651 Empirical Economics: Assignment 3 Theory Solutions

Per Hjertstrand* Andrew Proctor†

October 2017

15.1 i PC ownership is likely correlated with unobserved determinants of GPA for a number of reasons, perhaps most centrally income/socioeconomic status, but also perhaps academic interest, major, etc.

ii PC ownership is likely related to parents' annual income because college students budget sets are likely determined in large part by parental income/wealth. All else being equal, greater parental income is then likely to predict greater PC ownership.

Parental income is not a good instrument for the role of PC ownership, however, because it also predicts a great many other things, such as access to other resources, parental education (and therefore perhaps student educational investment before college, family "preferences for education", etc"), and even things such as outside options beyond doing well in school.

iii You can specify a first-stage as follows:

$$PC_i = \gamma_0 + \gamma_1 granted_i + v_i,$$

where granted indicates a student was randomly assigned to receive a computer. Because of this randomization, *granted* is therefore a valid (exogenous) predictor of computer ownership, but also inherently a relevant predictor of computer ownership.

15.2 i This likely depends on the housing regulations of the university, but most likely not. Supposing that university students could live off-campus, then living in on-campus vs off-campus (and thus more distant) housing is likely predictive of the relative prices of the two types of housing. Morever, among those living in off-campus housing, depending on where the university is located and where "desirable" neighborhoods are located, distance is likely a predictor of wealth and potentially other preferences

*Per.Hjertstrand@ifn.se

†Andrew.Proctor@phdstudent.hhs.se

& characteristics. Even in strictly on-campus housing, if housing assignment is non-random, housing distance may be a predictor of class-standing, wealth, or participation in groups/fraternities, etc.

ii In the vocabulary of the seminars, the exogeneity assumption is equivalent to validity. Here, Wooldridge uses validity to mean both relevance and validity, hence we require the relevance assumption that the instrument has a nonzero effect on class attendance.

iii Since $priGPA$ is assumed to be exogenous, then $priGPA * dist$ should be a good IV for $priGPA * atndrte$.

15.3 Equation 15.10 states:

$$\hat{\beta}_1 = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})}$$

Taking first the numerator, we have:

$$\sum(z_i - \bar{z})(y_i - \bar{y}) = \sum z_i(y_i - \bar{y}) - \sum \bar{z}(y_i - \bar{y}) = \sum z_i(y_i - \bar{y}) - \bar{z}\sum(y_i - \bar{y})$$
$$\sum(z_i - \bar{z})(y_i - \bar{y}) = \sum z_i(y_i - \bar{y}) - \bar{z}(n\bar{y} - n\bar{y}) = \sum z_i(y_i - \bar{y}) - 0 = \sum z_i(y_i - \bar{y})$$

Now using the z_i is binary and that $\bar{y_1}$ and $\bar{y_0}$ are the averages for $z_i = 1, 2$, respectively, we have:

$$\sum(z_i - \bar{z})(y_i - \bar{y}) = \sum z_i(y_i - \bar{y}) = \sum_{z_i=0}(z_i = 0)((y_i|z_i = 0) - \bar{y}) + \sum_{z_i=1}(z_i = 1)((y_i|z_i = 1) - \bar{y})$$

$$\sum(z_i - \bar{z})(y_i - \bar{y}) = \sum_{z_i=0} 0(y_{i0} - \bar{y}) + \sum_{z_i=1} 1(y_{i1} - \bar{y}) = \sum_{z_i=1}(y_{i1} - \bar{y}) = n_1(\bar{y}_{i1} - \bar{y}),$$

where n_1 is the number of observations for which the instrument is equal to 1.

Now, note: $\bar{y}_i = \frac{n_1\bar{y}_{i1} + n_0\bar{y}_{i0}}{n_1 + n_0}$
Hence:

$$\sum(z_i - \bar{z})(y_i - \bar{y}) = n_1(\bar{y}_{i1} - \bar{y}) = n_1\Big(\bar{y}_{i1} - \frac{n_1\bar{y}_{i1} + n_0\bar{y}_{i0}}{n_1 + n_0}\Big)$$
$$\sum(z_i - \bar{z})(y_i - \bar{y}) = n_1\Big(\frac{(n_1 + n_0)\bar{y}_{i1}}{n_1 + n_0} - \frac{n_1\bar{y}_{i1} + n_0\bar{y}_{i0}}{n_1 + n_0}\Big)$$
$$\sum(z_i - \bar{z})(y_i - \bar{y}) = n_1\Big(\frac{n_0\bar{y}_{i1} - n_0\bar{y}_{i0}}{n_1 + n_0}\Big) = \frac{n_1 n_0}{n_1 + n_0}\Big(\bar{y}_{i1} - \bar{y}_{i0}\Big)$$

Proceeding to the denominator, it is clear that $\sum \bar{z}(x_i - \bar{x})$ also sums to zero analogously to the numerator. Hence:

$$\sum(z_i - \bar{z})(x_i - \bar{x}) = \sum z_i(x_i - \bar{x}) = \sum_{z_i=0} 0(x_{i0} - \bar{x}) + \sum_{z_i=1} 1(x_{i1} - \bar{x}) = n_1(\bar{x}_{i1} - \bar{x})$$

Again following the same procedure as for y_i, we get:

$$\sum(z_i - \bar{z})(x_i - \bar{x}) = \frac{n_1 n_0}{n_1 + n_0}\Big(\bar{x}_{i1} - \bar{x}_{i0}\Big)$$

