Political Science 100a/200a

Fall 2001

Regression analysis, part IV[1]

1 Substantive versus statistical significance

- The fact that an estimated regression coefficient is "statistically significant" (i.e., you can reject the null hypothesis that the true β is 0 with a high level of confidence) *does not mean* that your independent variable is *substantively* important.
- Illustrate with diagram and reference to **ethfrac**.
- This is why it is good practice not to simply report "significance levels," but also give a sense of substantive impact of a change in an X variable. Illustrate.

2 More on dummy variables

- I thought I should give you a bit more on *dummy* or *dichotomous* independent variables, since these are so common in political/social science and they may seem initially puzzling or hard to interpret.
- What a dummy variable does in a regression model can be interpreted like this: It gives a *difference of means* test, very similar to what we saw a few weeks ago.
- Consider a question like this: Did subSaharan (SSA) countries grow significantly less rapidly from 1960 to 1980 than did other countries?
- You already know how to address this with a difference of means test: Consider whether you can reject the hypothesis that $\bar{y}_{SSA} - \bar{y}_{\sim SSA} = 0$, where $\bar{y}_{SSA}$ is the average 1960-80 growth rate for SSA countries, and $\bar{y}_{\sim SSA}$ is the average for the other countries.
- But what if you tried OLS regression, using **grw6080** for your dependent variable, and **ssafrica**, the dummy variable that marks the SSA countries with a '1' as the independent variable?
- Let's just do it: **use lifeexp**, **graph grw6080 ssafrica**. What are we looking at here? Interpret.
- What happens if we try **reg grw6080 ssafrica**?

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

- So here we have done a bivariate regression of 1960-80 growth rate of per capita income (the dependent variable) on a *dummy variable* marking countries in subSaharan Africa. We are estimating a and b for the line

$$\mathbf{grw6080}_i = a + b\mathbf{ssafrica} + e_i$$

- What do we get? How do interpret the coefficient on b? Note that the independent variable takes only two values here, 0 and 1. So, for countries NOT in SSA, the estimated equation is saying that

$$grw6080_i = a + b * 0 + e_i = a + e_i$$

- So what is a then, if we assume (in our regression assumptions) that $E(\epsilon_i) = 0$? **sum grw6080 if ssafrica == 0** to check.

- And for countries that ARE in SSA, **ssafrica** $= 1$, so for these countries we have

$$grw6080_i = a + b * 1 + e_i = a + b + e_i$$

- So how to interpret b here? It is the *difference* in the average growth rate of the subSaharan countries versus the non-SSA countries. The average grwoth rate of the SSA countries will be $a + b$. (**sum grw6080 if ssafrica == 1**.)

- Look back at the graph, **graph grw6080 ssafrica**. Interpret ... Draw regression line: **predict yhat**, **graph grw6080 yhat ssafrica, s(..) c(.l)**

- Recall that what regression does is to draw a line through the mean values of the dependent variable y, *conditional on* or *for each* x value. With a dummy variable we have a particularly simple case of this, where the mean conditional on $x = 0$ is just the average growth rate for non-SSA countries, and the mean conditional on $x = 1$ is the average growth rate for SSA countries.

- So, again, with a dummy X variable in a simple bivariate regression (Y on one X),

  1. the estimated constant a is just going to be the mean of the dependent variable when $x = 0$, and
  2. the estimated slope coefficient b is just the difference $\bar{y}|(x = 1) - \bar{y}|(x = 0)$.
  3. Thus, the mean of y when $x = 1$ is just $a + b$.

- Notice that we could just as easily have approached this problem as a hypothesis test of a difference of means. Suppose your question were: Did SSA countries grow significantly more slowly between 1960 and 1980 than other countries?

- Can you recall how to do this? What we want to know is if we can reject the null hypothesis

$$H_0 : \bar{y}_{SSA} - \bar{y}_{\sim SSA} = 0$$

where $\bar{y}_{SSA}$ is the averate growth rate for SSA countries and $\bar{y}_{\sim SSA}$ is the average growth rate for the others.

- To test the hypothesis, we needed to get an estimate of the standard error of this difference between two variables. Remembering that $var(X - Y) = var(X) + var(Y)$, we have

$$\begin{aligned} var(\bar{y}_{SSA} - \bar{y}_{\sim SSA}) &= var(\bar{y}_{SSA}) + var(\bar{y}_{\sim SSA}) \\ &= \frac{\sigma^2_{SSA}}{n_{SSA}} + \frac{\sigma^2_{\sim SSA}}{n_{\sim SSA}} \end{aligned}$$

- If we assumed further that the variance in growth rates is the same in both SSA and not SSA, call it σ^2, we would have that

$$var(\bar{y}_{SSA} - \bar{y}_{notSSA}) = \sigma^2 \left(\frac{1}{n_{SSA}} + \frac{1}{n_{\sim SSA}} \right) \quad (1)$$

- This is essentially what Stata computes with **ttest grw6080 ,by(ssafrica)**. Discuss. Compare to regression.

