

## Conduct, Advantages, and Disadvantages of Quantitative Analysis

In order to understand the advantages of quantitative analysis, several concepts need to be introduced. After these concepts are introduced, I will provide examples of statistical analyses one might conduct, to illustrate how to use the information one might obtain from a statistical analysis. Two points need be made at the outset. First, the emphasis here is on reading the results of a statistical analysis. At the end of this section you should be able to read at least some of the publications you find that use some statistical techniques. Second, this does not begin to cover the variety of things one can do with statistical analyses. But the basic principles are all in evidence in the examples I provide and, further, the example I provide is the modal statistical technique used in sociology.

### Populations and Parameters

The aim of quantitative analysis is to estimate population parameters. A **population parameter** is a characteristic of a population. A few examples will serve to make the point. Think of all the people who worked for pay outside the home in the United States in 2006. What is the average pay received for all of these workers? That value is a population parameter.

But population parameters can be more complex than the above example suggests. Think of all the black and white adults in the United States who were paid to work outside of their home in 2006. The difference between the mean income for blacks and the mean income for whites is a

population parameter--it tells us how much less (or more) blacks earn compared to whites.

But population parameters can be more complex than the above example suggests. Think again of all the black and white adults in the United States who were paid to work outside of their home in 2006. Now think of how much each of these people earns in work outside of the home, as well as how many years each of these persons spent in school. The relation between years of schooling and earnings is positive for both blacks and whites (that means the more years of school, the more money you make). It is also true that whites spend slightly more years in school than blacks. So, the white earnings advantage may be because whites spend more time in school. We can use statistical procedures to compare the average amount of earnings for a black with the average amount of earnings for a white, when both have spent the same number of years in school. This value is a population parameter.

A **parameter estimate** is what we typically discuss because it is very difficult to find out the value of the true population parameter. We obtain an estimate of a population parameter. One can use any procedure that one wants to use in order to estimate a population parameter. For example, I decide to use the following procedure to estimate the population parameter, mean earnings for 2006 workers in the United States:

Obtain a six-sided die and a ten-sided die.

Roll the six-sided die. Whatever number comes up, roll the ten-sided die that many times.

Each time the ten-sided die is rolled, record the number that comes up, moving from left to right.

When the ten-sided die has been rolled as many times as required, one has the "Dice Estimator Estimate" for mean earnings in the United States in 2006.

I just provided a procedure to obtain a parameter estimate, what I call the Dice Estimator. Procedures for obtaining a parameter estimate are called estimators. The burden of proof is upon me to argue that my procedure (my estimator) makes sense. I doubt it does. The point of this example is that an estimator is just a procedure, and you should know that there are good procedures and bad procedures.

Social researchers, statisticians, and more have worked hard to formally establish the basis for using particular procedures (such as probability sampling) to obtain parameter estimates. So distinguish the following:

**population parameter** -- a *characteristic* of a population

**estimator** -- a *procedure* for obtaining an estimate of a population parameter

**parameter estimate** -- an *estimate* of a population parameter derived from some estimator

Regression Analysis

For now, let's assume that we do have information about the earnings of all those who worked outside the home for pay in the United States in 2006, as well as other information, such as their race/ethnicity, years of schooling, and more. If so, we could use **Ordinary Least Squares Regression (OLS regression)** to figure out how much blacks earned compared to whites, once we account for the role of education in earnings. OLS regression, also known simply as regression, is the most commonly used statistical technique in sociology.

There are many requirements for the intelligent use of regression analysis. To learn those requirements you should take an Applied Regression course. Here, we only care that regression requires that the dependent variable is interval-level. Earnings are an interval variable. Regression coefficients (also called slope coefficients) are produced when one estimates, or "runs" a regression analysis. The regression coefficients are estimates of the association between the independent variable and the dependent variable, when all other independent variables are held constant. If I "ran" a simple regression model, I might obtain a result such as that listed below:

	Dependent variable-Income in dollars
	Estimate
Constant	3000
Years Education	1700

How do we interpret these results? To interpret the results one needs

to know the level of precision of the independent variables. If the independent variable is an ordinal or an interval level variable, then a one unit difference in the independent variable is associated with a difference in the dependent variable equal to the regression coefficient. In the example above, years of education is an interval-level variable, and thus the result above tells us that when you compare two people who are the same on everything else in the model, but differ by **one** year of schooling, the person with the additional year of schooling will be expected to earn \$1700 more than the other person. Because nothing else is in the model (except the constant, to which I will return below), this is a simple regression model.

Note what the result above does **not** tell us. It does not tell us that going back to school for a year will increase one's income by 1700 dollars. One reason we cannot make the latter inference is the latter inference is a causal argument, and regression coefficients measure association, not causation. It is very important to state the findings precisely, as stated above, and not to lapse into causal language. In this context, causal language is casual language.

