Statistics Crash Course

Much of social science research is designed to test hypotheses about causation. Such hypotheses take the form of an assertion that is something occurs, then something else will happen as a result.

In a causal hypothesis, the phenomenon that is explained is called the **dependent variable**.

It is called a variable because we are conceiving of something that can "vary" across a set of cases. It is called "dependent" because of the assertion of causation: its value is hypothesized to be dependent on the value of some other variable.

The other variable in the hypothesis – the one that is expected to influence the dependent variable – is called the **independent** (or explanatory) **variable**.

The relationship between a dependent and an independent variable is usually determined by a mathematical technique knows as **Ordinary Least Squares Regression**. Many of the quantitative studies you will read this semester use OLS Regression of another technique which is closely related to it.

An independent variable may be either **dichotomous**, **ordinal**, or **interval**.

A dichotomous, or "**dummy**," variable is one that can take on only two possible values.

One example of this is gender, where individuals can be either male of female. Dummy variables are usually coded as "0" or "1" (e.g. male=0, female=1)

An interval variable is one for which any one-unit difference in numerical scores (e.g. that between 1 and 2, or 87 and 88) reflects the same difference in the amount of the property being measured. One example would be income, in which the difference between $1 and $2 is the same as the difference between $87 and $88.

For an ordinal variable, we cannot make the claim that any one-unit difference in numerical scores reflects the same difference in the amount of the property being measured. In short, ordinal variables tend to be measures of "fuzzier" concepts. Ordinal variables tend to be measures of things which are harder to quantify, such as attitudes about some kind of political issue.

A common example would be support for the war in Iraq. A surveyor might ask you,

"How do you feel about the war in Iraq? Do you 1) Strongly Oppose It 2) Oppose It 3) Neither Support or Oppose It 4) Support It or 5) Strongly Support It"

In this case, there is no reason to believe that a one-unit difference on the scale always reflects the same difference for the various values.

In other words, the difference between "Strongly Oppose" and "Oppose" is not necessarily the same in magnitude as "Oppose" and "Neither Support or Oppose."

A Regression Model ($Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \epsilon$) consists of at least three parts:

1) An **intercept** (or constant) (α)

The intercept of a regression equation can be interpreted as the expected value of Y for cases having a score of zero on X. In a graph, it is the value at which the regression line intersects with the vertical axis.

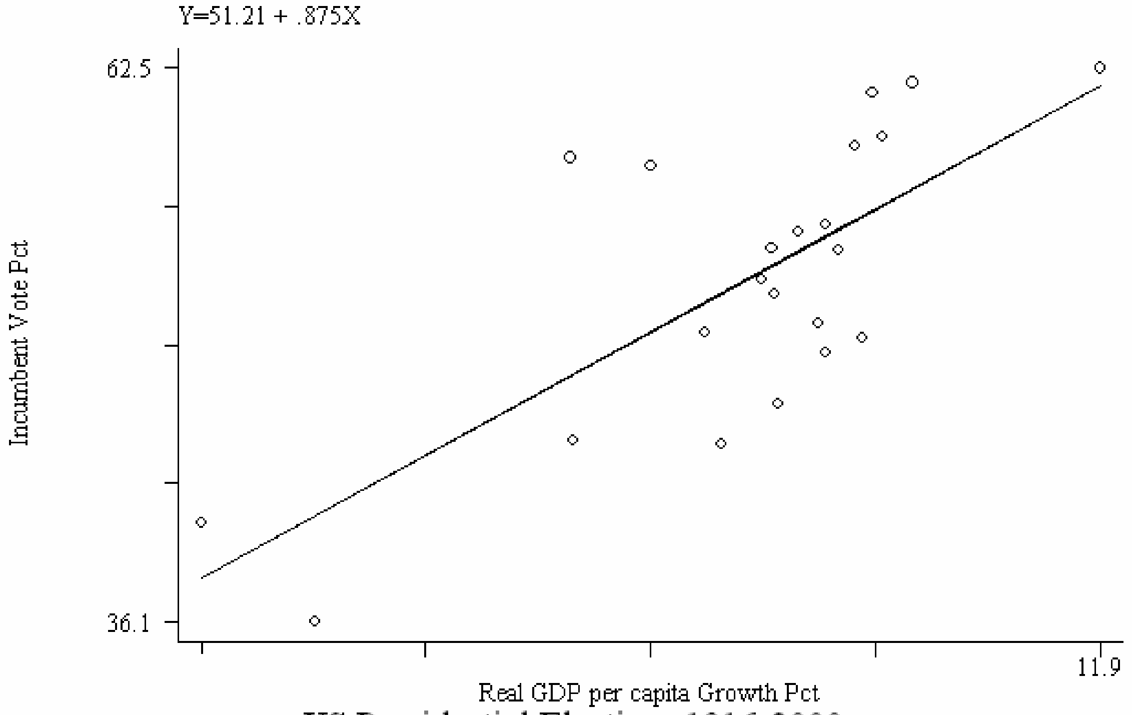
2) The Slope **Coefficient** (β)

Sometimes the intercept has a substantive meaning, sometimes it does not. However, the slope coefficient is ALWAYS relevant. In short, the slope coefficient can be viewed as a measure on the effect of X on Y.

It tells us the expected (or average) value of Y resulting from a one-unit increase in X.

3) The **Error** or Disturbance Term (ϵ)

The error or disturbance term represents the combined effect of all other variables not included in the regression model, plus any "inherent randomness" in the determination of the value of the dependent variable.



US Presidential Elections 1916-2000

Some illustrations of different Slopes….

Some notes about OLS Regression:

For the purposes of regression, the dependent variable must always be an interval variable! Different statistical techniques must be used for dichotomous or ordinal variables. I'll return to that at the end of class.

