Political Science 100a/200a

Fall 2001

Regression analysis, part II[1]

1 The OLS estimates a and b as random variables

- The last step in showing that if the key assumption holds, the OLS estimate b will be a good estimate of β is usually done a bit differently.
- Now we will work through this more usual approach, which is to think about the estimator b *as a random variable*, just as we thought of the sample mean $\bar{x}$ as a random variable.
- How do we do this in this case? A random variable, remember, is a number that summarizes some result of a chance procedure (Friedman). What is the chance procedure that generates b?
- There is a typical story that statisticians and econometricians tell at this point, that goes like this. They say, Imagine it's like this:
  - Imagine this is like an experiment where we can assign specific x_i values *at random* to different cases (here, countries, but in say a medical experiment, dosage levels to different people). We do this.
  - Then various other causes (which are now necessarily random with respect to x_i) work their effects ("are drawn"), the ϵ_i's, and added to $\alpha + \beta x_i$ for each case. So we observe $y_i = \alpha + \beta x_i + \epsilon_i$ and x_i for each case, but not the parameters or the error terms.
  - We use the data to generate the OLS estimates a and b.
  - Imagine (they say) you could do this over and over again, assigning *the same* x_i levels to the same cases each time, and then drawing anew from the "box of tickets" that gives us the random other causes, the ϵ_i's. Because of the sampling variability from the box of tickets, each time we will get slightly or somewhat different estimates for the α and β.
  - For each one of these hypothetical repetitions of the experiment, we have a new set of estimates for a and b. Thus you can imagine building up a sampling distribution for a and b, just as we did for the sample mean $\bar{x}$ earlier on.

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

  - This is the mental experiment used to explain how to think about a and b as random variables.
  - The key move is thinking about the other causes, the ϵ_i's, as *in effect* like numbers drawn from a box of tickets. We only *observe* the results of the one set of draws – the y_i's, which are produced by $y_i = \alpha + \beta + \epsilon_i$ – but we imagine that we could have observed different y_i's in a hypothetical replication, where the other causes came out a differently.

- Illustrate with Stata:

  1. **set obs 50**
  2. **gen x = invnorm(uniform())**
  3. **gen e1 = invnorm(uniform())**
  4. **gen y1 = 2 + 3*x + e1**
  5. **reg y1 x**
  6. **gen e2 = invnorm(uniform())**
  7. **gen y2 = 2 + 3*x + e2**
  8. **reg y2 x**

- There are some cases – mainly experimental designs – where this story of repeatedly "drawing" ϵ_i's for fixed x_i values is not so hypothetical. e.g.: assigning different levels of fertilizers to different acres of land to estimate the yield curve for the fertilizer. You really can imagine doing this repeatedly and getting different results due to random variability in the other causes of crop yields across the plots each time.

- But with observational data, the story is more hypothetical, in two main respects: (1) we didn't actually randomly assign x_i values to each case; and (2) we are imagining counterfactual worlds where we perform this procedure over and over, and "draw" different values on the other causes each time.

2 b is an unbiased estimate of β: $E(b) = \beta$

- Under the story that imagines the data we see to have been generated *as if* the other causes ϵ_i are random variables that were drawn from a distribution, we can treat the estimated parameter b as a random variable. This means that we can take its expectation:

$$E(b) = E\left(\beta + \frac{\sum(x_i - \bar{x})\epsilon_i}{\sum(x_i - \bar{x})^2}\right)$$

$$
\begin{aligned}
&= \beta + E\left(\frac{\sum(x_i - \bar{x})\epsilon_i}{\sum(x_i - \bar{x})^2}\right) \\
&= \beta + E\left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\epsilon_i\right) \qquad (1)
\end{aligned}
$$

We can take the expectation "through" β why? Note I also brought the variance term inside the sum in the numerator for later convenience.

- Recall that in the mental experiment described above, we imagined that we (or Nature) assigned given, fixed values of x_i to each case, and then in subsequent experiments we could draw ϵ_i values given these x_i values. (i.e., we imagine that the particular ϵ_i values that were realized and helped produce growth rates for these countries from 1960 to 1980 could have turned out differently, they were "drawn" from a distribution of possible worlds – see KKV).
- This is typically described as the assumption of *fixed* X, or the assumption that X *is fixed in repeated sampling*, which basically means that we are treating the independent variable X as nonstochastic (not a random variable).
- This may seem arbitrary to you, except in a case where X really is fixed by experimental control, but it turns out that this is pretty much an assumption for convenience in working through the math. We can allow X to be stochastic (i.e., imagine that we might draw different values, from a distribution, if we could do the experiment over and over), and almost nothing would change.
- *Given* the assumption of fixed X, the x_i values in equation (1) above are *constants*. So the sum in (1) is a sum of constants times the ϵ_i's. To see this, let

$$c_i \equiv \frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{x_i - \bar{x}}{nvar(X)}.$$

With the x_i values treated as fixed in repeated sampling, this is just a fixed number, a constant, for each case i.

