8. Regression basics

Regression analysis, like most multivariate statistics, allows you to infer that there is a relationship between two or more variables. These relationships are seldom exact because there is variation caused by many variables, not just the variables being studied.

If you say that students who study more make better grades, you are really hypothesizing that there is a positive relationship between one variable, studying, and another variable, grades. You could then complete your inference and test your hypothesis by gathering a sample of (amount studied, grades) data from some students and use regression to see if the relationship in the sample is strong enough to safely infer that there is a relationship in the population. Notice that even if students who study more make better grades, the relationship in the population would not be perfect; the same amount of studying will not result in the same grades for every student (or for one student every time). Some students are taking harder courses, like chemistry or statistics, some are smarter, some will study effectively, some will get lucky and find that the professor has asked them exactly what they understood best. For each level of amount studied, there will be a distribution of grades. If there is a relationship between studying and grades, the location of that distribution of grades will change in an orderly manner as you move from lower to higher levels of studying.

Regression analysis is one of the most used and most powerful multivariate statistical techniques for it infers the existence and form of a functional relationship in a population. Once you learn how to use regression you will be able to estimate the parameters—the slope and intercept—of the function which links two or more variables. With that estimated function, you will be able to infer or forecast things like unit costs, interest rates, or sales over a wide range of conditions. Though the simplest regression techniques seem limited in their applications, statisticians have developed a number of variations on regression which greatly expand the usefulness of the technique. In this chapter, the basics will be discussed. In later chapters a few of the variations on, and problems with, regression will be covered. Once again, the t-distribution and F-distribution will be used to test hypotheses.

What is regression?

Before starting to learn about regression, go back to algebra and review what a function is. The definition of a function can be formal, like the one in my freshman calculus text: "A function is a set of ordered pairs of numbers (x,y) such that to each value of the first variable (x) there corresponds a unique value of the second variable (y)".[3] More intuitively, if there is a regular relationship between two variables, there is usually a function that describes the relationship. Functions are written in a number of forms. The most general is "y = f(x)", which simply says that the value of y depends on the value of x in some regular fashion, though the form of the relationship is not specified. The simplest functional form is the linear function where

$$y = \alpha + \beta x$$

[3] George B. Thomas, Calculus and Analytical Geometry, 3rd ed., Addison-Wesley, 1960.

α and β are parameters, remaining constant as x and y change. α is the intercept and β is the slope. If the values of and are known, you can find the y that goes with any x by putting the x into the equation and solving. There can be functions where one variable depends on the values of two or more other variables:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2$$

where x_1 and x_2 together determine the value of y. There can also be non-linear functions, where the value of the dependent variable ("y" in all of the examples we have used so far) depends on the values of one or more other variables, but the values of the other variables are squared, or taken to some other power or root or multiplied together, before the value of the dependent variable is determined. Regression allows you to estimate directly the parameters in linear functions only, though there are tricks which allow many non-linear functional forms to be estimated indirectly. Regression also allows you to test to see if there is a functional relationship between the variables, by testing the hypothesis that each of the slopes has a value of zero.

First, let us consider the simple case of a two variable function. You believe that y, the dependent variable, is a linear function of x, the independent variable—y depends on x. Collect a sample of (x, y) pairs, and plot them on a set of x, y axes. The basic idea behind regression is to find the equation of the straight line that "comes as close as possible to as many of the points as possible". The parameters of the line drawn through the sample are unbiased estimators of the parameters of the line that would "come as close as possible to as many of the point as possible" in the population, if the population had been gathered and plotted. In keeping with the convention of using Greek letters for population values and Roman letters for sample values, the line drawn through a population is

$$y = \alpha + \beta x$$

while the line drawn through a sample is

y = a + bx.

In most cases, even if the whole population had been gathered, the regression line would not go through every point. Most of the phenomena that business researchers deal with are not perfectly deterministic, so no function will perfectly predict or explain every observation.

