Department of Economics UN3412
Columbia University Fall 2021

SOLUTIONS TO Problem Set 7
Introduction to Econometrics
Seyhan Erden and Tamrat Gashaw
for all sections.

1. (33p) Do you think attendance to lectures affects performance on final exam? A model to explain the standardized outcome on a final exam (*stndfnl*) in terms of percentage of classes attended, prior college grade point average, and ACT score is

$$stndfnl = \beta_0 + \beta_1\, atndrte + \beta_2\, priGPA + \beta_3 ACT + u$$

where variables are given in the following table:

Variable	Definition
stndfnl	Standardized final exam score
atndrte	Percentage of classes attended
priGPA	Prior college grade point average
ACT	Achievement Test score
priGPAxatndrte	Prior GPA times attendance rate

(a) (10p) Let *dist* be the distance from the students' living quarters to the lecture hall. Do you think *dist* is uncorrelated with u?

Sol: dist would probably be correlated with u. The students' attitude towards their study would influence their stndfnl a lot, and also it is generally the case that students who care more about their study would be more inclined to live near to the lecture hall.

if students argue that students do not have control over the location of their living quarters (e.g., if the school assigns dorms without student input), then we could argue that dist is uncorrelated with u.

(b) (10p) Assuming that *dist* and u are uncorrelated, what other assumption must *dist* satisfy in order to be a valid IV for *atndrte*?

Sol: To be a valid IV for atndrte, dist also needs to be correlated with atndrte.

(c) (13p) Suppose we add the interaction term *priGPAxatndrte:*

$$stndfnl = \beta_0 + \beta_1\, atndrte + \beta_2\, priGPA + \beta_3 ACT + \beta_4\, priGPAxatndrte + u$$

If *atndrte* is correlated with u, in general, so is *priGPAxatndrte*. What might be a good IV for *priGPAxatndrte*? [Hint: if E(u | *priGPA,ACT,dist*)=0, as happens when *priGPA, ACT* and *dist* are all exogenous, then any function of *priGPA* and *dist* is uncorrelated with u]

*Sol: we could use priGPA*dist as the instrument variable.*

2. (34p) The purpose of this question is to compare the estimates and standard errors obtained by correctly using 2SLS with those obtained using inappropriate procedures. Use the data file WAGE2.dta, variables are:

Variable	Definition
wage	Monthly earnings
educ	Years of education
exper	Years of working experience
tenure	Years with current employment
black	=1 if the person is African-American, 0 otherwise
sibs	Number of siblings

(a) (12p) Use a 2SLS routine to estimate the equation

$$\text{Log}(wage) = \beta_0 + \beta_1\, educ + \beta_2\, exper + \beta_3\, tenure + \beta_4\, black + u$$

using *sibs* as the IV for *educ*. Report your results

sol: The IV (2SLS) estimates are

*log(wage)_hat = 5.22 + 0.0936*educ +0.0209*exper + 0.0115*tenure -0.183*black*

(0.54) (0.0337) (0.0084) (0.0027) (0.050)

n=935; $R^2 = 0.169$

(b) (12p) Now, manually, carry out 2SLS. That is, first regress *educ* on *sibs, exper, tenure* and *black,* obtain the fitted values *educ_hat,* then run the 2nd stage regression log(*wage*) on *educ_hat, exper, tenure* and *black.* Verify that estimated sample coefficients are identical to those obtained form part (a) but that the standard errors are somewhat different. The standard errors obtained from the second stage regression when manually carrying out 2SLS are generally inappropriate. Are the standard errors larger or smaller than those from Part a? Explain.

sol: The coefficient on educ_hat in the second stage regression is, naturally, 0.0936. But the reported standard error is 0.0353, which is slightly too large.

