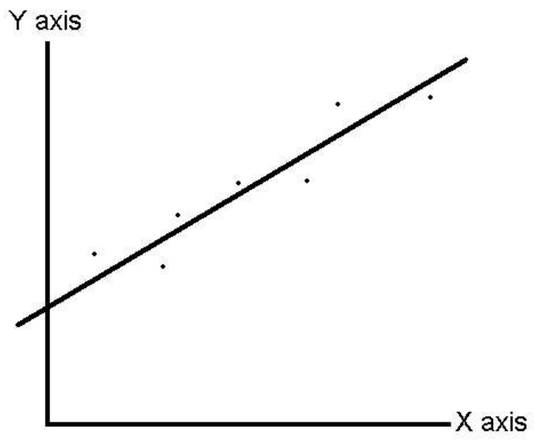
Simple Regression Theory I

Regression analysis lets you use data to explain and predict.

A simple regression line drawn through data points



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."

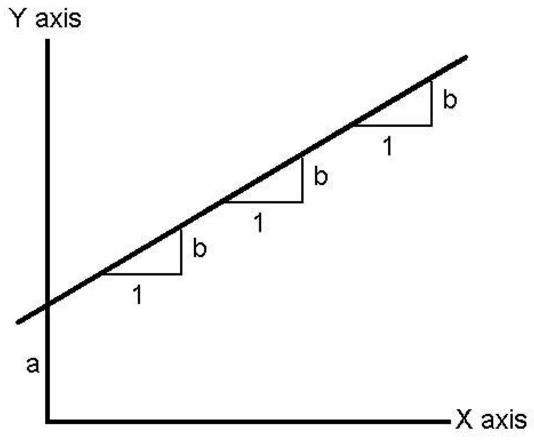
"Simple" means that we are working in two dimensions, on a flat piece of paper. It doesn't mean that the theory here is simple.

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



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, no matter how small or how big X is.

Once you have drawn a line on a graph, you can therefore derive its equation by doing this:

1. Measure the vertical distance from where the line hits the Y axis to where the X- and Y-axes cross. That is your **a**.

2. Pick another point on the line – the further to the right the better – and measure its X and Y values. Subtract **a** from the Y value and divide by X. That is **b**.

The Assignment 1 document gives an example with numbers.

Using a regression line for explanation

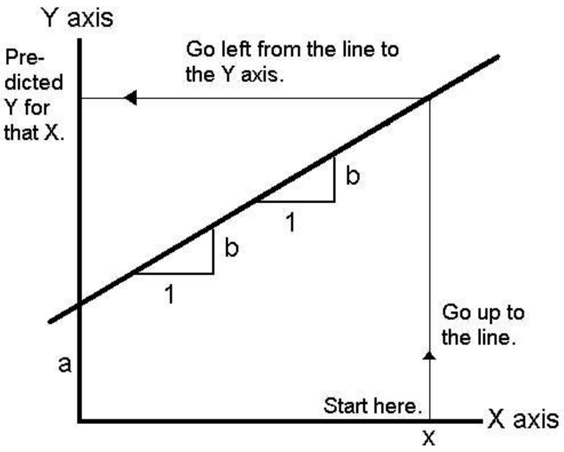
A regression line tells you your estimate of the effect on Y of a change in X. That estimated effect is **b**, the slope of the line. A change in X of 1 changes Y by **b**, on average. You can build this into an explanation for whatever phenomenon it is that Y represents.

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.

(Strictly speaking, if your simple regression line comes out horizontal, you have found that there is no linear relationship between X and Y. It is possible that there could be an uphill and downhill relationship, which would be a non-linear relationship. Later in the course, we will look at how to find non-linear relationships.)

The predicted value of Y for the chosen value of X.



