André Filipe Silva – 26005
João Seixo - 40510
Márcia Serra - 411221

EXERCISE 1

a)

exper – $\beta_1$and β_2 must be interpreted jointly, as it is a polynomial of the form $\beta_1 exper + \beta_2 exper^2$.

β_1 *and* β_2*:* For a male, single individual, one extra year of experience leads to a (3.56215-0.1538**exper*)% variation in hourly wage on average, all else constant. This means that there are decreasing marginal returns for experience after reaching a certain level – more experience increases hourly wage, but in a deaccelerating fashion.

β_3: A male, married individual, will gain on average 31.49537% more in hourly wage than a male, single individual, all else constant.

β_4*:* A female, single individual, will earn -14.43168% less in hourly wage on average than a male, single individual, all else constant.

β_5*:* A married and female individual will receive, on average, -35.87692% less in hourly wage than other gender/marital status counterparts, all else constant. This means that there is a constant hourly wage gap between these two groups.

b)

To find this, we simply take the derivative of *lwage* in order to *exper*. The result is, rounded to an integer: 23.

c)

$H_0: \beta_1 = \beta_2 = 0$

$H_1: \exists \beta_j \neq 0, j = 1,2$

F- Test following an F distribution.

F~(2,250)=19,39 > 3,03 → We reject the null (5% significance).

As such, we can conclude, for a 5% significance level, that experience is statistically significant in explaining the logarithm of wages.

André Filipe Silva – 26005
João Seixo - 40510
Márcia Serra - 411221

d)

We use White's test. This consists on regressing the estimated squared errors on the original regression variables, the squared terms of the variables and the cross-terms of the variables. We can't include squares or cross-terms of dummy variables, as they would be collinear with the original variables.

$$(e_i)^2 = \beta_0 + \beta_1 exper + \beta_2 exper^2 + \beta_3 married + \beta_4 female + \beta_5 female * married + \beta_6 exper^3 + \beta_7 exper^4 + \beta_8 exper * married + \beta_9 exper * female + \beta_{10} exper^2 * married + \beta_{11} exper^2 * female + \beta_{12} exper * (married * female) + \beta_{13} exper^2 * (married * female) + \mu_i$$

$H_0: \beta_1 = \beta_2 = \cdots = \beta_{13} = 0$

$H_1: \exists \beta_j \neq 0, j = 1,2, \ldots ,13$

Test – statistic: n*(R^2) ~ $\chi 2$ [13] , 5% significance

We reject if $\chi 2$ observed > $\chi 2$ critical

$\chi 2$ observed = 24.53 > $\chi 2$[13], 5% significance

We reject the null hypothesis. So, we conclude for the presence of heteroskedasticity.

e)

This is not a good idea as having the squared fitted values of the initial regression model and the fitted values of the initial regression model will lead to a multicollinearity problem. With that said, we could find better results by following White's methodology of using cross-products in addition, instead of the predicted values of the initial model.

André Filipe Silva – 26005
João Seixo - 40510
Márcia Serra - 411221

f)

We have to correct for the problem of heteroskedasticity because, when're estimating a regression by OLS, there are certain assumptions we must respect in order for the estimations to be Best, Linear, Unbiased and Efficient (BLUE). One of those assumptions is the assumption of homoskedasticity of the error variance.

What this assumption states, is that the variance of the error term has a homogeneous distribution throughout all values of the independent variable. Now, what usually happens is that the variance of the error is not homogenous throughout all values of the independent variable. The variance-covariance matrix is now different, and our estimations by OLS will be inefficient.

The resulting standard errors will be either under or over-estimated and inference will become impossible. As such, we have to fix this problem. One of the possible procedures is Feasible Generalized Least Squares.

The FGLS estimation method is a method in which we try to estimate the variance-covariance matrix empirically by the use of several procedures. After estimating this matrix, we use it in the OLS estimation process instead of the original variance-covariance matrix. In here we will present two procedures. One that was taught and presented in class and another that we found on the following source, (Yamano, 2009).

Following the method presented in class, these were the steps we took in order to estimate the variance-covariance matrix and then use it to correct for heteroskedasticity in the model.

We started by regressing the original regression and then predicting the residuals. We then followed by transforming the predicted residuals by the logarithm of the squared predicted residuals. From this transformation, we proceeded to an auxiliary regression. This auxiliary regression's dependent variable was the logarithm of the squared residuals and the independent variables were the original OLS dependent variables.

After this, we predicted the residuals of the auxiliary regression and applied an exponential transformation to get the weights.

With these weights, we then applied the WLS methodology and transformed the original model. Finally, we ran OLS through the transformed model and indeed our outputs and standard errors changed.

Nonetheless we did a heteroskedasticity test and we found out that our transformation did not correct for heteroskedasticity.

