Assignment 5

Michael Cai

March 22, 2016

1a) Consider the regression $Y_i = \beta_0 + \beta_1 X_{1i} + ui$. Do you think that $E[u_i|X_i] = 0$? Is the OLS estimator of β_1 unbiased and consistent? Explain.
$E[u_i|X_{1i}] \neq 0$ because there is important information contained in the u_i that is not controlled for by random selection to create the classes, namely X_{2i}. If current students and new students were randomly and evenly placed into small/regular sized classes, then there would be no bias and there would be consistency in the estimation of β_1 ;however, because 20% of the new students are assigned to the small-sized classes and 80% are assigned to the regular-sized classes, there are more new students in regular sized classes than there are in small classes. This is a case of sample selection bias, and thus the OLS assumption 1 is violated and consequently the OLS estimator of β_1 is biased and inconsistent.

1b) Consider the regression $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ui$. Do you think that $E[u_i|X_{1i}, X_{2i}]$ depends on X_1? Is the OLS estimator of β_1 unbiased and consistent? Explain. Do you think that $E[u_i|X_{1i}, X_{2i}]$ depends on X_2? Will the OLS estimator of β_2 provide an unbiased and consistent estimate of the causal effect of transferring to a new school (that is, being a newly enrolled student)? Explain.
The first two questions basically ask: $E[u_i|X_{1i}, X_{2i}] = E[u_i|X_{2i}]$ and $\hat{\beta}_1 \rightarrow_p \beta_1$. Or rather, if you were to control for whether or not a student was newly enrolled, does the class size have explanatory power over test scores. The answer is yes, and thus the conditional mean independence does not hold. The class size, as we spoke about during lecture, still proxies for many other aspects of a school district such as income level, which is positively correlated with test scores and has a non-zero covariance with class size (assuming wealthier school districts can afford more instructors and thus more personalized classroom experiences resulting in a smaller student to teacher ratio). Therefore, the OLS estimator of β_1 will be biased and inconsistent.

The second two questions ask: $E[u_i|X_{1i}, X_{2i}] = E[u_i|X_{1i}]$ and $\hat{\beta}_2 \rightarrow \beta_2$. Or rather, if you were to control for the class size, does the fact that a student was newly enrolled have explanatory power over test scores. The answer is likely no, assuming that these new students enrolling are not all coming from say a non-English speaking background. Therefore, the conditional mean independence holds, and we can thus use the following derivation to prove that $\hat{\beta}_2$ is a consistent and unbiased estimator of β_2.

Assuming these are true:
$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$
$u_i = E[u_i|X_{1i}, X_{2i}] + v_i$
Therefore:
$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E[u_i|X_{1i}, X_{2i}] + v_i$
And using the conditional mean independence assumption that: $E[u_i|X_{1i}, X_{2i}] = E[u_i|X_{1i}]$
$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + E[u_i|X_{1i}] + v_i$
Assuming that $E[u_i|X_{1i}]$ follows the form $a + bX_{1i}$ we get:
$Y_i = (\beta_0 + a) + (\beta_1 + b)X_{1i} + \beta_2 X_{2i} + v_i$
Therefore, we have proven that although the intercept and β_1 estimator will be biased, the estimator for β_2 will be an unbiased and consistent estimator for the causal effect of transferring to a new school (being a newly enrolled student).

2. (All written)

3a) Estimate three regression models and construct 95% confidence intervals for the estimated effect of smoking on birth weigh using each of the three regressions.

$[\beta_1 - 1.96(s.e.\beta_1), \beta_1 + 1.96(s.e.\beta_1)]$

1. [-305.83636, -200.61964]
2. [-268.851348, -166.308852]
3. [-228.075224, -122.678576]

3b) Does the coefficient on smoker in the first and second regressions suffer from omitted variables bias? Explain.

The coefficient on smoker in the first regression suffers from omitted variables bias because knowing that a person, who is pregnant with a child, is also a smoker should also imply that that person is engaged in other activities, such as drinking during pregnancy (as shown in the second regression), and other reckless behaviors that would likely adversely affect birthweight. The second regression, which does include alcohol consumption and also pre-natal visits, which I am assuming is a proxy for how responsible the mother is in regards to checking on her baby's health, still might suffer from the similar issue of OVB due to other behaviors that may not have been taken into account, for instance junk food or unhealthy food consumption, but the strongest indicators for reckless behavior are more or less accounted for in the second regression.

3c) Consider the coefficient on unmarried in the third regression. A family advocacy group notes that the large coefficient suggests that public policies that encourage marriage will lead, on average, to healthier babies. Do you agree?

I do not necessarily agree, and I do believe that the unmarried variable acts as a control variable as opposed to a regressor. Whether or not a person is unmarried - all else held constant - should not affect the babies health or birthweight. Unmarried is more likely a proxy for other important life circumstances that the mother is going through during the pregnancy, for instance stress levels would likely be higher in an unmarried pregnancy because the mother is likely a single mother, or is dealing with the stress of figuring out a pregnancy with a boyfriend (which I am assuming is on average a less stable relationship than one with a legal spouse). Unmarried can also be controlling for income as well, since on average single pregnant women tend to be of lower income and social status than married pregnant women. And lastly, unmarried could also be a proxy for age and thus related to the health of the mother since it is more likely that younger people will be unmarried as opposed to older people.