Problem Set 3:
Expectations, Linear Projection, and OLS

Mitch Downey
Econometrics I

February 9, 2024

1 Problems

1. *Law of iterated expectations.*

   (a) *Pure math.* Assume that $E|y| < \infty$. Prove that $E\big(E(y|x_1)|x_1, x_2\big)$ and $E\big(E(y|x_1, x_2)|x_1\big)$ both equal $E(y|x_1)$. Explicitly state within the proof where you use the assumption.

   (b) *OLS as conditional expectations.* An intuitive example to understand why $E\big(E(y|x_1, x_2)|x_1\big) = E(y|x_1)$. Simulate 500 observations according to the following data generating process (DGP):

   - $u_0 \sim N(0, \sigma^2)$
   - $u_1 \sim N(0, 1)$
   - $u_2 \sim N(0, 1)$
   - $x_1 = u_0 + u_1$
   - $x_2 = u_0 + u_2$
   - $\varepsilon \sim N(0, 1)$
   - $y = x_1 + \beta_2 x_2 + \varepsilon$

   i. Start with $\sigma^2 = 1$ and $\beta_2 = 5$. What is the correlation between x_1 and x_2?

   ii. Assume that it is easy for researchers to get access to x_1, but that y and x_2 are difficult to get access to (they requires special permissions, or are confidential data, etc.). Ana has all three variables, while Björn only has x_1. For his project, Björn needs an individual-level prediction of y. Ana cannot provide the data, but is happy to share any estimates that Björn asks for. Predict y as a function of x_1 only, and save the fitted values, which we will call $\hat{y}^{(1)}$. This is your feasible prediction of y given only information on x_1: $E(y|x_1)$. What is the mean squared error of this prediction (which Björn cannot calculate but Ana can), which is given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i^{(1)}\right)^2$$

iii. Given how important x_2 is for y, Björn is pretty sure he can do better if he can use that information somehow. Estimate:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Ana provides these estimates to Björn, and Björn calculates

$$\hat{y}^{(2)} \equiv= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 \bar{x}_2$$

where $\bar{x}_2$ is the sample mean of x_2 (which Ana can provide). Calculate the MSE for $\hat{y}^{(2)}$.

iv. Björn realizes that he's just adding a constant, and that's not improving his estimate of y. But since x_1 and x_2 are correlated, knowing x_1 actually tells him quite a bit about x_2 at an individual-level, and he can use this information (which varies across people) to improve his estimate of y. He asks Ana to estimate:

$$x_2 = \gamma_0 + \gamma_1 x_1 + \nu$$

He then calculates the fitted values of y as:

$$\hat{y}^{(3)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 \hat{x}_2$$

where $\hat{x}_2 \equiv \hat{\gamma}_0 + \hat{\gamma}_1 x_1$. Calculate the MSE for $\hat{y}^{(3)}$.

v. Calculate the correlation coefficients of $\hat{y}^{(1)}$, $\hat{y}^{(2)}$, and $\hat{y}^{(3)}$. Relate these correlations to the result from the Law of Iterated Expectations that you proved in part (a). Given that Björn only observes x_1, can you come up with a better estimate of $\hat{y}$ for him?

(c) **(This question is optional.)** *Implications for empirical work.* You are interested in how a grant awarded to municipalities affects wages. We'll simulate grant receipt that is correlated with average education in the community, and we'll simulate that in a way that leads to differences in average education across municipalities (since we already know that if education is iid across municipalities then the WLLN implies there will be no variation in municipality-level average education). Ultimately, wages (y) will be a function of grants (g), education (x_1), some other factor that we don't observe (x_2), and an idiosyncratic iid individual-level error term. Simulate data according to the following multilevel DGP, with 50 municipalities and 100 individuals in each municipality:

- Municipality-level stuff:
  - $\mu_{1,m} \sim N(0, \sigma_1^2)$. This will be education.
  - $\mu_{2,m} \sim N(0, \sigma_1^2)$. This will be some other thing we don't observe.
  - $g_m \sim binom(p_m)$ (i.e., $g_m \in \{0, 1\}$ with $Pr(g_m = 1) = p_m$), where $p_m = \frac{\mu_{1,m} - min(\mu_1)}{max(\mu_1) - min(\mu_1)}$. This ensures that the probability of g_m is linear in $\mu_{1,m}$.
- Individual-level stuff:
  - $x_{1,i,m} \sim N(\mu_{1,m}, 1)$. This is the education of individual i living in municipality m.

