Political Science 100a/200a

Fall 2001

Regression analysis[1]

- By the time you are bothering to collect some data to examine, in any form (be it a large-N survey of citizen opinions, a medium N sample of countries, or case study or historical narrative of some sort), you almost certainly have some suspicions or conjectures that one thing (an X) is likely to be related to another (a dependent variable, a Y).
- In the best case, you may have theory that implies a specific sort of functional relationship between two variables of interest.
  - e.g.: In macroeconomics, theory suggests that aggregate national consumption should be a linear function of aggregate national income (i.e., $C_t = a + bY_t$ where C_t is aggregation consumption in period t and Y_t is GDP).
  - e.g.: In political science, theoretical work by Gary Cox predicts that the number of viable candidates in election in a district that has M seats in the legislature will be $M + 1$. ("Viable" means getting a non-trivial number of votes).
- But far more typically in social science, if you have a theory or conjecture at all it takes the generic form "more of this (X) should be associated with more (or less) of that (Y)."
  - e.g.: PR systems (like Germany) should be associated with larger numbers of political parties than plurality rule systems (like the U.S.). (Duverger)
  - Greater ethnic heterogeneity should be associated with lower rates of economic growth (Easterly and Levine).
  - Democracies should be less likely to back down in militarized international disputes than nondemocracies (Fearon, Schultz).
  - Small U.S. states should receive greater per capita federal transfers because they effectively overrepresented in the Senate (AER article, can't recall authors)
  - The performance of regional governments in Italy should vary positively with measures of the vitality of civil society by region (Putnam).
  - Protestants should be more prone to suicide than Catholics, etc etc. (Durkheim).
  - Life expectancy should increase with per capita income.
  - and so on ...

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

- We have now seen some very simple ways to test hypotheses and conjectures of this sort.
  - IF you can represent your independent variable in the form of *two categories*, then we can perform a difference of means test to see if we can reject the null hypothesis that there is no significant difference across values of the independent variable.
  - OR, if the independent variable comes in the form or can be presented in the form of *two or more categories*, then we could use a χ^2 test to see if we can reject the null hypothesis that X and Y are stochastically independent.
  - This will work reasonably well for a case like: H_1 = Protestants commit suicide at a significantly higher rate than Catholics, since we can compare the mean suicide rate for Protestants in a sample versus that of Catholics (i.e., categorical independent variables).
- But what about a case like: H_1 = growth rate of per capita income should decrease with ethnic heterogeneity in a country? (Easterly and Levine)
  - **use lifeexp**, **graph grw6080 ethfrac**.
  - We could divide the sample into two groups, with high and low ethnic fractionalization, and then test to see if average growth rates are significantly different across the two.
  - But this seems a bit arbitrary, and also (even intuitively) doesn't seem to take advantage of all the information we have.
  - Eyeballing the scatterplot, it looks like there is something of a (somewhat) steady downward trend in 1960-80 growth rate as ethnic fractionalization increases. But two important questions arise:
    1. How can or should we characterize this (possible) downward trend?
    2. How can we decide whether a downward trend of this magnitude constitutes relatively strong or relatively weak evidence in favor of the initial hypothesis or conjecture?
  - Regression analysis is very helpful for answering these questions, though it is not a magic bullet and must be used with great self-awareness.
  - Notice, by the way, that these two questions parallel those we have been asking in the last few weeks:
    1. "What is a good estimate of the population mean μ?" parallels the question about how to characterize the downward trend, and
    2. "How much uncertainty is there around our estimate (σ^2/n)?" parallels the second question of whether the observed downward trend is significantly different from no trend.

- Here is how bivariate regression would typically be deployed to answer these questions in this case.
- Our theory or conjecture says that ethnic fractionalization influences the economic growth rate.
- In Easterly and Levine, the idea is that ethnice diversity causes problems of political coordination and fighting over the distributional consequences of macroeconomic policies; I think the argument is bogus, but whatever.
- But Easternly and Levine's, or any other plausible theory, doesn't imagine that ethnic fractionalization is the ONLY determinant of growth rates – there are certainly other, and probably much more important factors.
- Our theory/hypothesis thus imagines that a country's growth rate over a period of time depends on its degree of ethnic fractionalization *and* on these other things, let's call them ϵ_i for country i:

  $$y_i = f(x_i, \epsilon_i),$$

  where $f(\cdot)$ is some function, y_i = country i's growth rate, x_i = its degree of ethnic fractionalization, and ϵ_i represents all the other causes of its economic growth rate.
- But what should the function $f(\cdot)$ look like? Most likely our theory is no help here – it probably gives us no precise guidance on the functional form of the relationship. (There are rare exceptions in political science, many more in economics.)
- Here is where most researchers are likely to "cheat," in effect, and look at the scatter plot to see what it looks like. It looks like something linear wouldn't be too bad. So how about