Imagine that you wanted to study household use of laundry soap. You decide to estimate soap use as a function of family size. If you collected a large sample of (family size, soap use) pairs you would find that different families of the same size use different amounts of laundry soap—there is a distribution of soap use at each family size. When you use regression to estimate the parameters of soap use = f(family size), you are estimating the parameters of the line that connects the mean soap use at each family size. Because the best that can be expected is to predict the mean soap use for a certain size family, researchers often write their regression models with an extra term, the "error term", which notes that many of the members of the population of (family size, soap use) pairs will not have exactly the predicted soap use because many of the points do not lie directly on the regression line. The error term is usually denoted as "ε", or "epsilon", and you often see regression equations written

$$y = \alpha + \beta x + \varepsilon$$

Strictly, the distribution of ε at each family size must be normal, and the distributions of ε for all of the family sizes must have the same variance (this is known as homoskedasticity to statisticians).

It is common to use regression to estimate the form of a function which has more than one independent, or explanatory, variable. If household soap use depends on household income as well as family size, then soap use = f(family size, income), or

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2$$

where y is soap use, x_1 is family size and x_2 is income. This is the equation for a plane, the three-dimensional equivalent of a straight line. It is still a linear function because neither of the x's nor y is raised to a power nor taken to some root nor are the x's multiplied together. You can have even more independent variables, and as long as the function is linear, you can estimate the slope, β, for each independent variable.

Testing your regression: does y really depend upon x?

Understanding that there is a distribution of y (soap use) values at each x (family size) is the key for understanding how regression results from a sample can be used to test the hypothesis that there is (or is not) a relationship between x and y. When you hypothesize that y = f(x), you hypothesize that the slope of the line (β in $y=\alpha+\beta x+\epsilon$) is not equal to zero. If β was equal to zero, changes in x would not cause any change in y. Choosing a sample of families, and finding each family's size and soap use, gives you a sample of (x, y). Finding the equation of the line that best fits the sample will give you a sample intercept, α, and a sample slope, β. These sample statistics are unbiased estimators of the population intercept, α, and slope, β. If another sample of the same size is taken another sample equation could be generated. If many samples are taken, a sampling distribution of sample β's, the slopes of the sample lines, will be generated. Statisticians know that this sampling distribution of b's will be normal with a mean equal to β, the population slope. Because the standard deviation of this sampling distribution is seldom known, statisticians developed a method to estimate it from a single sample. With this estimated s_b , a t-statistic for each sample can be computed:

$$t = \frac{b - \beta}{estimated\ \sigma_b} = \frac{b - \beta}{s_b}$$

where n = sample size

m = number of explanatory (x) variables

b = sample slope

β= population slope

s_b = estimated standard deviation of b's, often called the "standard error".

These t's follow the t-distribution in the tables with n-m-1 df.

Computing s_b is tedious, and is almost always left to a computer, especially when there is more than one explanatory variable. The estimate is based on how much the sample points vary from the regression line. If the points in the sample are not very close to the sample regression line, it seems reasonable that the population points are also widely scattered around the population regression line and different samples could easily produce lines with quite varied slopes. Though there are other factors involved, in general when the points in the sample are farther from the regression line s_b is greater. Rather than learn how to compute s_b , it is more useful for you to learn how to find it on the regression results that you get from statistical software. It is often called the "standard error" and there is one for each independent variable. The printout in Exhibit 19 is typical.

Variable	DF	Parameter	Std Error	t-score
Intercept	1	27.01	4.07	6.64
TtB	1	-3.75	1.54	-2.43

Exhibit 19: Typical statistical package output for regression

You will need these standard errors in order to test to see if y depends upon x or not. You want to test to see if the slope of the line in the population, β, is equal to zero or not. If the slope equals zero, then changes in x do not result in any change in y. Formally, for each independent variable, you will have a test of the hypotheses:

$$H_o: \beta = 0$$

$$H_a: \beta \neq 0$$

if the t-score is large (either negative or positive), then the sample b is far from zero (the hypothesized β), and H_a : should be accepted. Substitute zero for b into the t-score equation, and if the t-score is small, b is close enough to zero to accept H_o :. To find out what t-value separates "close to zero" from "far from zero", choose an α, find the degrees of freedom, and use a t-table to find the critical value of t. Remember to halve α when conducting a two-tail test like this. The degrees of freedom equal n - m -1, where n is the size of the sample and m is the number of independent x variables. There is a separate hypothesis test for each independent variable. This means you test to see if y is a function of each x separately. You can also test to see if β> 0 (or β< 0) rather than simply if $\beta \neq 0$ by using a one-tail test, or test to see if his some particular value by substituting that value for β when computing the sample t-score.