- The OLS regression is in fact doing the same thing (under the assumption that the variances are the same across the two 'treatments,' i.e., homoscedasticity). The slope coefficient b is precisely a *difference of sample means* (show again). So the t-test given by Stata is a test against the null hypothesis that b, the difference of means, is zero.

- You might ask, Is there a difference between the s.e. calculated as in the difference of means test above, and the s.e. as calculated for a regression coefficient? The answer turns out to be no. Recall that the equation we got for the variance of the OLS estimator b was

$$var(b) = \frac{\sigma^2}{nvar(X)},$$

where σ^2 is the variance of the residuals (make graphical connection to σ^2 in difference of means test).

- You should be able to show – and I encourage you to do so – that this will be exactly the same as (1) above.

- In sum: *The slope coefficient on dummy independent variable X in an OLS regression represents the difference in the average values of the dependent variable for cases where the dummy variable equal 1 and cases where the dummy variable equals zero.*

- What about in a multiple regression, where we are adding a dummy variable as a "control" as we did in the last lecture? **reg grw6080 ethfrac ssafrica**. Discuss ...

3 An alternative approach: bivariate regression as a descriptive technique

- The last thing I want to do is to link up the FPP approach to what we have been doing (the more political science-y, causal models approach).
- Up to this point, this presentation of simple bivariate regression has followed the standard approach in political science (you should have seen it in the assigned texts, e.g., Achen's book, or KKV). Our point of departure has been the notion that we are interested in empirically evaluating *a causal hypothesis* about the relationship between a dependent variable of interest (the thing to be explained) and an independent variable.
- To sustain such an approach, we have to be willing to make and defend assertions such as "the other causes of growth rates were not systematically related to ethnic fractionalization in this period." We have to able to argue that even though the data we have was not generated by random assignment, we can treat at *as if* it had been.
- Even if one is not willing to do this, however, regression is still a potentially quite useful tool for the more modest goal of *describing* relationships between variables, rather than testing causal hypotheses about them.
- This is the approach taken in FPP. They present regression in connection to correlation, merely as a way of the describing how two variables are related.
- To understand this (rather simpler) approach, let's go back with a review of the idea of *correlation*, which we considered as a simple way of describing the relationship between two variables.
- Recall that the correlation coefficient

$$r \equiv \frac{cov(X,Y)}{sd(X)sd(Y)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

  contains TWO basic pieces of information:

  1. The SIGN of r – positive, negative, or zero – indicates whether the two variables X and Y covary positively, negatively, or not at all. That is, do higher values of X tend to be associated with higher or lower values of Y (or does Y vary independently of X, in the case of $r = 0$).

2. The MAGNITUDE of r – how close it is to 1 or -1 – indicates how tightly clustered the points in a scatterplot would be around a line. Substantively, this means that correlation answers the question, If you know the value of one variable (either X or Y), how well can you predict the value of the other?

- Recall also that the correlation of X with Y is the same as the correlation of Y with X: $r(X, Y) = r(Y, X)$.
- However, one important question that the correlation coefficient does NOT really answer is the following:

  How rapidly does Y change, on average, if you look at different values of X?

  and likewise,

  How rapidly does X change, on average, if you look at different values of Y?
- Or, in other words (still avoiding any hint of a causal relationship):

  On average, how big a change in Y is associated with a one unit change in X?
- You should already see how OLS regression fits a line to a cloud of data that can answer this question. But let's look again at how the correlation coefficient *fails* to answer it.
- Let's generate some correlated random variables in Stata. I am going to "cook" the numbers to make a point: **set obs 1000**, **gen x = invnorm(uniform())**, **gen e = invnorm(uniform())**, **gen y1 = 2 + 3*x + e**, **gen y2 = 2 + x + .3*e**. **corr x y1 y2**
- Notice that X has approximately *the same correlation* with both Y_1 and Y_2. Now, **graph y1 y2 x**. What do you notice?
- Not surprisingly (if we consider the structural equations that generated the y variables from x), these look like rather different relationships. In particular, notice that the *slope* of the cloud of points for $y1$ and x is steeper than that for y_2 and x. What should we find if **reg y1 x** and **reg y2 x**?
- Some questions:

  1. Why aren't the regression coefficients exactly equal to structural parameters for α and β? (this is an illustration of the hypothetical experiment justifying the causal interpretation of regression, by the way).
  2. If the correlations are the same, why does the cloud of points look more tightly clustered around a line for y_2 and x than for y_1 and x? (this is hard question at first, though the answer is straightforward; see FPP, pp. 144-145).
- The main point though: Correlation is not a reliable measure of the slope of the cloud of points of the scatterplot of two variables. (Except in one case, i.e., for one value of r; can you guess which?)