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:

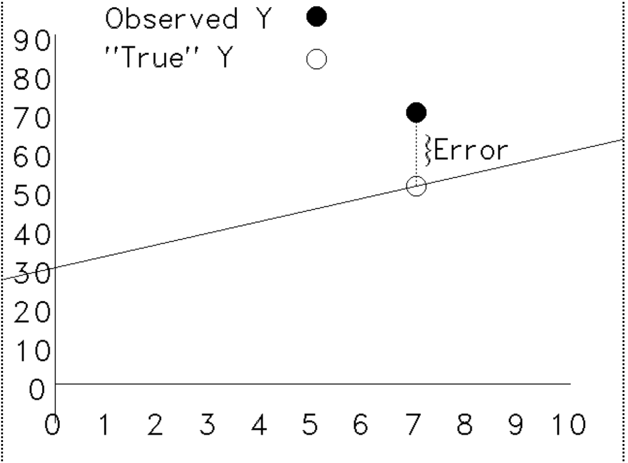
Assumption 1: There is a true line, and the observed data points differ from that line due to random error.

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.

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. When X is 7, Y should be about 50. We observe a Y of 70. The difference, 20, is random error. The vertical distance between the observed point and the true line is called the "error."

The true line and an observed data point



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.)

To draw a regression line, we need more than one point. Typically, we will have lots of points.

In algebra terms, we give each data point a number, from 1 up to however many points we have. The equation that generates each observation -- each (X,Y) data point -- can then be written this way:

$Y_i = \alpha + \beta X_i + e_i$ 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. We say that 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.

The true line has the equation $Y = \alpha + \beta X$. No e's in the equation for the true line.

α and β are called the **parameters** of the true line. They do not have *i* subscripts. They are the same for every point. That is how we put into our algebra the assumption that all the data points come from the same true line.

If we knew what the α and β 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.

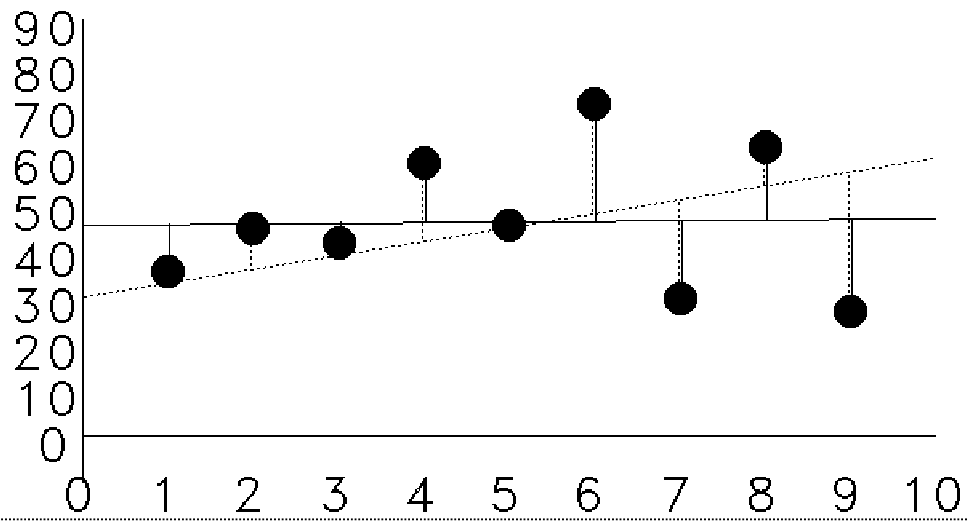
However, we do not know the true values of the parameters α and β. All we have to go on is a bunch of data points. We can only *estimate* what the true parameters are, and then use those estimates of α and β to make our predictions.

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

The true line (dotted) and the regression line (solid)



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 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,” the deviations of the points from the regression line.

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.

Paralleling the distinction between the true line and the regression line is the distinction between the errors and the 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. As said, 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 ^ 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."

Another way to think about the residuals: They 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 "Y-hats" lie along the regression line. The Y's don't.

