ESCOLA BRASILEIRA DE ECONOMIA E FINANÇAS - EPGE

Luciano Fabio Busatto Venturim

Econometrics 1 - Problem Set 6

Rio de Janeiro

4th Quarter - 2021

Question 1:

1. We have $\hat{\beta}_{IV} = (A'Z'X)^{-1}A'Z'Y$. If $J = K$, then $Z'X$ and A' are both $K \times K$ full rank matrices and, therefore, invertible. We then have $(A'Z'X)^{-1} = (Z'X)^{-1}(A')^{-1}$, which implies that $\hat{\beta}_{IV} = (Z'X)^{-1}(A')^{-1}A'Z'Y = (Z'X)^{-1}Z'Y$ does not depend on A. Hence, all IV estimators are the same.

2. Since $P_Z = Z(Z'Z)^{-1}Z'$, we have that $Z^* = Z(Z'Z)^{-1}Z'X = ZA$, where $A = (Z'Z)^{-1}Z'X$ is a $J \times K$ with full column rank, because $Z'X$ is full rank. That is, Z^* is of the form above, so if take $\tilde{Z} = Z^*$ we see that $\hat{\beta}_{2SLS}$ is indeed an IV estimator.

Question 2:

1. Since $worked$ is a dummy variable, the coefficient of $morekids$ represents the effect of $morekids$ on the probability of $worked = 1$. The estimated coefficient is -0.142284 which means that a woman with more than two kids has almost 15% more chance of being unemployed.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.582216   0.001322  440.30   <2e-16 ***
morekids    -0.142284   0.002295  -61.99   <2e-16 ***
```

Figure 1: Linear Regression of $worked$ on $morekids$

Since there may be other factors that influence the decision of work and is correlated with the number of kids the woman has, we should worry about omitted variable bias. Moreover, it may be the case that women that are not working are more likely to have more kids, so the variable $morekids$ is also correlated with the error.

2. For $samesex$ to be a good instrument for $morekids$, we need it to be exogenous and correlated with $morekids$. Since the sex of a child is randomly assign, there is no reason for $samesex$ to be endogenous. To assess the correlation between $morekids$ and $samesex$, we can run the first stage regression and test for the significance of the variable. As we see in Figure 2, the coefficient is roughly 0.6 and is significant at the usual levels.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.302144   0.001463  206.56   <2e-16 ***
samesex     0.058868   0.002056   28.64   <2e-16 ***
```

