314 Econometrics — MT 2018
Problem Set Computer Class 3 [‡]

Carefully answer the following questions, giving attention not only to "the solution" but also —and especially— to interpretation and motivation of your proposed solution. This will also give you an opportunity to learn and improve on your writing skills with respect to Econometrics. Hence, consciously consider the way you express yourself while writing down your answer.

Some of the following questions are voluntarily open-ended. Use your knowledge of econometrics to structure your answers. Please restrict yourself to clear and succinct answers showing that you know what you are doing and that you understand the tasks at hand. Remember the wise words of Antoine de Saint-Exupéry: «It seems that perfection is attained not when there is nothing more to add, but when there is nothing more to remove».

Some additional instructions:

- Unless the question specifically states otherwise, use a 5%-significance level when a test is required.
- In order for you all to work on different dataset, start your do-file with the following command:

  `set seed` BIRTHDATE

  where BIRTHDATE is of the format DDMMYYYY. Make sure this command is there, or the results obtained in Stata/R will change every time you run your program!
- Do not include raw regression output in your report!

QUESTION 1: FAMILY ECONOMICS: FAMILY SIZE VS FEMALE LABOUR SUPPLY ..... *47 points*

Two important trends observed in many countries during the last century are the increase in participation of women in the labour market and the decrease of the number of children per family. We can think of many reasons why these two trends would be related. On the one hand, we may think that because women now have access to the labour market, the opportunity cost of having children is higher, and hence, they decide to have fewer children. On the other hand, we may think that because women have fewer children, they have more time to work. Logically, both statements make sense.

In this question, we try to disentangle these two relations using an approach proposed by Angrist and Evans (1998). We are interested in the effect of the size of the family on the decision of married women to participate in the labour market.

The data we use come from the U.S. Census Bureau, and is part of the 1990 *Public Use Micro Survey* (PUMS). It is a very large micro-survey of American families. As its name implies, the data are available to use freely on the web site of the agency. However, the dataset we use is a subset of the data available via Angrist's website. The sample consists of so-called "traditional families" composed of one working man, and one woman, working or not. We store all variables in a dataset named `pums80.dta`. They are described in Angrist & Evans 1998 (`https://faculty.smu.edu/millimet/classes/eco7321/papers/angrist%20evans.pdf`). The dependent variable is the number of hours worked by married women, $workedm$. The independent variable of interest is the number of children. For the sake of this example, and for reasons that will become clear later, we will focus on the decision to have two children *versus* having more than two children. Hence, we will create a dummy variable $MoreThan2$, taking the value 1 if a woman has more than two children, and the value 0 if a woman has exactly two children. Families with one child only will be dropped.

(a) (1 point) Load the dataset into Stata/R

Stata	**R**
clear all	library(foreign)
set mem 50M	data <- read.dta("pums80.dta")
use pums80.dta	
compress	

[‡] This version: November 12, 2018.

Use the function `sample` to take a subsample of 50% of the observations. Don't forget to set the seed first. Select the observations where the mother has at least two children. We follow the approach of Angrist and Evans (1998), and focus on the sample of mothers aged between 21 and 35, and for whom the second child is at least 1 year old. Select this sample. Then, generate the variable $Morethan2$. How large is the sample that you retain?

(b) (3 points) Present a descriptive analysis of the variables relevant for this exercise, including the variables described in subsequent questions. Delete, if necessary, observations with problematic values (e.g. mothers who had a kid before 15 or implausible working hours). Explain the criteria you use in selecting your sample. How many observations did you drop and how many do you retain?

(c) (5 points) Let us start from the following model:

$$workedm_i = \alpha_0 + \alpha_1 MoreThan2_i + \boldsymbol{x}_i^\top \boldsymbol{\beta} + u_i, \tag{1}$$

where $\boldsymbol{x}$ is a vector containing characteristics of the mother and $\boldsymbol{\beta}$ is a vector of parameters. Let us include in $\boldsymbol{x}$ the following characteristics: the age of the mother, her age when she first gave birth, her number of years of schooling, and dummies for ethnicities. For the ethnicity of the respondent, we use caucasian as the reference class. The dummy variables we include refer to African-American and to "other race", when a respondent is in neither African-American or Caucasian. Additionally, we include a dummy if the mother is of Hispanic origin. Estimate this model by OLS and present the results of the estimation. Discuss these results. Make sure to mention the assumptions under which the OLS method would estimate consistently the parameters.

(d) (12 points) It is possible that the relationship between $agem1$ and $workedm$ is quadratic rather than linear. It could also be that an interaction effect between age and race is needed (African-American and "Other"). Estimate an extended model that captures this quadratic effect as well as the interaction effect. Interpret the coefficients, giving special attention to those on age, age squared and the interaction terms, and to potential differences between the estimates in the extended and the simple model in (c). Test whether the extended model performs significantly better than the model used in (c). Independently of the result of your previous test, compute the expected difference in the partial effect of $agem1$ on $workedm$ between an Afro-American woman with at age 29 and a Caucasian woman with same observable characteristics.

(e) (7 points) There are many ways to select the independent variables to be included in a model. These different approaches can lead to different models. In this question, we compare two different approaches of the "general-to-specific" type, meaning that we start from a model with a lot of regressors, and will remove some of them following a given procedure. Compare the following two approaches:

1. Start from the full model in part (d). Test jointly the significance of all the independent variables that are not individually significant at a 5%-level. Use a 5%-significance level for this test. If you find that the variables are not jointly significant, remove them and re-estimate the model. Repeat this procedure until you cannot remove variables anymore.
2. Start from the full model in part (d). Remove the insignificant variable with the largest p-value, and re-estimate the model. Repeat this procedure until there are no variables that do not have a significant effect.

Consider the following excerpt from the article of Angrist and Evans (1998):

> On one hand, papers on labor supply often treat child-status variables as regressors in hours of work equations, while on the other hand, economic demographers and others discuss regressions and models that are meant to characterize the impact of wages or measures of labor-force attachment on fertility.

Stated differently, there might be an endogeneity problem in model (1). In their paper, Angrist and Evans propose to use the sex of the first two children as an instrument for having more than 2 children. The idea is that parents have preferences for mixed-gender families (i.e. with at least one boy and one girl), and hence, will be more tempted to have another kid if they have two girls or two boys. On the other hand, the gender of your first two children is not influenced by how much you work. The last three questions discuss the use of this instrument to consistently estimate α_1 in (1).

(f) (5 points) Generate the variable $SameSex$ taking the value 1 if the first two children have the same gender, and 0 if they are of different genders. We will use this as an instrumental variable for $MoreThan2$. State the assumptions needed for this estimation by instrumental variable to yield a consistent estimate of α_1. In this case, do you need to use two stage least squares (why or why not)?

(g) (3 points) Regress $MoreThan2$ on $SameSex$ and $\boldsymbol{x}$. Test whether $SameSex$ is a significant explanatory variable in the this regression.

(h) (7 points) Estimate model (1) by the proposed instrumental variable approach. Include as exogenous regressors all the variables included in Question (d). Discuss your results. Compare them with what you previously obtained in (d).

(i) (4 points) Perform the Hausman test for α_1. State the null and the alternative. Summarize the steps you follow to perform this test.