(c) (10p) Now, use the following two-step procedure, which generally yields inconsistent parameter estimates of β_j and not just inconsistent standard errors. In step one, regress *educ* on *sibs* only and obtain the fitted values, say *educ_tilde* (Note that this is an incorrect first stage regression). Then, in the second step, run the regression of log(*wage*)

on *educ_tilde*, *exper*, *tenure* and *black*. How does the estimate from this incorrect, two-step procedure compare with the correct 2SLS estimate of the return to education?

when instead we (incorrectly) use educ_tilde in the second stage regression, its coefficient is 0.0700, and the corresponding standard error is 0.0264. Both are too low. The reduction in the estimated return to education from about 9.4% to 7.0% is not trivial. This illustrates that it is best to avoid doing 2SLS manually.

Do file for question #2
ivregress 2sls lwage exper tenure black (educ = sibs), r

reg educ sibs exper tenure black
predict educ_hat
reg lwage educ_hat exper tenure black

reg educ sibs
predict educ_tilde
reg lwage educ exper tenure black

3. (33p) By studying the probability limit (plim) of the IV estimator we can see that when Z and u are possibly correlated, we can write

$$plim\ \hat{\beta}_{1,IV} = \beta_1 + \frac{Corr(Z,u)}{Corr(Z,X)}\frac{\sigma_u}{\sigma_x} \qquad (1)$$

where σ_u and σ_x are the standard deviation of u and X in the population, respectively. The interesting part of this equation involves the correlation terms. It shows that, even if Corr(Z,u) is small, the inconsistency in the IV estimator can be very large if Corr(Z,X) is also small. Thus, even if we focus only on consistency, it is not necessarily better to use IV than OLS if the correlation between Z and u are smaller than that between X and u. Using the fact that $Corr(X,u) = Cov(X,u)/(\sigma_u.\sigma_x)$ along with the fact that plim $\hat{\beta}_1 = \beta_1$ + Cov(X,u)/Var(X) = β when $Cov(X,u) = 0$, we can write the plim of OLS estimator – call it $plim\ \hat{\beta}_{1,OLS}$ - as

$$plim\ \hat{\beta}_{1,OLS} = \beta_1 + Corr(X,u)\frac{\sigma_u}{\sigma_x} \qquad (2)$$

Assume that $\sigma_u = \sigma_x$, so that the population variance in the error term is the same as it is in X. Suppose the instrumental variable, Z, is slightly correlated with u: $Cov(Z,u) = 0.1$. Suppose also that Z and X have somewhat stronger correlation: $Cov(Z,X) = 0.2$.

(i) (17p) What is the bias in the asymptotic IV estimator?

(ii) (16p) How much correlation would have to exist between X and u before OLS has more asymptotic bias than TSLS?

(i) From equation (1) with $\sigma_u = \sigma_x$, plim $\hat{\beta}_{1,IV} = \beta_1 + (.1/.2) = \beta_1 + .5$, where $\hat{\beta}_{1,IV}$ is the IV estimator. So the asymptotic bias is .5.

(ii) From equation (2) with $\sigma_u = \sigma_x$, plim $\hat{\beta}_{1,OLS} = \beta_1 + \text{Corr}(x,u)$, where $\hat{\beta}_{1,OLS}$ is the OLS estimator. So we would have to have Corr(x,u) > .5 before the asymptotic bias in OLS exceeds that of IV. This is a simple illustration of how a seemingly small correlation (.1 in this case) between the IV (Z) and error (u) can still result in IV being more biased than OLS if the correlation between Z and X is weak (.2).

The following questions will not be graded, they are for you to practice and will be discussed at recitation:

1. SW Exercise 12.2

(a) When there is only one X, we only need to check that the instrument enters the first stage population regression. Since the instrument is Z = X, the regression of X onto Z will have a coefficient of 1.0 on Z, so that the instrument enters the first stage population regression. Key Concept 4.3 implies
corr(X_i, u_i) = 0, and this implies corr(Z_i, u_i) = 0. Thus, the instrument is exogenous.

(b) Condition 1 is satisfied because there are no W's. Key Concept 4.3 implies that condition 2 is satisfied because (X_i, Z_i, Y_i) are i.i.d. draws from their joint distribution. Condition 3 is also satisfied by applying assumption 3 in Key Concept 4.3. Condition 4 is satisfied because of conclusion in part (a).