So then:

$$\hat{\beta}_1 = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})(x_i - \bar{x})} = \frac{\frac{n_1 n_0}{n_1 + n_0}\Big(\bar{y}_{i1} - \bar{y}_{i0}\Big)}{\frac{n_1 n_0}{n_1 + n_0}\Big(\bar{x}_{i1} - \bar{x}_{i0}\Big)} = \frac{\bar{y}_{i1} - \bar{y}_{i0}}{\bar{x}_{i1} - \bar{x}_{i0}}$$

15.7 i We are not told much about the type of school choice program. Broadly speaking, families choose schools for a number of reasons, including not only income, but preferences/tastes for education, distance, school composition, etc. Schools themselves may also have defined income criteria. Hence, without knowing more about the school choice program, we have no reason to assume that attendance in a given school of a school choice program is random conditional on income.

ii Yes, to see this: consider the grants as a function of income:
First,to express the grants as a function of income, we can write:

$$grant = \begin{cases} G_1, & \text{for } 0 \leq faminc \leq c_1 \\ G_2, & \text{for } c_1 \leq faminc \leq c_2 \\ ... & \\ G_n, & \text{for } c_{n-1} \leq faminc \leq c_n \end{cases}$$

Lets denote the grant function as a function of the $faminc$ variable: $grant = G(faminc)$.
Question 15(ii) can then be formulated as:

$$Cov\big(u, G(faminc)\big) = 0$$

Note: $Cov\big(u, G(faminc)\big) = E[uG(faminc) - E(u)E(faminc)$
And, as always, $E(u) = 0$, so:

$$Cov\big(u, G(faminc)\big) = E[uG(faminc) - 0E(faminc) = E[uG(faminc)]$$

The Law of Iterated Expectations states that for any two variables, x and w, with g(x) some function of x:

$$E[wg(x)] = E[E[wg(x)|x] = E[g(x)E[w|x]]$$

Applying that to this question, we have:

$$Cov\big(u, G(faminc)\big) = E[uG(faminc)]$$

$$Cov\big(u, G(faminc)\big) = E[E[uG(faminc)|faminc]]$$

$$Cov\big(u, G(faminc)\big) = E[G(faminc)E[u|faminc]] = E[G(faminc)E[0]] = 0$$

Hence, $Cov\big(u, G(faminc)\big) = Cov(u, grant) = 0$.

iii It is worth noting that Wooldridge's terminology here departs from more typical usage and can be a bit confusing. By saying the reduced form for choice, what Wooldridge means is the first stage equation:

$$choice = \pi_0 + \pi_1 grant + \pi_2 faminc + v_i$$

iv *This* is what is generally referred to as the reduced form equation:

$$score = \gamma_0 + \gamma_1 grant + \gamma_2 faminc + \eta_i$$

15.8 i There are a great many possible answers. Some factors that you might want to control for include whether the school is private, school revenues, perhaps geographical location (eg county or region), student demographics and parental wealth, class sizes, etc.

Of course, all of these variables are also problematic in that they are likely endogenous to unobserved characteristics.

ii $score = \alpha + \beta_1 girlhs + \beta_2 private + \beta_3 schoolrev + \beta_4 pcrev + \beta_5 classsize + \beta_6 county + \beta_7 black + \beta_8 hispanic + \beta_9 nonnative + \beta_{10} parentalinc + u_i$

iii It is quite likely that parental support and motivation are correlated with $girlhs$, as thse factors concern both the accessibility of the non-standard placement into a girls' high school, and the desire to seek out a schooling type considered to be better aligned to the student's needs.

iv One needs to assume the that the availability of girls' high schools in a 20 mile radius is unrelated to the unobserved factors in the structural equation. This again seems somewhat unlikely as distance to girls' high schools is likely affected by issues like urbanity and choice of neighborhood, which is likely related to parental SES and preferences in ways beyond what we can reasonably control for.

v Probably not. The operative theory that motivated this choice of IV is that nearby availability of girls high schools makes it easier to attend a girls high school (therefore increasing the propensity to enroll), in a manner unrelated to unobserved factors in the structural equation. When nearby availability is estimated to have a negative effect on enrollment, this contradicts the logic behind our instrument and suggests it may be related to the decision to enroll in other ways (which may be relevant unobserved predictors in the structural equation).

SW1 a Relevance: $Cov(Z_i, X_i) \neq 0$

b Identification: There are at least as many excluded instruments as endogenous regressors.

c No perfect collinearity.

d Validity (or exogeneity): $Cov(Z_i, X_i) = 0$

SW2 a Let $Z_i = X_i$. Then by assumption MLR.4, $Cov(Z_i, u_i) = 0$.

b If we write down the first-stage, we have:

$$X_i = \pi_0 + \pi_1 X_i + v_i$$

Clearly then, we have $\pi_1 = 1 \neq 0$ (relevance) [we also have $\pi_0 = 0$ and $v_i = 0$].

c There are multiple ways to show this: Perhaps the easiest is:

$$\beta^{IV} = \frac{Cov(Y_i, Z_i)}{Cov(Z_i, X_i)} = \frac{Cov(Y_i, X_i)}{Cov(X_i, X_i)} = \frac{Cov(Y_i, X_i)}{Var(X_i)} = \beta^{OLS}$$