ADVANCED ECONOMETRICS

HOMEWORK 3

Fbio Marcaurelio (599027)

1 Introduction

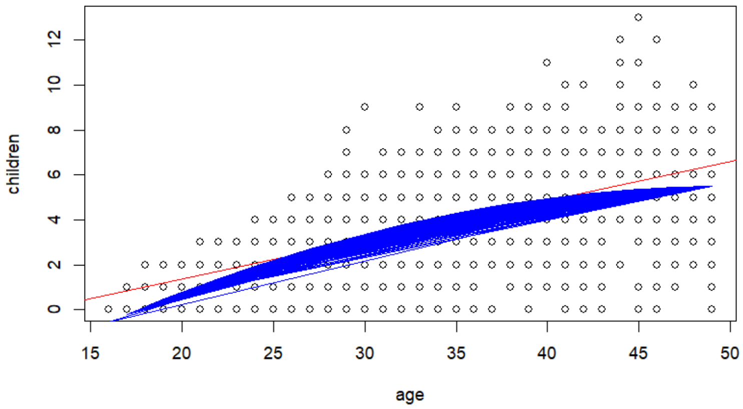
In this framework we are considering a sample of 1570 individuals from population of Botswana. The dataset includes, for women in Botswana during 1988, information on number of children, years of education, age, and other variables.

2 Question 1 - OLS

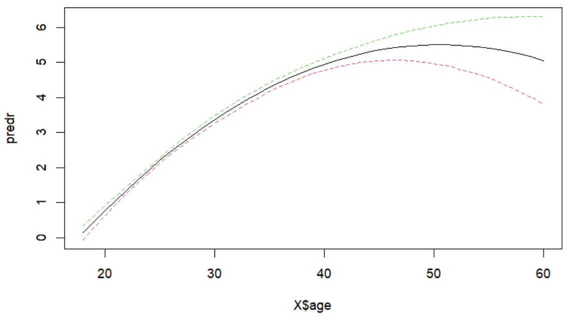
In principle we have to consider a model:

$$children = \beta_1 + \beta_2 educ + \beta_3 age + \beta_4 age^2 + u$$

Linear regression assumes that the relationship between the predictor X and the outcome Y is linear, in this framework this assumption does not holds and a linear estimation may not fit. then, we add a quadratic term.



we can see in the picture that the number of children for a women have a non linear dynamic, it increase but not linearly. in general we don't need to specify this thing but for an African country like Botswana with **4,75** fertility rate in 1988 this is an important evidence. we can do that also with the predictors, in this framework we can also estimate the turning point in wich the number of children per women start to decrease, it is 50.53982.



we can run the model in R and try to do some considerations:

Table 1: OLS

	Dependent variable:
	children
educ	-0.119^{***}
	(0.011)
age	0.507^{***}
	(0.045)
I(ageˆ2)	-0.005^{***}
	(0.001)
Constant	-6.573^{***}
	(0.732)
Observations	1,570
R^2	0.416
Adjusted R^2	0.415
Residual Std. Error	1.760 (df = 1566)
F Statistic	372.076^{***} (df = 3; 1566)
Note:	*p<0.1; **p<0.05; ***p<0.01

2.1 A-Discuss the interpretation of the coefficients

In this quadratic model we have a different interpretation for variable *age*, since we know that β_i is the derivative of Y with respect to X then in this case it became:

$$\frac{\partial Y}{\partial X} = \beta_3 + 2\beta_4 age$$

the direct interpretation is *the change in Y associated with a 1 unit change in X* but in this framework for example we have:

$$\frac{\partial children}{\partial age} = 0.507 - 0.005 \cdot age$$

for example for age 25 we have:

$$\frac{\partial children}{\partial age} = 0.507 - 0.005 \cdot 25 = 0.382$$

but for age equal to 45

$$\frac{\partial children}{\partial age} = 0.507 - 0.005 \cdot 45 = 0.282$$

then for a women with 25 years old 0.382 is the number of children associated with a 1 unit change in age, but for 45 years old this is 0.282. Each additional year of age reduces the slope by 0.005 unit, relation is concave. The other thing to interpret would be the intercept but you can only interpret that if it makes sense that our predictor variable be zero. In our case the children can be zero so we can interpret the intercept as the predicted number of children when all the dependent variable are zero.
about the interpretation of the coefficient *educ*, in this case the relation is negative, this means that a change in 1 unit of education (1 year more of education) implies in mean less children for 0.119 unit, in this case the relation is linear. all coefficients are statistically significant, this means that we reject the null hypothesis in which coefficients are equal to zero.