Figure 2: Linear Regression of $morekids$ on $samesex$

3. Figure 3 shows the results of the regression of $worked$ on $morekids$ using $samesex$ as an IV.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.56315    0.01226  45.949   <2e-16 ***
morekids    -0.08484    0.03678  -2.307   0.0211 *
```

Figure 3: Linear Regression of $worked$ on $morekids$ using $samesex$ as IV

See that the qualitative result is the same, i.e., women that have more than two kids is more likely to be unemployed, but now the effect is smaller, of only about 8.5%. For us to argue that this estimate is the $LATE$, we also need that there are no ”defiers”. That is, we cannot have a woman that decides to have more kids if she has two of different sex, but decides to not have more if she has a couple of girls or boys, which seems to be a very strong assumption, since for cultural reasons one might want two children of the same sex and therefore be inclined to have one more child only in the case of kids of different sex.

Question 3

1. We use the function group_by together with summarise to create the means. The next figures show the plots of each variable mean as a function of psu. As we can see, only $entercollege$ has a jump at $psu = 475$.

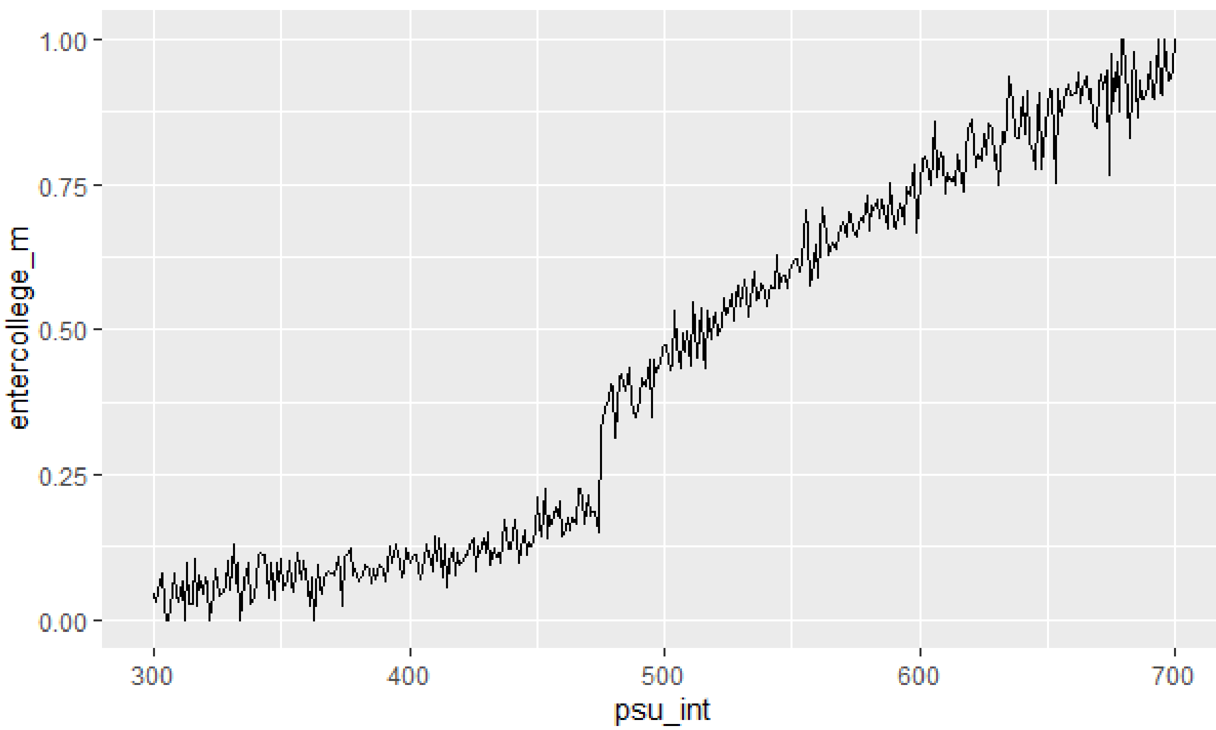

Figure 4: $entercollege$ mean on psu

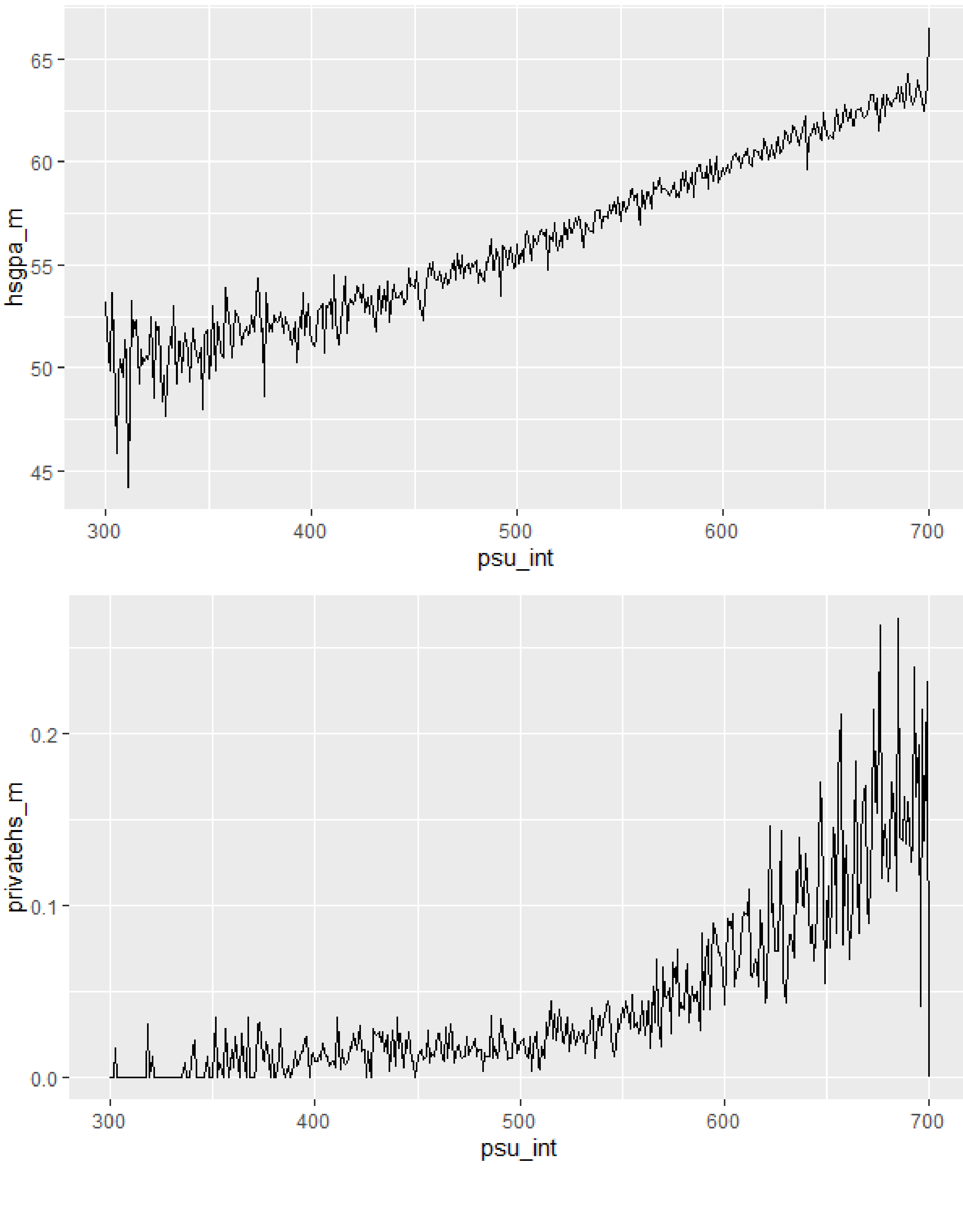

Figure 5: $hsgpa$ and $privatehs$ mean on psu

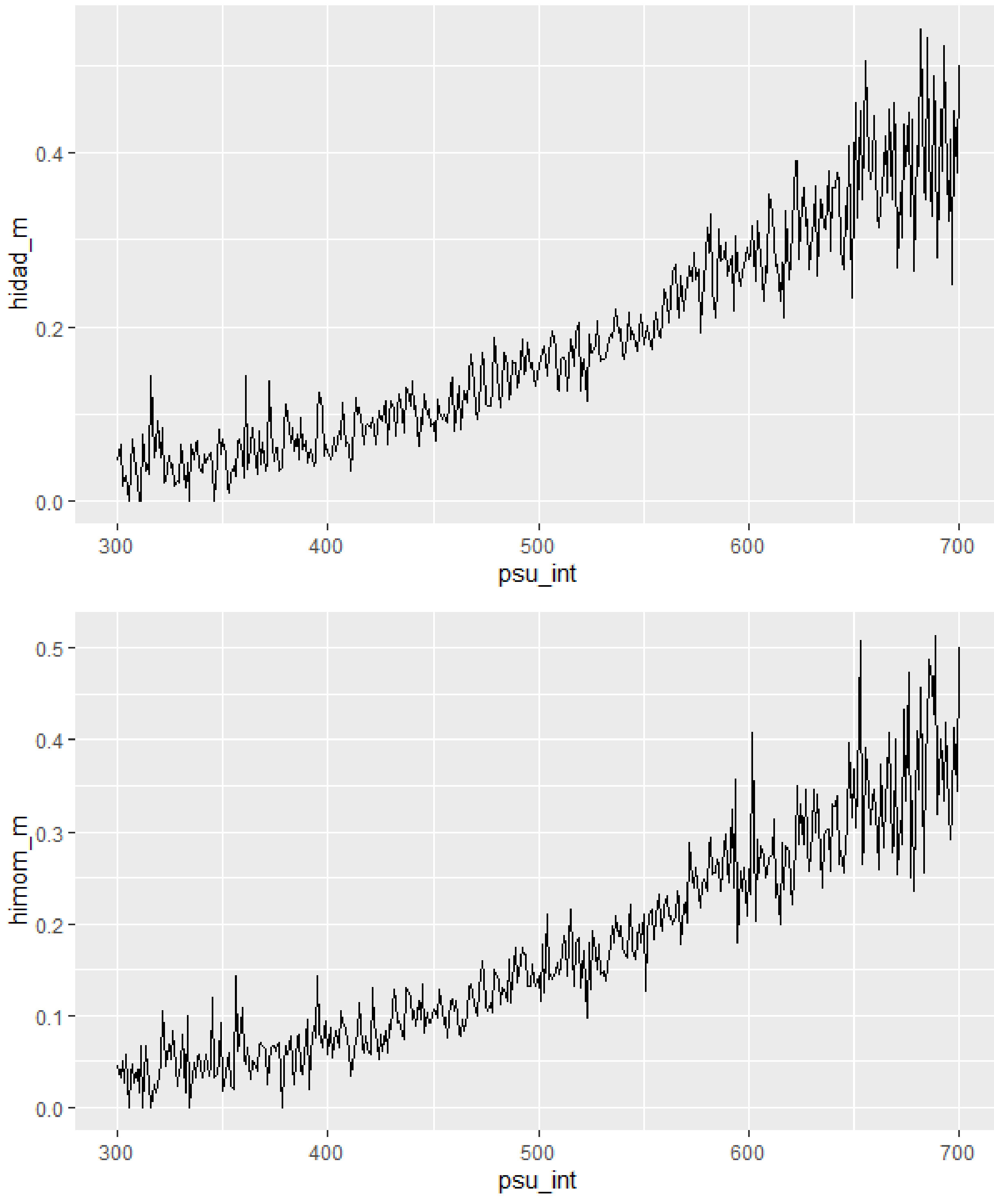

Figure 6: $hidad$ mean and $himom$ on psu

3a. The following table summarises the results of the regressions with bandwidth of 10 points. See that the "discontinuity" in $entercollege$ is 0.176 and significant.

Table 1:

	Dependent variable:			
	entercollege_m	hidad_m	himom_m	hsgpa_m
	(1)	(2)	(3)	(4)
psu_line	−0.004	0.002	0.005**	0.066
	(0.003)	(0.003)	(0.002)	(0.050)
I(psu_line >= 0)	0.176***	−0.024	−0.033*	−0.119
	(0.026)	(0.025)	(0.016)	(0.405)
psu_line:I(psu_line >= 0)	0.011**	0.002	−0.003	−0.111
	(0.004)	(0.004)	(0.003)	(0.067)
Constant	0.170***	0.141***	0.148***	55.096***
	(0.020)	(0.019)	(0.012)	(0.313)
Observations	21	21	21	21
R^2	0.932	0.135	0.316	0.142
Adjusted R^2	0.920	−0.018	0.195	−0.009
Residual Std. Error (df = 17)	0.029	0.028	0.018	0.458
F Statistic (df = 3; 17)	77.887***	0.881	2.615*	0.937

Note: *p<0.1; **p<0.05; ***p<0.01

