

Simple Regression Theory I

© 2006 Samuel L. Baker

Regression analysis lets you use data to explain and predict.

In Assignment 1, I will ask you to plot some data points on graph paper and draw a line through them to indicate their general trend. That action is called Simple Regression. The line that you draw is called a “regression line.”

“Simple” means that we are working in two dimensions, on a flat piece of paper. It doesn't mean that the theory here is simple.

Once you have your line drawn, I will ask you to derive an equation from the line you drew. Let us go over the theory you need for that.

A line and its equation, $Y=a+bX$

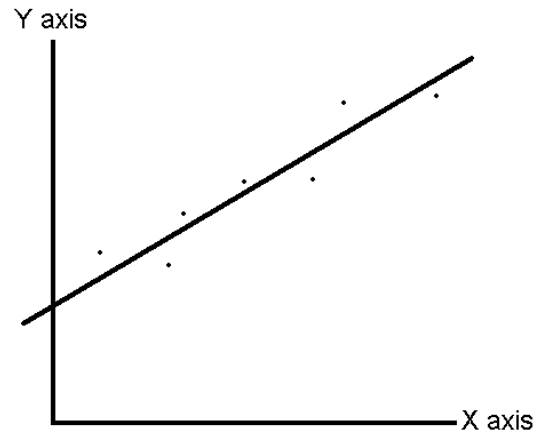
$Y=a+bX$ is the general form of an equation for a line, shown in the diagram to the right. In this diagram, each point on that line has a Y value that is calculated by multiplying the point's X value by **b**, and then adding **a**.

The line's **intercept** is the distance **a**, measured vertically, from the origin (the point where the X and Y axes meet) up to the point where the line crosses the Y axis.

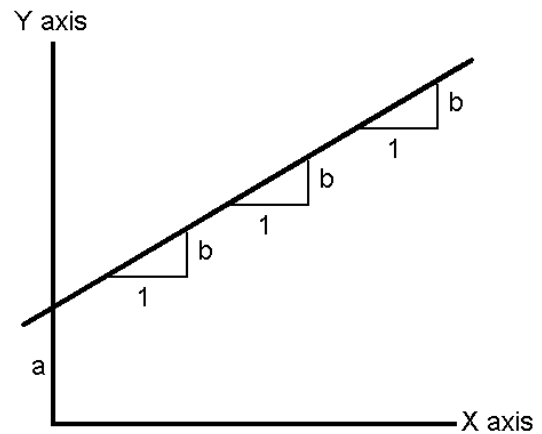
The **slope** of this line is **b**. The slope is how much the line rises for each unit of distance we move to the right. The line goes up by **b** for each 1 unit we move to the right.

It is the nature of a line that the amount it rises for each unit of horizontal distance is the same no matter where you measure along the line. If X and Y have a linear relationship, then the effect on Y of a change in X is the same at all levels of X.

A simple regression line drawn through data points



*The line $Y=a+bX$.
Its intercept is **a**. Its slope is **b**.*



Using a regression line for explanation

The regression line tells you your estimate of the effect on Y of a change in X. That estimated effect is **b**, the slope of the line. A change in X of 1 changes Y by **b**, on average. You can build this into an explanation for whatever phenomenon it is that Y represents.

Drawing a regression line does not prove that changes in X cause changes in Y. That is an idea that you have to bring to the analysis, based on your understanding of the situation that the data represent. If you have reason to believe that there is an effect, the regression line tells you how big that effect is.

While the regression line cannot prove that changes in X cause changes in Y, it can disprove it. If your regression line comes out horizontal, with a 0 slope, changes in X have no effect on Y.

Using a regression line for prediction

The regression line lets you calculate a predicted Y value that corresponds to any particular X value. To do this on a graph, pick the X value for which you want a corresponding predicted Y value. Start on the X axis, at that X value. Go straight up from your X value to the line. Then turn left and go straight left to the Y axis. This is your predicted Y value.

Algebraically, the prediction is calculated by substituting your chosen X value into the equation $Y=a+bX$.