The u_i's are the residuals. The u's are your estimates of the e's. Again, 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
$\hat{\alpha} + \hat{\beta}X_i$
in for
$\hat{Y}_i$
in the above equation, we get:

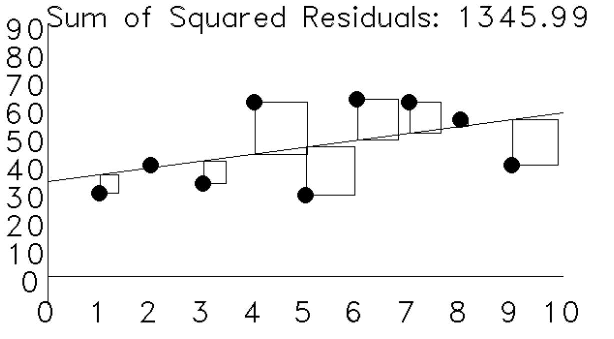
$Y_i = \hat{\alpha} + \hat{\beta}X_i + u_i$ the regression equation

The regression equation looks like the true equation, but with these differences:
1. The regression equation's Greek letters have hats to indicate that they are estimates, and
2. 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. They didn't ask me!

The least squares regression line

A regression line with the residuals "squared".



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 get 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 coefficients (alpha-hat and beta-hat) from formulas. That is one reason why least squares is a popular method.

Here is the formula for β, 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.

$\sum_{i=1}^{N}$ 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:

$\bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X}$ (The ¯ symbol signifies the mean. $\bar{Y}$ 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

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

Assignment 1 uses these formulas. First, you will draw a regression line by eye. Then, you will calculate and draw a least squares regression line. The two lines will probably be fairly close, but not the same. That does *not* mean that the least squares line is right and the eyeball line is wrong. Both are wrong, in the sense that neither coincides with the true line, unless you are very lucky. Later in the course, we will see evidence that the least squares line is probably closer to true than the eyeball line, because of the way that the assignment 1 data program generates your data. Even so, the best we can say is “probably closer.” You never know for sure.

Interpretation of simple regression results

Getting the formulas and computer work right means you are part way done. You must be able to state what the results mean.

The slope coefficient $\hat{\beta}$ means: A change in X of 1 makes Y change by $\hat{\beta}$.

A change in X of any amount ΔX makes Y change by $\hat{\beta}\Delta X$.

(ΔX means “change in X.”)

The intercept coefficient $\hat{\alpha}$ means: If X is 0 then Y is $\hat{\alpha}$.

The regression line crosses the Y-axis at $\hat{\alpha}$.

(Remember, “coefficient” means “estimated parameter.” A coefficient is an estimate of the corresponding true value.)

For example, suppose your X variable is the number of people who attend the state fair on any one day. Your Y variable is the number of elephant ears sold. (Elephant ears are a pastry. So far as I know, no elephants are harmed to make them.) You can theorize that there is a linear relationship between the number of elephant ears sold and the number of people at the fair, so Y = α + βX.

You collect data for several days on attendance and elephant ears, then draw or calculate a regression line. Suppose that the slope of that line, your estimated $\hat{\beta}$, which you can also call your X coefficient, is 3. Suppose that the intercept of that line, your estimated $\hat{\alpha}$, which you can call your estimated intercept, is 50.

Your equation is Y = 50 + 3X. If X goes up by 1, Y goes up by 3. This means that, on average, each extra person attending the fair increases the number of elephant ears eaten by 3. If 100 more people attend the fair, 300 more elephant ears are eaten.

The intercept of 50 means that if nobody attends, 50 elephant ears are eaten, presumably by people working at the fair.

When you make a prediction, you use the equation $\hat{Y} = \hat{\alpha} + \hat{\beta}X$.

To generate a specific prediction, plug the X value you want into that equation, and calculate $\hat{Y}$.

For example, if 10,000 people attend the fair, you predict that 30,050 elephant ears will be sold. 30,500 = 50 + 3*10,000.

You have now read enough theory to do Assignment 1.