- Typically, in social science, we don't just want to know

  1. how well you can predict one variable if you know the value of another (correlation). You also want to know
  2. how big a change in one variable is associated, on average, with a change in the other.

- (2) is what OLS regression is good for, *even in the absence of any story about how a change in X* **causes** *a change in Y.*

- Why? Because OLS gives you the equation of a line that goes through the predicted or "smoothed" average value of Y for each value of X. Clear?

- That said, there *are* interesting relationships between the correlation coefficient between and the results of a bivariate regression. (As you have seen, FPP develop regression "from" correlation, as it were.) Let's consider what the relationship is by inspecting the formula for r and comparing it to the formula for b, the slope parameter estimated by OLS:

$$r \equiv \frac{cov(X,Y)}{sd(X)sd(Y)}$$

  and

$$b = \frac{cov(X,Y)}{var(X)} = \frac{cov(X,Y)}{sd(X)sd(X)}$$

- So, to express b in terms of r, we have

$$b = \frac{cov(X,Y)}{sd(X)sd(X)} = \frac{cov(X,Y)}{sd(X)sd(Y)}\frac{sd(Y)}{sd(X)} = r(X,Y)\frac{sd(Y)}{sd(X)}. \quad (2)$$

- In words, the OLS estimate of the slope of the line that goes through the average value of Y for each x_i is just the correlation coefficient times the ratio of the standard deviation of Y to the standard deviation of X.

- FPP have a nice graphical interpretation of the meaning of this: $sd(Y)/sd(X)$ is the slope of a line that they call *sd line*, the line such that a one standard deviation increase in X is associated with a one sd change in Y. Draw illustration.

- What equation (1) implies is that the average increase in any one variable (the Y) associated with a one standard deviation change in another variable (the X) is *always* less than one standard deviation, by an amount that depends on how closely correlated the two variables are.

1. If the two variables are perfectly correlated – that is, they fall in a straight line on the scatterplot – then a one sd change in X is associated with a one sd change in Y.
2. If the two variables are completely uncorrelated – that is, $r(X, Y) = 0$ – then a one sd change in X is associated with a $0 * sd(Y)/sd(X) = 0$ sd change in Y.
3. Draw both cases, and an intermediate case.

- This is actually where the term "regression" comes from, the idea of "regression to the mean" (of Y).

1. On average, children of parents who are above average in height will be not as far above average as their parents were, whereas children of parents who were below average will be less far below average than their parents were.
2. And if I had given you a math test at the start of the term, and another one at the end of the quarter, I would find that the scores of those who were below average on the first test will have tended to improve in the second test, and those who did above average on the 1st will tend to have gotten worse on average on the second test. (Draw graph, lines. has obviously important implications for policy/social science research, e.g., voucher analyses)

- What is going on here intuitively? Students who did particularly badly on the first test tended on average to get worse random draws on the "chance error" part of their score (i.e., the part independent of ability). In the second test, their average "chance error" will be zero, so they will do better (though still worse than average, since there was a systematic component, in ability or preparation, as well).

- Back to the main arguments for today: The relationship and differences between correlation and regression.

- We have seen that in a bivariate regression, the OLS estimate for the slope coefficient b is just the correlation coefficient r times the ratio of the sd's, $sd(Y)/sd(X)$.

- There is another relationship, which is useful to develop because it allows us to explicate the last important bit of information provided by the Stata output for a typical regression, the R^2, which is often described as *the percentage of variance in Y "explained" by the model.*

- Recall that one of the two pieces of information contained by a correlation coefficient is: How tightly clustered about a straight line are the points in the scatterplot? Substantively, this means that r tells you how accurately you can predict Y is you know the value on X, and vice versa.

- If we think about the picture of an OLS regression line fit to a cloud of data, what would this correspond to? That is, what is the regression equivalent for "how tightly

clustered are the points around the line" and thus "how well can you predict Y if you know x_i?"

- The mean square error, $s^2 = \frac{1}{n-2}\sum e_i^2$ is a measure of how far, on average, the data points are from the regression line (see FPP p. 182). The square root of this is like the standard deviation of the estimated residuals or errors. Show graphically ...