If we add just one more independent variable to the model, we will transform the simple regression model into a multiple regression model. Here is where the value of regression becomes real. When one estimates a multiple regression model, then each regression coefficient reveals the association between the independent variable and the dependent variable,

*when everything else is controlled.* In other words, in multiple regression, the coefficient for one variable reveals the association between that variable and the dependent variable, after accounting for the association between the dependent variable and *all other independent variables in the model.*

	Dependent variable-Income in dollars
	Estimate
Constant	3000
Criminal Convictions	-1200
Years Education	1800

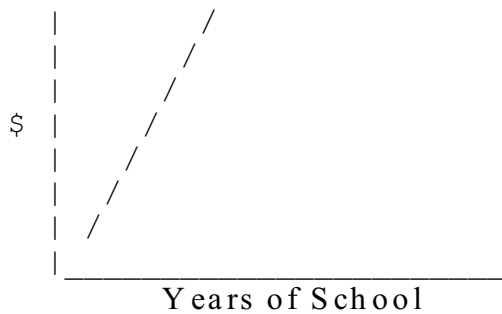
In this example we have added the number of crimes the person was convicted of to the model. Although few people have been convicted of more than a handful of crimes, we can assume here that criminal convictions are interval-like. We interpret the coefficient for years of education as before: when you compare two people who are the same on everything else in the model, but differ by **one** year of schooling, the person with the additional year of schooling will be expected to earn \$1800 more than the other person. This means that if you compare two people, both who have not been convicted of any crime, but differing by one year of schooling, the person with more schooling will be expected to be paid \$1800 more than the other. Also, if two people have been convicted of two crimes, but differ by one year in their total years of schooling, you would expect the person with more years of schooling to earn \$1800 more than the other person.

Similarly, if two people have the same years of schooling, but one

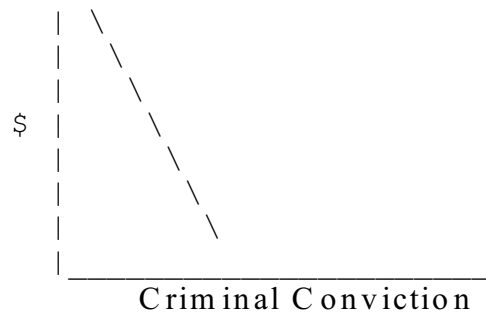
person has been convicted of one crime and the other person has not been convicted of a crime, we would expect the person with the criminal conviction to earn \$1200 less than the person who has never been convicted.

Note that when the higher the person's score on the independent variable, the higher the person's expected score on the dependent variable, the coefficient is positive. When the coefficient is negative, the higher the person's score on the independent variable, the *lower* the person's expected score on the dependent variable. It may pay to graph these differences:

Positive Coefficient



Negative Coefficient



Note that regression requires an interval-level dependent variable. However, it does not require interval-level independent variables. Nominal-level variables can be used, but their interpretation is different. The next example shows the difference:

Dependent variable-Income in dollars	
	Estimate
Constant	3000
Black	-2000
Asian	450
White Latino/a	50
Black Latino/a	-1450
Other non-white	-1150
Criminal Convictions	-1200
Years Education	1500

One could have a 6 category nominal-level variable to measure race/ethnicity. However, above you see that the 6-category nominal-level variable has been turned into 5 variables. Thus, above there are 5 variables to capture race/ethnicity: Black, Asian, White Latino/a, Black Latino/a, and Other non-white. Each of these variables is dichotomous, i.e., it has two values or categories.

Note that one racial group is not included in the table above: White.

This **omitted category** or **omitted group** is not being ignored. Instead, the omitted category is the category to which all other race/ethnicities are compared in the model. Before anyone goes off on a critique about racism in the above example because some group is omitted or, contrastingly, because some group has been made the standard, please note that this is an innocuous statistical manipulation. If we deleted the category for blacks, and put in the category for whites, we would end up making all comparisons to blacks. However, the reality would be the same; only the statistical numbers that reflect that reality would change.

Dependent variable-Income in dollars		
	Old Estimate	New Estimate
Constant	3000	3000
Black	-2000	-----
Asian	450	2450
White Latino/a	50	2050
Black Latino/a	-1450	550
Other non-white	-1150	850
White	-----	2000
Criminal Convictions	-1200	-1200
Years Education	1500	1500

In the first column the numbers for race show how much more or less different racial groups earn compared to whites. In the second column blacks are the comparison. But, as a check of the table will show, the order and the distances amongst the groups is unchanged.

In selecting an omitted category, people often select whites because they are the dominant group. But, results will not differ if a different group

is used (except, perhaps, slightly, due to rounding).

