Problem Set 4.1: Applied Economics

1. (Based on Wooldridge, pg 409) The data set in **fertil.gdt** comes from the National Opinion Research Center's General Social Survey for the even years from 1972 to 1984. We use these data to estimate a model explaining the total number of kids born to a woman ($kids$).

   i) First we want to know what has happened to fertility rates over time, after controlling for other observable factors. The factors we control for are: years of education ($educ$), age (age) and age squared ($agesq$), and race ($black$). Write an equation to estimate the evolution of fertility between 1972 and 1984 (use 1972 as the base year).

   ii) Estimate the equation using OLS. Do you find a significant drop in fertility in 1982 with respect to 1972? Do you think that the drop could be due to the increase in average education levels between 1972 and 1982?

   iii) The model estimated in (i) assumes that the effect of each explanatory variable, particularly education, has remained constant in this period. This may or may not be true; add the variables needed to investigate if there has been a change in the effect of education over time. Test the hypothesis that the return to education is different in 1982 with respect to 1972.

   iv) Now use the information for only two periods: 1972 and 1982. Perform a Chow test to see if there has been an structural change between 1982 and 1972.

2. For this exercise we use a sample of working women in a developing country. We have the following variables in the file **hours.gdt**: $hours$ (the number of hours each woman works), $educ$ (years of education), age (age, in years), $child$ (number of kids), $marr$ (a binary variable that takes on the value 1 if the woman is married and 0 otherwise).

   We will use time dummies, and their interactions with the number of children:

   $y80$ (a binary variable that takes the value 1 if the observation corresponds to the year 1980 and 0 otherwise);

   $y82$ (a binary variable that takes the value 1 if the observation corresponds to the year 1982 and 0 otherwise);

   $y84$ (a binary variable that takes the value 1 if the observation corresponds to the year 1984 and 0 otherwise).

   We analyze the evolution of hours and the impact of the number of kids on the hours worked. The following models are considered for this first analysis:

$$hours = \delta_0 + \delta_1 age + \delta_2 age^2 + \delta_3 black + \delta_4 child + \varepsilon_1 \quad (1)$$

$$hours = \delta_0 + \delta_1 age + \delta_2 age^2 + \delta_3 black + \delta_4 child + \delta_5 y82 + \delta_6 y84 + \varepsilon_2 \quad (2)$$

$$\begin{aligned} hours = {} & \delta_0 + \delta_1 age + \delta_2 age^2 + \delta_3 black + \delta_4 child + \delta_5 y82 + \delta_6 y84 \\ & + \delta_7 y82 \times child + \delta_8 y84 \times child + \varepsilon_3 \end{aligned} \quad (3)$$

$$\begin{aligned} hours = {} & \delta_0 + \delta_1 age + \delta_2 age^2 + \delta_3 black + \delta_4 y82 + \delta_5 y84 \\ & + \delta_6 y80 \times child + \delta_7 y82 \times child + \delta_8 y84 \times child + \varepsilon_4 \end{aligned} \quad (4)$$

i) Assume that Model (1) verifies the assumptions of the classical regression model. Consider two women interviewed in the same year, both of them are white and have the same number of kids, but one is 40 years old and the other one is 30 years old. Express the difference in hours between both women (in terms of the parameters of the equation).

ii) Assume that Model (1) verifies the assumptions of the classical regression model. Estimate the model using OLS. Explain if the marginal effect of age on the total number of working hours is constant. Is the effect positive for all ages? Explain.

iii) Assume that Model (2) verifies the assumptions of the classical regression model. Estimate the model using OLS. What was the evolution of working hours between 1980 and 1982 and between 1982 and 1984?

iv) Assume that Model (3) verifies the assumptions of the classical regression model. Estimate the model using OLS and analyze if the impact of the number of kids on hours worked remained constant in the whole period. In addition, analyze if the number of kids has a significant impact on working hours.

v) If possible, repeat the previous question using Model (4). Compare the coefficients from both models.

vi) Which of the following statements is true?

   a) Equation (1) is the most general since time is not important to explain hours.
   b) Equation (4) is more general than equation (3), since only in equation (3) we can estimate the effect of the number of kids on hours in 1980.
   c) Equation (4) cannot be estimated since it suffers from perfect multicollinearity.
   d) If we want to evaluate the evolution of the impact of the number of kids on hours, we need to use equation (3) or equation (4).

3. (Based on Wooldridge, pg 418) Meyer, Viscusi, and Durbin (1995) studied the length of time that an injured worker receives workers' compensation. In 1980, Kentucky raised the cap on weekly earnings that were covered by workers' compensation. An increase in the cap has no effect on the benefit for low-income workers, but it makes it less costly for a high-income worker to stay on workers' compensation. Therefore, the control group

is low-income workers, and the treatment group is high-income workers. Using random samples both before and after the policy change, the authors test whether more generous workers' compensation causes people to stay out of work longer (everything else fixed). Let's define $log(durat)$ the time spent on workers' compensation (in logs), $afchnge$ a dummy variable for observations after the policy change and $highearn$ a dummy variable for high earners.

i) Compute the mean duration of benefits before and after the policy change for the high-income earners of Kentucky. Is there a statistically significant increase in duration of benefits? Repeat for low-income earners.

ii) Perform a difference-in-differences analysis to estimate the effect of the benefit change. Write an equation that gives you the diff-in-diff estimator for the impact of the policy on the duration of benefits (use $log(durat)$ as the dependent variable). Compute all the variables that you need in order to perform the analysis. Use only the observations from Kentucky. Did the average length of time on workers' compensation change with the new cap?

iii) What's the interpretation of the coefficient on $afchnge$?

iv) What do you make of the small R-squared from part (ii)? Does this mean the equation is useless?

v) The authors also added a variety of controls for gender, marital status and industry. Reestimate the equation in part (ii) adding as explanatory variables $male$, $married$, and a full set of industry dummy variables. How does the estimate on the interaction $(afchnge * highearn)$ change when these other factors are controlled for? Is the estimate still statistically significant?

vi) Estimate the equation in (ii) using the data for Michigan. Is the Michigan estimate on the interaction term statistically significant?

4. A government from a developing country carried out an irrigation project. This irrigation project affected the areas at a certain distance to the river, and it did not affect areas too far away. The agency in charge of the project collected data on crop yields for farms in both areas, those close enough to the river to get irrigation and those too far away to be irrigated.

   i) If you have information of production in both areas for the season after the project was finished, which regression would you run to estimate the effect of irrigation? Do you think that this regression will give you consistent estimates?

   ii) What if the agency had collected two waves of data, one before the project was built and one after it was built? How would you change the regression in (i) to get an estimate of the causal effect of the project? Write the regression and give an interpretation of the parameters.