Political Science 100a/200a

Fall 2001

Regression part V, and review of course topics[1]

Some final points about OLS

- You'll see more of these in 100b/200b, but a very basic introduction is warranted at this point.
- *Measurement error*: Variables of interest in social science are very often hard to measure accurately. Sometimes the difficulty arises from practical constraints in obtaining the data (e.g., number killed in a civil war), and often from conceptual imprecision (e.g., democracy scales).
- What is the effect of having measurement error present on a simple bivariate OLS analysis?
- It turns out that the effects are quite different depending on whether the measurement error is in the dependent or independent variable.
- Suppose first that the dependent variable, Y, has random measurement error in it. That is, we observe values y_i^* that are produced as

  $$y_i^* = y_i + \delta_i$$

  where δ is the (unobserved) measurement error for case i, and y_i is the true value of the dependent variable for this case. Assume that the measurement error δ has mean zero and variance σ_δ^2.
- What will happen if we use OLS to try to estimate the model

  $$y_i = \alpha + \beta x_i + \epsilon_i \qquad (1)$$

  but where we are forced to use y_i^* in place of y_i (which we don't know)?
- In particular, we want to know how the measurement error affects our estimates a and b for α and β respectively. Are they still unbiased? Are their standard errors affected?
- It is not too difficult to see that the measurement error in the *dependent variable* will not bias our estimates for α and β. If we were to start with the true model,

[1] Notes by James D. Fearon, Dept. of Political Science, Stanford University, November 2001.

$$y_i = \alpha + \beta x_i + \epsilon_i$$

we can "derive" the model to be estimated by adding the measurement error noise to each side:

$$\begin{aligned} y_i + \delta_i &= \alpha + \beta x_i + \epsilon_i + \delta_i \\ y_i^* &= \alpha + \beta x_i + \epsilon_i^* \end{aligned}$$

where $\epsilon_i^* = \epsilon_i + \delta_i$.

- Since both ϵ and δ are random variables with means of zero, so is their sum. So this is really just like estimating α and β with more variance in the random "other causes." (It is not too difficult to show formally that b remains an unbiased estimate of β).
- The effect of random measurement error in the dependent variable is just to increase the "unexplained" (and unexplainable) part of the regression.
- So while the estimates a and b for α and β will remain *unbiased* with OLS, the estimate will become *less precise* – their standard errors will increase because the standard error of the regression (the r.m.s.e.) will increase due to the "noise" in Y. Clear?
- Construct e.g. in Stata: **gen grwerr = grw6080 + invnorm(uniform())**, then do **reg grwerr ethfrac** and compare to standard case ... do with bigger variance in measurement error.
- NOTE: The above presumes that the measurement errors in Y are *random* with respect to any independent variables. If this is not so, what do you think would happen?
- Next, what if there is measurement error in an *independent* variable? i.e., what if we observe $X^* = X + \delta$ instead of just X?
- Here, the problem introduced by measurement error is more severe. The easiest way to see it is to imagine the case where the measurement error in X^* is really *big*, so that it practically swamps any "signal" from X. What would happen, then, if we tried to estimate

$$y_i = \alpha + \beta x_i^* + \epsilon_i$$

- If the x_i^*'s are practically all just random noise (measurement error), what do think we will get?
- OLS will estimate a b close to zero, since from its "point of view" X^* is just a lot of random noise with respect to Y.

- Thus (important conclusion), *measurement error in an independent variable will tend to bias its estimated slope coefficient towards zero in OLS.*

- More generally, what this argument brings to light is another crucial assumption required for the earlier conclusion we reached about how OLS estimates for α and β are unbiased if certain assumptions hold: We need to add now the assumption that the independent variables are measured without error.

- In practice of course, almost nothing is measured without error. However note that the effect of measurement error on an independent variable is to tend to bias our estimate of its effect downwards, towards zero, reducing the likelihood that measurement error will cause to think a variable has an effect significantly different from zero when in fact it does not. (with multiple regression, matters are more complicated, as measurement error in one variable can bias the estimates of the effects of other independent variables in unpredictable directions.)

Some central ideas and course topics worth reviewing

- To "explain" something means to give reasons why something is the case *rather than* something else. Explanations necessarily explain *variation*, even if the variation is implicit ("counterfactual").

- Many if not all explanations in social science explain the occurence of one thing by reference to another thing that caused or helped "produce" it. Empirical methods in social science assist in evaluating and exploring claims of the sort "More of this thing causes more (or less) of that thing" or "The presence of this thing will be associated with the presence of that thing because ..."

- Two fundamental problems must be addressed when assessing such a claim by looking at empirical evidence:

  1. How do you know that X is causing Y and not something else that is associated with X and causes Y? This is the "other things equal" problem (discuss Mill's "Method of Difference"), or the problem of "spurious correlation." In regression analysis it appears in the form of the question of whether the independent variables are correlated with the unmeasured, unobserved "other causes" that are summarized in the residual.

     - e.g.: Are Head Start programs effective?
     - e.g.: Does being a democracy make a state a better at winning wars?
     - e.g.: Does a "civic culture" cause democracy?

2. Even if the "other causes" are random with respect to the independent variable X, couldn't it be that just by chance in the sample we have, the random other causes are associated with X? This is the question of *statistical significance*. How do we know if the observed association between X and Y reflects a causal relationship or if it is just *due to chance*.

