Some Notes on Regression

- What is regression?

Bivariate regression. Slope:

$$\widehat{\beta} = \frac{\frac{1}{n}\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\frac{1}{n}\sum_i (X_i - \overline{X})^2} \tag{1}$$

Intercept:

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}\overline{X} \tag{2}$$

Multivariate regression.

$$\widehat{\beta} = \left\{ \frac{1}{n}\sum_i X_i X_i' \right\}^{-1} \frac{1}{n}\sum_i X_i Y_i \tag{3}$$

In matrix notation, we write

$$\widehat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y} \tag{4}$$

but there is no real difference among these ideas. They are all just different ways of writing the same basic notion. In many ways the first way of writing it conveys the essence of the thing most simply.

- What does regression measure?

Consider the first formulation (i.e., equation (1)). Averages measure expectations in the population. So it seems reasonable to guess that $\widehat{\beta}$ and $\widehat{\alpha}$ measure

$$\beta = \frac{C[X_i, Y_i]}{V[X_i]} \tag{5}$$

$$\alpha = E[Y_i] - \beta E[X_i] \tag{6}$$

That turns out to be right. What do I mean by "measures"? I mean that if you have a really honking big sample, $\widehat{\beta}$ would be really close to β and that if the sample got yet bigger and yet bigger, $\widehat{\beta}$ would be drawn inexorably closer and closer to β. This is what is meant by the notion of a probability limit. It is like the limits you learned about in high school (e.g., the limit as n goes to infinity of $1/n$ is zero), except that the definition is more complicated because $\widehat{\beta}$ is a random variable. But the essential idea is the same. Here is the definition. Let Z_n be a random variable. We say that Z_n has probability limit μ when for every $\varepsilon > 0$

$$\lim_{n\to\infty} P(|Z_n - \mu| < \varepsilon) = 1 \tag{7}$$

That is, there is always a sample size big enough that Z_n will get within any pre-specified tolerance of μ with probability 1. When Z_n has probability limit μ, we write

$$\text{plim } Z_n = \mu \tag{8}$$

Usually the easiest way to show that Z_n has probability limit μ is to establish that the expectation of Z_n is μ and that the variance of Z_n goes to zero as n grows. This relies on a fancy result, proved in some textbooks and not in others, that convergence in mean square implies convergence in probability. Sometimes one has to resort to a direct proof, but if we need to worry about that I'll let you know.

The leading case is the sample mean, $\overline{X} = \frac{1}{n}\sum_i X_i$. The expectation of $\overline{X}$ is $E[X_i]$ and the variance is $V[X_i]/n$. Since the variance collapses to zero, $\overline{X}$ has probability limit of its expectation, which is $E[X_i]$. The same thing turns out to be true of the sample covariance: it has probability limit of the population covariance. Since the sample variance is a sample covariance, the same must be true of the sample variance. This means that the numerator and denominator in (1) are converging to the population covariance and population variance, respectively. But what about the ratio?

Here is a really useful fact about probability limits: they are continuous, in the following specific sense. Suppose $g(z)$ is a function that is continuous at $z = \mu$ and suppose that Z_n is a random variable with probability limit μ. Then the probability limit of $g(Z_n)$ is $g(\mu)$. This is not true of expectations, i.e., it is not generally true that $E[g(Z_n)]$ is equal to $g(E[Z_n])$.

Using plims and the knowledge that the ratio a/b is continuous in a and b as long as b isn't zero, we know that $\widehat{\beta}$ has probability limit β and then $\widehat{\alpha}$ has probability limit α.

• Do we like what it measures?

Often people fight about the meaning of regressions. One camp tends to write down a model for the outcome

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{9}$$

where we assume that the data are independent and identically distributed and moreover that $E[\varepsilon_i|X_i] = 0$ (and possibly that $V[\varepsilon_i|X_i] = \sigma^2$). If you haven't seen this before, this is called the "classical linear regression model" by most authors.

Note what this model assumes. The assumption that $E[\varepsilon_i|X_i] = 0$ implies immediately that

$$E[Y_i|X_i] = \alpha + \beta X_i \tag{10}$$

This is not really a "result". This is simply another interpretation of the assumption. What does it mean? It means that not only are you running the regression, you believe that you have the correct functional form. In particular, you believe the relationship between Y_i and X_i is linear. You don't believe in any curvature. Maybe you didn't realize you believed that when you wrote down the classical linear regression model. Rule # 17 of being a Ph.D. student is this: know what you are assuming. For the rules prior to # 17, see me in office hours.

So this first camp will say things like "my estimates are consistent for the parameters of the conditional expectation..." Sounds good!

The second camp feels guilty about assumptions. They all feel like an old joke of Jim Powell's, which goes like this. Setup line: What is the estimator consistent for? Punchline: It is consistent for its plim. So this second camp tends to assume that they are just running a regression and that we have all agreed to be interested in what regression measures. Apparently, if you are in the first camp, you are also in the second camp, since you like what regression measures. And yes, people from the two camps do break bread together. On a typical day, I am grumpy, and I am in the second camp.