2.2 B-Are individual coefficients statistically significant?

For test the level of significance in the model we have to consider all p-values, they are all close to zero, this means that coefficients is this model are all statistically significant and then we can reject the null hypothesis which allows us to conclude that there is a relationship between children and other variables. at the same time i compute the F-test that allow us to confirm that these variable are statistically significant since the F is very high.

2.3 C-Perform a suitable test to assess significance of a woman's age on the number of children.

a suitable test can be made with a restriction as:

$$H_0 : age + age^2 = 0$$

and the result allows us to reject the null hypothesis .

2.4 D-Add the variable urban to the regression, i.e. estimate the parameters of the model

from this first model we can conclude that we have a good model, the F-statistic is very large, the R^2 even if is not relevant is almost 0.5. at the same time we have a problem, we don't know if we are omitting some variable or if we have a problem of endogeneity in one of the independent variables. now we consider a new model adding the variable *urban*:

$$children = \beta_1 + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 urban + u$$

we are interested to see if there are some differences in the value of the parameter *educ* and *age* and if we run the model and we compare it with the old one we get:

Table 2:

	Dependent variable:
	children
educ	-0.108^{***}
	(0.011)
age	0.509^{***}
	(0.045)
I(ageˆ2)	-0.005^{***}
	(0.001)
urban	-0.335^{***}
	(0.093)
Constant	-6.457^{***}
	(0.730)
Observations	1,570
R^2	0.421
Adjusted R^2	0.419
Residual Std. Error	1.754 (df = 1565)
F Statistic	284.371^{***} (df = 4; 1565)
Note:	*p<0.1; **p<0.05; ***p<0.01

2.5 E-Is the coefficient of urban statistically significant?

in the model we see that *urban* has a p-value close to zero and is statistically significant, under the restriction we notice that this variable give to us a F-statistic equal to 12, close to the critical level of 10.

2.6 F-Compare the results obtained by estimating the two models. Is there evidence of omitted variable bias when urban is not included? [...]. Are the empirical results in line with your conjecture?

Table 3:

	Dependent variable:	
	children	
	(1)	(2)
educ	-0.119^{***} (0.011)	-0.108^{***} (0.011)
age	0.507^{***} (0.045)	0.509^{***} (0.045)
I(ageˆ2)	-0.005^{***} (0.001)	-0.005^{***} (0.001)
urban		-0.335^{***} (0.093)
Constant	-6.573^{***} (0.732)	-6.457^{***} (0.730)
Observations	1,570	1,570
R^2	0.416	0.421
Adjusted R^2	0.415	0.419
Residual Std. Error	1.760 (df = 1566)	1.754 (df = 1565)
F Statistic	372.076^{***} (df = 3; 1566)	284.371^{***} (df = 4; 1565)
Note:		*p<0.1; **p<0.05; ***p<0.01

looking at the results we can see that there are some small difference in the two models, in particular the variable educ, remaining significant, change from -0.119 to -108. from this first analysis we can say that adding urban affect the other coefficient (*educ* in particular). we can say that there exist an omitted variable bias when urban is not include, in particular omitting urban from the model could result in an overestimation of the effect of educ on the number of children. from the definition of omitted variable bias we know that the first OLS without *urban* can be biased if this variable is related both to children and educ.

now we have to say somethings about the variable *educ*. Let's contextualize, we have data about the years of education in Botswana in 1988, we don't know nothing about the quality of education and we have to take in mind that this type of country, probably, stands out for high heterogeneity among individuals, in particular we can imagine that there are more educated people in the city instead of in the rural villages and this can explain the result above. the problem of exponential fertility growth in the African country, in fact, comes from the incapacity of manage the rural villages or the small cities. given this reason we can justify the negative sine of *urban* in the second model, this tell us that living in a city affect negative the number of children per women in Botswana. coming back to the education problem, we suspect that this variable is correlated with urbanization, we can imagine that schools are likely diffused in the cities instead of rural village. at the same time, i my opinion, even if i suspect a relation between *educ* and *urban*, i do not think that *urban* can be used as an instrumental variable since i suspect that it is correlated with error term of the initial model.

