MIT 14.32 **J. Angrist (angrist@mit.edu)**
Spring 2023

Problem Set 5

Due: Thursday, April 27

I. 'Metrics Theory

1. Recall that regression R^2 is defined as the sample variance of fitted values, $\hat{Y}_i$, divided by the sample variance of the dependent variable, Y_i. Consider R^2 for a multivariate regression model (that is, more than one right-hand side variable). Prove that multivariate R^2 equals the square of the *coefficient of multiple correlation*, that is, $\left[CORR(\hat{Y}_i, Y_i)\right]^2$, as noted in LN8.

2. Suppose you're interested in the economic returns to college (C_i) estimated with an ability control, A_i. Specifically, you'd like to know the coefficient ρ in the long regression:

$$\ln Y_i = \alpha + \rho C_i + \gamma A_i + \eta_i, \tag{1}$$

where η_i is the long-regression residual. Alas, A_i is unobserved.

(a) Use the OVB formula to explain why a regression without control for A_i is likely to be a biased measure of the coefficient you seek.

(b) Suppose next that you've got data on Z_i, a dummy variable indicating people who were randomly awarded college scholarships. Assume that:

$$E[C_i|Z_i = 1] \neq E[C_i|Z_i = 0]$$
$$E[\gamma A_i + \eta_i|Z_i = 1] = E[\gamma A_i + \eta_i|Z_i = 0]$$

i. Why are these assumptions plausible in this case?

ii. Define $y_i = \ln Y_i$. Show that the assumptions above imply:

$$\rho = \frac{E[y_i|Z_i = 1] - E[y_i|Z_i = 0]}{E[C_i|Z_i = 1] - E[C_i|Z_i = 0]},$$

in other words, the long regression coefficient you seek is *identified* from the joint distribution of y_i, C_i, Z_i.

(c) Now, consider average y_i and C_i in each *sample* group (call these $\bar{y}_1, \bar{c}_1$ for $Z_i = 1$ and $\bar{y}_0, \bar{c}_0$ for $Z_i = 0$). Define the following *estimator* of ρ:

$$\hat{\rho}_w = \frac{\bar{y}_1 - \bar{y}_0}{\bar{c}_1 - \bar{c}_0}.$$

This is called an instrumental variables (IV) estimator, where the scholarship offer dummy, Z_i, is said to be an instrument for college attendance, C_i. IV is introduced in LN11 and MM Chapter 3. Simple IV estimators like $\hat{\rho}_w$, which can be expressed as ratios of differences in means, are also called *Wald estimators.*

i. Recall than an estimator is said to be *consistent* when it converges in probability to the thing it's meant to estimate. Briefly explain why, given the assumptions used to derive the formula for ρ in (b), $\hat{\rho}_w$ to be a consistent estimator of ρ.

ii. Suppose you have information on additional covariates, collected in the vector X_i. How might this information help you validate one of the assumptions listed in (b)?

II. 'Metrics Practice

1. This problem asks you to replicate and extend the Krueger (1993) study of the effects of computer use on wages using k93.dta, a CPS extract similar to that used in the paper and posted on the Pset5 Canvas page. Many of the relevant variables are constructed for you, but the K93 paper applies sample restrictions you'll also need to apply (e.g. regarding age and work status for all tables, plus maximum/minimum wages allowed in the sample for Table 2 onward; see Appendix A of K93 for details, though the $999 vs. $1,923 top-coding issue is taken care of). Note also that the hourly wage data used for regressions in Table II and beyond are available only for one quarter of the sample, those in CPS outgoing rotation groups.

   Include Stata logs with your solutions, while also reporting replication results alongside the original results in tables formatted like the originals, with added columns for the replication. Your replication won't be perfect. Still, your estimates should mostly be close. (For example, a coefficient of 0.090 with a standard error of 0.025 is close to a coefficient of 0.095 with a standard error of 0.029).

   (a) Replication tables

      i. Reproduce Table I with the exception of the occupation means (the definition of these variables is unclear). The K93 part-time variable differs from ours, but everything else should match closely.

      ii. Reproduce Table II. You should be able to match this closely except for the "other race" coefficient and the intercept.

      iii. Reproduce Table III. You should be able to match this reasonably closely.

   (b) Re-estimate the regressions report in column 6 of Table II without region dummies. Use Table I to explain why region dummies matter little for estimated effects of computer use. Use the R^2 version of the F-test to test the joint significance of region effects (check this with Stata's `test` command).

   (c) Compare old-fashioned (regular) and heteroskedasticity-consistent (robust) standard errors for the estimates in Columns 1 and 4 of Table II.

      i. How much does heteroskedasticity matter for inference?

      ii. Is serial correlation likely to be an issue for the estimates in Tables II and III? Why or why not?

   (d) Estimate the *change* in the returns to schooling and returns to computer use by pooling data for 1984 and 1989 and adding interactions with survey year to the models in columns 2 and 5 of Table II. Include a *year effect*, that is, a dummy for survey year, but restrict effects of regressors other than schooling and computer use to be the same in each year. Test the joint significance of your estimated year-to-year changes in effects of computer use and schooling returns. Compare your estimated changes to those implied by the results reported in Table II.

   (e) Finally, note that columns 3 and 6 of Table VII allow the returns to schooling to vary with computer use (columns 2 and 5 in Table VII report the same thing reported in columns 2 and 5 of Table II). Estimate a version of these models that allows the schooling coefficient to vary freely with both computer use and sex (this model has two second-order interactions and one third-order term). Use Stata `lincom` to estimate and contrast the returns to schooling for male and female computer users implied by this model.

2. Download the fish data (fish.dta) used in LN10 to illustrate the consequences of serial correlation for standard errors.

   (a) Replicate the OLS and CORC estimates and standard errors for the Asian ethnicity effect on wholesale fish prices reported in LN10.

   (b) Manually implement your own version of CORC by quasi-differencing. That is,

      i. Compute OLS estimates and use the residuals from these to estimate the relevant AR(1) coefficient, $\hat{\rho}$

ii. Lag the data and use $\hat{\rho}$ to quasi-difference the dependent and right-hand-side variables

iii. Re-estimate the model using quasi-differenced data; compare these results with Stata's CORC results (they should be close)

3. The Tennessee Student/Teacher Achievement Ratio experiment (another Project STAR!) randomly assigned kindergarten students and their teachers to classes of different size. This question uses data from Krueger (1999), a study that uses STAR data to estimate class size effects on achievement.

   (a) Download data file k99.dta from the Pset5 Canvas page. This file contains information on STAR participants, including data on their test scores in kindergarten (*pscore*), class size (*cs*), and an identifier for each class (*classid*). Regress test scores in kindergarten on kindergarten class size. Interpret the results.

   (b) Re-estimate the relationship between test scores and class size, using Stata to cluster standard errors at the class level. Why should you cluster? What happens to coefficient estimates and standard errors when you do?

   (c) Re-estimate the relationship between test scores and class size by using Stata to collapse student-level data to class-level means. How do the standard errors from the grouped-data regression compare to those estimated in part (b)?

   (d) Econometricians sometimes worry that regression models for group averages are especially likely to have heteroskedastic residuals. We can, of course, address this by using robust standard errors. An alternative (old-fashioned but still sensible approach) is to weight the grouped data by group size. Use Stata's `aweight` option to estimate the regression using class-level means, weighting by class size. How does weighting affect your estimated class size effects and standard errors? (MHE Section 3.4.1 discusses weighted least squares in detail).