3b. The following table summarises the results of the regressions with bandwidth of 20 points. See that the "discontinuity" in $entercollege$ is 0.178 and is approximately equal to the one with bandwidth of 10 points. For the other regressions, the estimates are also similar.

Table 1:

	Dependent variable:			
	entercollege_m	hidad_m	himom_m	hsgpa_m
	(1)	(2)	(3)	(4)
psu_line	0.001	0.002**	0.002***	0.033
	(0.001)	(0.001)	(0.001)	(0.022)
I(psu_line >= 0)	0.178***	−0.019	−0.020*	−0.191
	(0.019)	(0.015)	(0.012)	(0.361)
psu_line:I(psu_line >= 0)	0.001	−0.0003	0.0001	−0.008
	(0.002)	(0.001)	(0.001)	(0.030)
Constant	0.189***	0.142***	0.133***	54.918***
	(0.014)	(0.011)	(0.009)	(0.267)
Observations	41	41	41	41
R^2	0.928	0.359	0.555	0.183
Adjusted R^2	0.922	0.307	0.519	0.116
Residual Std. Error (df = 37)	0.030	0.025	0.018	0.576
F Statistic (df = 3; 37)	158.093***	6.895***	15.405***	2.755*

Note: *p<0.1; **p<0.05; ***p<0.01

3c. Finally, we plot the coefficient β_3 in the regression with dependent variable $entercollege$, that measures the ATE at the discontinuity point $psu = 475$, for different choices of bandwidth. See that for smaller bandwidth, the coefficient is close to 1.85, and start to decrease with the window size.