Econometrics PS 2

Himani Pasricha (20201143)
Dimitrios Argyros (20201790)
Tatiana Bezdenezhnykh (20201138)
Matthew Abagna(20201136)
Manuel Rodriguez(20201144)

February 2021

1 Exercise 2

1. Table 1 presents the OLS estimates of earnings on schooling. Controlling for fourth-order polynomials of age and year of birth, the results reveal a positive effect of the number of years of schooling on earnings. Keeping all other factors constant, one additional year of schooling increases average earnings by 15.8 per cent.

   The results above, even if they are troubled by an endogeneity problem, will be addressed by an Instrumental variables approach. Notwithstanding, the results are consistent with most findings in the literature suggesting a positive effect of schooling on future earnings (e.g. Angrist & Krueger, 1994; Ashenfelter & Krueger, 1994; Devereux & Hart, 2010). It must be pointed out though a smaller number of studies (e.g. Pischke & Watcher,1994; Grenet, 2013;) find zero or close to zero returns of schooling on future earnings.

   The example examined here treats endogeneity between schooling and future earnings by using an instrumental variables approach. Education is endogenous to schooling as unobservable characteristics of the individual, such as family background, affect both education and earnings, leading to an upward bias of the results as we will see later.

   The use of instrumental variables to tackle unbiased treatment effects is common practice in the literature examining the effects of schooling on various outcomes (e.g. Angrist & Krueger, 1994; Card, 1995; Acemoglu & Angrist, 2001).Several studies have used a change in compulsory schooling laws to instrument education (e.g. Oreopoulos, 2006; Black et al., 2007) and that is the approach that this example follows. Using a policy that induced the treatment group to have at least one more year of schooling that the control group (those not subject to the policy), we implicitly consider that the treatment group has on average more years of schooling

Table 1: OLS Regression Results

	Dependent variable:
	log(Earnings)
Schooling	0.158***
	(0.002)
Age^4	0.00000
	(0.00000)
Year of birth^4	-0.0001
	(0.0002)
I(ageˆ2)	0.009
	(0.017)
age	-0.254
	(0.540)
I(yobˆ4)	-0.00000
	(0.00000)
I(yobˆ3)	0.0003
	(0.0002)
I(yobˆ2)	-0.018
	(0.012)
yob	0.441*
	(0.252)
Constant	1.489
	(6.176)
Observations	30,801
R^2	0.147
Adjusted R^2	0.147
F Statistic	591.542*** (df = 9; 30791)
Note:	*p<0.1; **p<0.05; ***p<0.01

than the control group. Whether this is the case has to be seen when we estimate the first stage.

2. The treatment assignment mechanism is the following:

$$S_i = 1(\gamma_0 + \gamma_1 Z_i > \eta_i), \text{with } E(Z_i \eta_i) = 0. \quad (1)$$

There is a first stage for two reasons. Firstly, the covariance between policy and schooling is not zero, given that policy is a compulsory law; it induces an increase in year of schooling. Secondly, there could have been an option that instead of going through the first stage, we computed the reduced form, regressing the instrument on the outcome. However, given that there might be no perfect compliance, the policy's covariance is not equal to 1. Therefore, a first stage is needed to capture the causal effect of schooling on earnings.

For an instrument to be valid following assumptions should be hold:

- Exogeneity Assumption: The instrument(policy) should be orthogonal to the any unobservables that affect the outcome(earnings).
- Exclusion Restriction: The instrument(policy) should affect the outcome(earnings) only through the treatment(schooling).
- Monotonicity: No presence of defiers.
- Strength: There should be a sufficiently strong correlation between instrument(policy) and treatment(Schooling).

Potential threats to validity:

- Potential threats to internal validity coming from violation of the above restrictions:
  Even though there is no reason to assume any violation of the exogeneity and monotonicity assumptions, the exclusion restriction might be violated. An example of a violation of the exclusion restriction might have been if the policy made one more year of schooling mandatory and had explicitly or implicitly an effect in the curriculum. In that case, future earnings might be affected by the policy through one more year of schooling and a more comprehensive or extensive curriculum biasing upward the results. At the same time a violation of the exclusion restriction might come from the fact that people having to stay one more year in school, who have otherwise choose to leave after 14 years of schooling, enter the labour market later and potentially result in lower future earnings. In case that the adverse effects from entering the labour market later are more significant than the positive effects of one more year of schooling the results are

downwards biased.

Concerning the instrument's strength, a weak instrument causes problem-related to unreliable inference, small sample bias, and the IV estimator's inconsistency. Whether these potential issues hold in the specific case will be examined later.

