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

Problem Set 2

Due: Thursday, March 9

Please submit solutions as a single PDF including Stata logs on Gradescope. Mark PDF pages with question numbers as explained in this video: https://www.youtube.com/watch?v=KMPoby5g_ nE and in Can and Margaret's Guide to Online Learning Tools posted on Canvas.

A. Regression Theory

1. Suppose you'd like to use a linear function of an explanatory random variable, X_i, to *predict* a dependent random variable, Y_i. Do this by choosing constants a and b to minimize mean squared *prediction error:* $MSE_p(a,b) = E\{(Y_i - [a + bX_i])^2\}$.

   (a) Show that the solution to this linear prediction problem is given by:

   $$a = \alpha \equiv E[Y_i] - bE[X_i]$$
   $$b = \beta \equiv \frac{C(Y_i, X_i)}{V(X_i)}$$

   (b) Consider instead use of a linear function to *approximate* the CEF, $E[Y_i|X_i]$. Approximate well by choosing constants a and b to minimize mean squared *approximation error*: $MSE_a(a,b) = E\{(E[Y_i|X_i] - [a + bX_i])^2\}$.

      i. Why is approximation error random? In other words, what distribution does the outer expectation in $MSE_a(a,b)$ use?

      ii. Show that the solutions to these linear prediction and linear approximation problems are the same. Hint: this is most easily done without calculus. Show directly that the same linear function minimizes both $MSE_p(a,b)$ and $MSE_a(a,b)$.

2. Suppose X_i is a Bernoulli (dummy) variable that equals one with probability p and is zero otherwise.

   (a) Prove that the $E[Y_i|X_i]$ is a linear function of X_i. What are the slope and intercept of this function in terms of CEF values?

   (b) Show that the intercept and slope in 2a equal α and β as defined in 1a.

3. The ordinary least squares (OLS) *estimator* of a bivariate regression slope coefficient is the sample analog of β, that is,

   $$\hat{\beta}_{OLS} = \frac{s_{XY}}{s_X^2},$$

   where s_{XY} denotes sample covariance and s_X^2 denotes sample variance. The sample least squares prediction problem finds the values of constants a and b that minimize the sample sum of squared errors:

   $$SSE(a,b) = \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2$$

   Show that $\hat{\beta}_{OLS}$ minimizes $SSE(a,b)$ (hence, the name "OLS").

4. Define bivariate regression residuals to be $\varepsilon_i = Y_i - \alpha - \beta X_i$, so that,

   $$Y_i = \alpha + \beta X_i + \varepsilon_i, \tag{1}$$

   where β and α are the regression slope and intercept.

(a) Assuming regressor X_i is fixed in repeated samples, use equation (1) to show that $\hat{\beta}_{OLS}$ an unbiased estimator of β. Hint: start by writing $\hat{\beta}_{OLS}$ in terms of second moments by deviating the regressor from its sample mean. Then use the linear model to substitute for Y_i in the formula for $\hat{\beta}_{OLS}$.

(b) (more challenging) Assume $E[Y_i|X_i]$ is a linear function of X_i, that is, $E[Y_i|X_i] = a + bX_i$, for some constants a and b. Also, assume X_i is random, meaning that in each new sample you get a new draw of both Y_i and X_i for each i. Assume also that observations are independent of one another.

   i. Show that $a = \alpha$ and $b = \beta$, where α and β are as defined in 1a. In other words, when the CEF is linear, regression is it!

   ii. Use this fact and the law of iterated expectations (iterating over all values of X_i in the sample) to prove that a linear CEF is a sufficient condition for $\hat{\beta}_{OLS}$ to be an unbiased estimator of β with random regressors.

B. Empirical Work

1. Regression practice

   (a) Return to the NHIS data used in Pset 1. As before, start by selecting the sample used to produce MM Table 1.1. Use regression to compare average health for husbands with and without health insurance. Construct confidence intervals for this comparison and comment on the precision of the estimated difference.

   (b) Differences in health between those with and without health insurance may be due to differences between the insured and uninsured population that arise even in the absence of insurance. Regression-adjust your comparison of husbands' health by insured status by sequentially adding controls for age, years of education, employment status, and income (these variables are named `age, yedu, empl,` and `inc` in the NHIS data set). Explain and interpret changes in the insurance coefficient as you add controls.

2. Regression in Practice

   (a) This question uses replication data from Angrist, Lang, and Oreopoulos (2009), posted on Canvas.

      i. Using Stata's `ttest` command, compare fall grades (`grade_20059_fall`) in the control group to fall grades in each of the 3 treatment groups (SSP, SFP, and SFSP). For each test, report the difference in means, the standard error of the difference in means, and the t-statistic. Which of the estimated treatment effects are significantly different from zero?

      ii. Use Stata's regression command (`regress`) to compute the same pairwise comparisons and the associated test statistics. Note that the samples differ for each pairwise comparison.

   (b) Estimate a multivariate regression of fall grades on all three treatment dummies.

      i. How do these coefficients compare to those obtained in part (a) above?

      ii. Compare your results here to the estimates reported in the first column (top panel) of Table 5 in ALO (2009). Study the table carefully: why do your estimates differ?

   (c) Dummy variables are sometimes said to be "indicator variables," because indicate when a logical condition is true. Code a dummy variable indicating students belonging to either the SFP group *or* the SFSP group. Call this dummy any-SFP. Regress fall grades on the SSP dummy and the any-SFP dummy. Interpret the magnitude of the coefficient on any-SFP in view of the estimates obtained in 2b.

   (d) Add controls for gender, high school GPA, and parents' education to the regression from part (c). How much does the addition of controls matter for estimated treatment effects from the STAR experiment? Briefly contrast and compare the importance of controls for the coefficients of interest in empirical questions 1 and 2.