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

Problem Set 4

Due: Thursday, April 13

I. Regression Theory

1. Consider the regression model:

$$\ln Y_i = \alpha + \rho C_i + \gamma X_i + \varepsilon_i, \tag{1}$$

where Y_i is worker i's weekly earnings at age 40, C_i is a dummy indicating college graduates and X_i is $i's$ family income when he or she was aged 16.

   (a) Show that ρ can be interpreted as measuring the percent change in Y_i as a function of C_i, conditional on X_i. Why might we want to condition on family income when measuring the economic returns to a college degree?

   (b) Consider a version of (1) that replaces X_i with $\ln X_i$. How should γ be interpreted in this case?

2. Consider the regression model:

$$\ln Y_i = \alpha + \beta_1 C_i + \beta_2 W_i + \beta_{12}\left(C_i \times W_i\right) + \gamma_1 X_i + \gamma_{12}(X_i \times W_i) + \varepsilon_i,$$

where W_i is a dummy for women. Explain how to use this model to test whether the economic returns to college are equal for men and women.

3. For a bivariate regression model, $Y_i = \alpha + \beta X_i + \varepsilon_i$, define the OLS estimated *fitted values*, $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$, where $\hat{\alpha}$ and $\hat{\beta}$ are the usual OLS estimates.

   (a) Prove that $\sum \hat{Y}_i e_i = 0$, where $e_i = Y_i - \hat{Y}_i$ and the sum is computed in the same sample used to compute $\hat{\alpha}$ *and* $\hat{\beta}$.

   (b) Use this fact to show that the sample variance of Y_i can be written as the sum:

$$s_Y^2 = s_{\hat{Y}}^2 + s_e^2, \tag{2}$$

   where $s_{\hat{Y}}^2$ is the sample variance of fitted values and s_e^2 is the sample variance of residuals. It's customary to label the variance of fitted values the "explained" variance associated with a particular regression model, while the variance of what's left over, e_i, is called the "unexplained" or *residual variance* associated with this model. The ratio of explained to total variance is called the regression *R-squared*, a number between zero and one (LN8 covers R-squared and related concepts in detail; see also the appendix to MM Chpt 2).

4. Assume the CEF of Y_i given a single regressor, X_i, is linear, that is, $E[Y_i|X_i] = a + bX_i$,so regression is it: $b = \beta = \frac{C(Y_i, X_i)}{V(X_i)}$ and $a = \alpha = E[Y_i] - \beta E[X_i]$. Typically, we estimate β with the OLS estimator, $\hat{\beta}_{OLS} = \frac{s_{XY}}{s_X^2}$. But there are many ways to fit a line. Here's one: split the data in half by dividing the sample into observations with values above and below median X_i. Compute above-median and below-median average Y_i and X_i; call these $\bar{y}_1, \bar{x}_1$ for means above and $\bar{y}_0, \bar{x}_0$ for means below. Finally, define an alternative slope estimator,

$$\hat{\beta}_w = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0}.$$

$\hat{\beta}_w$ is called a Wald estimator, after Hungarian mathematician Abe Wald, who introduced it in 1940.

   (a) Assuming the CEF is linear and treating X_i as fixed in repeated random samples (throughout this problem), show that $\hat{\beta}_w$ is an unbiased estimator of β.

(b) Assuming homoskedastic residuals, derive a formula for the sampling variance of $\hat{\beta}_w$ as a function of the variance of residuals and the denominator of $\hat{\beta}_w$.

(c) Continuing to assume residuals are homoskedastic, determine whether $\hat{\beta}_{OLS}$ or $\hat{\beta}_w$ is a more precise estimator of β. (Hint: the answer here is a consequence of a well-known theoretical result. No actual math needed.)

(d) More challenging: Using the same assumptions as in Q4c, compare formulas for the sampling variance of $\hat{\beta}_w$ and $\hat{\beta}_{OLS}$ directly, rather than relying on the theorem you used for part (c), (Hint: this requires math! In Q4b, you saw that the sampling variance of $\hat{\beta}_w$ is inversely proportional to the square of the denominator of $\hat{\beta}_w$. Next, note that the term $(\bar{x}_1 - \bar{x}_0)^2$ is proportional to the variance of fitted values from a regression of X_i on a dummy variable that indicates values of X_i above the median. Finally, complete the argument using the variance decomposition for regression established in Q3b, above.)

