Econometrics for Data Science (JBM045) Fall 2022– Q1

Assignment

This is an individual assignment. Each student has to submit an individual answer.

- a complete assignment includes the relevant R code, the output as well as your answers to the questions.
- please submit a PDF of completed assignment on Canvas.
- for questions regarding the assignment, contact Yanhong Lin (y.lin@tilburguniversity.edu) or Minh Nguyen (m.nguyen_1@tilburguniversity.edu).

The due date for this assignment is October 22, at 23:59 h.

Suppose you are a data analyst at the federal office for policy analysis. The latest figures indicate that fertility rates have dropped over the past years. You are asked to analyze whether the observed pattern is related to an increase in female education.

You have access to bi-annual data on women's fertility and schooling for the years 2002–2014. Besides, the data contain a set of additional characteristics and background information. The following Table provides you with an overview on the available information in the data set.

variable	*description*
educ	years of schooling
kids	total number of children
black	= 1 if woman is African-American
east	= 1 if woman lived in East at age 16
northcen	= 1 if woman lived in North-Central region at age 16
west	= 1 if woman lived in West at age 16
farm	= 1 if woman lived on farm at age 16
othrural	= 1 if woman lived in rural area at age 16
town	= 1 if woman lived in town at age 16
smcity	= 1 if woman lived in small city at age 16
year	years in which woman was observed, 2002-2014
meduc	mother's education
feduc	father's education

Use the data set `fertility.csv` and open the data in `R`. The data file can be found on Canvas. To carry out the assignment you need the following libraries: `AER`, `lmtest`, `Hmisc`, `dplyr`, `tibble`, `ivreg`, `modelsummary`, `summarytools`.

Further notes:

- The most convenient way of producing a pdf from R is to use the R markdown/some notebook environment.
- Most questions require some coding. Make sure that you provide your solution code in the document.
- Provide all your output with at least two digits after the decimal.
- When asked for the interpretation of a coefficient: A complete interpretation of a coefficient does not only include a statement about the coefficient magnitude but also about its statistical significance. You can use the 5% significance level as a benchmark.

1. (24pts) First, you would like to get familiar with the data.

   (a) (4pts) Provide a table with descriptive statistics, showing the the number of observations, mean, standard deviation, minimum and maximum for *kids*, *educ*, *age* and *black* in your data. Briefly describe the statistics in the table in words.

   (b) (6pts) Your outcome variable is *kids*, the total number of children a woman has in 2002–2014. Plot the mean number of children against years and show the time trend. Label your axes, select an appropriate scale limit and provide a title for your graph. Interpret the pattern shown by the figure.

   (c) (6pts) Now plot the mean years of schooling, *educ*, against years and show the time trend. Label your axes, select an appropriate scale limit and provide a title for your graph. Interpret the pattern shown by the figure.

   (d) (8pts) Compare the figures in 1(b) and 1(c). What relationship between fertility and schooling over time do you see? Is it intuitive? Explain.

2. (30pts) Now investigate the relationship between fertility and schooling in more detail.

   (a) (9pts) Write down a bivariate, linear model to specify the influence of the years of schooling on the number of children. Explain all terms in the model. Estimate this model using OLS and interpret the estimated intercept and slope coefficient. What do you conclude?

   (b) (7pts) Now, estimate a richer specification of the model in 2(a), by including the following control variables: *black*, *east*, *northcen*, *west*, *farm*, *othrural*, *town*, *smcity*. Interpret the estimated coefficient on *educ* and compare it that in 2(a). Are they different? Briefly explain.

   (c) (4pts) You want to add dummies for *year* to specification 2(b) to control for a potential time trend. Could you include dummies for all years? Why (not)?

   (d) (10pts) Estimate the specification controlling for *year* dummies and covariates in 2(b) using OLS. What pattern do the estimated coefficients on the year dummies suggest? Explain by interpreting jointly the estimated coefficients. Does adding year dummies improve the R^2 compared to your model in 2(b)? How could you assess formally whether the model has improved significantly?

3. (32pts) Your colleague claims that the relationship between schooling and fertility is non-linear. *Note: For estimation, always include the same set of controls (including year dummies) as in 2(d).*

   (a) (4pts) You first would like to estimate a model which uses educational attainment categories rather than years of schooling as a predictor. Therefore, you specify a variable with three categories for educational attainment:

      - low:0–9 years of schooling

- medium: 10–12 years of schooling
- high: >12 years of schooling

Formulate a model which illustrates the potential non-linearity in the impact of educational attainment on fertility.

(b) (12pts) Estimate the model in 3(a) using OLS. Interpret the estimated coefficients on the categories of educational attainment. Does the impact of education on fertility seem to be constant across categories? Why (not)?

(c) (16pts) Your colleague is convinced that the impact of schooling on fertility differs by race because, in the US, black women more often go to low quality public schools than white women do. Estimate the differential impact of years of schooling *educ* by race *black* on *kids* using OLS. Interpret all estimated coefficients on schooling and race. What is the average difference in fertility between black and white women with 12 years of education? (*Hint: For calculation, ignore covariates*)

4. (64pts) Your OLS results have indicated that there is a relationship between fertility and years of schooling. In the following you further investigate this relationship. *Note: For estimation, always include the same set of controls (including year dummies) as in 2(d).*

(a) (8pts) Consider again the estimated coefficient on *educ* in 2(d). Do you think that it can be interpreted as a causal effect? Discuss this using the OLS assumptions.

(b) (10pts) A potential instrument for years of schooling is mother's years of schooling, *meduc*. Use the IV conditions to discuss if *meduc* could be a good instrument for *educ*.

(c) (9pts) Run a first stage regression of *educ* on *meduc*. Interpret the estimated coefficient on *meduc*. Perform a first-stage F-test. What do you conclude from this test? Explain.

(d) (8pts) Apply two-stage least squares (TSLS) to estimate the impact of *educ* on *kids*. Use *meduc* as instrument for *educ*. Interpret the estimated coefficient on *educ* and compare to the OLS estimate in 2(d). Explain.

(e) (12pts) Conduct the correct formal test to compare the OLS estimates in 2(d) and the TSLS estimates in 3(d). Formulate the null hypothesis, the alternative hypothesis and write down the test statistic. On the 5% significance level, what does the test result tell you?

(f) (5pts) Now use father's years of schooling, $feduc$, as additional instrument and estimate TSLS. How does the estimated coefficient on educ change?

(g) (12pts) Perform an appropriate test on the joint validity of *meduc* and $feduc$. Formulate the null hypothesis, the alternative hypothesis and write down the test statistic. On the 5% significance level, what does the test result tell you?