# Econometrics II - Assignment 1

Floris Holstege, Stanislav Avdeev

## Exercise 1

### A

The only statistically significant variable (5% threshold) is the measure for schooling. On average, an increase of one year of schooling leads to an 22% increase in earnings.

Table 1:

| | *Dependent variable:* |
|---|---|
| | Log of Wage |
| Schooling | $0.22^{***}$ |
| | (0.032) |
| Age | $-0.342$ |
| | (0.52) |
| $\text{Age}^2$ | $-0.011$ |
| | (0.01) |
| Intercept | $26.409^{***}$ |
| | (8.06) |
| Observations | 416 |
| $R^2$ | 0.815 |
| Adjusted $R^2$ | 0.813 |
| Residual Std. Error | 1.499 (df = 412) |
| F Statistic | $604.261^{***}$ (df = 3; 412) |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

### B

The problem is that there is selection on unobservables and observables. We have a selected sample, excluding the unemployed. This means we have no info on their potential earnings, if they were employed. We cannot just extrapolate our findings in the selected sample to the non-employed, since the selection influences both unobservable variables (for example; the unemployed might have less motivation) and observable variables (for instance; the effect of years of schooling on earnings might be less strong for people who have not worked for some years).

We model this selection bias in two steps. First, we indicate if the dependent variable is observed as follows:

$$I_i^* = Z_i'\gamma + V_i$$

$I_i$ takes value 1 if $I_i^* > 0$, and 0 if $I_i^* \leq 0$. Second, we define our latent variable $Y_i^*$.

$$Y_i^* = X_i'\beta + U_i$$

We use these two definitions to define our observed dependent variable, $Y_i$.

$$Y_i = \begin{cases} Y_i^* & \text{for } I_i = 1 \\ Missing & \text{for } I_i = 0 \end{cases}$$

If we try to estimate $Y_i$ with OLS, it will be consistent under one of two conditions. First, if $U_i$ and $V_i$ are independent. The intuition here is that in this case, there isis random sampling. But this is not the case here, since the observations are removed from the selected sample based on a criterion (unemployment). Second, if $X_i$ and $Z_i$ are uncorrelated. The intuition here is that when this holds, the variables that determine if one falls outside of the selected sample ($Z_i$) are unrelated to the the independent variables used in the selected sample ($X_i$), and thus the zero mean condition ($E(U_i|X_i) = 0$) still holds. If neither of these two conditions hold, the OLS is inconsistent in this context.

## C

The variable for the exclusion restriction should fulfill two conditions. First, the variable should have explanatory power when determining $I_i$. In this case, the variable should realistically influence the likelihood someone is unemployed. Second, the variable should be unrelated to the dependent variable $Y_i$, in this case earnings. From the variables available, our best candidate appears to be the dummy variable of married (1 if married, 0 if not). Regarding the first condition; When someone is married, this affect their willingness to continue working, since they share income with their partner. Regarding the second condition; if one assumes that wages are reflective of someones productivity, the fact that someone is married, should not meaningfully change one's productivity. However, there are some reasons to believe marriage still impacts wages - for one, married couples are more likely to have a baby, which likely affects ones earnings. But given the available variables, this appears the one that comes closest to fulfilling both criteria. Looking at the correlations, these confirm our intuitions; the correlation between marriage and unemployment is 0.17, but only 0.02 with logged wages.

## D

We first estimate without the exclusion restriction, in which $X_i = Z_i$ for the Heckman estimator. Because these terms are equal, there is perfect collinearity in the second step of the Heckman estimator, biasing the estimates. To explain this a bit more formally, consider the two steps in the Heckman estimator:

$$\text{Step 1: Using a probit model, estimate } I_i \text{ with: } \hat{I}_i = Z_i'\hat{\gamma}_i + V_i$$

$$\text{Step 2: Estimate } \beta \text{ and } \rho\sigma \text{ with:} Y_i = X_i\beta + \rho\sigma\frac{\phi(Z_i'\hat{\gamma}_i)}{\Phi(Z_i'\hat{\gamma}_i)} + U_i^*$$

# Exercise 2

##A

A key condition for OLS being the best linear unbiased estimator (BLUE) is that the errors are uncorrelated with the independent variables - the so-called exogeneity condition, or more formally, $plim(\frac{1}{n}X_i'\epsilon_i) = 0$. There are several reasons why this might not hold - one is that there is another variable that influences both $X_i$ and $y_i$. In the case of schooling ($X_i$) and earnings ($y_i$), there are several variables that affect both; for instance, racist sentiments in society both affect people's ability to obtain schooling and find a job, and socio-economic privilege of one's family likely make both schooling and the job search easier. Given this, it seems implausible that this condition of OLS is satisfied.

## B

## C

If all the independent variables are exogenous, then the OLS estimator is the most efficient consistent estimator - the IV estimator will still be consistent, but according to the Gaus-Markov theorem less efficient than OLS. We use the Hausman test to test for the exogeneity of the independent variable (schooling). (. . . ). In the Hausman test with distance and subsidy as the instrument variables, the Hausman test rejects the null hypothesis (exogenous independent variables, OLS is consistent) at a 5% threshold. This means we should prefer the IV estimator with distance and subsidy as the instrument variables.