Assumptions required to justify using a regression line to explain or predict

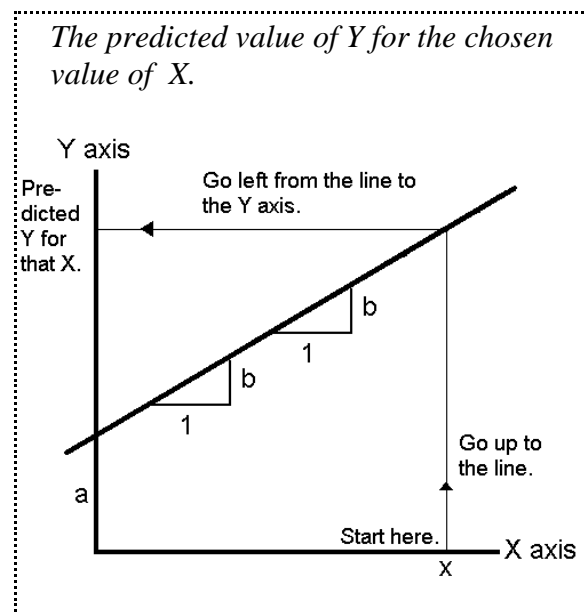
To use a regression line for these purposes, you have to make certain assumptions about the process by which the data were generated. To develop the theory of regression analysis, we have to make these assumptions explicit, so here they are:

Assumption 1: There is a true line, and the observed data points differ from that line due to random error.

The first assumption's first half is the idea that there is a true line -- an underlying linear relationship between our X variable and our Y variable. To predict using a regression line, we have to assume that the straight line relationship between X and Y existed in the past, when we got our data, and will continue to exist in the future.

This linear assumption means that when X changes, we are assuming that Y always tends to change by an amount that is a certain multiple of the change in X. Regardless of how big or small Y is, the effect of a change in X is the same.

If there is this true line, why don't our data points line up perfectly along it? This is where the second half of the assumption comes in. Our data points do not line up because there is random error in each

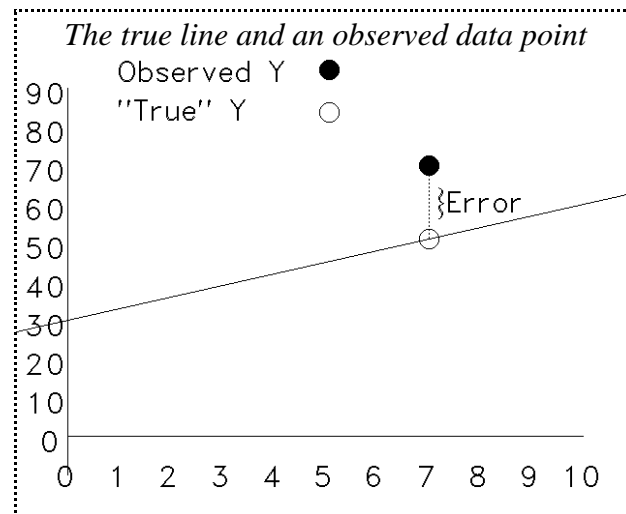


observation.

This figure shows one observation, and how we assume it relates to the true line. The vertical distance between the observed point and the true line is the “error.”

We assume that something that we cannot predict is causing the error. (If we could control or predict the error, we could make a better prediction that we would get from a simple linear regression. We will come back to this idea in assignment 2.)

For regression analysis, we need at least two data points. (With one point, we can't even draw a line!) The more data points we have, the better idea we should have of where the true line is.



Algebraically, we treat the points in our data as if they were numbered, from 1 up to however many points we have. The equation that generates each observation -- each (X, Y) data point -- can be written this way:

$$Y_i = \alpha + \beta X_i + e_i \quad \text{the "true equation"}$$

The subscript i is the number of the observation. For example, if your data set has 20 observations, i goes from 1 to 20. (X_1, Y_1) is the first observation. (X_{20}, Y_{20}) is the twentieth. X_i and Y_i are the X and Y values of the i th observation.

α is the intercept of the true line. β is the slope. I have switched to Greek letters, because most textbooks use them.

e_i is the random error of the i th observation. It is what is labeled “Error” in the diagram above. e_i is the vertical distance from the i th observed point to the corresponding point on the true line. If the i th observed point is below the true line, e_i is negative.

To summarize,

the true line has the equation $Y = \alpha + \beta X$,

and the i th data point has an X value of X_i and a Y value of $Y_i = \alpha + \beta X_i + e_i$.