Casper Gains has noticed that various stock market newsletters and services often recommend stocks by rating if this is a good time to buy that stock. Cap is cynical and thinks that by the time a newsletter is published with such a recommendation the smart investors will already have bought the stocks that are timely buys, driving the price up. To test to see if he is right or not, Cap collects a sample of the price-earnings ratio (P/E) and the "time to buy" rating (TtB) for 27 stocks. P/E measures the value of a stock relative to the profitability of the firm. Many investors search for stocks with P/E's that are lower than would be expected, so a high P/E probably means that the smart investors have discovered the stock. He decides to estimate the functional relationship between P/E and TtB using regression. Since a TtB of 1 means "excellent time to buy", and a TtB of 4 means "terrible time to buy", Cap expects that the slope, β, of the line $P/E = \alpha + \beta * TtB + \epsilon$ will be negative. Plotting out the data gives the graph in Error: Reference source not found.

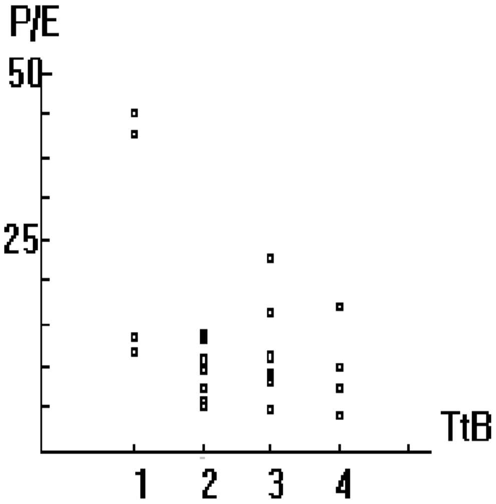

Exhibit 20: A plot of Cap's stock data

Entering the data into the computer, and using the SAS statistical software Cap has at work to estimate the function, yields the output given above.

Because Cap Gains wants to test to see if P/E is already high by the time a low TtB rating is published, he wants to test to see if the slope of the line, which is estimated by the parameter for TtB, is negative or not. His hypotheses are:

$$H_o: \beta \geq 0$$

$$H_a: \beta < 0$$

He should use a one-tail t-test, because the alternative is "less than zero", not simply "not equal to zero". Using an $\alpha = .05$, and noting that there are n-m-1, 26-1-1 = 24 degrees of freedom, Cap goes to the t-table and finds that he will accept H_a : if the t-score for the slope of the line with respect to TtB is smaller (more negative) than -1.711. Since the t-score from the computer output is -2.43, Cap should accept H_a : and conclude that by the time the TtB rating is published, the stock price has already been bid up, raising P/E. Buying stocks only on the basis of TtB is not an easy way to make money quickly in the stock market. Cap's cynicism seems to be well founded.

Both the laundry soap and Cap Gains's examples have an independent variable that is always a whole number. Usually, all of the variables are continuous, and to use the hypothesis test developed in this chapter all of the variables really should be continuous. The limit on the values of x in these examples is to make it easier for you to understand how regression works; these are not limits on using regression.

Testing your regression. Does this equation really help predict?

Returning to the laundry soap illustration, the easiest way to predict how much laundry soap a particular family (or any family, for that matter) uses would be to take a sample of families, find the mean soap use of that sample, and use that sample mean for your prediction, no matter what the family size. To test to see if the regression equation really helps, see how much of the error that would be made using the mean of all of the y's to predict is eliminated by using the regression equation to predict. By testing to see if the regression helps predict, you are testing to see if there is a functional relationship in the population.