- $x_{2,i,m} \sim N(\mu_{2,m}, 1)$. This is the other characteristic for individual i living in municipality m.
- $e \sim N(0, \sigma_e^2)$.
- $y_{i,m} = \beta_0 + \beta_g g_m + \beta_1 x_{1,i,m} + \beta_2 x_{2,i,m} + e_{i,m}$. This is the wage equation.

i. Simulate the data, starting with $\beta_0 = \beta_g = \beta_1 = \beta_2 = \sigma_e^2 = \sigma_1^2 = \sigma_2^2 = 1$. Regress wages on g, x_1, and x_2, and verify that $\hat{\beta}_g$ is close to 1. Note that we are assuming you cannot run this regression because you don't observe x_2.

ii. Regress wages on g_m only, and verify that $\hat{\beta}_g$ is biased.

iii. Regress wages on g_m and $x_{1,i,m}$. Verify that this $\hat{\beta}_g$ is closer to β_g than what you got in ii above.[1]

iv. Calculate $\bar{x}_{1,m} \equiv \frac{1}{n_m} \sum_{i=1}^{n_m} x_{1,i,m}$ as the average level of education in the municipality. Regress wages on g_m and $\bar{x}_{1,m}$. Compare the coefficient $\hat{\beta}_g$ to what you got in parts ii and iii. Note the R^2 from this regression. Clearly individual-level education is an important determinant of individual-level wages, but does controlling for municipality-level education only really under-perform controlling for the individual-level variable? It's often helpful to re-run the simulation a few times over and over to get an informal sense of the distribution of the estimates?

v. Play with the parameter values for $\beta_0, \beta_g, \beta_1, \beta_2, \sigma_e^2, \sigma_1^2, \sigma_2^2$. Change the values a bit, and for some set of values, run 50 iterations of the simulation under each value. Create one figure showing a result that you consider interesting.

vi. Set $\beta_g = 0$ and keep all other values as they were in i. Run the full simulation 100 times. How often is the estimated coefficient $\hat{\beta}_g$ statistically significant at the 5% level? Note that the true $\beta_g = 0$, so $\hat{\beta}_g$ "should" only be significant 5% of the time.[2]

vii. Now calculate averages wages and average education, and estimate

$$\bar{y}_m = \beta_0 + \beta_g g_m + \beta_1 \bar{x}_{1,m} + \nu_m$$

How often is the estimated coefficient $\hat{\beta}_g$ statistically significant at the 5% level?

2. *FWL, OV'B', and LATE's.* You are interested in the causal effect of parental income on children's outcomes. You will simulate the data, so you know the truth and can compare that with regression results. Simulate 500 observations according to the following data generating process (DGP):

- Earnings $\sim N(19, 1)$. Note that if I later refer to "labor market earnings," I am referring to this variable.
- Capital gains $\sim N(1, 1)$

[1] If you're curious, you could simulate g_m with $p_m = \frac{\mu_{1,m}^2 - min(\mu_1^2)}{max(\mu_1^2) - min(\mu_1^2)}$. Now, the probability of a grant is not simply a linear function of x_1, and you can verify that controlling for x_1 makes less of a dent in the bias in $\hat{\beta}_g$.

[2] This question was inspired by Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. "How much should we trust differences-in-differences estimates?" *The Quarterly Journal of Economics* 119.1 (2004): 249-275.

