Political Science 100a/200a

Fall 2001

Regression analysis, part III[1]

- There is one more general theoretical result concerning the OLS estimators that I would be remiss not to tell you about. It is usually called the Gauss-Markov Theorem, which establishes that the OLS estimators are "BLUE" – the Best Linear Unbiased Estimates.

Th$\underline{\text{m}}$: Suppose the process that generates data (y_i, x_i) is $y_i = \alpha + \beta x_i + \epsilon_i$, where ϵ_i satifies the assumptions given above. Define a *linear estimator* for α or β as an estimator that can be expressed as a linear combination of the y_i's. Then among the class of linear and unbiased estimators for α and β, the OLS estimators a and b have the smallest variance (i.e., are most precise, and make the most efficient use of the data).

- To see that the OLS estimators are "linear" in this sense, remember that the equation for b is

$$b = \frac{cov(X,Y)}{var(X)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

  Using c_i for the parts of this involving the x_i's, this can be written as

$$b = \sum c_i(y_i - \bar{y}),$$

  so it is clear that the least squares estimator for β is a linear function of the y_i's.

- What the theorem says it that among all possible estimators that are unbiased and that can be expressed as linear functions of the y_i values, you cannot do better than the OLS estimators regarding their precision (i.e., you can get any other estimator with smaller standard errors around the estimates).
- This is an argument in favor of this approach. You might ask, however, what other possibilities are there? What other estimators for α and β might you try? In principle there are lots. e.g.:

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

1. Consider taking the data point for the smallest x_i and the data point for the largest x_i, and drawing a line through them (illustrate). The slope of this line will be an unbiased estimate of β (not bigger or smaller than β on average across replications where we draw new ϵ_is). And it can be expressed as a linear function of y_i values. The Gauss-Markov theorem asserts that the OLS estimates will be better in the sense of being more precise. (It is not hard here to get an intuition for why, given that this approach clearly doesn't use much of the data.)
2. A more plausible rival to the OLS approach would to choose a and b to minimize not the *squared* difference between residual and the line $\hat{y}_i = a + bx_i$, but rather the *simple distance*. I.e., choose a and b to minimize

$$\sum_{i=1}^{n} |y_i - a - bx_i|.$$

   Again it is possible to show that this is a linear, unbiased estimator, so the GM theorem implies that it will not be as good as the OLS estimates of a and b re precision (actually, the theorem just says you can't do better than the OLS estimates on this score).

- This last case – the approach of "least absolute deviations" as it is sometimes called – is worth dwelling on for another minute because it helps make another point about the OLS approach: *sensitivity to outliers*.
  - Think way back to the 3rd week of the course, when we were talking about descriptive statistics, and in particular about measures for the central tendency of a distribution of a variable. We considered two main contenders, both of which have merits, the mean and the median.
  - What was the main liability of the mean relative to the median?
  - In defending the mean as a measure of central tendency, I said that, well, it has nice "statistical properties." You have now seen the most important of these: the sample mean is a sum of random variables, and so has an approximately normal distribution and so allows ready estimates of variance, for example.
  - But the sensitivity to outliers *is* a liability, and it carries over into the OLS approach to estimating structural parameters of a model. To see how, recall that in week 3 I showed (briefly) that for any variable Y, the number a that minimizes

$$\sum_{i=1}^{n} (y_i - a)^2$$

    is in fact the mean of Y, μ.
  - That is, the mean minimizes the sum of squared deviations from a variable.

- But if you look again, this is exactly what we are doing with OLS, except that we are now minimizing the sum of squared differences of Y with a linear function of another variable X.
- So OLS is going to be sensitive to outliers, since in effect OLS weights *squared* distance from the proposed regression line, not the absolute distance.
- Not surprisingly, the method of minimizing the sum of the absolute difference of the residuals produces estimators that are more *robust* to outliers – that is, the presence of a mismeasured or "wacky" observations (perhaps an observation that follows some idiosyncratic process) may screw up OLS estimates more than they would the estimators based on absolute distance.
- What is the practical lesson here?

  When using OLS, and especially when the n is relatively small (say less than 150, in my experience), you need to check for highly influential outliers that may be driving the results you get.
- How do you do this? Scatterplots are valuable, or you can use Stata to generate a variable that stores the residuals, and then list the most extreme ones. (do example with **reg grw6080 ethfrac**, **predict e, resid**, **list country grw6080 e**, **sum e ,d**. Try regress dropping Oman: **reg grw6080 ethfrac if country = "OMAN"**).

- Checking for outliers and influential cases is just one of several important reasons, in practice, to study the residuals from a regression analysis. Here are a couple of other things you can and should do after running the regression. To start, as we did above, use Stata to generate the (estimated) residuals: **predict e, resid** will store them in a new variable called **e**.