Imagine that you have found the mean soap use for the families in a sample, and for each family you have made the simple prediction that soap use will be equal to the sample mean, $\bar{y}$. This is not a very sophisticated prediction technique, but remember that the sample mean is an unbiased estimator of population mean, so "on average" you will be right. For each family, you could compute your "error" by finding the difference between your prediction (the sample mean, $\bar{y}$) and the actual amount of soap used.

As an alternative way to predict soap use, you can have a computer find the intercept, α, and slope, β, of the sample regression line. Now, you can make another prediction of how much soap each family in the sample uses by computing:

$$\hat{y}=\alpha+\beta(familysize)$$

Once again, you can find the error made for each family by finding the difference between soap use predicted using the regression equation, ŷ, and actual soap use, $\bar{y}$. Finally, find how much using the regression improves your prediction by finding the difference between soap use predicted using the mean, $\bar{y}$, and soap use predicted using regression, ŷ. Notice that the measures of these differences could be positive or negative numbers, but that "error" or "improvement" implies a positive distance. There are probably a few families where the error from using the regression is greater than the error from using the mean, but generally the error using regression will be smaller.

If you use the sample mean to predict the amount of soap each family uses, your error is $(y-\bar{y})$ for each family. Squaring each error so that worries about signs are overcome, and then adding the squared errors together, gives you a measure of the total mistake you make if you use to predict y. Your total mistake is $\sum(y-\bar{y})^2$. The total mistake you make using the regression model would be $\sum(y-\hat{y})^2$. The difference between the mistakes, a raw measure of how much your prediction has improved, is $\sum(\hat{y}-\bar{y})^2$. To make this raw measure of the improvement meaningful, you need to compare it to one of the two measures of the total mistake. This means that there are two measures of "how good" your regression equation is. One compares the improvement to the mistakes still made with regression. The other compares the improvement to the mistakes that would be made if the mean was used to predict. The first is called an F-score because the sampling distribution of these measures follows the F-distribution seen in the “F-test and one-way anova” chapter. The second is called R^2 , or the "coefficient of determination".

All of these mistakes and improvements have names, and talking about them will be easier once you know those names. The total mistake made using the sample mean to predict, $\sum(y-\bar{y})^2$, is called the "sum of squares, total". The total mistake made using the regression, $\sum(y-\hat{y})^2$, is called the "sum of squares, residual" or the "sum of squares, error". The total improvement made by using regression, $\sum(\hat{y}-\bar{y})^2$ is called the "sum of squares, regression" or "sum of squares, model". You should be able to see that:

Sum of Squares Total = Sum of Squares Regression + Sum of Squares Residual

$$\sum(y-\bar{y})^2=\sum(\hat{y}-\bar{y})^2+\sum(y-\hat{y})^2$$

The F-score is the measure usually used in a hypothesis test to see if the regression made a significant improvement over using the mean. It is used because the sampling distribution of F-scores that it follows is printed in the tables at the back of most statistics books, so that it can be used for hypothesis testing. There is also a good set of F-tables at http://www.itl.nist.gov/div898/handbook/eda/section3/eda3673.htm. It works no matter how many explanatory variables are used. More formally if there was a population of multivariate observations, $(y, x_1, x_2, \ldots, x_m)$, and there was no linear relationship between y and the x's, so that y ≠ $f(x_1, x_2, \ldots, x_m)$, if samples of n observations are taken, a regression equation estimated for each sample, and a statistic, F, found for each sample regression, then those F's will be distributed like those in the F-table with (m, n-m-1) df. That F is:

$$F=\frac{\dfrac{\sum \mathit{of\,Squares\,Regression}}{m}}{\dfrac{\sum \mathit{of\,Squares\,Residual}}{(n-m-1)}}$$

$$=\frac{\dfrac{\mathit{improvement\ made}}{m}}{\dfrac{\mathit{mistakes\ still\ made}}{n-m-1}}$$

$$F=\frac{\dfrac{\sum(\hat{y}-\bar{y})^2}{m}}{\dfrac{\sum(y-\hat{y})^2}{(n-m-1)}}\ .$$

where: n is the size of the sample

m is the number of explanatory variables (how many x's there are in the regression equation).