Using column 1 we infer that when you compare blacks and whites who have the same years of schooling, the black will receive 2000 dollars less in pay than the white person. You can construct similar statements comparing other groups with whites.

### The Rare Value of the Constant and the Calculation of Predicted Values

Now, as for the **constant**, which is sometimes called the **intercept**, it is of technical concern, and requires a statistical discussion to clarify. Typically, however, we are not at all concerned about the constant. As you will see in published research, it is usually ignored. However, if you want to calculate the predicted value for the dependent variable for a particular person, you need the constant.

We rarely calculate predicted values, but sometimes it can be helpful. The results of the model can be used to calculate predicted values. To do so, one needs to know a person's scores on each of the independent variables. So, let's say we know a black person, who has no criminal convictions, graduated from high school but did not go to college. Here is what that means in terms of their scores on the independent variables:

Constant	1
Black	1
Asian	0
White Latino/a	0

Black Latino/a	0
Other non-white	0
Criminal Convictions	0
Years of Education	12

Everyone has a score of 1 for the constant. Then, you fill in the values on the other independent variables. Note that a black person would have 1 in the category for black, and zero in all the other race/ethnicity categories. To calculate predicted values, you multiply the person's value on the variable by the coefficient for the variable. The "spreadsheet" below shows the calculations. Note that wherever the person has a zero on the variable, the coefficient does not matter. Here I have included those rows, to reveal what is happening in the calculation.

Variable	Value	Coefficient	Total
Constant	1	x \$ 3000	\$3000
Black	1	x - \$ 2000	-\$2000
Asian	0	x \$ 450	0
White Latino/a	0	x \$ 50	0
Black Latino/a	0	x - \$ 1450	0
Other non-white	0	x - \$ 1150	0
Criminal Convictions	0	x - \$ 1200	0
Years of Education	12	x \$ 1500	\$18000
Total	Not Applicable	Not Applicable	\$19000

So, based on the model, we would predict this person would be paid \$19,000. Please note that this prediction is only as good as the information used to make it. If key variables are not included in the model, then we should not expect a very accurate prediction.

Consider a white high school graduate with 1 criminal conviction. The table below shows their values on the variables and the necessary calculations. Note that for this person all of the race/ethnicity categories have 0s for their values. This is correct, because a white non-Latino/a is non-Black, non-Asian, non-Latino/a, and non-Other non-White. Thus, they should have 0s on all the race/ethnicity variables in the model:

Variable	Value	Coefficient	Total
Constant	1	x \$ 3000	\$3000
Black	0	x - \$ 2000	0
Asian	0	x \$ 450	0
White Latino/a	0	x \$ 50	0
Black Latino/a	0	x - \$ 1450	0
Other non-white	0	x - \$ 1150	0
Criminal Convictions	1	x - \$ 1200	-\$1200
Years of Education	12	x \$ 1500	\$18000
Total	Not Applicable	Not Applicable	\$19800

On the basis of this prediction we could infer (roughly) that a black high school graduate with no criminal convictions will be paid less than a white high school graduate with 1 criminal conviction. Sometimes such

comparisons are substantively illuminating, and so you should know how to construct predictions using the results of an OLS regression model.

### Accounting for Uncertainty

Above I posited that we had information on the full population of 2006 workers. However, analysts rarely have data on the population. Instead, they have data on a sample. Thus, in the typical analysis the results above would not be population parameters but, instead, estimates of population parameters. Because they are usually estimates (i.e., educated guesses) we are not certain our numbers are correct. In what follows, I lay out the logic whereby quantitative analysts account for uncertainty in their estimates. This is a major advantage of quantitative work.

If you repeatedly use an estimator you will obtain an estimate every time. But the estimates will not always agree. (Remember my graph of repeated samples from a population with 11.3% of the people living in poverty. The range of estimates of percent in poverty was from 4% to 19%, which means that many samples were **wrong**.) However, sampling theory tells us that, if the procedure is a good one (if the estimator is good), in the long run the estimates will cluster around the true population value. (Recall that as I drew more and more samples, the samples tended to cluster around 11% in poverty--still wrong, but pretty close).

As researchers we have neither the time nor the inclination to obtain repeated samples forever. We obtain one sample, and use estimators to

estimate population parameters. In actuality, we really do want to know the population parameter. However, we settle for an estimate of the population parameter. But, because we know the estimate is very likely to be wrong, we'd like to have some measure of how close our estimate is likely to be.

When researchers use appropriate techniques (good estimators) they obtain a measure of closeness. This figure of closeness is a standard error. Thus, good estimators provide two things: 1) a parameter estimate, and 2) an estimate of how likely the procedure is to produce an estimate that is near the true value of the population parameter. We can use the standard error to figure out how likely it is that the parameter value falls within some bound around the parameter estimate.