in general, *educ* is a case of endogeneity for many studies, in the next question we will try to use some instruments for judge if *educ* can bias the OLS. in particular we have two main instruments by which we can assume $cor(Z'\epsilon) = 0$, $frsthalf$ and *bicycle*.

about *bicyle* we can think about a relation with *educ*, even if in a not very strong form. i suspect that bicycle can explain educ not as a causal effect (i have a bike *implies* more years of education) but in a simple correlation form or inverse relation, (i am educated *then* i have a bike or i go to school *implies* that i need a bike). this idea comes from the fact that during 1988-1990 GDP reached its historic high until that moment driven by investments which we know are mainly infrastructural (transport) in African countries and has been studied in the paper below.

A. V. LIONJANGA AND v. RAMAN, 1990. Transportation and Economic Development in Botswana: A Case Study. Transportation Research Record Journal of the Transportation Research Board

about $frsthalf$, in principle we can think that this variable is useless. how can we justify the relation between $frsthalf$ and years of education? in reality there exist a realistic answer. it's nothing new if I say that the entry age at school is a crucial aspect for analyse the quality of education and in Botswana until 1993-1994 there was a law that required entry to school at the age of 6, this law changed in 1993 and has been the subject of studies. in particular before 1993 who was born after January has to enroll to school for the next year, this simple shift of one year can have strong effects on the education level of an individual, in particular in an African county. in particular if we divide the sample in first six mouth and after six month, as in the dataset, we can compare the effects of enter in the next year schools older than 6 months at least and the opposite side. under these assumptions we can study this instruments suspecting that it can be valid. as in the *bicyle* case i find a very interesting paper for justify this idea.

Noam Angrist, Owen Pansiri AND Gabatshwane Tsayang, 2017. EFFECT OF ENTRY AGE ON SCHOOL PERFORMANCE: EVIDENCEFROM BOTSWANA. Lonaka Journal of Learning and Teaching.

3 Question 2 - IV

in this section we proceed to study the bias of the previous model, in particular we suspect that the variable *educ* is endogenous, for this reason we take into account two possible instruments, $frsthalf$ and *bicycle.*

3.1 A-Discuss (no statistical test at this stage) whether you think that frsthalf and bicycle are valid instruments.

in reality i discuss part of the main idea in question **1.F**. then starting from these proposals we have to consider what does it mean *Valid Instruments*. starting from a generic model

$$Y = \beta_1 + x_2\beta_2 + x_3\beta_3 + \epsilon$$

and assume that variable x_3 is endogenous, a variable that depends on other variables in a statistical and/or economic model. now we can define another regression with x_2 as a dependent variable with π_i the coefficients, this is called First Stage:

$$\hat{x}_3 = \pi_1 + x_2\pi_2 + Z_1\pi_3 + r$$

this linear combination represents the part of x_2 that is explained by Z. now let's proceed to the Second Stage, from the First we collect the fitted values and we run the following regression:

$$y_i = \beta_1 + x_2\beta_2 + \hat{x}_3\hat{\beta}_3 + u_i$$

this OLS estimator corresponds to the IV estimator but standard error are wrong.

with this methods we can see if the Instrument is *Relevant*, then if $E[Z_1 x_3] \neq 0$ holds Z is related to the suspected endogenous variable.

this is not enough, we have to consider another property. an instrument is Valid if and only if is both *Relevant* and *Exogenous*, this last is more complicated and it depends from cases to cases. in particular, we want that $E[Z_1\epsilon] = 0$, this means that we are searching for a variable that is irrelevant for the starting generic model (but correlated with an independent variable). i said that this in more complicated since in our example of exactly-identified case (number of instruments = number of endogenous variables) we can not test it and we can only assume the orthogonality by construction with economic intuition, this is the reason why i speak a lot about the idea behind these instruments in question **1.F**. in the over-identified case (number of instruments ¿ number of endogenous variables) we can say somethings about validity, in particular we can test if $\sum z_i' \hat{u}_i$ is sufficiently close to zero:

$$J = (\sum z_i \hat{u}_i) \cdot \hat{S}^{-1} \cdot (\sum z_i' \hat{u}_i)$$

under the null hypothesis that the over-identifying restrictions are valid, the test is asymptotically distributed as a chi-square $L - K$. take in mind that the test can not give an answer for the validity of the instruments but if the test rejects the null, we can argue that the instruments are not valid (exogenous but not informative on which instruments are valid) and if the test does not reject the null, we cannot argue that all the instruments are valid (we must assume that at least k are valid).
now if we consider these properties and what we said in **1.F** i can ASSUME that instrument $frsthalf$ is exogenous with respect to the error of the original OLS, then if this instrument is Relevant we can say, by construction, that $frsthalf$ is a Valid instrument. for *bicycle* i suspect that this variable is not so exogenous for the model.