  1. Graph them with a histogram: **graph e ,bin(12)** for instance. How do they look? Fairly Normal. Is it important that they have a Normal-like distribution? (think back to results above) Not essential, but nice. Also, if our "story" holds that there are lots of other uncorrelated "other causes" of country growth rates, rather than one or two big ones, then we should expect the residuals to look Normal due to the central limit theorem. If they looked highly not-Normal, then we would have to wonder why, and we would want to investigate what sort of cases were driving the skewness, e.g.
  2. Graph the residuals against the independent variable, the x_is. **graph e ethfrac ,s([cn3]) yline(0)**.
     - What are we looking at? Can someone explain in words?
  3. What can/should we do with this? First off, remember that we made two additional assumptions in deriving estimates for the *standard errors* of our estimates for the model's parameters, a and b. We assumed (1) that the random "other causes" had

the same variance across different values of X (homoscedasticity); and (2) that the covariance of the other causes case-by-case was zero (no serial or autocorrelation).

4. We can evaluate the plausibility of these assumptions by looking at the plot of residuals against X.
5. If homoscedasticity was not a good assumption, we would see a pattern of some sort in the variance of the errors as X changes (note it could be a wavy pattern). No very striking problem here.
6. If autocorrelation were a problem, we might see a pattern of errors "following" each other from case to case.
   - This is purely cross-sectional data (has no time component), so there can be no problem of serial correlation (will show example of this in a minute).
   - But there can be other forms of correlation between the residuals, such as *spatial autocorrelation*. E.g., let's sort the data by region **sort region**, and then create a variable that represents this ordering of the data, **gen caseno = _n**, so we can plot the residuals in this order: **graph e ethfrac, s([cn3]) yli(0)**. Here there seems to be some pattern; discuss. Indicates also omitted variables ...
   - Let's look at a more conventional case of autocorrelation, with some time series data: Presidential approval ratings, inflation, and unemployment. **use approval**, **graph approve caseno**, **reg approve infl3**, discuss, **predict e ,resid**, **graph e caseno ,s(.)**. What do you notice? What is the substantive interpretation? So the OLS estimates will be unbiased (conditional on the other assumptions, like $E(\epsilon_i) = 0$), but the standard errors will be wrong (too small here). (Some of the correlation in the errors here is due to the fact that different presidents had consistently different average approval levels. With multiple regression we can control for this effect, by adding "dummy variables" that mark each different president. Stata lets you take a categorical variable like **pres** and covert it very simply into a set of dummy variables for a regression as follows: **xi: reg approve infl3 i.pres**. Interpret ...
   - One approach in a case like this is to change question slightly, to ask about the effect of a *change* in the inflation rate on the *change* in the president's approval rating. I generated these change variables, **chapprov** and **chinfl**. consider **reg chapprov chinfl**. ...

- Ok, so these are basic diagnostics you should always do after an OLS regression: (1) check for outlying, highly influential cases; (2) check the distribution of the estimated errors; (3) check for violation of homoscedasticity; and (4) check for autocorrelation (especially in time-series data). Questions?
- Two important notes:

1. By looking at the residuals, we can evaluate whether the regression assumptions about homoscedasticity and autocorrelation are warranted. However, our "warrant" in turn depends on the validity of the core assumption that $E(\epsilon_i) = 0$.
2. With observational (nonexperimental) data, the assumption that the other causes are uncorrelated with the independent variable, that $E(\epsilon_i) = 0$, *cannot be tested* with the data points (y_i, x_i). The only thing we can do here is to speculate and think about possible confounds. If we can develop measures for these, we can go further, using *multiple regression* to control for their effects to see if doing so alters our estimate of the effect of the independent variable that we were hypothesizing about in the first place. *This process of thinking through and testing possible confounds is one of the central activities social science research, whether it is explicitly statistical or not.*
   - e.g.: Reconsider the hypothesis that ethnic fractionalization causes lower economic growth rates. In the bivariate analysis **reg grw6080 ethfrac** we found some support for this hypothesis. However, we also noted a possible source of confounding factors: subSaharan African states tended to have very low growth rates in this period, and they also are measured as the most ethnically fractionalized in the sample (show **table region ,c(mean ethfrac median ethfrac n ethfrac)**).
   - So this raises the question, What if it is not ethnic fractionalization that causes poor economic performance, but some other factor that particularly influenced subSaharan states?
   - With multiple regression (which you will explore in 200B), we can begin to assess this possibility by adding a *dummy variable* to the regression equation that takes a value of '1' for the SSA countries and is '0' otherwise. **reg grw6080 ethfrac ssafrica**
   - Interpret ... Some of the relatively homogenous SSA states nonetheless had very poor economic performance, casting some doubt on the initial hypothesis. We can't discard it, but we have moved to a new level of investigation ...
   - An alternative approach, which you have seen everything you need to understand and implement, would be simply to run the bivariate regressions *by region.* e.g.: **reg grw6080 ethfrac if ssa == 1** and**reg grw6080 ethfrac if asia == 1**. Interpret ... not looking good for the hypothesis ...