- So we can rewrite (4) as

$$E(b) = \beta + E(\sum_{i=1}^{n} c_i\epsilon_i) = \beta + E(c_1\epsilon_1 + c_2\epsilon_2 + \ldots + c_n\epsilon_n)$$

Clear?

- Expectations can "pass through" a constant to the random variable ϵ_i, so we have

$$E(b) = \beta + \sum_{i=1}^{n} c_i E(\epsilon_i)$$

- SO, as I said at the outset, if we are willing to make and defend the assumption that $E(\epsilon_i) = 0$ for every i, we can conclude that

$$E(b) = \beta$$

Very Important Result: If the true model generating the y_i's is $y_i = \alpha + \beta + \epsilon_i$, and the other causes ϵ_i can be treated as random variables with zero mean ($E(\epsilon_i) = 0$), then *the least squares estimator b is an unbiased estimate of* β.

- That is, if these assumptions hold, then the distribution of the least squares estimator b will be centered on the true value of the underlying "structural" slope parameter β, and so in this sense, it will be a "good" estimator for β.
- Incidentally we could also show that a is an unbiased estimator for the intercept term α.

3 Interpretation and some implications

- Two sets of observations about the meaning of this result:
  1. First, notice that to get unbiasedness of our estimator b the ONLY thing we needed to assume about the distribution of the random other causes was that $E(\epsilon_i) = 0$. Here are some important things we did NOT need to assume.
     - We didn't have to assume anything about the *shape* of the distributions which the ϵ_i's are drawn. e.g., Normal, not normal, etc.
     - We didn't need to assume that they were all drawn from the *same* distribution. e.g.: It could be that the variance of the other causes ϵ_i is larger for some values of x_i than for others, i.e., $var(\epsilon_i|x_i)$ varies systematically with the x_i's in some way. (This is what we will soon describe as *heteroscedasticity*.) For instance, it could be that the variability of growth rates for country's with high levels of ethnic fractionalization is smaller than that of countries with low levels of ethnic fractionalization, and still our estimate of β will be unbiased. (**graph grw6080 ethfrac**, draw pictures ...)
     - We didn't need to assume that the random other causes ϵ_i were *independent* across cases. That is, we didn't have to assume that $cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$. What would violation of this assumption mean substantively in a case like this one? Perhaps there are factors that affect growth rates for whole regions, so that the growth rates of all the countries in a particular region are influenced by a common factor. Then the ϵ_i's for the countries in this region will have

some commonality that will induce correlation between them. BUT, our result says that even so, our estimate of the effect of ethnic fractionalization on growth rates will be unbiased. (This general issue is called *autocorrelation*, either *spatial* as in this example, or *serial* (or *temporal*), to be discussed later.)

2. Second, if $E(\epsilon_i) = 0$, then $\alpha + \beta x_i$ is the expectation of the dependent variable Y *conditional on* x_i.
   - If we treat the other causes ϵ_i as a random variable drawn from some distribution, then this implies of course that our dependent variable Y is also a random variable produced by adding realizations of the random variable ϵ to $\alpha + \beta x_i$:

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

   - Thus, we can take expectations of both sides, and expectations pass through the α (a constant) and the βx_i (fixed X) parts:

$$\begin{aligned} E(y_i) &= E(\alpha + \beta x_i + \epsilon_i) \\ E(y_i) &= \alpha + \beta x_i + E(\epsilon_i) \\ E(y_i) &= \alpha + \beta x_i \end{aligned}$$