Furthermore, given the data available, we cannot control for regional or other characteristics that might affect the policy's implementation and thus the strength of the instrument. These characteristics might be related to the enforcement of the policy.

- Potential threats to external validity coming from the fact that IV estimates Local Average Treatment Effects:
  LATE captures schooling's effects on those who complied with the policy at a specific time and place and excludes never-takers and always takers. If the outcomes differ substantially between the three groups, it is difficult to extrapolate the whole population's results from LATE. A way to assess external validity is to examine different populations that receive the same treatment and compare how LATE might differ.
  Also, the results might be driven by specific groups among the population for which the gains from one additional year of education might have very high returns while for most other groups is low.

Compliers, never-takers and always-takers: (By assumption, there are no defiers).

Compliers: Compliers are the ones who react to the policy (instruments). In this case, students who have 15 or more years of schooling after the implementation of policy and would have 14 or fewer years of schooling if the policy was not implemented are the compilers.
Never-Takers: These are those who never take up the treatment irrespective of instrument. In this case, they are those who have 14 or less years of schooling irrespective of whether they are affected by the policy.
Always-Takers: These are those who always take up the treatment irrespective of instrument. In this case, they are the students who have 15 or more years of schooling irrespective of when they are affected by the policy.

As mentioned, the IV estimate does not capture the average treatment effect(ATE) but rather the local average treatment effect(LATE). This is because of non-perfect compliance; thus, the causal effect of schooling on earnings is not comparable for compilers and non-compliers. If there were no non-compliers (always- takers and never-takers), the IV estimator

would have converged to the ATE. The difference between ATE and LATE increases as the proportion of compliers in the population decreases.

3.
   - Figure 1 shows the probability of leaving schooling before 15 against the year of birth. The probability of those born before 1933 and not subject to the policy was between 0.8 and 0.6. For those born after 1933 is less than 0.1. From that discontinuity, we see that the policy had an effect on year of schooling. The above/below graph indicates that the correlation between the policy and schooling is positive, indicating a strong first stage. The probability of dropping school before 15 for people born after 1933 is not zero suggests non-perfect compliance. (add continuous drop before)

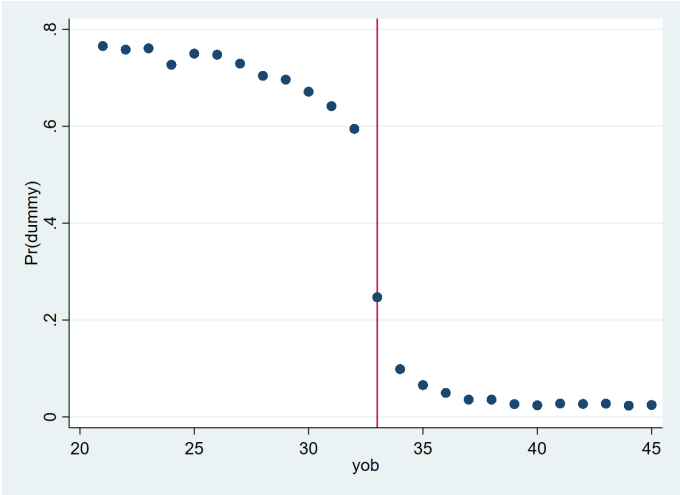
   - Figure 2 shows that people born before 1933 have less schooling years than the people born after 1933. Due to policy implementation, after the 1933 cutoff level, the jump in the years of schooling at this specific point of time can be observed. Like the previous graph, this figure suggests a positive correlation between years of schooling and policy..

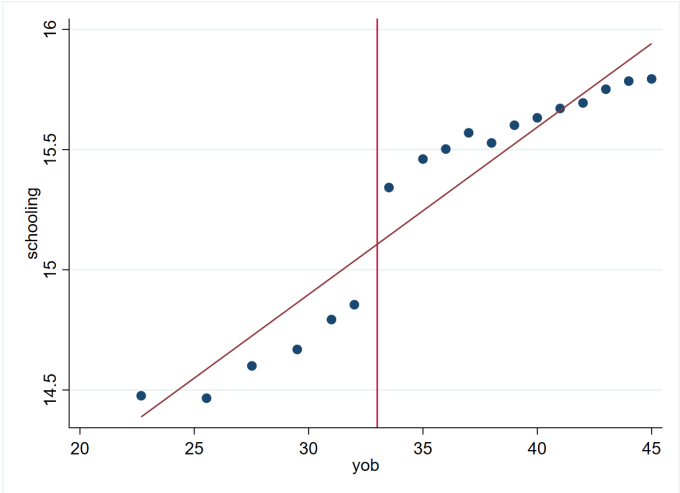
   - Figure 3 shows the relationship between log earnings(outcome) and year of birth (a proxy for policy/instrument). In general, there is a positive relationship between earnings and years of schooling. However, there are no notable differences between the two sides of the cutoff level as those shown on the previous figure. This may suggest that the policy did not have a large significant effect on future earnings.