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

- This says that a country's growth rate from 1960-80 is a linear function of ethnic fractionalization, plus the total impact of all other causes. (Note that β can be negative, indicating negative slope, as can α, indicating a negative intercept, in general. Review slope/intercept ... If increase ethnic fractionalization by .10, would imply that country growth rate increases by $\beta(.10)$.)
- Draw line through graph ...
- We would like to estimate values for α and β, in order to see if ethnic fractionalization in fact has a causal impact on economic growth. (If we estimate $\beta = 0$ and have reason to believe this a good estimate, then we would be concluding that the scatterplot indicating some suggestion of a negative relationship was misleading.)

- But here we run in a major problem: We observe the growth rates (the y_i values) and the countries' ethnic fractionalization levels (the x_i's), *but we do not observe the impact of the other causes for each country, the ϵ_i's.*
- Mathematically speaking, we have n equations – one for each observation (list) – but $n + 2$ unknowns, an ϵ_i for each country, plus α and β. (Draw ...)
- This means that we have no hope of pinning down estimates for α and β unless we make further assumptions.
- In fact, IF we are willing to make one crucial assumption about the relationship between the x_i's and the other causes, the ϵ_i's, then we can come up with estimates for α and β. But of course whether these estimates are valid/good estimates depends entirely on whether this assumption is a good one.
- The assumption: That the other causes of country growth rates *do not vary systematically with ethnic fractionalization.*
  - That is, we need to assume that it is not the case that
    1. the ϵ_i's, the other causes of growth rates tend to be bigger when when ethnic fractionalization is large, or
    2. the ϵ_i's tend to be smaller when ethnic fractionalization is larger.
  - Draw both cases on board ...
  - Mathematically, a sufficient condition for this assumption is simply that the average or expected value of the other causes is zero for any given value of x_i. Formally,

    $$E(\epsilon_i) = 0.$$

    Draw a vertical "band" and show how this assumption would rule out (1) and (2) above.
- To repeat, IF we are willing to make this assumption – i.e., if we can defend it as plausible, arguing that likely candidates for omitted "other causes" are unlikely to be systematically related to our independent variable x_i – then we have a way of generating estimates for α and β that have some nice properties.
- Recall: The best and really only way to *guarantee* that the other causes are not systematically related to the independent variable would be to randomly assign x_i values to different cases. The problem is that for most interesting social science questions, we can't do this first best approach (randomization), so we are forced to do our best in "controlling for" possible confounds.

- The estimators for α and β, call them a and b, are the famous *least squares estimators* that solve the following problem:

$$\text{Choose } a \text{ and } b \text{ to minimize } SSR = \sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

(SSR stands for "sum of squared residuals").

- Explain in terms of scatterplot – a and b give the equation for the line

$$\hat{y}_i = a + bx_i$$

that minimizes the sum of the squared vertical difference between the line and each y_i. These differences, $\hat{e}_i = y_i - a - bx_i$ are estimates of the *residuals* ϵ_i under the assumption that they are not systematically related to the x_i's.

- $\hat{y}_i$ is the *predicted value* for country i's growth rate, 1960-80, based on these estimates.

- So what are a and b? We need a little calculus to derive them. First, take the derivative of SSR with respect to a, the intercept:

$$\begin{aligned}
\tfrac{\partial SSR}{\partial a} &= \tfrac{\partial}{\partial a}\textstyle\sum_{i=1}^{n}(y_i - a - bx_i)^2 \\
&= \tfrac{\partial}{\partial a}\left[(y_1 - a - bx_1)^2 + (y_2 - a - bx_2)^2 + \ldots + (y_n - a - bx_n)^2\right] \\
&= -2(y_1 - a - bx_1) - 2(y_2 - a - bx_2) - \ldots - 2(y_n - a - bx_n)^2 \\
&= -2\textstyle\sum_{i=1}^{n} y_i - a - bx_i \\
&= -2(n\bar{y} - na - nb\bar{x}).
\end{aligned}$$

- So this is the slope of the sum of squared residuals as we increase a, holding b constant. To find the minimum with respect to a (note second derivative w/r/to a and cross-partial w/r/to b are positive, thus probably a minimum), we set this equal to zero and solve for a:

$$\begin{aligned}
\tfrac{\partial SSR}{\partial a} &= -2(n\bar{y} - na - nb\bar{x}) = 0 \\
a &= \bar{y} - b\bar{x}.
\end{aligned}$$

- Rewriting slightly, this says $\bar{y} = a + b\bar{x}$, which means what? *The regression line goes through the point of averages for x_i and y_i.* Draw ...