Regression models assume that the effect of the independent variable, X, on the dependent variable, Y, is **linear**. That is, the effect which a one-unit change in X has on Y is identical for all values of X (e.g. whether we are moving X from 1 to 2 or from 87 to 88).

Problems that can occur in regression models:

Sometimes the independent variables in a regression model are related to one another. This is called **multicollinearity**.

There will be some multicollinearity in virtually every regression model, but if the relationships are not strong, the OLS regression procedure for estimating slop coefficients can proceed without difficulty.

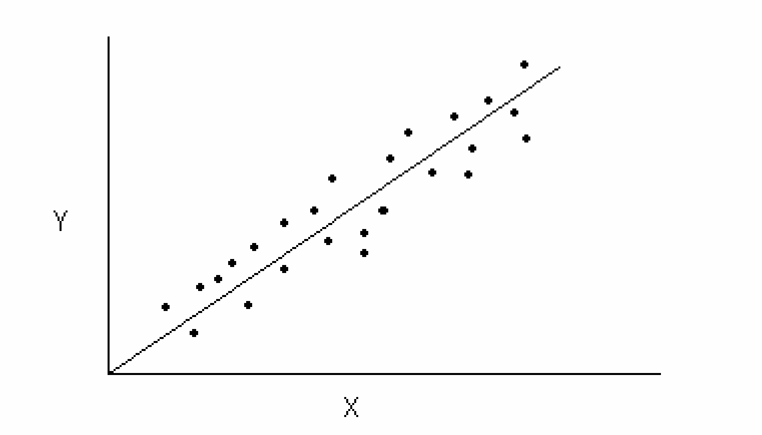
However, strong multicollinearity poses serious problems for regression analysis. If two independent variables are strongly related to one another (that is, they always vary together), OLS cannot tell which of the independent variables is causing the observed change in the dependent variable.

Measures of Fit

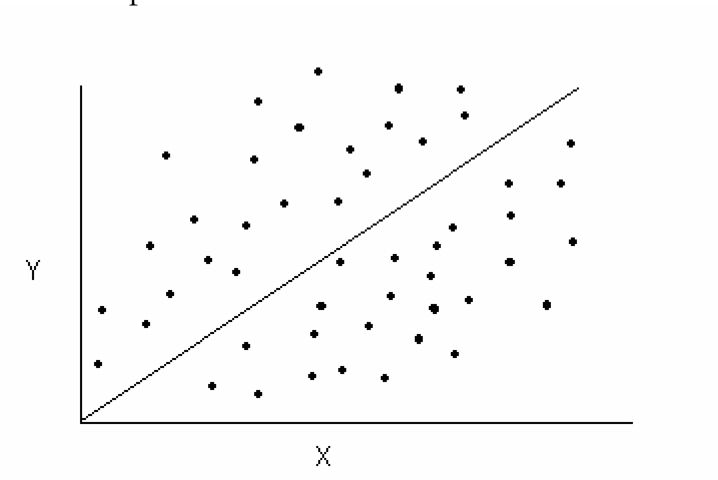
The most commonly reported measure of fit is the $\mathbf{R^2}$. R^2 is always between 0 and 1. A higher value means a better "fit," and a value of 1 indicates a perfect fit.

A common definition of an R^2 is the proportion of the variance in the dependent variable explained by the independent variables in the model. Lower values mean that the independent variables have little explanatory power, and higher values indicate greater explanatory power.

An example of a high R^2



An example of a low R^2



What would this graph look like if $R^2 = 1$?

!!!Statistical Significance!!!!

In any random sample from a population, the relationships between the independent and the dependent variables are not exactly the same relationships between them in the population. Indeed, some samples are unusual enough in their character that relationships appear in the sample even when there are no relationships in the underlying population…..This is bad.

The concept of statistical significance helps us to express a degree of confidence that a relationship we detect is more than a chance occurrence in the one sample that we are able to observe.

Usually, we want something to be statistically significant at the .05 level. By this we mean that if the true (population) value of β were zero (e.g. X had NO effect on Y), then a random sample from the population could be expected to yield an estimate of β at least as large as the observed estimate less frequently than t times out of 100.

It is standard practice for social scientists to present evidence concerning statistical significance of their coefficient estimates alongside the estimates themselves.

They may do so in a number of different ways. Researchers may report each slope coefficient's **standard error**.

Don't worry about what a standard error is or how it is calculated; just know that if a coefficient is twice as large as its standard error it is significant at the .05 level.

There is also something called a **t-statistic**. This is simply the coefficient divided by the standard error. If it is larger than 2 (1.96 to be specific) or less than -2, then the coefficient is usually considered "statistically significant."

Researcher often mart coefficients that are statistically significant with asterisks (*) to indicate that they are statistically significant.

Some Examples…..

What do we do our dependent variable is NOT an interval variable?

Sometime we want to study variables which are not interval. For example, we might want to know what causes someone to vote. Voting is a dichotomous variable. A person either votes or does not vote.

To study dichotomous dependent variables we use one of two techniques: **logit** or **probit**.

When we use logit or probit the coefficients we get do NOT tells us the expected value of Y resulting from a one-unit increase in X. Rather, they tell us something about the change in the PROBABILITY that Y will be 1.

Statistical significance, however, still has the same meaning.

For ordinal dependent variables we rely on a procedure called **ordered probit** or **ordered logit**.

Lastly, sometimes we hypothesize that the effect of X_1 on Y varies depending on the value of X_2. Thus, we might multiply X_1 and X_2 in order to test whether or not relationship between X_1 is contingent upon X_2. We call this an **interaction term**.