Econometrics II. Assignment 1: Sample selection model, weak instrument problem, and quantile regression

Due date: Tuesday, January 11, 11.59 am (pay attention it's midday and not midnight). Hand in your solutions as a **single .pdf file** including your code via Canvas. Include your R (or any other language) code by using R Markdown (preferred) or by using the package "minted" in your .tex file (see a template on Canvas). Each team has to come up with a unique name (without names or student numbers). Both teammates have to submit solutions via Canvas.

Question 1: Sample selection model

The dataset `assignment1a.csv(\.dta)` contains information on earnings of workers, their characteristics, an indicator based on the distance between the secondary school and the residence of an individual while at school-going age, and an indicator depending on regional subsidies of families for covering school expenses. A higher value for the first indicator implies a higher distance. Similarly, a higher indicator for the latter indicator implies higher amount of regional subsidies.

Variable	Description
logwage	log of earnings
age	age
agesq	age squared
married	1 if married
schooling	years of schooling
distance	distance between secondary school and residence
subsidy	regional subsidy for school expenses

A researcher aims to gain insight in the potential earnings of the non-employed (in the data, the non-employed can be identified by a missing value for the earnings variable). She realizes that the sample of observed wages may be subject to sample selection.

(i) Run an OLS regression for log-earnings on schooling, age, and age squared. Present the results and comment on the estimates.

(ii) Briefly discuss the sample selection problem that may arise in using these OLS estimates for the purpose of predicting the potential earnings of the non-employed. Formulate the sample selection model. In your answer, include an explanation why OLS may fail in this context.

(iii) Choose a suitable candidate as an exclusion restriction for the sample selection model. Motivate your choice. Estimate the sample selection model with the Heckman two-step estimator both with and without the exclusion restriction and compare the outcomes.

(iv) Estimate the sample selection model with maximum likelihood both with and without the exclusion restriction and compare the outcomes.

(v) On the basis of your results, how would you specify the distribution of potential earnings for the non-employed?

Question 2: Weak instrument problem

The same researcher is interested in estimating the causal effect of schooling on earnings for employed individuals only. As a consequence, she performs the subsequent analysis on the (sub)sample of employed individuals. Use the dataset from the previous exercise.

(i) Discuss the estimation of the causal effect of schooling on earnings by OLS. In particular, address whether or not it is plausible that regularity conditions for applying OLS are satisfied.

The researcher has collected data on two potential instrumental variables *subsidy* and *distance* for years of schooling.

- *distance* measures the distance between the school location and the residence of the individual while at school going age.
- *subsidy* is an indicator depending on regional subsidies of families for covering school expenses.

The researcher has the option to use only *distance* as an instrumental variable, or to use only *subsidy* as an instrumental variable, or to use both *distance* and *subsidy* as instrumental variables.

(ii) Perform instrumental variables estimation for these three options. Which option do you prefer? Include in your answer the necessary analyses and numbers on which you base your choice.

(iii) Compare the IV estimates with the OLS outcomes. Under which conditions would you prefer OLS over IV? Perform a test and use the outcome of the test to support your choice between OLS and IV. Motivate your choice.

Question 3: Quantile regression

In this exercise you will use a dataset on medical expenditures `assignment1b.csv(\.dta)`. The dataset includes the following variables:

Variable	Description
lntotexp	log of total medical expenditure
age	age
female	1 if female
white	1 if white
totchr	number of chronic problems
suppins	1 if has a supplementary private insurance

Consider the following quantile regression model for q^{th} quantile:

$$Y_q(LnTotExp_i|X_i) = \beta_{0q} + \beta_{1q}TotChr_i + \beta_{2q}SuppIns_i + \beta_{3q}Age_i + \beta_{4q}Female_i + \beta_{5q}White_i + u_i$$

(i) Create a quantile plot for the log of total medical expenditure. Draw vertical lines to indicate the median, the 10^{th} percentile, and the 90^{th} percentile. Describe your plot.

(ii) Estimate the model for the quantiles $q = 0.1$, 0.25, 0.5, 0.75 and 0.9. Briefly explain your result. Compare quantile regression results to OLS estimates.

(iii) Graph the estimated coefficients from the quantile regressions for q from 0.05 to 0.95 in increments of 0.05, together with their 95% confidence interval and the corresponding estimates from a linear regression (and their 95% confidence interval). Discuss your findings.