(c) The TSLS estimator is $\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}}$ using Equation (12.4) in the text. Since $Z_i = X_i$, we have

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} = \frac{s_{XY}}{s_X^2} = \hat{\beta}_1^{OLS}.$$

2. SW Exercise 12.5

(a) Instrument relevance. Z_i does not enter the population regression for X_i

(b) Z is not a valid instrument. $\hat{X}^*$ will be perfectly collinear with W. (Alternatively, the first stage regression suffers from perfect multicollinearity.)

(c) W is perfectly collinear with the constant term.

(d) Z is not a valid instrument because it is correlated with the error term.

3. SW Exercise 12.7

(a) Under the null hypothesis of instrument exogeneity, the J statistic is distributed as a χ_1^2 random variable, with a 1% critical value of 6.63. Thus the statistic is significant, and instrument exogeneity $E(u_i|Z_{1i}, Z_{2i}) = 0$ is rejected.

(b) The J test suggests that $E(u_i|Z_{1i}, Z_{2i}) \neq 0$, but doesn't provide evidence about whether the problem is with Z_1 or Z_2 or both.

4. SW Exercise 12.8

(a) Solving for P yields P = $\frac{\gamma_0-\beta_0}{\beta_1}+\frac{u_i^d-u_i^s}{\beta_1}$; thus $Cov(P,u^s)=\frac{-\sigma_{u^s}^2}{\beta_1}$

(b) Because Cov(P,u) ≠ 0, the OLS estimator is inconsistent (see (6.1)).

(c) We need a instrumental variable, something that is correlated with P but uncorrelated with us. In this case Q can serve as the instrument, because demand is completely inelastic (so that Q is not affected by shifts in supply). γ0 can be estimated by OLS (equivalently as the sample mean of Qi).

5. SW Exercise 12.10

$$\hat{\beta}_{TSLS}=\frac{Cov(Z_i,Y_i)}{Cov(Z_i,X_i)}=\frac{Cov(Z_i,\beta_0+\beta_1X_i+\beta_2W_i+u_i)}{Cov(Z_i,X_i)}=\frac{\beta_1Cov(Z_i,X_i)+\beta_2Cov(Z_i,W_i)}{Cov(Z_i,X_i)}$$

(a) If $Cov(Z_i,W_i)=0$ the IV estimator is consistent.

(b) If $Cov(Z_i,W_i)\neq 0$ the IV estimator is not consistent.

6. SW Empirical Exercise 12.1

(Results using full dataset)

Regressor	**Estimation Method**		
	OLS	**IV**	**IV**
Morekids	−5.387	−6.313	−5.821
	(0.087)	(1.275)	(1.246)
Additional Regressors	*Intercept*	*Intercept*	*Intercept, agem1, black, hispan, othrace*
First Stage *F*-Statistic		1238.2	1280.9

(a) The coefficient is −5.387, which indicates that women with more than 2 children work 5.387 fewer weeks per year than women with 2 or fewer children.

(b) Both fertility and weeks worked are choice variables. A women with a positive labor supply regression error (a women who works more than average) may also be a woman who is less likely to have an additional child. This would imply that *Morekids* is positively correlated with the regression error, so that the OLS estimator of $\beta_{Morekids}$ is positively biased.

(c) The linear regression of *morekids* on *samesex* (a linear probability model) yields

$$\widehat{morekids} = 0.346 + 0.066 samesex$$

$$(0.001)\ (0.002)$$

so that couples with *samesex* = 1 are 6.6% more likely to have an additional child that couples with *samesex* = 0. The effect is highly significant (*t*-statistic = 35.2)

(d) *Samesex* is random and is unrelated to *any* of the other variables in the model including the error term in the labor supply equation. Thus, the instrument is exogenous. From (c), the first stage *F*-statistic is large (*F* = 1238) so the instrument is relevant. Together, these imply that *samesex* is a valid instrument.

(e) No, see the answer to (d).

(f) See column (2) of the table. The estimated value of $\beta_{Morekids} = -6.313$.

(g) See column (3) of the table. The results do not change in an important way. The reason is that *samesex* is unrelated to *agem1, black, hispan, othrace,* so that there is no omitted variable bias in IV regression in (2).