Bjarne Strøm. Assessment guidelines SØK3001 V2021-60% exam (3hour) May 19

The grades are based on an overall assessment of the answers

Question 1.

Ageing of the population due to declining fertility of females is a major challenge in modern societies. Declining fertility is often attributed to increased education of females. Several researchers have tried to establish a causal link between fertility and education. The questions below are connected to an analysis of the relationship between women's fertility and education level on Norwegian data. The authors use individual data from birth cohorts from 1947 to 1958 to estimate several versions of the following relationship:

(1)$\boldsymbol{y_i = \beta_0 + \beta_1 ED_i + \beta_2 COHORT_i + \beta_3 MUNICIPALITY_j + u_{ij}}$

where $\boldsymbol{y}$ is fertility outcome (explained below), ED is the number of years of education obtained. COHORT refers to a full set of birth year dummy variables, $\boldsymbol{MUNICIPALITY}$ refers to a full set of dummy variables for the municipalities where the women were born. $\boldsymbol{u_{ij}}$ is an error term. Subscript i refers to individual, while subscript j refers to municipality. All outcomes are measured in 2002, when the youngest of these women were 44 years old.

The following fertility outcome variables were used:

The number of children in total

The timing of births represented by the following dummy variables:

1 if first birth is at age 15-20, 0 otherwise

1 if first birth is at age 20-25, 0 otherwise

1 if first birth is at age 25-30, 0 otherwise

1 if first birth is at age 30-35, 0 otherwise

They also use an indicator for childlessness defined as the dummy variable

1 if childless, 0 otherwise

The authors are concerned that ED is an endogenous variable, and in addition to OLS, they estimate the model by 2SLS using the introduction of a compulsory school reform in Norway during the 1960's and early 1970's as an instrumental variable. Before the reform, the Norwegian school system required children to attend school through the 7'th grad, while after the reform, this was extended to the ninth grade, thus adding two years of required schooling. This reform was implemented at different points in time in different municipalities. They use as an instrument the dummy variable $\boldsymbol{REFORM_j}$ equal to 1 if the individual born in municipality j was affected by the reform and zero otherwise

Table 1 reports the estimated β_1 coefficient, where each column shows the result for each outcome and thus the results from separate regressions. Estimated standard errors are in parenthesis. N is the number of observations.

a) Discuss why it is important to include the variables COHORT and MUNICIPAL in these regression equations. Discuss reasons why the education level may be endogenous in this type of regression model and the consequences for OLS estimation.

b) Interpret the estimated coefficients reported in Table 1 and comment on the difference between the OLS and 2SLS results.

c) Explain the identification assumptions behind the 2SLS estimations, and how the 2SLS estimations are conducted.

The authors report that the estimated coefficient in front of the variable REFORM in an OLS regression between EDU, REFORM and other variables equals 0.116 with an estimated standard error 0.017. Interpret this coefficient and explain what this result means for the credibility of the 2SLS results in Table 1. What other variables must be included in this regression? Explain.

d) What is your conclusion regarding the relationship between fertility and education based on these results?

e) A fellow student of you suggests enthusiastically that a similar analysis can be conducted in the future using the 6 year school-start reform that was introduced in all municipalities in Norway in 2004 and extended the required years of schooling from 9 to 10 years. Comment on this suggestion.

Table 1. Estimation results. Estimated standard errors in parenthesis. Additional control variables in the estimated equations are $COHORT_i$ and $MUNICIPALITY_j$

Outcome	Childless	Number of children	First birth age 15-20	First birth age 20-25	First birth age 25-30	First birth age 30-35	First birth 35-40
OLS	0.006 (0.001)	-0.013 (0.004)	-0.032 (0.001)	-0.024 (0.001)	0.030 (0.0001)	0.015 (0.0002)	0.005 (0.0001)
2SLS	0.011 (0.018)	-0.009 (0.087)	-0.080 (0.039)	0.044 (0.032)	0.012 (0.028)	-0.008 (0.018)	0.021 (0.009)
N	290 596	290 604	290 604	290 604	290 604	290 591	289 057

