 is .547 The more spread out the points from the line, the lower the R^2.

Nonlinear Imperfect Functions

Suppose your scatter plot of data does not approximate a line, but a parabola, or an exponential function? There are advanced techniques for fitting a quadratic or exponential (or other type) curves. Your calculator requires that you specify the type of curve, and it will fit the best equation to your data based on the sum of squared distances of data points from curve as in the linear case above.