Substantively, sociologists often want to know whether something is associated with something else. For example, a sociologist might want to know whether being black is associated with earnings. That is, if two people are the same on all the relevant aspects, would their earnings be about the same even if one was black and one was white. Or, would their earnings likely differ, such that the black is likely to have lower or higher earnings than the white. If we obtained a population parameter rather than a parameter estimate, we could just look at whether the population parameter is zero, and we would have our answer. If the population parameter were zero, we would conclude that blacks and whites earn the same.

However, we do not have and usually cannot obtain the population

parameter. Instead, we have an estimate of the population parameter. Thus, we need to formally include our uncertainty in our appraisal of the estimate. In doing so we ask is it likely that one would obtain an estimate of the size we obtained, if the true population parameter were zero? Practically speaking, sociologists often use a particular level of certainty that means that if we have probability samples we may simply multiply the standard error by 1.96 and 1)add the result to the estimate, and 2)subtract the result from the estimate. If zero falls in the bounds, researchers say that the estimate is not statistically significant. What this means is that one cannot tell whether the true population parameter is zero. I call this the **multiplication trick**; if you have taken statistics you will recognize the procedure above as constructing a confidence interval. For our purposes the name is not important, the procedure is.

Consider the results from the previous regression model. This time I have included the standard errors.

Dependent variable-Income in dollars		
	Estimate	Standard Error
Constant	3000	500.73
Black	-2000	743.01
Asian	450	228.75
White Latino/a	50	481.03
Black Latino/a	-1450	193.48
Other non-white	-1150	918.37
Criminal Convictions	-1200	214.04
Years Education	1500	435.69

Let's do the multiplication trick for the parameter estimate for black.

The estimate is -2000, meaning that blacks earn, on average, 2000 less than whites with the same number of years of schooling. The standard error is 743.01. For ease of use I modify the multiplication trick slightly by using 2 rather than 1.96. If it's close I go back and use 1.96 exactly.

$$\text{Part 1) --> } -2000 + (2 \times 743.01) = -2000 + 1486.02 = -513.98$$

$$\text{Part 2) --> } 2000 - (2 \times 743.01) = -2000 - 1486.02 = -3486.02$$

So, it is quite likely that the true population parameter (difference between earnings for blacks and whites with the same amount of schooling) is somewhere between -\$3486.02 and -\$513.98. That is:

$$-\$3486.02 < \text{-----True Pop Parameter-----} > -\$513.98$$

Because zero is not in the interval above, we can be pretty confident that there is an association. And, because the interval does not include any positive numbers, we can be pretty sure that the association is negative. However, we are not completely sure about either statement. Recall that I once obtained a sample that showed only 4% of the people were in poverty just by chance, when the true population parameter was 11%. The same could happen here; I could have obtained an atypical sample just by chance!

Note the by-the-bootstraps logic we have employed:

- 1) We know, from statistical theory, that if you conduct the same analysis on several different samples using good estimators and compare the results, the results will cluster around the true population parameter.
- 2) We have one sample, and, because most samples are not atypical, we assume we did not obtain an atypical sample.
- 3) We use the multiplication trick on the results from our analysis of a single sample (construct the confidence interval) and look to see whether zero is inside the interval. If it is not, we infer that the population parameter is unlikely to be zero, i.e., there is an association.

That's the logic. And, much more often than not it will lead you to accurate inferences.

I suggest you do the multiplication trick for some of the other variables in the model, and discuss your results.

The multiplication trick, or the confidence interval construction, will work whenever analysts give you the standard errors and the parameter estimates. However, sometimes you may be given the t-ratios rather than standard errors. This is the multiplication trick in another guise. If the t-ratio is greater in absolute value than 1.96, then zero is not in the interval and so we may distinguish between zero and our parameter estimate.

### Concluding Remarks

For purposes of understanding social research, you need to be able to understand how to use the standard error to assess whether an association is

likely to exist or not (the multiplication trick). Also, for purposes of understanding social research you need to be able to evaluate whether it makes sense to run a regression. Regression is a technique of use when the dependent variable is interval-level or ratio-level. (There are variants of regression one may use for ordinal and nominal-level variables, (for example, probit models, logit models, multinomial logit models, etc.) These variants are beyond this course--but the same multiplication trick works for those models, too!).

Some advantages of quantitative analyses are as follows:

- 1) Allows one to formally include uncertainty in one's appraisal of the evidence
- 2) Can allow one to consider multiple factors simultaneously
- 3) Because of 2, allows one to tease out relationships that might be masked by other factors.
- 4) The systematic nature of statistics can ease efforts at replication and extension

Some disadvantages of quantitative analyses are as follows:

- 1) Can encourage researchers to act as if they can ignore whatever they have yet to learn how to measure.
- 2) Can encourage researchers to act as if their findings are useless unless they include everything that might be related to the phenomenon of interest.