  1. If all the data points fall perfectly on a line, then this will be zero, indicating that you can perfectly predict Y if you know X.
  2. What if there is no relationship between Y and X at all, so that the results of the OLS regression is that $b \approx 0$? (Draw picture.)
     - Then what will s^2 be?

$$\begin{aligned} s^2 &= \frac{1}{n-2}\sum e_i^2 \\ &= \frac{1}{n-2}\sum (y_i - a - bx_i)^2, \text{ but } b \approx 0, \text{ so} \\ &= \frac{1}{n-2}\sum (y_i - a)^2, \text{ but } a = \bar{y} - b\bar{x} = \bar{y} \\ &= \frac{1}{n-2}\sum (y_i - \bar{y})^2. \end{aligned}$$

     - What is this? Approximately, the variance of Y.
     - So in this case of no relationship between Y and X, the residual sum of squares (RSS) is just the variance of Y. By contrast, if there is a perfect relationship, the RSS is zero.

- This suggests a regression-based measure of the overall *fit* of the model, or how tightly clustered the data is around the regression line: How close are we to the case where the RSS = 0 (a perfect fit)?

- Here is how we can derive such a measure. Begin with the OLS equation for observation i:

$$y_i = a + bx_i + e_i.$$

- Remember that the OLS intercept is $a = \bar{y} - b\bar{x}$, so

$$\begin{aligned} y_i &= \bar{y} - b\bar{x} + bx_i + e_i \\ y_i - \bar{y} &= b(x_i - \bar{x}) + e_i \end{aligned}$$

- Square both sides:

$$(y_i - \bar{y})^2 = b^2(x_i - \bar{x})^2 + e_i^2 + 2b(x_i - \bar{x})e_i$$

- And take the sum over all observations, $i = 1$ to $i = n$:

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

- It is not difficult to show that the last term on the right hand side is zero (the intuition: the OLS line *is constructed* so that the covariance of the errors and the x_i's is zero):

$$\begin{aligned} 2 \sum b(x_i - \bar{x})e_i &= 2b \sum (x_i - \bar{x})(y_i - \bar{y} - b(x_i - \bar{x})) \\ &= 2b(\sum (x_i - \bar{x})(y_i - \bar{y}) - b \sum (x_i - \bar{x})^2) \\ &= 2b(b \sum (x_i - \bar{x})^2 - b \sum (x_i - \bar{x})^2) \\ &= 0. \end{aligned}$$

- So we have the interesting conclusion that

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

- What does this say? You can think of this a decomposition of the variation in the dependent variable, Y.
- It says that the variation in Y (the left hand side) equals *the variation "explained" by the X variable* (the $b^2 \sum (x_i - \bar{x})^2$) PLUS *the unexplained variation* that is left over in the residuals (the $\sum e_i^2$).
- Draw picture ... (see FPP, p182ff)
- Returning to the question of how tightly clustered around the OLS line the points are (and thus how well our regression "fits"), consider the following question:

  What proportion of the variation in Y is "explained" by the X variable?

- To answer this, we only need to divide both sides by the total variation in Y, $\sum(y_i-\bar{y})^2$, getting

$$1 = \frac{b^2\sum(x_i-\bar{x})^2}{\sum(y_i-\bar{y})^2} + \frac{\sum e_i^2}{\sum(y_i-\bar{y})^2}$$

- The first term on the right hand side is the proportion of the variation in Y "explained" by the regression, which is usually called

$$R^2 = \frac{b^2\sum(x_i-\bar{x})^2}{\sum(y_i-\bar{y})^2} = 1 - \frac{\sum e_i^2}{\sum(y_i-\bar{y})^2}$$

- Lo and behold:

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

  where r is the simple correlation coefficient.

- SO: In a bivariate regression, the square of the correlation coefficient gives the proportion of variation in Y that is "explained" by variation in the independent variable X.

- illustrate with **reg grw6080 ethfrac**.

- Some comments:

  1. Why did I keep putting "explained" in quotes? Because we are only talking about association here, not necessarily causation.
  2. The expression derived above,

$$\sum (y_i - \bar{y})^2 \;=\; b^2 \sum (x_i - \bar{x})^2 + \sum e_i^2$$

can be rewritten

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

I would advise you to take note of this, because it gives you a way to compute the residual sum of squares (and thus s^2, and thus s.e.'s for b) using ONLY the variance of Y, the variance of X, and their covariance (in b).
On the exam (which will not require Stata), you will be asked to calculate regression results from variances and covariances, so you will need this. Questions?