II. Empirical Work

1. The Canvas assignment tab for Pset4 contains a CSV file (PS4.csv) with observations on log weekly wages, log hourly wages, age, sex (1=male), race (1=White, 2=Black, 3=Native American, 4= Asian or Pacific Islander, 5=Other), and years of schooling for men and women aged 25-50 in the March 1992 CPS.

   (a) As mentioned in class, labor economists often model wages as a function of *potential experience*, which approximates years employed by adjusting for time out of the labor force while in school. Potential experience is defined as $potex = age - years\ of\ education - 6$. Use PS4.csv to compute *potex* and check its distribution. Set implausible values to missing, or to a plausible value that seems consistent with the underlying data.

   (b) Regress log weekly wages ($\ln uwe$) on race, potential experience and its square, and years of schooling. Use `i.race` so that Stata include dummies for all race categories.

      i. Why might you want a quadratic term in this model? Is your estimate of the quadratic term significantly different from zero? What explains the small magnitude of the estimated coefficient on experience squared?

      ii. Re-estimate this model with robust standard errors - does this change your conclusions regarding statistical significance?

   (c) Assuming the regression discussed above is a model for the relevant CEF, compute the derivative of the CEF with respect to potential experience. How does this derivative vary with potential experience?

   (d) Econometricians call the average derivative of a CEF with the respect to an independent variable the variable's *average marginal effect*. Compute the average marginal effect of potential experience on wages. Use Stata to compute a standard error for this, treating the value of average experience as non-random.

   (e) Using your estimates of the model fit in empirical Q1b, above, compute the age at which weekly wages are predicted to peak, separately for college and high school graduates. For whom do wages peak first?

   (f) More challenging:

      i. Use Stata to obtain standard errors for the two estimated peak-earnings ages computed above. How precisely are these values estimated? (Hint: find a Stata routine that computes standard errors for nonlinear functions of regression estimates).

      ii. What's the standard error of the *difference* between them?

2. Focus here on hourly wages ($\ln ahe$).

(a) Estimate a version of the model explored in empirical Q1b allowing the relationship between schooling and $\ln ahe$ to differ for men and women, including a female main effect (i.e., allow the intercept to differ for men and women). Use this model to test the hypothesis that the returns to schooling are the same for men and women.

(b) Estimate a version of the model explored in Q1b that allows the relationship between potential experience and $\ln ahe$ to differ for men and women. Include a female main effect while also allowing for sex differences in both the linear and quadratic experience terms.

i. What happens to the female main effect when experience returns differ by sex? Why is this result of economic interest?

ii. The relationship between labor market experience and wages is called an *experience profile*. Use the model you've just estimated to construct an F-test comparing male and female experience profiles. How many restrictions does the test evaluate?

3. Replication and exploration of results in Angrist, Oreopoulos, Williams (2014; AOW2014). Start by re-reading this paper.

(a) Our Canvas site has replication data for AOW2014.

i. Replicate the all-applicant balance estimates in cols 9-10 of Table 1. Report your replication results in a table showing the original published findings and your replication side-by side; the replication table therefore has four columns.

ii. Which covariates play the role of "strata controls" needed to ensure balance? Re-estimate the balance regressions without strata controls, reporting your results alongside the replication results (this adds a 5th column to your table). Does the omission of strata controls matter for balance?

iii. Replicate the all-applicant impact estimates for effects on full-year grades reported in Panel C of AOW2014 Table 4a (your replication generates 3 columns, so this replication table has 6 columns).

iv. Which covariates are included in the Table 4a models solely to increase precision? Re-estimate models corresponding to those used for (iii), above, omitting these covariates. Report these additional results below the replication results. Do covariates matter for the precision of estimated treatment effects?

(b) More challenging: Section IVC of AOW2014 (titled "Additional Results") describes a set of estimates not included in the published study. Specifically, this section starts by noting that:

*We might expect OK incentives to have been more powerful for financially constrained students. But treatment effects come out similar in subgroups defined by whether students expressed concerns about funding [*claim 1*]. Effects are somewhat larger in the subsample of students whose parents had not been to college than among those with college-educated parents, but the gap by parents' schooling is not large or precisely estimated [*claim 2*].*

Use the replication data to check these two claims. Notes: The variable `s_highfundsconcern` indicates students worried about funding. Dummies for students' parents' college degree status are are labeled `s_mothercolldegree, s_fathercolldegree,` etc.