Worksheet **4**
Regression analysis with qualitative information

Note: In those problems that include estimations and have a reference to a data set the students should check the outputs obtained with Gretl.

1. [Based on problems 4.2 and 7.13 in Wooldridge textbook] In order to explain the salary of a CEO, *salary*, the following equation was estimated with the data in the file `CEOSAL1.GDT`:

$$\begin{aligned} \widehat{\log(salary)} &= \underset{(0,294)}{4{,}362} + \underset{(0,033)}{0{,}275}\log(sales) + \underset{(0,0039)}{0{,}0179}roe \\ n &= 209, \quad R^2 = 0{,}282 \end{aligned}$$

where *sales* is annual sales and *roe* is the nominal return on equity of a share

a) Interpret the coefficient of $\log(sales)$ and test whether it is positive and significant.

b) We decide to include a new dummy variable, *rosneg*, which equals 1 when *ros* is negative and 0 if *ros* is positive or zero, where *ros* is the real return on equity of a share, using the specification

$$\begin{aligned} \log(salary) &= \beta_0 + \beta_1\log(sales) + \beta_2 roe + \beta_3 rosneg + \beta_4\log(sales) * rosneg \\ &\quad + \beta_5 roe * rosneg + \varepsilon \end{aligned}$$

and obtaining the following OLS estimates,

$$\begin{aligned} \widehat{\log(salary)} &= \underset{(0,307)}{4{,}074} + \underset{(0,035)}{0{,}314}\log(sales) + \underset{(0,004)}{0{,}017}roe \\ &\quad + \underset{(1,009)}{2{,}094}rosneg - \underset{(0,112)}{0{,}258}\log(sales) * rosneg - \underset{(0,0178)}{0{,}00343}roe * rosneg \\ n &= 209, \quad R^2 = 0{,}315, \end{aligned}$$

we obtained the following estimated variances for the coefficients of $\log(sales)$, *roe*, *rosneg*, $\log(sales) * rosneg$ and $roe * rosneg$, respectively

$$\widehat{V}\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \\ \hat{\beta}_5 \end{pmatrix} = \begin{pmatrix} 0{,}001236 & 1{,}47E-05 & 0{,}010414 & -0{,}001236 & -1{,}47E-05 \\ 1{,}47E-05 & 1{,}64E-05 & 0{,}000410 & -1{,}47E-05 & -1{,}64E-05 \\ 0{,}010414 & 0{,}000410 & 1{,}018975 & -0{,}109247 & -0{,}003193 \\ -0{,}001236 & -1{,}47E-05 & -0{,}109247 & 0{,}012544 & -0{,}000116 \\ -1{,}47E-05 & -1{,}64E-05 & -0{,}003193 & -0{,}000116 & 0{,}000318 \end{pmatrix}.$$

Test whether it is necessary to distinguish firms in terms of the sign of *ros* in the model.

c) Test whether for firms with *ros* negative, an increase in sales leads to an increase in the CEO's salary, else constant.

d) Explain how you would test the hypothesis that for a CEO of a firm with negative *ros*, $\log(sales) = 10$ and $roe = 20$, is the same to *(a)* increasing sales until $\log(sales) = 11$, and *(b)* making *ros* become positive (everything else constant).

2. Consider the following models to explain the weight of a newborn, *bwght*. The file `BWGHT.GDT` contains data for the USA on *bwght* the weight of babies at birth (in ounces), *cigs* is the daily number of cigarettes smoked by the mother during the pregnancy, $faminc$ is the annual family income in thousands of dollars, *male* is a constructed variable indicating if the newborn is male ($male = 1$) or female ($male = 0$) and *white* is another constructed variable indicating if the newborn is white ($white = 1$) or not ($white = 0$). and $nowhite = 1 - white$.