Here is a standard argument that β and α, defined by

$$\beta = \frac{C[X_i, Y_i]}{V[X_i]} \tag{11}$$

$$\alpha = E[Y_i] - \beta E[X_i] \tag{12}$$

are interesting. One can show that an equivalent definition of α and β is

$$(\alpha, \beta) = \arg\min_{a,b} E\left[(Y_i - a - bX_i)^2\right] \tag{13}$$

Now decompose

$$(Y_i - a - bX_i)^2 = (Y_i - m(X_i))^2 + (m(X_i) - a - bX_i)^2 + 2\,(Y_i - m(X_i))\,(m(X_i) - a - bX_i) \tag{14}$$

where $m(X_i)$ is the conditional expectation of Y_i given X_i. I will take expectations of both sides of this expression. Since expectation is linear, that means the expectation of the left is the sum of the expectations of the 3 terms on the right. The third term, it turns out, has zero expectation. To see why, recall the law of iterated expectations: $E[Y_i] = E[E[Y_i|X_i]]$. Then we have

$$E\left[(Y_i - m(X_i))\,(m(X_i) - a - bX_i)\right] = E\left[E\left[(Y_i - m(X_i))\,(m(X_i) - a - bX_i)\,|X_i\right]\right] \tag{15}$$

$$= E\left[(m(X_i) - a - bX_i)\,E\left[(Y_i - m(X_i))\,|X_i\right]\right] \tag{16}$$

$$= E\left[(m(X_i) - a - bX_i) \times 0\right] = 0 \tag{17}$$

This is analogous to the Pythagorean Theorem you learned about in high school, but where (like probability limits being like "regular" limits) everything is a bit fancier. Because of this, we have the following fundamental conclusion

$$E\left[(Y_i - a - bX_i)^2\right] = E\left[(Y_i - m(X_i))^2\right] + E\left[(m(X_i) - a - bX_i)^2\right] \tag{18}$$

Going back to (12), this means that one way to view regression is that it gives predictions that are the best linear approximation to the conditional expectation $m(X_i)$.

• Inference, or the Artist Formerly Known as "How Do I Get Good Standard Errors"?

Often we spend so much time talking about how to get coefficients and how to interpret them that we forget to talk about how to measure their precision. This is not good. A lot of the reported estimates that you see in the literature have an exaggerated sense of their self-worth. This leads cynics (you know who you are) to require t-ratios to be above 5 before they pay any attention, unless they detect that the student is concerned enough to be careful in computing standard errors. See `notes2.pdf` for more on this topic.

- Why are we using regression?

One answer to this question is that we are approximating the conditional expectation. Another answer to this question is that we are trying to estimate the coefficient β, because (1) we are interested in the causal question of how much a hypothetical one unit increase in X_i will increase Y_i, on average, and (2) we believe that X_i is "exogenous" in the sense that X_i causes Y_i, Y_i does not cause X_i and there is no common factor causing both. One way to guarantee that X_i is exogenous is to have X_i be randomized. But most of the time we are willing to run regressions where X_i is not randomized.

What happens if X_i is measured with error? Consider just the bivariate regression. Suppose the following model for the measurement errors

$$Y_i = \alpha + \beta X_i^* + \varepsilon_i \tag{19}$$

$$X_i = X_i^* + u_i \tag{20}$$

where X_i^* is the "true" covariate and X_i is the observed covariate, u_i and ε_i are independent of one another and independent of X_i^*. That is, the measurement error is here conceptualized as something like a transcription error, or a machine that has some glitches to it that are idiosyncratic. Situations are sometimes more complicated and u_i can then be systematically related to X_i^* in which case conclusions are more contingent on the exact nature of the relationship. You only observe Y_i and X_i. The classical conclusion is this:

$$\text{plim } \widehat{\beta} = \frac{C[X_i, Y_i]}{V[X_i]} = \frac{C[X_i, \alpha + \beta X_i^* + \varepsilon_i]}{V[X_i]} = \beta \frac{C[X_i, X_i^*]}{V[X_i]} = \beta \frac{C[X_i^* + u_i, X_i^*]}{V[X_i]} = \beta \frac{V[X_i^*]}{V[X_i]} \tag{21}$$

which is generally smaller in magnitude that β. This is often called "attenuation bias".

Suppose we are using regression for prediction or forecasting. Then attenuation bias is a feature: we discount variation in a covariate because it contains measurement errors. Suppose we are interested in estimating the causal effect of X_i on Y_i. Then attenuation bias is a bug: we want to measure β and are sad that $\widehat{\beta}$ measures something smaller than it.

- Linear Approximations and Discrete Covariates

Suppose X_i is like education in that it takes on a small number of values but could also be viewed as "continuous". One great way to estimate the conditional expectation of Y_i given X_i in such a situation (where I am thinking you have lots of data) is to compute averages of Y_i separately for each value of X_i. Call those estimates $\widehat{\pi}$. We can obtain those estimates from a regression with a series of dummies for the values of X_i. If X_i takes on J values, then we need J dummies. Run a regression of Y_i on the J dummies and exclude the constant. Then we have $\widehat{\pi}$ in the coefficient vector and if, for example, the data are iid, then we have the variance matrix for $\widehat{\pi}$. Let the estimated variance matrix for $\widehat{\pi}$ be denoted $\widehat{\Sigma}$. Then we can view the $\widehat{\pi}$ as belonging in their own data set with J observations. Associated with the jth row of this data set is $\widehat{\pi}_j$, the estimated coefficient for the jth value of X_i, and x_j, the actual value. We could then run a regression of $\widehat{\pi}_j$ on a constant and x_j. The constant and slope from this regression *measure* the same thing as if we had run the regression of Y_i on X_i in the original microdata. Interesting. An important difference is that in the grouped data with J observations, we know the variance of the observations (from the first step regression). So we can actually view this as a GMM problem that is overidentified. We are fitting J estimates to 2 parameters (α and β). In particular, one might consider being efficient and estimating

$$\arg\min_{a,b} \left(\widehat{\pi} - a - bx\right)' \widehat{\Sigma}^{-1} \left(\widehat{\pi} - a - bx\right) \tag{22}$$

where here x denotes the column vector with jth element x_j. The minimized value of this objective function is distributed chi-square with $J-2$ degrees of freedom under the null hypothesis of linear conditional expectation.

We give a fuller treatment to this kind of idea after we have reviewed GMM, later in the term.