Proposed solution

a)The MUNICIPAL dummies accounts for all observable and unobservable time-invariant variables at the municipal level that may affect fertility. This includes permanent differences in norms, culture and income across municalitie. The COHORT dummies account for aggregate variables affecting different cohorts over time independent of municipality. Examples are availability and costs of contraceptives or societal norms or cultural factors that varies over time (cohorts) and affect women born in different municipalities in the same way. The ED variable may be endogenous because unobserved variables are likely to affect both the decision to invest in education and the number of children and timing of birth, even when controlling for MUNICIPAL and COHORT dummies, i.e. there is a potential omitted variable problem. Similar arguments can be found in the Woolridge textbook (ch.15) in the relationship between wages and education (return to education equations). See also ch 15-8 in Woolridge.

b)The only continuous dependent variable is the number of children in third row. The estimated coefficients in front of this variable measure the effect of years of education on the number of children of a women. For example, the OLS estimate, -0.013, imply that one year of education decreases the expected total number of children with 0.013, i.e. it is estimated that average fertility falls by 0.013 given one more year of education (see also ch 7-7 in Woolridge for a similar interpretation). According to the OLS estimate, the parameter is significantly different from zero. Note that the estimate of this coefficient is lower (-0.009) and not significantly different from zero in the 2SLS estimation. The other dependent variables are dummy variables and can be interpreted in terms of probabilities, see ch 7-5 in Woolridge. For example, the interpretation of the estimated 2SLS coefficient in the third column, -0.08, is that an additional year of education reduces the probability to have first birth in the age 15-20 by 0.08, i.e by 8 percentage points. The interpretation of the other coefficients in front of dummies is similar.

c)The identification assumptions for the 2SLS is that the REFORM dummy affects the level of education, but does not affect fertility directly. The assumption is that women belonging to cohorts that went to school in municipalities with 9 year compulsory school has a different level of education than cohorts born in the same municipality but starting compulsory school before the REFORM was implemented. The key to identification is *that different cohorts within the same municipality* experienced different compulsory schooling regimes. Those cohorts starting before the reform was implemented in municipality j had 7 years of schooling,while later cohorts had 9 years of schooling In terms of 2SLS, the REFORM dummy is assumed to be uncorrelated with the error term in the fertility equation, (1) but correlated with the education variable in the first stage regression equation which can be formulated as

(2) $ED_i = \alpha_0 + \alpha_1 REFORM_j + \alpha_2 COHORT_i + \alpha_3 MUNICIPALITY_j + v_{ij}$

All rhs variables in (2) are assumed to be exogeneous and uncorrelated with the error term in the structural equation (1). The 2SLS estimation is conducted by estimating (2) by OLS in the first stage since the error term v_{ij} in the first stage (or reduced form) education equation is exogeneous. The predicted ED variable from first stage replaces the endogenous ED variable in the second stage which is also estimated by OLS.

The identification assumption implies that the he $\alpha_1$coefficient in (2) should be clearly different from zero, i.e. the null hypothesis that $\alpha_1 = 0$ in eq (2) should be rejected by a sufficiently large margin, The estimated coefficient of 0.116 means that a women going to school with 9 year compulsory school has 0.1 more years of education than a women going to school in the same municipality with 7 years compulsory school. The t-value of the coefficient is 0.116/0.017=6.8 corresponding to an F-value of ≈ 46 ,which is clearly above the rule of thumb of F-value of 10.

d)According to the 2SLS results it is found that education does not affect the number of children a woman give birth to (the coefficient in front of number of children is small and insignificant), but it *changes the timing* of birth. This is illustrated by the fact that the coefficient in front of the ED variable for the outcome to have birth when 15-20 is reduced by 8% points per year of additional education. The 2SLS coefficient 0.021 means that an additional year of education gives a statistically significant 2.1% increase in the probability to have child at age 35-40

e) In principle it is possible to estimate the effect of a reform that affects all females in the same cohort which is the case for the 2004 reform. But since the 2003 reform was implemented in all municipalities in the same year, there is no variation over time within the same municipality. Thus the ability to identify the causal effect of education is less than in the reform considered above, see also the comments on 1c.

Question 2.

A real estate economist collects information on 1000 house price sales from two similar neighborhoods called "University Town" bordering a large public university and one neighborhood about three miles from the university.

He wants to quantify how house location relative to the university affects house prices and estimates a house price equation with results reported in Table 2 below.