- Next, we take the derivative of SSR with respect to b, the slope coefficient, which has not been pinned down yet (neither of course has a, we need two equations to pin down these two unknowns).

$$\begin{aligned}
\frac{\partial SSR}{\partial b} &= \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - a - bx_i)^2 \\
&= \sum_{i=1}^{n} 2(y_i - a - bx_i)(-x_i) \\
&= -2(\sum_{i=1}^{n} x_i y_i - a x_i - b x_i^2) \\
&= -2(\sum x_i y_i) + 2an\bar{x} + 2b\sum x_i^2.
\end{aligned}$$

- Setting this equal to zero (to find the minimum with respect to b) and dividing out the -2 yields

$$\begin{aligned}
\frac{\partial SSR}{\partial b} &= (\sum x_i y_i) - an\bar{x} - b\sum x_i^2 = 0 \\
\sum x_i y_i &= a\sum x_i + b\sum x_i^2.
\end{aligned}$$

- So now we have two equations and two unknowns (a and b). These equations are often called *the normal equations*.

$$\bar{y} = a + b\bar{x} \tag{1}$$
$$\sum x_i y_i = a\sum x_i + b\sum x_i^2. \tag{2}$$

- Notice that these are all quantities that we can compute with our sample data, our list (y_1, x_1), (y_2, x_2), etc.
- Solving for b yields

$$b = \frac{(\sum x_i y_i) - n\bar{x}\bar{y}}{(\sum x_i^2) - n\bar{x}^2}$$

- Does either numerator or denominator look familiar? We showed earlier on (week 3) that these are ways of writing

$$\begin{aligned}
b &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\
b &= \frac{cov(x_i, y_i)}{var(x_i)}.
\end{aligned}$$

- So, the slope of the regression line equals the ratio of the covariance of X and Y to the variance of X. This is worth remembering. From this we can substitute back in easily to get the estimator for the intercept term, a:

$$\begin{aligned} a &= \bar{y} - b\bar{x} \\ a &= \bar{y} - \frac{cov(x_i, y_i)}{var(x_i)}\bar{x}. \end{aligned}$$

- These are the *ordinary least squares* estimators for α and β in the case of simple bivariate regression (two variables).
- This should be exactly what computer programs like Stata compute: **correlate grw6080 ethfrac, covariance** (this gives the variance-covariance matrix for these two variables), **di ../..**, **regress grw6080 ethfrac**. Discuss Stata output for **regress**.
- To plot the regression line in Stata, we first need to generate the predicted values of growth for each country in the sample given its level of ethnic fractionalization, i.e., $\hat{y}_i = a + bx_i$. To do this, type **predict yhat**, which computes $\hat{y}_i$ using the estimates for a and b and stores this in a new variable called **yhat**. **list country grw6080 yhat** ... Then **graph grw6080 yhat ethfrac ,symbol([cn3].) connect(.l)**.
- Let's recap:

  1. We entered with a theory or just a conjecture, holding that countries with higher levels of ethnic fractionalization should be expected to have lower growth rates on average. We proposed the model

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

     where y_i was country i's 1960-80 growth rate, x_i its (1960) ethnic fractionalization score, ϵ_i the contribution to i's growth rate of all other causes, and α and β the "structural parameters" that relate ethnic fractionalization to growth rates.
  2. We wanted to come up with estimates for α and β, but faced a fundamental problem: We don't observe the impact of the other causes, making it highly problematic to use the data we have (values on y_i and x_i for n countries) to sort out the impact of x_i (the α and β) from the impact of the other causes.
  3. I claimed that if we were prepared to make (and plausibly defend) one big assumption about the relationship between ethnic fractionalization and the other causes of growth rates, the x_i's and the (unobserved) ϵ_i's, then we could come up with good estimates for the structural parameters α and β.
  4. The assumption was that the unobserved other causes are not systematically related to independent variable x_i, which is implied by the assumption that $E(\epsilon_i) = 0$.

