Empirical Economics - Data exercise 1

Introduction:

- Deadline: Thursday 18-11-2021 (23:59)
- Submission: submit one PDF file through the Canvas assignment.
- Page limit: maximum 3 pages excluding tables/appendix.
- Note that this assignment should be made in groups of 3. Only 1 member of the group needs to upload your answers on Canvas.

Part I:

One of the policies at the welfare office in Amsterdam is to give applicants a search period at the moment that they apply for welfare benefits. This means they are sent away for four weeks, with the instruction to apply for at least 20 jobs. If they still have not found a job after four weeks they can return to the agency and start collecting welfare benefits. Benefits will be paid retrospectively, so from the moment of application on. The agency is interested in the effect of such a search period on how much welfare benefits an individual receives ($beta_1$):

$$benefits = \beta_0 + \beta_1 searchperiod + u$$

Normally, it is up to the caseworker to decide whether an applicant gets a search period or not and about 50% of the applicants get a search period. Caseworkers likely base their decision on (un)observed characteristics of the applicant, so that *searchperiod* is endogenous resulting in a biased estimate for β_1.

To solve this, the agency decided to run an experiment. During the experiment, caseworkers received a so-called "default option". If caseworkers were assigned the default option "normal" they continued as before, and decided themselves whether an applicant gets a search period or not. For the default option "never", caseworkers were instructed to not assign a search period to any of their applicants. Applicants are randomly assigned to caseworkers.

The dataset data_assignment1.csv that you will use for this assignment contains data from this experiment. The dataset contains some information on the applicants (gender, age, partner, children, education, aggregate income 24 months before the application) and information on the location of the application (part of the city of Amsterdam). Furthermore, for each applicant the dataset contains the default option of their caseworker (equal to normal or never) and the variable "search period" which indicates whether an individual received a search period or not.

Finally it contains three outcomes: the total number of weeks that an individual received benefits in the first 6 months after application (min 0, max 26), the number weeks an individual had a positive wage income in the first 6 months after application (min 0, max 26) and the total income in the first 6 months after application (benefits + wage, in euro's).

1. Create a dummy that is equal to one if the default option that the caseworker got is equal to "normal". We want to know whether caseworkers followed the default option the agency assigned them. Compute the average of the dummy variable "sp" for applicants with the default "normal" and applicants with the default "never". Did all caseworkers comply with their default option?

2. Now just look at applicants with the default "normal". For this group, create a table that compares the average characteristics of individuals with a search period ($sp = 1$) with those without a search period ($sp = 0$). Include the characteristics gender, age, partner, children, education, and previous income. What do you conclude? Are applicants that get a search period similar to those who don't?

3. Next, create a so-called balancing table for the groups in the default groups "normal" and "never". Your table should include the means of the covariates for each of the treatment groups, and the p-value of whether the difference between the groups is significantly different from zero. Include the same applicant characteristics as in the previous question. What do you conclude, was the treatment randomly assigned? Hint: you can use the option "tableby" from the package "arsenal". Other ways of creating a balancing table are also correct.

In the remainder, you will use Instrumental Variables to estimate the effect of a search period on the number of weeks that someone receives benefits. You can use the default option of the caseworker as an instrument for whether an applicant received a search period.

4. State both the instrument exogeneity and instrument relevance condition, and indicate whether you think they hold in this case. Explain your answer.

5. Run first stage regression (without controls). How do the numbers compare to the table from the first question? Given an exact interpretation of the coefficients.

6. Is there a weak instrument problem or not? Explain your answer

7. Use the predicted values from your first stage to estimate the second stage equation. Next, estimate a model that estimates your Instrumental Variables in one step, using the ivreg() function from the AER package. Compare the results of the second stage with your 2SLS estimates. What is the difference?

8. Based on the results, what do you conclude about the effect of a search period on the number of weeks an individual receives welfare benefits?

9. Explain the type of treatment effect that you estimated, and under which additional assumption your estimated coefficient is consistent for this treatment effect. Argue whether you believe this additional assumption is likely to hold or not.

10. From the data, determine the four elements that make up the WALD estimator. Next, compute the WALD estimator using these elements. Compare your results to those of question 7.

11. Can you test whether *searchperiod* is indeed an endogenous variable? If yes, perform this test. Clearly state the null-hypothesis of the test. What do you conclude?

12. Finally, if a searchperiod reduces the likelihood to receive benefits, that's only good news if individuals actually find jobs instead. Using the ivreg() function, determine the effect of a search period on the number of weeks individuals have a job (i.e. a positive wage) and their total income in the first 26 weeks after application. What do you conclude?

13. Include your script as an appendix.

Part II:

Suppose you are interested in the effect of smoking cigarettes on health. The amount of cigarettes an individual smokes is likely related to many other things that may affect an individuals health, for example the whether or how often an individual exercises, making cigarettes smoked an endogenous variable. For each of the following potential instruments, briefly discuss whether you would use them by discussing both instrument relevance and the instrument exogeneity.

14. (a) The number of units of alcohol consumed by the individual.
    (b) An increase in cigarette taxes.
    (c) The introduction of smoking bans in bars and restaurants.