The variable $PRICE$ is given in 1000 US$

House size $SQFT$ is measured in number of hundreds of square feet.

House age, AGE is measured in years.

Location is measured by dummy, $UTOWN$=1 for houses near the university.

Additional house characteristics are measured by dummy variables, $POOL$=1 if a pool is present, and $FPLACE$=1 if a fireplace is present.

Table 2. Estimated house price equation. Dependent variable is $PRICE$. Method: OLS

Variable	Coefficient	Estimated standard error
Constant	24.5000	6.1917
UTOWN	27.4530	8.4226
$SQFT$	7.6122	0.2452
$SQFT \times UTOWN$	1.2994	0.3320
AGE	-0.1901	0.0512
$POOL$	4.3772	1.1967
$FPLACE$	1.6492	0.9720

R-square=0.8706

a)Write down the equations for predicted house prices for houses located near the university and away from the university, respectively.

b)What is the change in expected price per additional square foot?

c) What is the depreciation in house prices per year? Construct a 90 percent confidence interval for the depreciation.

d) What is the expected price effect of a pool and a fireplace, respectively?

e) A commentator suggests that the variable AGE, the age of the house, contains large amount of measurement error. How would that likely affect the estimated house depreciation in c)? Explain.

f) Another commentator suggest that the house price depreciation is likely to depend on the location of the house and the size of the house. Suggest how you would change the house price equation to account for the commentator's proposal and how you could test his proposal.

Proposed solution

a)Straight forward to answer. The estimated population equation can be written as

$Price = \beta_0 + \delta_1 UTOWN + \beta_2 SQFT + \gamma SQFT \times UTOWN + \beta_3 AGE + \delta_2 POOL + \delta_3 FIREPLACE + u$ where u is a random error term.

Thus the predicted price for a house nearby university (UTOWN=1) is

$$\widehat{Price} = (24.5 + 27.453) + (7.6122 + 1.2994)\, SQFT - 0.1901 AGE + 4.3772 POOL + 1.6492 FIREPLACE$$

The predicted price for a house in other areas (UTOWN=0) is

$$\widehat{Price} = 24.5 + 7.6122 SQFT - 0.1901 AGE + 4.3772 POOL + 1.6492 FIREPLACE$$

b) Students should keep in mind that PRICE is measured in $1000, while the size variable is measured in 100 square feets. Thus, the change in the expected price per additional square feet is $89.12 for houses near the university (UTOWN=1). The change in the expected price per additional square feet is $76.12 for houses located in other areas (UTOWN=0).

c)Depreciation per year is measured by the estimated coefficient in front of AGE in Table 2. Since PRICE is measured in 1000$, the coefficient of -0.190 in front of AGE means that houses are expected to depreciate by $190 per year. The 90% confidence interval around the coefficient is($-0.1901 \pm 1.645 * 0.0512$)

d) Since PRICE is measured in 1000$, a pool increases the expected value of a house by $4377.20. A fireplace increases the expected value of a house by $1649.20

e) Measurement error in rhs variable, see ch 9-4-b in Woolridge. Can start the discussion of this problem within a stripped down version of the model with AGE as the single independent variable. According to the classical errors in variable model (p.311 in textbook), and the corresponding bias formula, measurement error leads to bias towards zero in the OLS coefficient in front of AGE. Student may also use the formula for bias to notice that the size of the bias towards zero increases in the noise to signal ratio in the AGE variable.

May extend the discussion to the case with more rhs variables, in addition to AGE, see Woolridge p. 312 and the basic result on bias towards zero is the same.

f) The proposal of the commentator can be formulated as an extended house price equation with interaction terms between AGE and UTOWN and SQFT and UTOWN.

$$\begin{aligned} Price = \ & \beta_0 + \delta_1 UTOWN + \beta_2 SQFT + \gamma SQFT \times UTOWN + \beta_3 AGE + \theta_1\, AGE \times UTOWN \\ & + \theta_2\, SQFT \times UTOWN + \delta_2 POOL + \delta_3 FIREPLACE + u \end{aligned}$$

The student should be able to see that the proposal of the commentator can be tested by testing the joint hypothesis $\theta_1 = \theta_2 = 0$ with an F-test using the extended house price equation.