Microeconometrics

Spring Semester 2019/20

Problem set 1: Solution topics

(Please ignore any obvious typos)

Exercise 1 (Ordinary Least Squares | Generalized Least Squares)

For this exercise please create a do-file and corresponding log-file to hand in with your solution.

Use the dataset **wage1.dta** to estimate the determinants of wages. Consider the following regression model:

$$lwage_i = \beta_0 + \beta_1 exper_i + \beta_2 exper_i^2 + \beta_3 married_i + \beta_4 female_i + \beta_5 married_i \times female_i + u_i,$$

where the dependent variable $lwage$ is the logarithm of hourly wages. The independent variables are the individuals' experience (in years) and squared experience ($exper$ and $exper^2$, respectively), a dummy variable equal to 1 for married individuals and 0 for single individuals ($married$), a dummy variable equal to 1 for females and 0 for males ($female$), and an interaction term between married and female ($married \times female$).

(a) Estimate this regression model by OLS. Provide a rigorous interpretation of the regression coefficients.

$$\widehat{lwage} = 1.348 + 0.0356 exper_i - 0.00077 exper_i^2 + 0.3150 married_i - 0.1443 female_i - 0.3588 married_i \times female_i$$

Therefore, the marginal impact of an additional year of experience is given by the expression $(0.0356 - 2 \times 0.00077 \times exper) \times 100\%$*, i.e., it depends on the number of years of experience itself.*

Everything else constant:

$$\text{Single males: } E[lwage|married = 0, female = 0] = \widehat{\beta}_0$$
$$\text{Married males: } E[lwage|married = 1, female = 0] = \widehat{\beta}_0 + \widehat{\beta}_3$$
$$\text{Single females: } E[lwage|married = 0, female = 1] = \widehat{\beta}_0 + \widehat{\beta}_4$$
$$\text{Married females: } E[lwage|married = 1, female = 1] = \widehat{\beta}_0 + \widehat{\beta}_3 + \widehat{\beta}_4 + \widehat{\beta}_5$$

Therefore,

$\widehat{\beta}_3 = 0.3149$: married males earn on average 31.49% more than single males, *ceteris paribus*.

$\widehat{\beta}_4 = -0.1443$: single females earn on average 14.43% less than single males, *ceteris paribus*.

$\widehat{\beta}_3 + \widehat{\beta}_4 + \widehat{\beta}_5 = 0.3150 - 0.1443 - 0.3588 = -0.1881$: married females earn on average 18.81% less than single males, *ceteris paribus*.

(b) Assuming the Gauss-Markov assumptions hold in the theoretical model and holding the other factors constant, what is the number of years of experience that maximizes the average logarithm of wages?

The number of years of experience that maximizes the average logarithm of wages is: $\frac{\partial lwage}{\partial exper} = 0 \Leftrightarrow exper^* = 0.0356/(2 \times 0.00077) \approx 23$

(c) Assuming the Gauss-Markov assumptions hold in the theoretical model, does experience have a statistically significant impact on the logarithm of wages? State clearly the test statistic, the null and alternative hypotheses, and the decision rule you use.

The null and alternative hypotheses are:

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_a : \text{Not } H_0 \text{ (this is equivalent to at least one of these coefficients is different from 0)}$$

The F-stat ($F(2, 520)$) equals 19.39 and the associated p-value equals 0.000. Therefore we reject the null hypothesis in favor of the alternative hypothesis at 1% significance level and conclude that experience has a statistically significant impact on wages.

(d) Perform a heteroskedasticity test in the estimated regression model. State clearly the null and alternative hypotheses, the test statistic you consider, and the rejection rule. What do you conclude?

The White test is a test of heteroskedasticity that allows for nonlinear forms of heteroskedasticity. The alternative White test is based on the following equation:

$$\widehat{u}_i^2 = \delta_0 + \delta_1 \widehat{lwage}_i + \delta_2 \widehat{lwage}_i^2 + v_i$$

where $\widehat{u}^2$ *denote the squared residuals and* $\widehat{lwage}$ *and* $\widehat{lwage}^2$ *denote the fitted values and squared fitted values, respectively, of the initial model.*

The null and alternative hypotheses are:

$$H_0 : \delta_1 = \delta_2 = 0 \text{ (Homoskedasticity)}$$

$$H_a : \text{Not } H_0$$

The F-stat under this null hypothesis $F(2, 523) = 3.56$ *and the associated p-value is 0.0291. Therefore, we reject the null hypothesis in favor of the alternative hypothesis at 5% significance level.*

(e) One possible heteroskedasticity test is to regress the squared residuals on the set of explanatory variables and the squared fitted values of the initial regression model. Someone suggests adding the fitted values of the initial regression model to this heteroskedasticity test. Is it a good idea? Explain.

This is not a good idea because the fitted values are a perfect linear combination of the explanatory variables and therefore they would be perfect collinear (the coefficient of the fitted values wouldn't be identified).