   - You will often see this written as:

$$E(Y|x_i) = \alpha + \beta x_i,$$

     which says that the expected value of the dependent variable Y conditional on knowing (ethnic fractionalization level) x_i is just $\alpha + \beta x_i$.
   - Compare this to the simple, unconditional expectation of Y, where μ_Y is the mean of the dependent variable Y:

$$E(Y) = \mu_Y$$

   - If Y is a random variable, then we can think it about, FPP-style, as a constant μ_Y plus a draw from a box of tickets, the "errors," that has mean zero:

$$y_i = \mu_Y + \epsilon_i$$

– Another way of thinking about what we are doing with regression is just proposing that the mean of the dependent variable depends on, or *is conditional on* another variable X in a particular way:

$$\mu_Y = \alpha + \beta x_i,$$

– Draw picture of Y distributed conditionally on X ...
– With our OLS estimators a and b, we are fitting a line that *is constructed* so that $E(e_i) = 0$, where e_i is the fitted residual for case i, i.e., $e_i = y_i - a - bx_i$.
– It is easy to show that with OLS, *by construction* $\sum e_i = 0$:

$$\begin{aligned}
e_i &= y_i - a - bx_i, \text{ so, taking sums of each side} \\
\sum e_i &= \sum y_i - \sum a - \sum bx_i \\
\sum e_i &= n\bar{y} - na - nb\bar{x} \\
\frac{1}{n}\sum e_i &= \bar{y} - a - b\bar{x} \\
\frac{1}{n}\sum e_i &= 0.
\end{aligned}$$

The last step follows from what saw in deriving the OLS estimators, that the OLS line always goes through the point of means $\bar{y} = a + b\bar{x}$.
– So the regression line $y_i = a + bx_i$ is fitted so that *it is going through the (predicted) average value of* Y *for a given* x_i.
– This is a nice way to think about it when we are interpreting the results of a regression like (in Stata) **reg grw6080 ethfrac**. It produces the estimated line

$$\begin{aligned}
\hat{y}_i &= a + bx_i \\
\hat{y}_i &= 3.41 - 1.96x_i
\end{aligned}$$

– What does this mean? It can be interpreted as follows: *On average*, a country with an ethnic fractionalization score of x_i had a 1960-80 annual growth rate of per cap income of $3.41 - 1.96x_i$. I.e., this is the mean of a country's growth rate conditional on having a level of ethnic fractionalization x_i. (Note: It is just an *unfortunate accident* that the estimated coefficient value is 1.96, the same as the distance to either side of μ that gets 95% of the area under a normal curve. Don't be confused by this ...)

– In other words: If you compare two countries, one of which has an ethnic fractionalization level .1 higher than the other, the more fractionalized country's growth rate will *on average* be about $.1(1.96) = .196 \approx .2$ percentage points lower.

– This last formulation isn't as substantively easy to assess as we would like. Here is a better way that is a typical practice when one of your variables is measured on a scale that is not incredibly intuitive (like **ethfrac**). Use **sum ethfrac ,d** to get the 10th and 50th and 90th percentiles for **ethfrac**: these are .036, .4, and .77 respectively. Thus the predicted (average) rate of 1960-80 income growth at these percentiles is

Pctile of **ethfrac**	Expected growth rate ($\hat{y}$)
10th	3.41 - 1.96*.036 = 3.34
50th	3.41 - 1.96*.4 = 2.63
90th	3.41 - 1.96*.77 = 1.90

– So we can say that a country at the 90th percentile in terms of ethnic fractionalization had, *on average* an annual growth rate about seven tenths of one per cent lower than a the median country in terms of ethnic diversity. (Over time this would actually a quite substantively significant difference, if estimate were valid.)

– The main points here are:

(a) OLS fits a line to data that gives you a "prediction" of the average value of Y given knowledge of x_i. As FPP say, the regression line is a "smoothed" graph of averages (i.e., if you divided the x variable into categories and plotted a line through the average Y for each group ... Draw ...) Under this interpretation, the estimated residuals $e_i = y_i - a - bx_i$ are prediction errors, and we would NOT say things like "if you made Botswana twice more ethnically heterogenous in 1960, the expectation of it 1960-80 growth rate would have been (this much lower)."

(b) IF we are willing to assume that there is a process that produced the data we observed that is "as if" by experiment, with the unobserved ϵ_i's representing random other causes, and IF we can assume that $E(\epsilon_i) = 0$, then OLS estimates parameters for a *causal model* that says that the mean of the dependent variable is a linear function of the independent variable.