If, $\sum(\hat{y}-\bar{y})^2$ the sum of squares regression (the improvement), is large relative to $\sum(y-\hat{y})^2$, the sum of squares residual (the mistakes still made), then the F-score will be large. In a population where there is no functional relationship between y and the x's, the regression line will have a slope of zero (it will be flat), and the ŷ will be close to y. As a result very few samples from such populations will have a large sum of squares regression and large F-scores. Because this F-score is distributed like the one in the F-tables, the tables can tell you whether the F-score a sample regression equation produces is large enough to be judged unlikely to occur if y ≠ $f(x_1, x_2, \ldots, x_m)$. The sum of squares regression is divided by the number of explanatory variables to account for the fact that it always decreases when more variables are added. You can also look at this as finding the improvement per explanatory variable. The sum of squares residual is divided by a number very close to the number of observations because it always increases if more observations are added. You can also look at this as the approximate mistake per observation.

$$H_o : y \neq f(x_1, x_2, \ldots x_m)$$

To test to see if a regression equation was worth estimating, test to see if there seems to be a functional relationship:

$$H_a : y = f(x_1, x_2, \ldots, x_m)$$

This might look like a two-tailed test since H_o : has an equal sign. But, by looking at the equation for the F-score you should be able to see that the data supports H_a : only if the F-score is large. This is because the data supports the existence of a functional relationship if sum of squares regression is large relative to the sum of squares residual. Since F-tables are usually one-tailed tables, choose an α, go to the F-tables for that α and (m, n-m-1) df, and find the table F. If the computed F is greater than the table F, then the computed F is unlikely to have occurred if H_o : is true, and you can safely decide that the data supports H_a :. There is a functional relationship in the population.

The other measure of how good your model is, the ratio of the improvement made using the regression to the mistakes made using the mean is called "R-square", usually written R^2. While R^2 is not used to test hypotheses, it has a more intuitive meaning than the F-score. R^2 is found by:

$$R^2 = \frac{\sum \text{ of Squares Regression}}{\sum \text{ of Squares Total}}$$

The numerator is the improvement regression makes over using the mean to predict, the denominator is the mistakes made using the mean, so R^2 simply shows what proportion of the mistakes made using the mean are eliminated by using regression.

Cap Gains, who in the example earlier in this chapter, was trying to see if there is a relationship between price-earnings ratio (P/E) and a "time to buy" rating (TtB), has decided to see if he can do a good job of predicting P/E by using a regression of TtB and profits as a percent of net worth (per cent profit) on P/E. He collects a sample of (P/E, TtB, per cent profit) for 25 firms, and using a computer, estimates the function

$$P/E = a + \beta_1 TtB + \beta_2 \, profit$$

He again uses the SAS program, and his computer printout gives him the results in Figure 3. This time he notices that there are two pages in the printout.

The SAS System

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	R Sq
Model	2	374.779	187.389	2.724	0.192
Error	23	1582.235	58.72		
Total	25	1957.015			

The SAS System

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t
Intercept	1	27.281	6.199	4.401
TtB	1	-3.772	1.627	-2.318
Profit	1	-0.012	0.279	-0.042

Exhibit 21: Cap's SAS computer printout

The equation the regression estimates is:

P/E = 27.281 - 3.772TtB – 0.012 Profit

Cap can now test three hypotheses. First, he can use the F-score to test to see if the regression model improves his ability to predict P/E. Second and third, he can use the t-scores to test to see if the slopes of TtB and Profit are different from zero.

To conduct the first test, Cap decides to choose an $\alpha = .10$. The F-score is the regression or model mean square over the residual or error mean square, so the df for the F-statistic are first the df for the model and second the df for the error. There are 2,23 df for the F-test. According to his F-table, with 2.23 degrees of freedom, the critical F-score for a = .10 is 2.55. His hypotheses are:

H_o: P/E ≠ f(Ttb,Profit)

H_a: P/E = f(Ttb, Profit)

Because the F-score from the regression, 2.724, is greater than the critical F-score, 2.55, Cap decides that the data supports H_a : and concludes that the model helps him predict P/E. There is a functional relationship in the population.