The other procedure, based on what we learnt in class, was the following:
We started with the initial regression and predicted the residuals from that regression. We also applied a transformation to get the squared predicted residuals and generated an extra variable of 1's, called uno.
We then transformed all the independent variables by getting the absolute value of all of them. From that point onwards, like in the previous procedure, we also did a auxiliary regression.
Now, the difference lies in the structure of the auxiliary regression. This regression's dependent variable was only the squared predicted residuals of the initial regression on an exponential transformation of the fitted values of the transformed absolute independent variables and the variable uno we created. We also applied some restrictions to this regression. It was non-linear and with no log.
From this auxiliary regression we predicted the fitted values of the dependent variable. This differs from the previous methodology. Our weights are different now.
The final step was also slightly different. In here we estimated the initial regression, transformed by the new weights we found out on the previous steps.
Despite the differences in methodology in determining the weights, in both cases the final model was not homoscedastic. This highlights the need to use robust standard errors with FGLS estimators.

g)

What we want to test is the following:

$$H_0: \beta_1 - 2\beta_2 = 0$$
$$H_A: \beta_1 - 2\beta_2 \neq 0$$

In order to test directly with a simple t-test, we have to rewrite our model. Thus, our new parameter of interest will be $\theta = \beta_1 - 2\beta_2$. We can transform our original model, as explained below.

$$lwage = \beta_0 + \beta_1\, exper + \beta_2\, married + \beta_3\, female + \mu_i$$
$$\Leftrightarrow \quad lwage = \beta_0 + (\theta + 2\beta_2)exper + \beta_2 married + \beta_3 female + \mu_i$$

$$\Leftrightarrow \quad lwage = \beta_0 + \theta_1 exper + 2\beta_2 exper + \beta_2 married + \beta_3 female + \mu_i$$

$$\Leftrightarrow \quad lwage = \beta_0 + \theta_1 exper + \beta_2 (2 * exper + married) + \beta_3 female + \mu_i$$

$$newvar = 2 * exper + married$$

$$\boxed{lwage = \beta_0 + \theta_1 exper + \beta_2 newvar + \beta_3 female + \mu_i}$$

It is important to note that this model is completely equivalent to the original model – it is just transformed in order to suit our purposes.

With this, we just need to test the following:

$$H_0: \theta_1 = 0$$
$$H_a: \theta_1 \neq 0$$

This follows a t-test, with a t-student distribution, and we reject the null if t-observed > t-critical.

We can look directly at the p-value on the regression, and it equals 0.000. As such, for a 5% significance level, we can conclude that the impact of one additional year of experience on the logarithm of wages is not double the impact of being married.

EXERCISE 2

a)

Firstly, used the command “describe” to see the number of observations and the description of the different variables. This database has 816 435 observations and 10 different variables.

Secondly, in order to have a broad overview on the data, we used the command “Summarize”. The average age of the individuals in the sample is 38,48 years and the average number of education years is 13,24.

After this, we divided the sample in two groups: individuals born in the 30’s and individuals born in the 40’s. After that, we divided the groups into quarter of birth. We concluded that there are no major differences on years of education depending on the

quarter of birth of the individual. For people born in the 30's and in the first quarter, the average number of school years was 12.79116, while for individuals also born in the 30's but in the last quarter was 12.80084, very similar as we have stated.

b)

According to the simple OLS model, an increase of one year in education, leads to a 7.10821% increase, on average, in wages, all else constant.

However, we suspect that Education is an endogenous variable that is determined by other omitted variables that are simultaneously correlated with lwages and education. If that is the case that $cov(education, u_i) \neq 0$, then it would violate the Gauss-Markov assumption of exogeneity and our estimates would be biased and inconsistent. If that is actually true, this model has nothing to say in what concerns to causation, the maximum it can do is state that the variables education and lwages are positively correlated, but it can't say anything about if it is education that is really leading the increase in wages.

c)

For an instrument to be valid it has to be relevant and exogenous. The first condition is possible to prove through a significance tests, however, the second condition is not provable in an objective way. In this case, quarter of birth seems to be a good instrument because it is related to the educational attainment. In the US, education is mandatory until the age of 16. Thus, people who were born in the first months of the year will be able to leave school earlier than those who were born closer to the end of the year. This seems to satisfy the relevance condition (QOB influences EDUC), but of course, further formal tests would be needed to effectively prove this. Regarding the exogeneity condition: if the moment when one is born is uncorrelated with unobservable factors then this second necessary condition for a proper IV will also be satisfied. This is expectable as births can be thought of being independently distributed throughout the year without any particular justification for occurring in any quarter. If quarter of birth was to be correlated with any factor then we could expect to find one or more quarters of the year with a significant difference in births, but that doesn't happen. Thus, it seems that the IV could also fulfill the exclusion restriction, but we will further discuss this issue in e) iv.

André Filipe Silva – 26005
João Seixo - 40510
Márcia Serra - 411221

d)

OLS Estimate: 0.070851

IV Estimate: 0.101995

The IV estimate is larger than the OLS estimate. There are several hypotheses that could explain this behaviour. On one hand, it could that OLS had a negative bias. On the other hand, it could be also that our instrument was poorly chosen and did not respect the exclusion restriction.

As we know, EDUC is correlated with several other variables on the error term such as ability, or the number of years the progenitors went to school. This created a problem of endogeneity which led the OLS estimation to be biased.