The errors e_i are random. They are why the points do not lie exactly on the true line.

α and β are called the **parameters** of the true line. If we knew what those parameters were, we would know what the true line was, and we could make the best possible prediction of what Y will be if X takes on some new value. We would not expect that prediction to come true exactly, because any new Y value will also have some random error in it. Still, the prediction from the true line would be more likely to be close to what actually happens than any other prediction.

However, we do know not the values of α and β . All we have to go on is a bunch of data points. We can

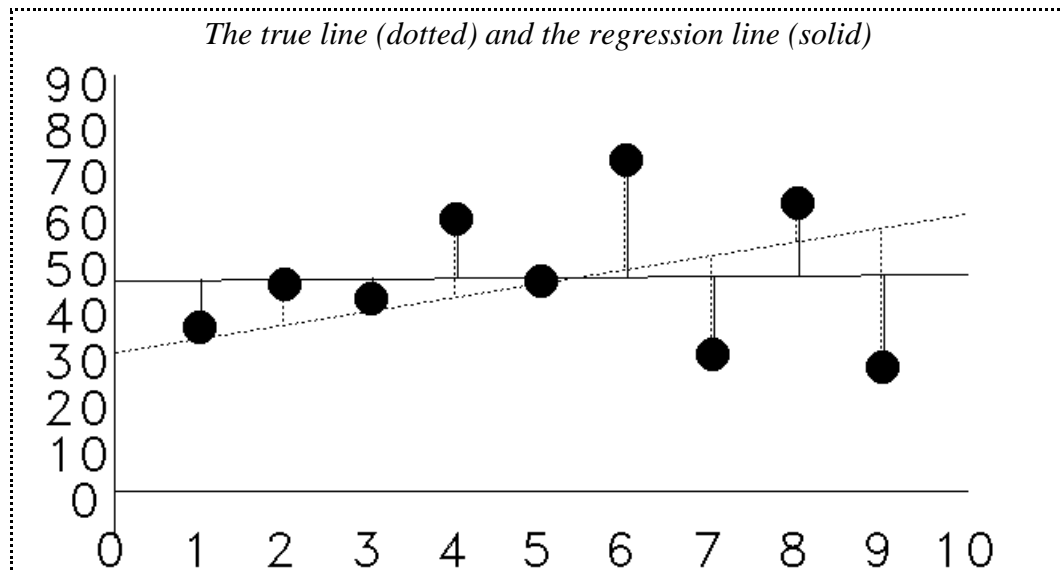
only *estimate* what the true parameters are, and then use those estimates of α and β to make our predictions.

The way we estimate the parameters α and β is to draw a **regression line**.

The regression line

Given a bunch of points on a graph, we can draw a line that seems to best represent the points' general trend. This line that we draw is called a **regression line**. The regression line and the true line are two different things.

In this figure, the true line that actually generated the points is shown dotted. It has an intercept just over 30 and tilts up. The dotted-line vertical distances from the points to the true line are the “errors.”



In real life, we would not know what the true line is. Only in cooked-up examples, like this one, do we know what the true line is.

In the figure, a regression line that we might draw is shown solid. It is the line we think best catches the trend of the points. In this example, the regression line has an intercept of about 50 and is close to horizontal, so the slope is close to 0. Solid lines in the diagram that run vertically from the points to the regression line represent the “residuals.”

In this example, the regression line and the true line are not very close to each other. The random errors are such that the general trend of the data points is more level than the true line. That happens sometimes.

Looking again at the above diagram, suppose we want to predict Y when $X=10$. Based on the regression line, our prediction for Y when $X=10$ is about 50. (Start at 10 on the X axis. Go up to the solid line. Go straight left. You should hit the 50 on the Y axis.) If we knew what the true line was, we would predict a higher Y value, about 60. (Start at 10 on the X axis. Go up to the dotted line. Go straight left. You should hit the Y axis a little above 60.)

In practice, we don't know what the true line is. We can only see the regression line we draw, so our prediction for Y is 50 when X is 10.

I need to emphasize the distinction between errors and residuals.

Errors are the vertical distances from the points to the true line.

Residuals are the vertical distances from the points to the regression line.

This distinction is crucial to understanding the theory using regression for prediction. It follows from the important idea is that the regression line and the true line are not the same.