- $u \sim N(0,1)$
- $e \sim N(0,1)$
- Occupational status = Earnings + u
- Child Outcomes = Earnings - Capital gains + e. Note that this means that labor market earnings are good for children (improve their outcomes), while capital gains are actually bad for them (perhaps because they are unearned and send a bad signal to them about the value of hard work.
- Income $\equiv$ Earnings + Capital gains

(a) On average, across all individuals in your simulated sample, what fraction of income comes from labor market earnings?

(b) Note that the average respondent will have income of 20. Given that you simulated the data, what would you say is the true causal effect on child outcomes of increasing the average person's income by 10%, from 20 to 22?

(c) Regress child outcomes on earnings and capital gains and verify that OLS recovers the correct coefficients. Verify that controlling for occupational status (which is not part of the DGP) does not affect these coefficients.

(d) A researcher is not interested in the potential distinction between earnings and capital gains, and she pools both into income. Regress child outcomes on income, and compare the coefficient to your answer to (b) above. Is income endogenous?

(e) Within your sample, what is the correlation between occupational status and income?

(f) The researcher is concerned that she should be controlling for occupational status: it is highly correlated with income and excluding it might cause omitted variable bias (OVB). In this univariate context, OVB is a function of three terms, and since you simulated the data, you know what all three terms are. Analytically (by hand, without a computer) calculate the OVB that results from excluding occupational status.

(g) Regress child outcomes on income, controlling for occupational status, and compare the coefficient to your answer to (f) above. Does the researcher conclude that there is OVB? Has she now recovered the causal effect of income on child outcomes? Compare your answer to your answer to (b) above, and discuss the role of endogeneity.

3. *Measurement error and indices.* Assume that x and y are two mean zero variables. Suppose that the true model is given by $y = \beta x + \varepsilon$ where $E(x'\varepsilon) = 0$. Let σ_x^2 and σ_ε^2 be the variance of x and ε, respectively.

(a) You do not observe the true x. Instead, you observe x only with error. You observe $\tilde{x} = x + \nu$ where ν is a mean zero white noise error term[3] with variance σ_ν^2. You regress y on $\tilde{x}$. Write $plim\hat{\beta}$ as a function of β, σ_ν^2, σ_x^2, and σ_ε^2 (not all of those terms will show up in the expression). Interpret the result. Note: This is called classical measurement error, and the bias in $\hat{\beta}$ is called attenuation bias.

[3] "White noise error term" means it is iid and independent of all other variables in the model.

(b) You **do** observe the true x, but you do not observe the true y. You observe y only with error: You observe $\tilde{y} = y + e$ where e is a mean zero white noise error term with variance σ_e^2. You regress $\tilde{y}$ on x. Write $plim\hat{\beta}$ as a function of β, σ_ν^2, σ_x^2, and σ_ε^2 (not all of those terms will show up in the expression). Interpret the result.

(c) (**Note**: I do not currently know how much algebra this one is. If it's crazy, please give up.) You observe x only with error. You observe $\tilde{x} = x + \nu$ where ν is a mean zero and has a correlation of ρ with ε. You regress y on $\tilde{x}$. Write $plim\hat{\beta}$ as a function of β, σ_ν^2, σ_x^2, σ_ε^2, and ρ (not all of those terms will show up in the expression). Interpret the result. Note: This is called non-classical measurement error because the error in your measure of x is systematically correlated with the dependent variable.

(d) Return to the setup of 3ai: Classical measurement error, in which only x is measured with error and that error is white noise. Suppose you observe some z, which is correlated with x but not correlated with ν.[4] You use z as an instrument for x. Let $\hat{\beta}_{IV}$ be the two-stage least squares estimate (from the second stage) of the coefficient on x. Write $plim\hat{\beta}_{IV}$ as a function of β, σ_ν^2, σ_x^2, σ_ε^2, and σ_z^2 (not all of those terms will show up in the expression).

[4]Note that our assumption that $E(x'\varepsilon) = 0$ and $E(x'z) \neq 0$ implies that $E(z'\varepsilon) = 0$.