(f) Estimate the model using a feasible generalized least squares estimator. Explain the steps of the estimation procedure. Why is it important to correct for the problem of heteroskedasticity?

It is important to correct for heteroskedasticity because OLS estimators are no longer BLUE in the presence of heteroskedasticity. This means that they are not necessarily the minimum variance unbiased estimators among the linear estimators even though they are still unbiased and consistent if the remaining Gauss-Markov assumptions hold.
In the presence of heteroskedasticity inference may be not valid (standard errors, t-statistics, F-statistics,... may be wrong).

We use FLGS when we don't know the expression of the variance. We assume that $var(u|\mathbf{X}) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + ... + \delta_k x_k)$. *The procedure is as follows:*

(i) *Run the regression of lwage on exper, expersq, married, female, and* $married \times female$ *and obtain the residuals* $\widehat{u}$

[Stata: reg lwage exper expersq married female married_female
predict uhat, res];

(ii) *Create the* $\log(\widehat{u}^2)$

[Stata: gen $uhatsq = uhat \times uhat$
gen $\log\_uhatsq = \log(uhatsq)$];

(iii) *Run a regression of the* $\log\_uhatsq$ *on exper, expersq, married, female, and* $married \times female$ *and obtain the fitted values,* $\widehat{g}$

[Stata: reg lwage exper expersq married female married_female

predict ghat, xb];

(iv) *Create the exponential of the fitted values:* $\widehat{h} = \exp(\widehat{g})$

Stata: gen hhat=exp(ghat);

(v) *Estimate the original model by WLS using as weights* $1/\widehat{h}$

reg lwage exper expersq married female married_female [aweight = hhat]

(g) Consider now a simpler version of the initial model:

$$lwage_i = \beta_0 + \beta_1 exper_i + \beta_2 married_i + \beta_3 female_i + u_i,$$

Rewrite the regression model given above in a way that allows you to test directly (i.e. by performing a simple t-test on a coefficient) whether the impact of one additional year of experience on the logarithm of wages is double the impact of being married. Perform the test describing carefully the null and alternative hypotheses of the test, the test statistic, and the rejection rule.

The null and alternative hypotheses would be $H_0 : \beta_1 = 2\beta_2$ *and* $H_a : \beta_1 \neq 2\beta_2$. *We can rewrite the null hypothesis as* $H_0 : \beta_1 - 2\beta_2 = 0$. *Therefore, we can define* $\theta = \beta_1 - 2 \times \beta_2$ *or equivalently* $\beta_1 = \theta + 2\beta_2$. *Then,*

$$\begin{aligned} lwage_i &= \beta_0 + (\theta + 2\beta_2) exper_i + \beta_2 married_i + \beta_3 female_i + u_i \\ &= \beta_0 + \theta exper_i + 2\beta_2 exper_i + \beta_2 married_i + \beta_3 female_i + u_i \\ &= \beta_0 + \theta exper_i + \beta_2(2exper_i + married_i) + \beta_3 female_i + u_i \end{aligned}$$

Then, $H_0 : \theta = 0$ *and* $H_a : \theta \neq 0$ *allows us to directly test the hypothesis stated above. The t-statistic associated with* $\widehat{\theta}$ *equals -4.79 and the associated p-value is 0.0000. Therefore we reject the null* *hypothesis in favor of the alternative hypothesis and we find no evidence in favor of the hypothesis that the impact of an additional year of experience is double the of being married.*

Exercise 2 (Instrumental variables)

For this exercise please create a do-file and corresponding log-file to hand in with your solution.

Consider the dataset **QOB.dta** that is comprised of a subset of the data used in the seminal paper by Angrist and Krueger (1991) "Does compulsory school attendance affect schooling and earnings?". In the paper the quarter of birth of individuals is used as an instrument for education in order to estimate the impact of compulsory school on earnings. The authors use samples from Census data for men born in 1920s, 1930s, and 1940s.

(a) Compute some descriptive statistics in order to get a general idea of the data. In particular, calculate the average number of years of schooling by each quarter of birth for men born in the 1920s, the 1930s, and the 1940s. What do you observe?

Stata commands: sum
sum EDUC if $YOB \geq 30$ *&* $YOB \leq 39$
sum EDUC if $YOB \geq 40$ *&* $YOB \leq 49$
bys QOB: sum EDUC if $YOB \geq 30$ *&* $YOB \leq 39$
bys QOB: sum EDUC if $YOB \geq 40$ *&* $YOB \leq 49$

For the next parts of the exercise consider the sub-sample of men born in 1930-1939.

(b) Estimate the returns to education by OLS using age and squared age as control variables. Interpret the results and explain why the estimated returns to education might not have a causal interpretation.

Stata commands: reg LWAGE EDUC AGE AGESQ

The marginal impact on wages of getting older by one year is given by $(0.0711 - 2 \times 0.0034 AGE) \times 100\%$.

One additional year of schooling increases predicted wages by approximately 7.1%, on average, ceteris paribus. This impact may not have a causal interpretation because there are concerns about omitted variable bias in this model. Perhaps, innate ability is clearly correlated with both years of schooling and wages.