4 Estimating the uncertainty attached to b

- OK, back to our analysis of the relationship of our estimate, b, to the unobserved true value, the structural parameter β.
  - Recall that we showed that first that if the x_i's and the other causes ϵ_i's are uncorrelated, we could expect to b to be a good estimate for β.
  - If we treat the other causes ϵ_i as realizations of a random variable (i.e., a random process), then we could say more specifically that the assumption $E(\epsilon_i) = 0$ for each i (which is sufficient for zero correlation) implies that $E(b) = \beta$.
- Under this last approach, where b is treated as a random variable (of which we see one realization, our estimate from the data we have), we can go further, to analyze the *variance* and *standard deviation* of our estimate.
- Why would this be a good thing? For just the same reason that we wanted an estimate of the variability of the sample mean $\bar{x}$ – to be able to gain a sense of whether we have a "good" estimate in the sense of a relatively precise estimate.
- In particular, an estimate of the variance (and s.d.) of b will allow us to test hypotheses such as: $H_1 : \beta < 0$, vs. $H_0 : \beta = 0$. (In words), Can we reject the null hypothesis that the unobserved structural parameter β relating ethnic fractionalization to country growth rates is actually zero rather than negative?
- Recall that, from algebra, we got that

$$b = \beta + \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \epsilon_i.$$

- Treating the other causes ϵ_i as draws of a random variable, we can also treat the OLS estimator b as a random variable also. Above, we took its expectation and found that the assumption that $E(\epsilon_i) = 0$ implied that $E(b) = \beta$.
- Now we want to ask about the *variance* of b. Since it is (being treated as) a random variable, we can take the variance

$$var(b) = var\left(\beta + \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \epsilon_i\right).$$

- β is a constant, so remembering that $var(c + X) = var(X)$ we have

$$var(b) \quad = \quad var\left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\epsilon_i\right).$$

- Remember that for notational ease and clarity we let

$$c_i \equiv \frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{x_i - \bar{x}}{nvar(X)}.$$

  so that

$$var(b) \quad = \quad var\left(\sum_{i=1}^{n} c_i\epsilon_i\right) = var(c_1\epsilon_1 + c_2\epsilon_2 + \ldots + c_n\epsilon_n).$$

- Now, again using the assumption that the x_i values are "fixed in repeated sampling" (i.e., fixed numbers rather than random variables that take new values with each hypothetical replication), this is just the variance of a sum of random variables. What is the variance of a sum of random variables?

- IF the random variables are independent, then it is just sum of their variances (recall, $var(X + Y) = var(X) + var(Y)$ provided that X and Y are like independent draws from the box).

- So, we have our next important assumption: Suppose that $E(\epsilon_i\epsilon_j) = 0$ for all $i \neq j$. That is, assume that knowing the value of any one error term would not help you predict the values of any of the others. (This would be violated in the regional-influences-on-country-economic-growth case, and (typically) in time series analysis where it is natural to assume that "other causes" in period t may still be acting in period $t + 1$.) In more typical other words:

  Assume that the other causes ϵ_i are **independent** *random variables (i.e., like independent draws from a box of tickets, so that knowing one draw doesn't help you predict any other).*

- If we are willing to (and can defend) this assumption as reasonable (doesn't have to be perfect), then the last expression can be rewritten

$$var(b) \quad = \quad \sum_{i=1}^{n} c_i^2 var(\epsilon_i) = c_1^2 var(\epsilon_1) + c_2^2 var(\epsilon_2) + \ldots + c_n^2 var(\epsilon_n).$$

- Time for the next assumption:

  Assume that the other causes ϵ_i all have the same variance, in particular, that

$$var(\epsilon_i) = \sigma^2 \text{ for all } i.$$

- What does this say? In (somewhat obscure) words, this assumption is the assumption that the errors or other causes are *homoscedastic*. "scedastic" means "scatter" in Greek, so this is the assumption that the errors have the same scatter or spread (i.e., variance). It rules out cases where different x values are systematically associated with higher or lower variance. Draw pictures of possible violations. In FPP terms, this rules out drawing the error terms from a different box for each x_i value, with the variances different across the boxes.
- If we are willing to grant/defend this assumption (and as we will see, both this assumption and the last are to an extent testable), then the last expression for $var(b)$ above becomes

$$var(b) = \sigma^2 \sum_{i=1}^{n} c_i^2 = \sigma^2(c_1^2 + c_2^2 + \ldots + c_n^2).$$

- Now we are down to figuring out what this sum of the c_i^2's is. Remembering that

$$c_i = \frac{x_i - \bar{x}}{nvar(X)},$$

  we have

$$\begin{aligned}
\sum_{i=1}^{n} c_i^2 &= \left(\frac{x_1 - \bar{x}}{nvar(X)}\right)^2 + \left(\frac{x_2 - \bar{x}}{nvar(X)}\right)^2 + \ldots + \left(\frac{x_n - \bar{x}}{nvar(X)}\right)^2 \\
&= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n^2 var(X)^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n^2 var(X)^2} \\
&= \frac{nvar(X)}{n^2 var(X)} \\
&= \frac{1}{nvar(X)}.
\end{aligned}$$

