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

Problem Set 3 (corrected 3-16-23)

Due: Tuesday, March 21 (Note the early due date)

Please submit solutions as a single PDF including Stata logs on gradescope.com.

A. Regression Theory

1. Consider the multivariate regression of Y_i on X_{1i} and X_{2i}, with slope coefficients, β_1 and β_2. In class (and in LN6), we show that
$$\beta_1 = \frac{C(Y_i, \tilde{x}_{1i})}{V(\tilde{x}_{1i})},$$
that is, multivariate β_1 is the bivariate slope coefficient from a regression of Y_i on $\tilde{x}_{1i}$, where $\tilde{x}_{1i}$ is the residual from an auxiliary regression of X_{1i} on X_{2i}.

   (a) Show that it's also true that:
$$\beta_1 = \frac{C(\tilde{y}_i, \tilde{x}_{1i})}{V(\tilde{x}_{1i})},$$
   so multivariate regression coefficients can be interpreted as removing the effect of control variables from <u>both</u> the dependent variable and the regressor of interest.

   (b) Show that, in general,
$$\beta_1 \neq \frac{C(\tilde{y}_i, X_{1i})}{V(X_{1i})}.$$
   In other words, it's not enough to remove the effect of control variables from the dependent variable <u>alone</u>.

2. Let $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ be the OLS estimates from a multivariate regression of Y_i on a constant, X_{1i}, and X_{2i}. For any nonzero constants, c_0, c_1, c_2, show that the OLS intercept and slopes from the regression of $c_0 Y_i$ on $c_1 X_{1i}, c_2 X_{2i}$ for $i = 1, 2, \ldots, n$ are given by $\tilde{\beta}_0 = c_0\hat{\beta}_0, \tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1, \tilde{\beta}_2 = (c_0/c_2)\hat{\beta}_2$. Conclude that given an initial set of estimates, the econometric consequences of changing the units in which a dependent variable or regressor is measured are known without further estimation. (Hint: This argument uses the OLS first order conditions, but does not otherwise require a lot of math.)

B. Empirical Work

1. An extract from the March 2013 Current Population Survey (CPS) is posted on Canvas. This data set contains observations on annual earnings, weeks worked last year, usual hours worked per week, age, race, and sex for men and women aged 25-59.

   (a) Limit the sample to people in their 30s and 40s. Use Stata to make a table of descriptive statistics for this sample. For the subsample of workers, construct a measure of average weekly earnings (AWE_i), average hourly earnings (AHE_i), $\ln(AWE_i)$, and $\ln(AHE_i)$, and provide descriptive stats for these as well. Report samples sizes for each variable.

   (b) Use Stata's t-test command to construct t-tests and 95 percent confidence intervals for differences in $\ln(AHE_i)$ and $\ln(AWE_i)$ by sex for white men and women aged 30-49. Interpret the magnitude of these estimates. (Tab the `race` and `sex` variables to see value labels for race and sex.)

   (c) Use Stata's regression command to generate the difference in means calculated for part b.

   (d) Denote respondents' age by A_i. Regress $\ln(AHE_i)$ on A_i and A_i^2, separately for male and female workers aged 30-49.

i. Explain why a quadratic term seems useful here.

ii. Plot the fitted values for men and women on the same axis (that is, fitted values on the y-axis, A_i on the x-axis, one plotted line for men, one for women in a different color). For whom is the estimated experience profile more steeply sloped, men or women? Verify this mathematically (Hint: this requires a simple calculation for each group).

(e) Continuing to limit the sample to workers aged 30-49, estimate the male-female wage gap using a single regression with and without controls for race, a quadratic function of age, and college graduation status. (Tab the `educ99` variable to see value labels for schooling; you'll need to decide how to code a dummy for college graduates based on these.) How do the extra control variables change your estimate of the female wage penalty?

2. Social distancing may lead to smaller classes, at least in the near term. Might be a good thing!

The Pset 3 assignment tab contains data from Angrist and Lavy (1999; AL99), a paper from Canvas Module F (the data set is called `final5`). AL99 uses the fact that Israeli class size is capped at 40 to estimate the effects of elementary school class size on pupil test scores by applying a combined instrumental variables and regression discontinuity research design. We'll learn about these advanced econometric methods later. For now, we use the AL99 data to explore regression basics.

(a) Read Angrist and Lavy (1999) through Section II (at least), download the data, and replicate the descriptive stats in Table I for 5th graders. A few helpful notes:

i. The unit of observation in this data set is a class average, but some variables such as grade-level enrollment and the pupil disadvantage index describe schools.

ii. Variables in the table are named as follows: *classize* (class size) *c_size* (grade-level enrollment) *tipuach* (disadvantage index) *verbsize* (number taking the verbal test) *mathsize* (number taking the math test) *avgverb* (verbal score) *avgmath* (math score). The data require a little clean-up and the sample should be limited as described in footnote 11 in the paper. Specifically, limit the sample to classes with 5-44 pupils and recode average scores as missing when the number of test takers equal zero. Also, recode a few scores greater than 100 by subtracting 100 when this happens. With these corrections made, and limiting the sample to the 2019 classes with non-missing math scores and class size in the relevant range, your stats should be very close to those in Table I.

(b) Economists and educators have long debated whether the high cost of extra teachers required to reduce class size is justified by student learning gains in smaller classes. What should the sign of the achievement/class-size relationship be if small classes are worthwhile? As an initial exploration, regress average math and verbal scores on class size. Comment on the sign and statistical significance of your results.

(c) A possible concern with the bivariate regression of test scores on class size is that bigger schools have larger classes and also better students. This may reflect the fact that students in Israel's dense urban centers are generally better off and hence higher-achieving than those in thinly-populated rural areas.

i. Check this by adding grade-level enrollment as a control variable. Use the OVB formula to explain precisely how and why the coefficient on class size changes when the model includes an enrollment control.

ii. Next, control for the percent of students who came from disadvantaged backgrounds instead of controlling for enrollment. (Recall that the precent disadvantaged variable name is *tipuach*.) How does this affect the class size coefficient? Use the OVB formula to explain why the coefficient changes more here than when you added enrollment.

iii. Finally, estimate the effects of class size in a model that includes both disadvantage and enrollment controls. Is the class size coefficient here closer to the one you saw in Part (i) or Part (ii) above? Why?

(d) All told, does the analysis in this question suggest that class size reductions are good, bad, or irrelevant for student learning?