Cap can also test to see if P/E depends on TtB and Profit individually by using the t-scores for the parameter estimates. There are (n-m-1)=23 degrees of freedom. There are two sets of hypotheses, one set for β_1, the slope for TtB, and one set for β_2, the slope for Profit. He expects that β_1, the slope for TtB, will be negative, but he does not have any reason to expect that β2 will be either negative or positive. Therefore, Cap will use a one-tail test on β_1, and a two-tail test on β_2:

$$H_o: \beta_1 \geq 0 \qquad H_o: \beta_2 = 0$$

$$H_a: \beta_1 < 0 \qquad H_a: \beta_2 = 0$$

Since he has one one-tail test and one two-tail test, the t-values he chooses from the t-table will be different for the two tests. Using $\alpha = .10$, Cap finds that his t-score for β_1 the one-tail test, will have to be more negative than -1.32 before the data supports P/E being negatively dependent on TtB. He also finds that his t-score for β_2, the two-tail test, will have to be outside ±1.71 to decide that P/E depends upon Profit. Looking back at his printout and checking the t-scores, Cap decides that Profit does not affect P/E, but that higher TtB ratings mean a lower P/E. Notice that the printout also gives a t-score for the intercept, so Cap could test to see if the intercept equals zero or not.

Though it is possible to do all of the computations with just a calculator, it is much easier, and more dependably accurate, to use a computer to find regression results. Many software packages are available, and most spreadsheet programs will find regression slopes. I left out the steps needed to calculate regression results without a computer on purpose, for you will never compute a regression without a computer (or a high end calculator) in all of your working years, and there is little most people can learn about how regression works from looking at the calculation method.

Correlation and covariance

The correlation between two variables is important in statistics, and it is commonly reported. What is correlation? The meaning of correlation can be discovered by looking closely at the word—it is almost co-relation, and that is what it means: how two variables are co-related. Correlation is also closely related to regression. The covariance between two variables is also important in statistics, but it is seldom reported. Its meaning can also be discovered by looking closely at the word—it is co-variance, how two variables vary together. Covariance plays a behind-the-scenes role in multivariate statistics. Though you will not see covariance reported very often, understanding it will help you understand multivariate statistics like understanding variance helps you understand univariate statistics.

There are two ways to look at correlation. The first flows directly from regression and the second from covariance. Since you just learned about regression, it makes sense to start with that approach.

Correlation is measured with a number between -1 and +1 called the correlation coefficient. The population correlation coefficient is usually written as the Greek "rho", ρ, and the sample correlation coefficient as r. If you have a linear regression equation with only one explanatory variable, the sign of the correlation coefficient shows whether the slope of the regression line is positive or negative, while the absolute value of the coefficient shows how close to the regression line the points lie. If ρ is +.95, then the regression line has a positive slope and the points in the population are very close to the regression line. If r is -.13 then the regression line has a negative slope and the points in the sample are scattered far from the regression line. If you square r, you will get R^2, which is higher if the points in the sample lie very close to the regression line so that the sum of squares regression is close to the sum of squares total.

The other approach to explaining correlation requires understanding covariance, how two variables vary together. Because covariance is a multivariate statistic it measures something about a sample or population of observations where each observation has two or more variables. Think of a population of (x,y) pairs. First find the mean of the x's and the mean of the y's, μ_x and μ_y . Then for each observation, find $(x-\mu_x)(y-\mu_y)$. If the x and the y in this observation are both far above their means, then this number will be large and positive. If both are far below their means, it will also be large and positive. If you found $\sum(x-\mu_x)(y-\mu_y)$, it would be large and positive if x and y move up and down together, so that large x's go with large y's, small x's go with small y's, and medium x's go with medium y's. However, if some of the large x's go with medium y's, etc. then the sum will be smaller, though probably still positive. A $\sum(x-\mu_x)(y-\mu_y)$ implies that x's above x are generally paired with y's above μ_y , and those x's below their mean are generally paired with y's below their mean. As you can see, the sum is a measure of how x and y vary together. The more often similar x's are paired with similar y's, the more x and y vary together and the larger the sum and the covariance.