   - e.g.: Does seeing 8 heads in 10 flips indicate that this coin is biased (i.e., weighted so as to *cause* heads to appear systematically more often than tails, or vice versa), or is it fair and the 8 heads appeared due to chance?
   - e.g.: Did Bush actually get more votes, or is his apparent margin in Florida due to chance errors in the machine counts?
   - e.g.: Volunteers are randomly assigned to canvass certain homes urging residents to vote on election day. Is the observed difference in rates across treatment and control groups due to chance (the homes randomly assigned to treatment were already on average more likely to have voters), or due to a causal effect of contacting?

- The only way to be highly confident that an observed association reflects a causal connection is if the values of the independent variable X are assigned to cases by randomization, so that we know that any other influences cannot be systematically related to them. (Though clever social scientists can sometimes think of circumstances that are like *natural experiments*.)

- With non-experimental data ("observational data"), there is no purely statistical way to ascertain if there are other causes of the phenomenon in question that are systematically correlated with the X in question.

- The clever researcher has to work through all plausible such possibilities, *controlling for their impact* as far as possible. The clever researcher will also come up with alternative implications of the theory that X causes Y that allow him or her to use different sorts of data to assess whether there is a problem with confounds.

- The *second* problem, of determining whether an observed association between one thing and another is *statistically significant* (i.e., unlikely to be *due to chance*), is one of the central contributions of statistical methods. Most of the course focused on this problem and the methods deployed to address it.

- Here is fairly full list of the main topics/issues that we covered and that, optimally, you would understand going into the exam. (Of course, not everything can be tested on the exam).

  1. Descriptive statistics:

     - comparisons between measures of central tendency of a variable, comparisons among measures of dispersion;

- the summation operator, properties of summations;
- what standard units are and how to convert to find them;
- identifying outliers with box plots;
- the ideas of covariance and correlation, their relation to each other, what they are good for and not good for;
- the idea of selection bias, what happens if you select on the dependent variable, on the independent variable;

2. Probability theory:
   - ideas and definition of a probability measure, a probability distribution, and a probability number;
   - the multiplicative and additive laws of probability and their main implications (i.e., the probability of events A and B both occuring is the product of their probabilities if they are stochastically independent; the prob that either A occurs or B occurs is the sum of their probabilities if they are mutually exclusive); simple probability problems involving these rules/ideas;
   - the idea of marginal probabilities and the definition of stochastic independence;
   - binomial coefficients, Bernoulli trials, and the binomial probability distribution;
   - conditional probability, Bayes' Rule;
   - the idea of a *random variable* (very important);
   - discrete and continuous probability distributions, cumulative distribution functions, joint probability distributions, density functions;
   - expected values, mathematical properties of expectations, variance as the expectation of the square of the difference between a random variable and its mean, expectation and variance of the sum of a sequence of Bernoulli trials (important);
   - important distributions – esp. uniform and normal.
3. The central limit theorem and the law of large numbers
   - idea of a simple random sample;
   - be able to show that the sample mean is an unbiased estimate of the population mean, and that the variance of the sample mean is σ^2/n, where n is the size of the sample;
   - be sure you understand how it makes sense to think of the sample mean as a random variable;
   - be sure you can find the standard deviation of a *sum* of a sequence of random variables (e.g., the *number* of Jewish respondents in a sample of 20,000) as well as the (related) question of the standard deviation of a *mean* or average (e.g., the *proportion* of Jewish respondents in a sample of 20,000);
   - the meaning and implications of the law of large numbers and the idea of convergence in probability;

- the meaning and implications of the central limit theorem, when the approximation will be exact, good, and bad;

4. Statistical inference, hypothesis testing
   - idea of using the sample variance $\sum(x_i - \bar{x})^2/(n-1)$ to estimate σ^2/n, why $n-1$ instead of n;
   - idea of a 95% confidence interval, why it is not correct to say that the true population mean is within the confidence interval with probability .95;
   - why the sample mean has a t distribution rather than a normal distribution, and when this matters;
   - the idea of a test statistic, i.e., $t =$ (observed value - expected value if null were true)/s.e. of difference;
   - how to test hypotheses about the difference between two means, difference between one- and two-tailed tests and how to apply them;
   - the meaning and method of the χ^2 test.
5. Bivariate regression
   - be sure you understand what OLS does in mechanical terms – i.e., fits the line through a cloud of data points that minimizes the sum of the squared vertical distances between each point and the line; interpretation of this (line goes through the ("predicted") average value of Y for each value of X; relation to correlation as a descriptive tool, other implications, such as $\sum e_i = 0$, line goes through point of averages, meaning of the r.m.s.e);
   - VERY IMPORTANT: be sure you understand the full set of assumptions necessary to justify or undergird the interpretation of OLS as a way of estimating unobserved structural parameters that govern some causal relationship; be sure you understand the main properties of the OLS estimators (unbiasedness, conditions under which the s.e. estimates hold) and where they come from; be sure you understand what are the implications of violations of the key OLS assumptions.
   - Be able to derive all the important parts of Stata output for a simple bivariate regression given knowledge of the means, variances and covariances of two variables.
   - Be able to interpret regression output in ordinary language.
   - Be able to conduct an analysis of residuals from an OLS regression.