5. We then derived the least squares estimators a and b for α and β. The approach was to choose a and b to minimize the sum of the squared (estimated) residuals $\sum \hat{e}_i^2$, where $\hat{e}_i = y_i - a - bx_i$.

- What remains to be established is *in what sense and under what conditions are the least squares estimators a and b good estimators for α and β, the parameters presumed/to be tested for the theory?*

- The answers to these questions are in a certain sense very many and very involved – filling out the answers will fill up much of PS200b, for example. But we will work through the basics in the simple bivariate regression case here.

- To answer the questions, we need to answer the following question:

  What is the relationship between the estimators a and b and the unobserved parameters α and β that we are trying to estimate?.

- The situation parallels exactly our former problem of having the sample mean $\bar{x}$ and asking about its relationship to the unobserved population mean μ.

- To establish the relationship between a and b and α and β, let's begin by considering the least squares slope estimate,

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}.$$

- According to our theory, $y_i = \alpha + \beta x_i + \epsilon_i$. We can substitute this in for y_i in the expression for b to write

$$\begin{aligned} b &= \frac{\sum(x_i - \bar{x})(\alpha + \beta x_i + \epsilon_i - \bar{y})}{\sum(x_i - \bar{x})^2} \\ &= \frac{\alpha \sum(x_i - \bar{x}) + \beta \sum x_i(x_i - \bar{x}) + \sum(x_i - \bar{x})\epsilon_i - \bar{y} \sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \end{aligned}$$

- But $\sum(x_i - \bar{x}) = 0$ (the sum of deviations from the mean is zero), so

$$\begin{aligned} b &= \frac{\beta \sum x_i(x_i - \bar{x}) + \sum(x_i - \bar{x})\epsilon_i}{\sum(x_i - \bar{x})^2} \\ &= \frac{\beta \sum x_i(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} + \frac{\sum(x_i - \bar{x})\epsilon_i}{\sum(x_i - \bar{x})^2}. \end{aligned}$$

- The first term, it turns out, is just β! This is because $\sum x_i(x_i - \bar{x}) = \sum(x_i^2 - x_i\bar{x}) = (\sum x_i^2) - n\bar{x}^2$, and you may recall (or you can check in your notes for week 3), that this last expression is just another way of writing $\sum(x_i - \bar{x})^2$. So top and bottom cancel there, leaving

$$b = \beta + \frac{\sum(x_i - \bar{x})\epsilon_i}{\sum(x_i - \bar{x})^2} \tag{3}$$

- If you found the algebra hard to follow, don't worry, but definitely tune back in now. This equation (3) is an important and instructive result.
- It says that our estimate b for the slope relating ethnic fractionalization x_i to income growth rate y_i equals the parameter value we are trying to estimate, β, PLUS the ratio of the covariance of x_i and ϵ_i to the variance of x_i.
- So, what will be the case if ethnic fractionalization is not systematically related to the unobserved other causes of a country's income growth rate? Then $cov(x_i, \epsilon_i)$ would tend to be close to zero, which would mean that the our estimate b would be very close to what we are trying to estimate, β.
- Thus the importance of the plausibility (and defensibility) of the assumption that our independent variable is not systematically related to the other causes of the dependent variable. If this assumption in not a good one, b will NOT be a good estimator for β, and we may be sorely mislead by using OLS.
    1. What if the truth is that $cov(x_i, \epsilon_i) > 0$? Then $b > \beta$, so we will be lead to believe that x_i has a bigger impact on economic growth y_i than it really does if β is positive, and a smaller (closer to zero) impact if β is negative. e.g., this is the case of a confounding factor that positively affects economic growth and by coincidence happens to occur more frequently in countries with higher levels of ethnic fractionalization, so that using OLS would attribute some of this positive affect of the other causes to ethnic fractionalization, thus masking the full effect of **ethfrac**.
    2. If the truth is that $cov(x_i, \epsilon_i) < 0$, then $b < \beta$. We will underestimate the impact of x_i when it is positive, and overestimate its impact when the impact is negative. explain ...
- And thus we can be justified in using OLS to estimate α and β even if we don't observe the other causes. This one assumption, if we can plausibly make it, allows us to estimate α and β with some accuracy even though we don't observe the impact of the other causes (we estimate these as well, with $\hat{\epsilon}_i = y_i - a - bx_i$).
- Questions?