- Substituting back in, we get the important result that:

$$var(b) = \frac{\sigma^2}{nvar(X)} = \frac{\sigma^2}{\sum(x_i - \bar{x})^2} \quad (2)$$

- If you look at the middle version, part of the expression might strike you as familiar. We can write it like this

$$var(b) = \frac{\sigma^2}{nvar(X)} = \frac{\sigma^2}{n}\frac{1}{var(X)}$$

  What does the σ^2/n remind you of? variance of the sample mean. So what will happen to the precision of your estimate of β as the sample size increases? And what happens as the variance (spread) of the x_i values increases? Why? draw pictures ... (why you want variability on your independent variable: precision).

- SO, with the help of several more assumptions about the other causes ϵ_i along the way, we have derived an expression for the variance of our estimate b of β. What good is this? If we take the square root, we have the standard error of b as

$$se(b) = \frac{\sigma}{\sqrt{n}sd(X)}.$$

- If we can come up with an estimator for σ, the (unobserved) variance of the (unobserved) other causes, then we can test hypotheses about the unobserved parameter we are trying to estimate, β. Suppose we can come up with an estimate for σ^2, and let's call it s^2. (This whole procedure should be feeling somewhat familiar at this point; it is the same approach we took in analyzing the sample mean $\bar{x}$ as an estimator for the unobserved population mean μ.)

- Then, if we wanted to test an alternative against the null hypothesis that $\beta = 0$, we could formulate the test statistic

$$\begin{aligned} t &= \frac{\text{observed } b - \text{expected } b|H_0 \text{ true}}{se(b)} \\ &= \frac{b - 0}{\frac{s^2}{\sqrt{nvar(X)}}} \\ &= \frac{b}{\frac{s^2}{\sqrt{nvar(X)}}}. \end{aligned}$$

- With some additional theory about the probability distribution of the OLS estimator b (which we will supply in a minute), we will be in a position to do this. But first, let's come back to the problem of estimating σ^2, the variance of the other causes of Y (the ϵ_i's).
- Recall that the true values, the ϵ_i's, are not observed. But under the crucial assumption that $E(\epsilon_i) = 0$ (the independent variable does not covary with the unobserved other causes),

  $$e_i = y_i - (a + bx_i)$$

  is an estimator of ϵ_i. (Show a particular example, e.g., Canada ..., in Stata).

  (Question to test understanding: Why isn't $e_i = \epsilon_i$ exactly if e_i is defined as above?)
- This naturally leads to the proposal to use something like the following for an estimator of the variance of the unobserved other causes, σ^2:

  $$\sigma_s^2 = \frac{1}{n}\sum_{i=1}^{n}(e_i - 0)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2.$$

- What was the problem last time with this same approach? To get an unbiased estimator of σ^2, we have to allow for the two degrees of freedom that are "used up" in estimating a and b. It turns out that you can show that

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

  is an *unbiased* estimate for σ^2, so this is what we (and everyone else) will use.
- This quantity, s^2, is called *the residual sum of squares.* Stata shows you the square root of s^2 in the **regress** output, calling it (like FPP) *the Root Mean Squared Error.*
- Ok, so now we have a proposal for a test statistic for testing hypotheses about the unobserved slope parameter β that relates ethnic fractionalization to income growth rates. And we can implement the proposal if we want, since everything that goes into the formula for the t statistic above is from data we actually have (the y_i's and the x_i's).
- All that is left is to say what is the probability distribution of the OLS estimator b. This we can learn by looking back at the expressions we derived for b as a random variable that is a function of the unobserved other causes ϵ_i:

  $$b = \beta + \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}\epsilon_i$$

- Recall that with the "fixed X" assumption, the x_i parts of this amount to a constant

$$c_i \equiv \frac{x_i - \bar{x}}{\sum_{j=1}^{n}(x_j - \bar{x})^2} = \frac{x_i - \bar{x}}{nvar(X)},$$

so this can be rewritten

$$\begin{aligned} b &= \beta + \sum_{i=1}^{n} c_i \epsilon_i \\ b &= \beta + c_1\epsilon_1 + c_2\epsilon_2 + \ldots + c_n\epsilon_n. \end{aligned}$$

- Written this way, what can you conclude about the distribution of b? It is the sum of a sequence of independent random variables, so as n gets large it has an approximately Normal distribution! (why?) Three results follow immediately:

  1. (as we just said) The distribution of the OLS estimators a and b are *approximately Normal* as the sample size gets large, regardless of the distribution of the other determinants of the dependent variable Y (the ϵ_i's).
  2. If $Y|x_i$ has a Normal distribution (i.e., if the ϵ_i's are drawn from a Normal distribution), then b is exactly Normally distributed. (This follows why?)
  3. Our test statistic for the null hypothesis $\beta = 0$,

$$t = \frac{b}{\frac{s}{\sqrt{n}sd(x)}}$$

  will have a t distribution with $n-2$ degrees of freedom.

