Econometrics II: Assignment 1, Sample selection models & Instrumental variables estimation

The dataset logEarnings.dta(\.csv) contains information on earnings of workers, the years of schooling, age, the square of age, marriage status of the worker, an indicator based on the distance between the secondary school and the residence of the 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 regional subsidies.

Variable name	Description
age	age
age2	square of age
distance	distance between secondary school and residence
married	1 if married, 0 if not
schooling	years of schooling
subsidy	regional subsidy for school expenses
logWage	log of earnings

1. **The sample selection model.** 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.

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

   (b) 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 wht OLS may fail in this context.

   (c) Which variable in your data may be a suitable candidate as an exclusion restriction for the sample selection model?

   (d) Estimate the sample selection model with the Heckman two-step estimator, both with and without the exclusion restriction and compare the outcomes.

(e) Estimate the sample selection model with Maximum Likelihood, both with and without the exclusion restriction and compare the outcomes.

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

2. **Earnings and schooling.** 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.

   (a) 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.

   (b) 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 the instrumental variable *subsidy*, or to use both *distance* and *subsidy* as instrumental variables. 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.

   (c) 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.