Now, the direction of the bias is a discussion out of the scope of this assignment, but it could very well be that the bias was negative and that the IV estimates are the true ones.

Nonetheless, we cannot exclude other possibilities related to the instrument itself. It could be that the average LWAGE for the sub-population group (QOB=1), is higher than the average LWAGE for the entire population set, thus pushing the IV estimates higher.

Another possible idea to keep in mind, and the most likely one, is that our instrument is poorly chosen. We could be in the presence of a weak instrument. QOB could in fact be weakly correlated with EDUC. The problem that arises from here is that, if the covariance of EDUC and QOB=1 is low, while our instrument is not exogenous, then a bias will arise.

If in fact, COV(EDUC, z1) is lower than COV(z1, u), then the bias will be even higher. The resulting estimations might be even worse than the original OLS estimations.

The question that arises is: Is there significant reason to suspect that the instrument is weakly correlated with EDUC? And, we have to argue that there might be. Educational attainment might in fact be more correlated with the level of education attainment of the individual's parents or the individual's ability than with the individual's quarter of birth.

Then, we can also suggest that the instrument does not respect the exclusion restriction as the QOB could be related with socio-economic factors such as Holidays, social background, personal beliefs, number of romantic travels throughout the year, amongst others. This would then show that $Cov(z1, u) \neq 0$, further affecting the validity of our IV estimates.

André Filipe Silva – 26005
João Seixo - 40510
Márcia Serra - 411221

e.i)

$$EDUC = \beta_0 + \beta_1 z_1 + \beta_2 z2 + \beta_3 z3 + \beta_4 AGE + \beta_5 AGE^2 + \beta_6 RACE + \beta_7 MARRIED + \beta_8 SMSA + \mu_i$$

In the first stage of the 2SLS estimation procedure, we regress the independent endogenous variable to the instruments and the remaining exogenous variables, which are our controls.

In this step we are aiming to find an estimate of EDUC which is explained by the instruments and the controls. This estimated EDUC will then be used on the second-stage of the process.

e.ii)

Validity of the instruments in IV requires two restrictions: relevance and exogeneity.

We can only test for the **relevance** of the instrumental variables. We can't test neither definitively prove exogeneity – we need argumentation and a "good story". Nonetheless, despite this limitation, if we have several instruments, we can test if the extra instruments are valid, under the assumption that our first instrument is valid. Let z1 be our first instrument. We added now z2 and z3 as extra instruments, instruments in which we can now test exogeneity.

We can test this either with the Sargan test or the J-statistic. According to the Sargan-test, under the assumption that z1 is a valid instrument, we found out that z2 and z3 are valid instruments. Our test statistic was 5.4734773 for a chi2 value of 0.06478128. We rejected the null hypothesis, so that means that our new instruments are exogenous.

However, if z1 is not a valid instrument, then the tests become irrelevant. In that sense, there is no definitive and certain way of proving that instruments are valid. We can only test relevance with certainty. According to our F-test, the instruments are relevant.

e.iii)

OLS: 0.0642292

2SLS: 0.1248746

Once more the new estimates are larger than the OLS ones. This can relate to the previous problem of the validity of the instruments. Now as we are using 3 instruments, we are inclined to believe that this is no longer a problem of a larger average from the sub-sample of the population due to the instruments chosen.

It seems that the problem might come from the validity of the instruments themselves. As in the previous question, we found out that our z2 and z3 instruments are exogenous and relevant, if z1 is valid, then the problem might lie in the validity of the z1 instrument.

We discussed before that the instrument could be weakly correlated with the endogenous variable, and it is not by adding extra instruments that are weakly correlated that we ought to fix the problem.

It seems very likely that indeed the problem lies in fact in the choice of the instrument and that we could achieve a better result by finding better instruments such as education attainment level of the parents.

Another final possibility is that we don't have a problem at all and that by correcting endogeneity using 2SLS or IV, gives us larger than OLS estimates. This is actually not uncommon.

e.iv)

Performing the Sargan test, we concluded that the instruments are exogenous. However, this only holds if Z1 is exogenous.

Z1 is exogenous if we can argument that it behaves as if it was randomly assigned. This means that the instrument cannot be correlated with any other variable than the treatment. There are some reasons that support this argument. Births can be thought of being independently distributed throughout the year without any variable influencing it. As we observe homogenous number of births throughout the year, this corroborates with our argument of exogeneity.

However, it could be the case that the level of education has impact on the part of the year a person is born. If we think that more educated people have more demanding jobs with lower free time, they will probably be more willing to think about kids during vacation time, typically during summer or Christmas time, then the babies will probably be born in February, March and April or in September, violating the hypothesis of strictly

exogeneity of the instrument. Besides this, natality policies implemented at a certain quarter of the year could also be more appealing for those less educated, and, because of this, affect the exogeneity of the instrument.

Nonetheless, weighting the arguments in favor and the arguments against the exogeneity of the IV, we think that it is relatively safe to assume that all the instruments are solid to the exclusion restriction, thus being exogenous.