3.2 B-Perform proper statistical procedures to test whether frsthalf and bicycle are valid instruments. Report the results of the auxiliary regressions (if any) that you are using, the statistical test(s) that drives your decision, and your conclusions.

we start defining a First stage regression:

$$educ = \pi_1 + \pi_2 educ + \pi_3 age + \pi_4 age^2 + \pi_5 urban + \pi_6 frsthalf + \pi_7 bicycle + r$$

in this step we want to check for relevance of the instrumental variables ($frsthalf$ and *bicycle*). below the result, we notice that both coefficients are statistically significant (p-value close to zero). if we check for Relevance of instruments we deduce that both are Relevant since, under constraint $instruments = 0$, the F statistics is greater than 10, since we are in a large sample (1580) the expected value of IV/2SLS estimator can be approximated by

$$\beta_k + \frac{\hat{\beta}_k^{OLS} - \beta_2}{E[F] - 1}$$

with $E[F]$ the expected value of the first-stage F.
we can comment now the effects of the instruments on *educ* in the first stage.

Table 4:

	Dependent variable:
	educ
age	-0.082
	(0.101)
I(ageˆ2)	0.0002
	(0.002)
urban	2.398^{***}
	(0.202)
frsthalf	-1.278^{***}
	(0.202)
bicycle	0.859^{***}
	(0.217)
Constant	6.677^{***}
	(1.644)
Observations	1,570
R^2	0.138
Adjusted R^2	0.135
Residual Std. Error	3.957 (df = 1564)
F Statistic	50.065^{***} (df = 5; 1564)
Note:	*p<0.1; **p<0.05; ***p<0.01

$frsthalf$ have a negative sine, we expected it by our previous discussion, this is not enough for check the validity of this instruments for the general model. *bicycle* is this framework enter positively, then having a bicycle can affect positively the years of education.
we can move on to the second stage, remember that at the end of the first stage we have to collect the fitted values. in this framework we define an OLS:

$$children = \beta_1 + \beta_2 educ + \beta_3 age + \beta_4 age^2 +$$

the $2SLS$ procedure allows exploiting in estimation only the part of the variability of *educ* that is uncorrelated with u (the one "captured" by the variables in z). we remember that the coefficients of the second stage and the IVreg ones are equal but Standard Error are wrong in the 2S.
now we can run an IVreg that give us the same result of the second stage but allow us to test the exogeneity of the instrumental variable. basically we use the "diagnostic" test for summary and we notice that we reject the null hypothesis in the $Sargan$ test, remember that this test is not always accurate. this means that, in general, probably instruments are not valid if we take them together but it can not tell us which instrument is valid, but we can reject the null also if the model is not correctly specified, this is the case in which one instrument have a direct effect on y. at this point we understand why a lot of words in question **1.F**. in fact, we are still convinced that first half is an exogenous variable for the number of children but bicycle can be correlated on y, we can think that if you have a lot of children a bicycle is needed.

3.3 C-Now let us assume that frsthalf is uncorrelated with u, and only consider this variable as instrument. Show that frsthalf is a relevant instrument for educ.

for testing the Relevance, as we said and did before, we have to test if the F-statistics is greater than 10 and this is satisfied. we said before that we define Valid an instrument when it is Relevant and Exogenous ($cor[Z'u] = 0$). in this framework we assume that the second holds, this is not casual since we can not test it in case of exactly-identified case. then, with some realistic economic discussion we can assume the zero correlation between the instrument and the error term of the starting model. we just discuss it during question **1.F** and **2.B** and we conclude that the exogeneity of first half can be a realistic assumption, we thought also that it can be a Relevant instrument and we confirm it. but with this assumption and the empirical Relevance we can say that $frsthalf$ is a Valid instrument, then by construction.