The term for a single observation, $(x-\mu_x)(y-\mu_y)$, will be negative when the x and y are on opposite sides of their means. If large x's are usually paired with small y's, and vice-versa, most of the terms will be negative and the sum will be negative. If the largest x's are paired with the smallest y's and the smallest x's with the largest y's, then many of the $(x-\mu_x)(y-\mu_y)$ will be large and negative and so will the sum. A population with more members will have a larger sum simply because there are more terms to be added together, so you divide the sum by the number of observations to get the final measure, the covariance, or cov:

$$population\ cov = \frac{\Sigma(x-\mu_x)(y-\mu_y)}{(N)}$$

The maximum for the covariance is the product of the standard deviations of the x values and of the y values, $\sigma_x\sigma_y$. While proving that the maximum is exactly equal to the product of the standard deviations is complicated, you should be able to see that the more spread out the points are, the greater the covariance can be. By now you should understand that a larger standard deviation means that the points are more spread out, so you should understand that a larger σ_x or a larger σ_y will allow for a greater covariance.

Sample covariance is measured similarly, except the sum is divided by n-1 so that sample covariance is an unbiased estimator of population covariance:

$$sample\,cov = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)}$$

Correlation simply compares the covariance to the standard deviations of the two variables. Using the formula for population correlation:

$$\rho = \frac{cov}{\rho_x \rho_y} \quad \text{or} \quad \rho = \frac{\sum(x-\mu_x)(y-\mu_y)/N}{\sqrt{\sum(x-\mu_x)^2/N}\sqrt{\sum(y-\mu_y)^2/N}}$$

At its maximum, the absolute value of the covariance equals the product of the standard deviations, so at its maximum, the absolute value of r will be 1. Since the covariance can be negative or positive while standard deviations are always positive, r can be either negative or positive. Putting these two facts together, you can see that r will be between -1 and +1. The sign depends on the sign of the covariance and the absolute value depends on how close the covariance is to its maximum. The covariance rises as the relationship between x and y grows stronger, so a strong relationship between x and y will result in r having a value close to -1 or +1.

Covariance, correlation, and regression

Now it is time to think about how all of this fits together and to see how the two approaches to correlation are related. Start by assuming that you have a population of (x, y) which covers a wide range of y-values, but only a narrow range of x-values. This means that σ_y is large while σ_x is small. Assume that you graph the (x, y) points and find that they all lie in a narrow band stretched linearly from bottom left to top right, so that the largest y's are paired with the largest x's and the smallest y's with the smallest x's. This means both that the covariance is large and a good regression line that comes very close to almost all the points is easily drawn. The correlation coefficient will also be very high (close to +1). An example will show why all these happen together.

Imagine that the equation for the regression line is y=3+4x, μ_y = 31, and μ_x = 7, and the two points farthest to the top right, (10, 43) and (12, 51), lie exactly on the regression line. These two points together contribute $\sum(x-\mu_x)(y-\mu_y)$ =(10-7)(43-31)+(12-7)(51-31)= 136 to the numerator of the covariance. If we switched the x's and y's of these two points, moving them off the regression line, so that they became (10, 51) and (12, 43), μ_x, μ_y, σ_x, and σ_y would remain the same, but these points would only contribute (10-7)(51-31)+(12-7)(43-31)= 120 to the numerator. As you can see, covariance is at its greatest, given the distributions of the x's and y's, when the (x, y) points lie on a straight line. Given that correlation, r, equals 1 when the covariance is maximized, you can see that r=+1 when the points lie exactly on a straight line (with a positive slope). The closer the points lie to a straight line, the closer the covariance is to its maximum, and the greater the correlation.

As this example shows, the closer the points lie to a straight line, the higher the correlation. Regression finds the straight line that comes as close to the points as possible, so it should not be surprising that correlation and regression are related. One of the ways the "goodness of fit" of a regression line can be measured is by R^2. For the simple two-variable case, R^2 is simply the correlation coefficient, r, squared.