In algebraic terms, we represent the distinction this way:

$Y = \alpha + \beta X$ is the true line.

$\hat{Y} = \hat{\alpha} + \hat{\beta}X$ is the regression line

The hats $\hat{}$ mean that these are numbers we calculate. $\hat{\alpha}$ (“alpha-hat”) is our estimate, or educated guess, of α , the true intercept. $\hat{\beta}$ (“beta-hat”) is our estimate of β .

\hat{Y} is the predicted value of Y . For every value of X , there is a corresponding value of \hat{Y} on the regression line.

$\hat{\alpha}$, the regression line's intercept, and $\hat{\beta}$, the regression line's slope, are not equal to the true line's intercept and slope α and β , unless, by extraordinary luck, the regression line and the true line happen to coincide.

To further distinguish the true line from the regression line that you draw or calculate, the $\hat{\alpha}$ and $\hat{\beta}$ numbers are called **estimated coefficients**, or just **coefficients** for short. We called the true α and β “parameters.”

The residuals are the differences between the observed Y values and the predicted \hat{Y} values.

For each observation, we can write:

$Y_i = \hat{Y}_i + u_i$ i is the observation number. It can be any number from 1 to N . N is the number of observations.

The u_i 's are the residuals. The u 's are your estimates of the e 's. To keep the theory of regression analysis straight, you must bear in mind that the u 's (the residuals) and the e 's (the errors) are not the same.

If we plug $\hat{\alpha} + \hat{\beta}X_i$ for \hat{Y}_i in the above equation, we get:

$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i + u_i$ the regression equation

The regression equation looks like the true equation, but with some differences. The regression's Greek letters have hats to indicate that they are estimates, and the vertical distances from the data points to the regression line are u 's (residuals), rather than e 's (errors).

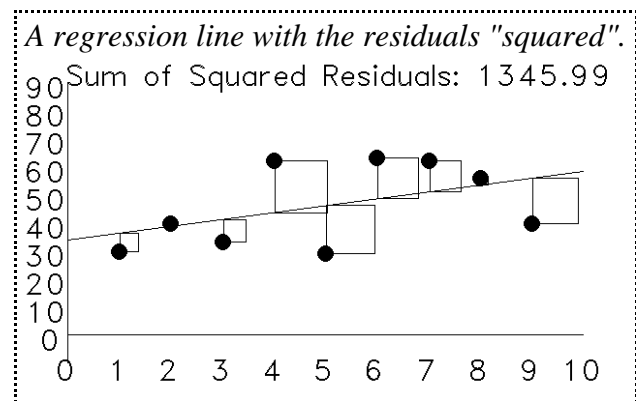
The distinction between errors and residuals can be hard to keep straight, partly because statisticians themselves sometime say “error” when they mean “residual.” For example, the upcoming Simple Regression Theory II chapter discusses what statisticians call the “standard error” of a regression. It's a

kind of an average size of the residuals, so it should be called “standard residual,” but it’s not.

The least squares regression line

The most popular method for drawing a regression line is “least squares.” This means finding the line that minimizes the sum of the squares of the residuals.

In this figure, I’ve drawn squares for each residual to symbolize this. Imagine moving that line up or down or changing its tilt. Each move would make the squares change their sizes. Some would get bigger and some would be smaller. The total area of the squares would change. Least squares regression finds the line that minimizes the total area of these squares.



Fortunately, you do not have to find the least squares line by trial and error, moving the line and then calculating how big the squares are. Instead, you can calculate the least squares line parameters from formulas. That’s one reason why least squares is a popular method.

Here is the formula for $\hat{\beta}$, the slope of the least squares regression line:

The slope of the least squares regression line

$$\hat{\beta} = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2}$$

\bar{Y} is the mean of the Y values. This symbol $\sum_{i=1}^N$ means evaluate the expression that follows it for each value of i from 1 up to N and then add them all up.

As for the intercept, $\hat{\alpha}$, it can be shown mathematically that the least squares line must go through the average of the data points (\bar{X}, \bar{Y}) . This means that:

$$\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X} \quad (\text{The } \bar{\ } \text{ symbol signifies the mean. } \bar{Y} \text{ is the mean of the Y values of all the points.})$$

Solving for $\hat{\alpha}$ gives the least squares estimate of α :

The intercept of the least squares regression

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$$