3.4 D-Estimate the parameters of the model by two stages least squares using frsthalf as an instrument for educ. Is the instrument valid?

as we introduce before, under the previous assumption and the relevance of the instrument we can say that this instrument is Valid, we can report below the coefficient of the 2SLS: as we said before, the coefficient in the two methods are the same but S.E in the second stage are wrong.

Table 5:

	Dependent variable:	
	children	
	instrumental variable	*OLS*
	(1)	(2)
educ	−0.193***	
	(0.073)	
frsthat		−0.193***
		(0.073)
age	0.503***	0.503***
	(0.046)	(0.046)
I(ageˆ2)	−0.005***	−0.005***
	(0.001)	(0.001)
urban	−0.129	−0.129
	(0.199)	(0.201)
Constant	−5.952***	−5.952***
	(0.858)	(0.865)
Observations	1,570	1,570
R^2	0.399	0.388
Adjusted R^2	0.397	0.387
Residual Std. Error (df = 1565)	1.786	1.803
F Statistic		248.156*** (df = 4; 1565)
Note:		*p<0.1; **p<0.05; ***p<0.01

3.5 E-On the basis of the last 2SLS regression, perform a test to assess whether educ is indeed endogenous and report your conclusions (use the regression-based version of the Hausman test to answer this question).

since we suspect that the variable *educ* is endogenous we have to check for exogeneity. we need this test since if we reject the null ($H_0 : E[x'u] = 0$) then OLS is biased and inconsistent and use $IV/2SLS$. this test is called Hausman test and under validity of the instruments we can say if in this model and with this instruments the OLS coefficients are biased or not, then if x is endogenous or not. before the test, we know that, under assumption of exogeneity ($cor[Z'u] = 0$) of the instrumental variable and empirical relevance we have a Valid instruments. in addiction, this test is robust to the presence of heteroskedasticity. starting again by the first stage with $frsthalf$ instruments:

$$educ = \pi_1 + \pi_2 educ + \pi_3 age + \pi_4 age^2 + \pi_5 urban + \pi_6 frsthalf + r$$

now collect the residuals of this model:

$$\hat{r} = educ - \hat{\pi}_1 - \hat{\pi}_2 educ - \hat{\pi}_3 age - \hat{\pi}_4 age^2 - \hat{\pi}_5 urban - \hat{\pi}_6 frsthalf$$

and if we consider the starting OLS we have to include $\hat{r}$ in the model:

$$children = \beta_1 + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 urban + \delta\hat{r} + u$$

now we obtain the result of the Hausman test for this model under this instrumental variable. since in our case $\delta = 0$, no statistically significant (p-value=0.22951) we do NOT reject the null hypothesis, then in this particular framework *educ* is NOT endogenous. then in this framework seems that OLS is the correct model instead of the IV.

3.6 F-What is your preferred specification? Write a comment on the effect of educ on the number of children on the basis of your preferred specification.

at this point there are a lot of open question. we started some question above by saying that $frsthalf$ can be a Valid instrument given by an empirical relevance and an assumption on the exogeneity of the instrument, we have added an economic intuition on the importance of this variable on the level of education by an article *1.F* (even if in this framework we are interested in the years of education but we can assume that these things can be correlated) but in the final test we do not reject the hypothesis of exogeneity of the variable *educ* in this model. i keep saying "in this model" since this test (Hausman) can not be generalised, if at the starting point we make any error such that OLS is biased we will have some problem also with the Hausman test. then, in the previous model assuming that it is correct for evaluate the impact of educ on the number of children per women we can say that *educ* is correctly specified in the OLS

case w.r.t. IV.
Recall that if a Hausman test fails to reject that the 2SLS slope is different from the OLS slope implies that the data does not contain sufficient evidence to reject the null hypothesis that the two slopes are identical. In general, there are two reasons why you could get a null result in hypothesis testing: the null hypothesis is true and you correctly failed to reject it, or the null hypothesis is false, but you incorrectly failed to reject it. the reason why we fail to reject the null hypothesis is that the null of no endogeneity is probably technically false, but the magnitude of violation is so small that it would be unreasonable to expect a test to pick up on this difference. then in these respects we can conclude that *educ* is not endogenos only if the starting model is the *right* model of the true world and we can for sure admit that $frsthalf$ can not be the unique instruments for explain how educ affect the number of children since it's effect alone is not strong, we can suppose that with another type of data, like about quality of education, institution and other cultural issues we can be able to explain educ as an endogenous variable and then we could have a real impact of the years of education in the number of children.