4. Table 2 displays the IV estimation model without employing the previously introduced controls. The policy reform is used as the instrument of the years of schooling. As it can be observed, an additional year of schooling leads to an increase in the logarithm of earnings by 16.3 percent on average. This estimation, similar to the OLS regression, is statistically significant at the one percent level. However, in this estimation no controls are employed, which may lead to an overestimation of the impact of returns to schooling on the dependent variable. Moreover, the IV standard errors are 0.06 which is twice as larger as in the previous OLS estimation. This may be reflecting that our instrument is only employing the exogenous variation in years of schooling.

   In order to calculate the Wald estimator without controls based on conditional averages, the following formula is employed:

$$\widehat{\beta}_{IV} = \frac{E[log(earnings)_i|Reform_i = 1] - E[log(earnings)_i/Reform_i = 0]}{E[Schooling)_i|Reform_i = 1] - E[Schooling_i/Reform_i = 0])}$$


Pr(dummy)
0
.2
.4
.6
.8
20
25
30
35
40
45
yob


schooling
14.5
15
15.5
16
20
25
30
35
40
45
yob

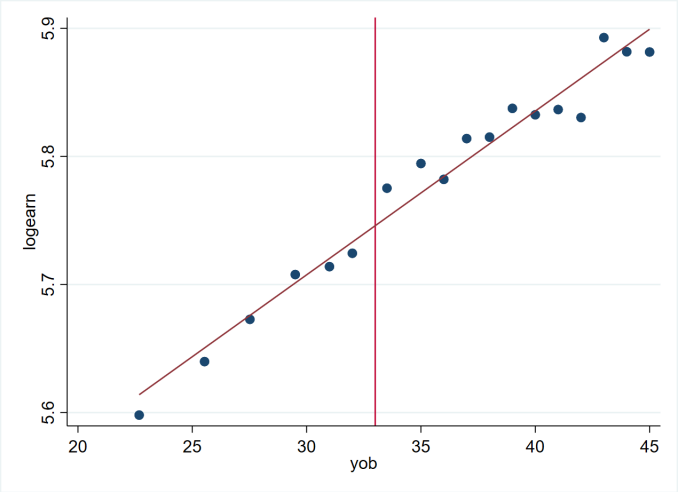


Table 2: IV Regression Results without controls

	Dependent variable:
	Earnings
Schooling	0.163***
	(0.006)
Constant	3.283***
	(0.095)
Observations	30,801
R^2	0.138
Adjusted R^2	0.138
Note:	*p<0.1; **p<0.05; ***p<0.01

Which in our case is = (5.833-5.671)/(15.615-14.624), leading to a final value of 0.163, which is, as expected, the same value as the one obtained in the previous IV estimation.

5. Table 3 reports the coefficients obtained on the estimation of the first stage and reduced form, respectively. The first stage is obtained by regression the endogenous variable schooling on the instrument Reform and all other controls to show whether the instrument chosen is relevant.

   The coefficient obtained (0.424) is both positive and statistically significant at the one per cent level.The obtained estimate can be interpreted as a person who had been born after 1993 having on average 0.424 years of schooling more then those born earlier. It signifies that the instrument is strong, although, the joint significance of the estimates (F-statistic) should be taken into account. It is worth nothing that the year of birth has both linear and non-linear negative impact on schooling while variable age shows no statistical significance at all. The reduced form estimates are obtained by regressing log earnings on the instrumental variable 'reform' and the set of control variables. The coefficient for reform is 0.029 and statistically significant at the ten percent level meaning that people born after 1933 have 2.9 percent higher annual earnings then those born before. The rest of the estimates inherit no statistical significance.

   The IV estimator is obtained from the ratio of the reduced form and the first stage, implying that individuals born in 1933 or after have on average 6.8 percent higher earnings than those individuals who were born before:

$$\widehat{\beta}_{IV} = \frac{\widehat{\lambda}_{reduced}}{\widehat{\delta}_{first}} = \frac{0.029}{0.424} = 0.068$$

   Table 4 shows a comparison between the results between the manual and the inbuilt estimation command. Column 1 replicates the first stage already discussed above. The second column represents the results of regressing log earnings on the fitted values of schooling obtained from the first stage and the set of control variables. Column 3 represents the results from *ivreg*, an inbuilt R function. The main finding is that the coefficient of the manual and inbuilt calculations coincide and are statistically significant at a ten and five percent level, respectively. Standard errors with *ivreg* are a bit higher meaning that we obtain slightly less precise estimates, although at a higher level of statistical significance in this case.