- (3) comes from the same logic that produced that conclusion that the test statistic for a sample mean $\bar{x}$ follows a t distribution. Remember that the t distribution has fatter tails than a Normal distribution. You can think of this coming from the fact that we only have an estimate of σ^2, s^2, rather than the real thing.

- This gets us to the point where we can test hypotheses about the unobserved structural parameters α and β.

- (Note: I won't show it (you actually can if you try), but the variance of the OLS estimate for the intercept is

$$var(a) = \frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{var(X)}\right).$$

Likewise, you can show that a is a sum of random variables and so has an approximately Normal distribution as the sample size gets larger, etc...)

5 Review and example

- Let's see how this plays out with our example from Stata: **reg grw6080 ethfrac**. Note that Stata estimates for you

  1. the coefficients a and b;
  2. the standard errors of these coefficients;
  3. the t statistics for $H_0 : \beta = 0$ and $H_0 : \alpha = 0$, which are just the coefficients divided by the appropriate standard error;
  4. the p values for the t statistics under the assumption of a two-tailed test. (What do you need to do if you are testing a one-tailed hypothesis?)
  5. 95% confidence intervals for the estimates a and b.

- This is all very convenient. Where is it coming from? Recall that we showed that the s.e. of b is

$$\frac{\sigma}{\sqrt{n}sd(X)},$$

  which we estimate with

$$\frac{s}{\sqrt{n}sd(X)},$$

  where

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - bx_i)^2}.$$

- Stata *also* reports s as the *Root Mean Squared Error*. (Be sure to read the FPP chapter on the RMSE.) We can check its calculations by generating the predicted errors ourselves: **predict e, resid** does it. **list country grw6080 e**. then **egen s2 = sum(e^2)/(136-2) if e(sample)**, **sum s2**.

- Knowing s we can also easily check on how Stata calculates the s.e. for b: **di ...**

- Again, to get the predicted values for each country, **predict yhat**. Now we can **list country grw6080 yhat ethfrac e**.

- The next practical stuff I wan't to work through is analyzing the estimated residuals, the e_i's. Recall that we made a couple of assumptions along the way about the true, unobserved other causes, the ϵ_is, and we can check and illustrate these better by looking at our estimated e_i's.

- But first, let's review the main results we've derived or stated, of which there were really quite a few:

Assumptions: IF we are prepared to assume

1. That values on the dependent variable Y are produced by a process that can be depicted as
$$y_i = \alpha + \beta x_i + \epsilon_i,$$
where
2. the ϵ_i's are random variables with $E(\epsilon_i) = 0$, and
3. are independent (i.e., $E(\epsilon_i \epsilon_j) = 0$ for all $i \neq j$), and
4. have the same variance σ^2,

THEN

Results: the OLS estimators a and b for α and β

1. are $b = cov(X, Y)/var(X)$ and $a = \bar{y} - b\bar{x}$;
2. have $var(b) = \frac{\sigma^2}{nvar(X)}$ and $var(a) = \frac{\sigma^2}{n}(1 + \frac{\bar{x}^2}{var(x)})$;
3. are *unbiased* ($E(a) = \alpha$, $E(b) = \beta$);
4. have an exactly Normal distribution when $\epsilon_i \sim N(0, \sigma^2)$;
5. have an approximately Normal distribution for large samples, regardless of the distribution of ϵ_i;
6. have standard errors that can be estimated using $s^2 = \frac{1}{n-2}\sum(y_i - a - bx_i)^2$ for σ^2;
7. have test statistics for the nulls $H_0 : \beta = 0$ and $H_0 : \alpha = 0$, $t_b = b\sqrt{n}sd(X)/s$ and $t_a = \frac{a}{s}\sqrt{\frac{nvar(X)}{\bar{x}^2 + var(X)}}$ that have t distributions with $n - 2$ degrees of freedom.