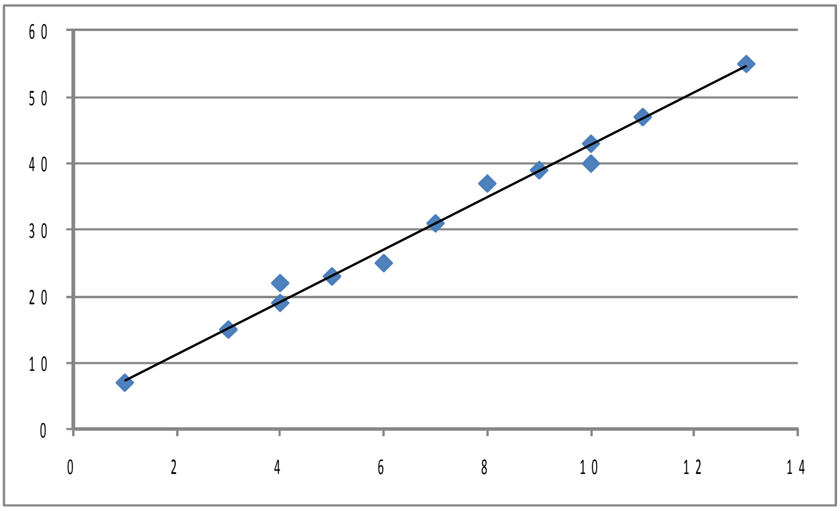

Exhibit 22: Plot of initial population

Correlation does not tell us anything about how steep or flat the regression line is, though it does tell us if the slope is positive or negative. If we took the initial population shown in Exhibit 20, and stretched it both left and right horizontally so that each point's x-value changed, but its y-value stayed the same, σ_x would grow while σ_y

stayed the same. If you pulled equally to the right and to the left, both μ_x and μ_y would stay the same. The covariance would certainly grow since the ($x- \mu_x$) that goes with each point would be larger absolutely while the ($y- \mu_y$)'s would stay the same. The equation of the regression line would change, with the slope, b, becoming smaller, but the correlation coefficient would be the same because the points would be just as close to the regression line as before. Once again, notice that correlation tells you how well the line fits the points, but it does not tell you anything about the slope other than if it is positive or negative. If the points are stretched out horizontally, the slope changes but correlation does not. Also notice that though the covariance increases, correlation does not because σ_x increases causing the denominator in the equation for finding r to increase as much as covariance, the numerator.

The regression line and covariance approaches to understanding correlation are obviously related. If the points in the population lie very close to the regression line, the covariance will be large in absolute value since the x's that are far from their mean will be paired with y's which are far from theirs. A positive regression slope means that x and y rise and fall together, which also means that the covariance will be positive. A negative regression slope means that x and y move in opposite directions, which means a negative covariance.

Summary

Simple linear regression allows researchers to estimate the parameters—the intercept and slopes—of linear equations connecting two or more variables. Knowing that a dependent variable is functionally related to one or more independent or explanatory variables, and having an estimate of the parameters of that function, greatly improves the ability of a researcher to predict the values the dependent variable will take under many conditions. Being able to estimate the effect that one independent variable has on the value of the dependent variable in isolation from changes in other independent variables can be a powerful aid in decision making and policy design. Being able to test the existence of individual effects of a number of independent variables helps decision makers, researchers, and policy makers identify what variables are most important. Regression is a very powerful statistical tool in many ways.

The idea behind regression is simple, it is simply the equation of the line that "comes as close as possible to as many of the points as possible". The mathematics of regression are not so simple, however. Instead of trying to learn the math, most researchers use computers to find regression equations, so this chapter stressed reading computer printouts rather than the mathematics of regression.

Two other topics, which are related to each other and to regression, correlation and covariance, were also covered.

Something as powerful as linear regression must have limitations and problems. In following chapters those limitations, and ways to overcome some of them, will be discussed. There is a whole subject, econometrics, which deals with identifying and overcoming the limitations and problems of regression.