   It must be noted that the instrument is sufficiently strong. This is evident by the first-stage F-statistic (637.53) which is significantly larger than the rule of thumb for an instrument to be strong enough (Stock et al., 2002)F¿10. However, there is a strong possibility of heteroskedasticity. In that case, the "conventional" (non-robust) F-statistic results in standard errors that are too small, resulting in a rejection of the null hypothesis. However, the robust Kleibergen Papp (2006) F-statistic is large enough

Table 3: First Stage and Reduced Form with controls

	Dependent variable:	
	Schooling	log(Earnings)
	(First Stage)	*(Reduced form)*
	(1)	(2)
Reform	0.424***	0.029*
	(0.039)	(0.017)
Age^4	-0.00000	-0.000
	(0.00000)	(0.00000)
Age^3	0.001	-0.00002
	(0.001)	(0.0003)
Age^2	-0.047	0.001
	(0.041)	(0.018)
Age	1.434	-0.006
	(1.264)	(0.549)
Year of birth^4	0.00001***	0.00000
	(0.00000)	(0.00000)
Year of birth^3	-0.002***	-0.0001
	(0.001)	(0.0003)
Year of birth^2	0.086***	0.006
	(0.028)	(0.013)
Year of birth	-1.930***	-0.116
	(0.619)	(0.281)
Constant	13.439	5.483
	(14.453)	(6.068)
Observations	30,801	30,801
R^2	0.150	0.034
F Statistic (df = 9; 30791)	603.247***	120.190***
Note:	*p<0.1; **p<0.05; ***p<0.01	

(116.78) and more significant than any of the critical values (ranging from 5.53 to 16.38), something that provides even more substantial evidence of the strength of the first stage. The fact that the coefficient of the *Policy* is relatively large (0.424) and significant corroborates, furtherly, the above findings. Given the strength of the first-stage potential issues related to the validity of the IV driven by a weak instrument can be discarded.

Comparing the results from OLS regression (Table 1) and those form IV estimation we can see the evidence of the upward bias. It happens due to the endogenous nature of schooling. The probable reason for the upward bias is that the OLS estimate suffers from omitted variable bias if, for example, we don't control for students ability. One can propose that ability is positively correlated with both schooling and earnings. Its omission from the OLS regression means that the effect of ability on earnings is picked up by the schooling variable, overestimating the direct effect of education on earnings.

Finally, the IV estimates from the regressions with and without controls also differ, with the Wald estimate uncontrolled for age and year of birth almost twice as higher then the one form IV controlled regression (0.163 versus 0.069). It reflects the fact that the ATE is not equal to LATE.

WAHT ARE THESE SENTENCES BELOW? SHALL WE DELETE THEM?

F statistic 116.86 in just the normal version

relatively strong and significant coefficient of the *Reform* 1st stage F-statistic equals 637.53

Table 4: Manual and In-built 2SLS Regressions

	Dependent variable:		
	Schooling	log(Earnings)	
	Manually	*Manually*	*Inbuilt*
	(1)	(2)	(3)
Reform	0.424***		
	(0.039)		
Fitted Schooling		0.068*	
		(0.038)	
Schooling			0.069**
			(0.040)
Age^4	-0.00000	0.00000	-0.00000*
	(0.00000)	(0.00000)	(0.00000)
Age^3	0.001	-0.0001	-0.0001
	(0.001)	(0.0003)	(0.000)
Age^2	0.047	0.004	0.004
	(0.041)	(0.018)	(0.001)
Age	1.434	-0.104	-0.111
	(1.264)	(0.553)	(0.550)
Year of birth^4	0.00001***	0.00000	0.00000
	(0.00000)	(0.00000)	(0.00000)
Year of birth^3	-0.002***	-0.00003	-0.00002
	(0.001)	(0.0003)	(0.00002)
Year of birth^2	0.086***	0.0004	0.0003
	(0.028)	(0.014)	(0.014)
Year of birth	-1.930***	0.016	0.016
	(0.619)	(0.317)	(0.304)
Constant	13.439	4.563	4.629
	(14.456)	(6.155)	(6.086)
Observations	30,801	30,801	30,801
R^2	0.150	0.034	0.111
F Statistic (df = 9; 30791)	637.53***		
Note:			*p<0.1; **p<0.05; ***p<0.01