(c) Explain why the quarter of birth might be a good instrument for education when estimating returns to education.

"If the fraction of students who desire to leave school before they reach the legal dropout age is constant across birthdays, a student's birthday should be expected to influence his or her ultimate educational attainment. This relationship expected because, in the absence of rolling admissions to school, students born in different months of the year start school at different ages. This fact, in conjunction with compulsory schooling laws, which require students to attend school until they reach a specified birthday, produces a correlation between date of birth and years of schooling.
Students who are born early in the calendar year are typically older when they enter school than children born late in the year." (Angrist and Krueger, 1991)

(d) Construct a dummy variable, *first_qob*, which equals one for men born in the first quarter of the year and zero otherwise. Compute the IV estimate $IV = (Z'X)^{-1}Z'Y$ of returns to education considering *first_qob* as instrumental variable. Compare the estimate with the OLS returns to education estimated in the regression of *lwage* on a constant and years of education.

Stata commands: gen first_qob= (QOB==1)
ivregress 2sls LWAGE (EDUC = first_qob)

In the IV setting one additional year of schooling is associated with an approximately 10.2% increase on wages, on average, ceteris paribus.

OLS regression suggests that one additional year of schooling is associated with an approximately 7.1% increase on wages, on average, ceteris paribus.

(e) Suppose in the following that we have three instrumental variables, $Z1$, $Z2$, and $Z3$ representing dummy variables for first-, second-, and third-quarter births, respectively. Generate these three instrumental variables.

gen Z1 = first_qob

gen Z2= (QOB==2)

gen Z3= (QOB==3)

(i) Describe and estimate the first-stage equation with multiple instruments. Include the following explanatory variables as additional control variables: *age*, *agesq*, *race*, *married*, and *smsa*.

In the first-stage equation, the dependent variable is the endogenous variable (EDUC) and the explanatory variables are the set of instruments and the exogenous variables.

[Stata: reg EDUC Z1 Z2 Z3 AGE AGESQ RACE MARRIED SMSA]

(ii) Compute an F-test under the null hypothesis that the quarter of birth dummy variables have no effect on the total years of education. Are these instruments valid?

Let δ_1, δ_2, and δ_3 denote the coefficients on $Z1$, $Z2$, and $Z3$, respectively. The null and alternative hypotheses in a test of joint significance of the three quarter of birth dummy variables are:

$$H_0 : \delta_1 = \delta_2 = \delta_3 = 0$$
$$H_a : \text{Not } H_0$$

The joint significance test yields an F-statistic equal to 12.28 and the associated p-value is 0.000. Therefore we reject the null hypothesis in favor of the alternative hypothesis and the instruments are jointly statistically significant (a rule of thumb for validity of the instruments in the first stage is an F-statistic greater than 10).

Stata: test Z1 Z2 Z3

(iii) Estimate the returns to education by 2SLS and compare the results to standard OLS estimates [consider the same set of control variables as in part (ii)].

[Stata: ivregress 2sls LWAGE (EDUC= Z1 Z2 Z3) AGE AGESQ RACE MARRIED SMSA]

In this regression model, an additional year of schooling is associated with a 12.49% increase in wages, on average, ceteris paribus. OLS regression yields an estimate of the returns to schooling equal to 0.0642, i.e., an additional year of schooling is associated with a 6.42% increase in wages, on average, ceteris paribus.

(iv) Are the instruments exogenous?

We can perform a test of instrument exogeneity. The overidentification test can be described as follows: obtain the residuals from 2SLS regression such that $uhat\_IV = y - x\widehat{\beta}_{IV}$ *; then run a regression where the dependent variable are the residuals and the independent variables are the instruments and the exogenous variables (*$uhat\_IV = z\gamma + \epsilon$*). Test the joint significance of the instruments: if the instruments are exogenous then we should fail to reject the null hypothesis that all the coefficients are equal to zero (*$H_0 : \gamma = 0$*). The test statistic is equal to mF and follows a* χ^2_{m-k}.

More specifically, the null hypothesis $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$ *in the following equation stands for exogeneity of the instruments, whereas the alternative hypothesis* H_a *: Not* H_0 *says that at least one instrument is not exogenous.*

$$uhat\_IV = \gamma_0 + \gamma_1 Z1 + \gamma_2 Z2 + \gamma_3 Z_3 + v$$

The F-statistic of global significance of this regression is equal to 1.80 (with associated p-value equal to 0.1442) and therefore we fail to reject the null hypothesis that the instruments are exogenous at 10% significance level.

Stata commands:

ivregress 2sls LWAGE (EDUC= Z1 Z2 Z3) AGE AGESQ RACE MARRIED SMSA

predict uhat_IV, res

```
reg uhat_IV Z1 Z2 Z3
```

Alternatively:

```
ivregress 2sls LWAGE (EDUC= Z1 Z2 Z3) AGE AGESQ RACE MARRIED SMSA, robust
estat overid
```