$$\begin{aligned}
\log(\widehat{bwght}) &= \underset{(,019)}{4{,}69} - \underset{(,00085)}{0{,}0042}cigs + \underset{(,0059)}{0{,}0084}\log(faminc) + \underset{(,01)}{0{,}026}male + \underset{(,014)}{0{,}053}white \\
R^2 &= 0{,}0416, \quad n = 1388 \\
\log(\widehat{bwght}) &= 4{,}687 - 0{,}0042cigs + 0{,}0083\log(faminc) + 0{,}028male + 0{,}054white \\
&\quad -0{,}002white * male \\
R^2 &= 0{,}0417, \quad n = 1388 \\
\log(\widehat{bwght}) &= 4{,}689 - 0{,}0042cigs + 0{,}0077\log(faminc) + 0{,}028male * nowhite \\
&\quad +0{,}0677white \\
R^2 &= 0{,}0381, \quad n = 1388
\end{aligned}$$

   a) For the first equation, interpret the coefficient of the variable *cigs*. Give a 95 % confidence interval for the effect of smoking more than 10 cigarettes a day on the weight of the newborn, else constant.

   b) Consider now the first two equations. How much more weight does each model predict for a white male newborn ($male = 1$) with respect to a non white male newborn, else constant? Is this difference significant?

   c) Using the second model, estimate the weight difference between a newborn girl and a newborn boy, both white, keeping all remaining factors constant. Is the weight difference significant?

3. The following wage equations have been estimated using data on workers from Bangladesh:

$$\log(\widehat{salario}) = \underset{(0{,}35)}{1{,}25} + \underset{(0{,}03)}{0{,}15}\,hombre + \underset{(0{,}004)}{0{,}02}\ experiencia, \tag{1}$$

$$\log(\widehat{salario}) = \underset{(0{,}48)}{1{,}55} + \underset{(0{,}05)}{0{,}10}\,hombre + \underset{(0{,}005)}{0{,}015}\,experiencia - \underset{(0{,}002)}{0{,}005}\,hombre{*}experiencia, \tag{2}$$

   where *salario* is measured in US dollars and *hombre* is a binary variable taking the value of 1 if the worker is male and 0 if the worker is female, *experiencia* measures the years of work experience. The numbers in brackets are the standard errors

   a) What is the estimated average difference between a man's salary with 5 years work experience and that of a woman's with 10 years work experience? Use equation (1)

   b) What is the estimated average difference between a man's salary with 5 years work experience and that of a woman's with 10 years work experience? Use equation (2)

   c) Test that the salary difference between men and women does not depend on experience.

4. Imagine you have survey data on wages, education, labor experience and gender. Also, you have answers to a question concerning the use of marihuana. The question is the following: "How many times have you smoked marihuana in the last month?"

a) Write down an equation that allow us to estimate the effect of marihuana consumption on wages, considering the effect of other factors. The objective is to be able to make statements of this sort: Ïncreasing the consumption of marihuana by 5, would change wages in x %"

b) Specify a model that allow us to test whether the consumption of drugs has different effects of males'wages and females'wages. How would you test for this difference to be non-existent?

c) Assume that marihuana consumption is measured by dividing people into 4 categories: no consumer, occasional consumer (1 to 5 times a month), moderate consumption (6 to 10 times a month) and regular consumer (more than 10 times a month). Write down a model that allows to estimate the effects of consuming marihuana on wages.

d) Using the model proposed in section *(c)*, explain in detail how you would test the null hypothesis that marihuana consumption does not affect wages. The answer should be specific and it should include a detailed list of the degrees of freedom.

e) What are the potential problems to be faced when carrying out casual inference with the available survey data?

5. Assume that we are interested in analyzing the potential differences of beer consumption across gender. For that we specify the following linear regression model

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i X_i) + u_i$$

where Y_i is individual $i's$ expenditure on beer, X_i is his/her income and D_i is a constructed variable that takes the value of 1 if the individual is a woman and 0 if the individual is a man. Using a sample of size $n = 34$, we have obtained the following results:

$$\begin{array}{rcll} \widehat{Y_i} & = & 186.47 - 126.00 D_i + 2.33 X_i - 1.29 (D_i X_i) & R^2 = 0.5055 \\ & & \quad (45.67) \quad (57.01) \qquad (0.86) \quad (1.02) & \end{array}$$

The numbers in brackets are the standard errors. Moreover, using the same sample, we have estimated the model $Y_i = \alpha_0 + \alpha_1 X_i + u_i$, obtaining a coefficient of determination of 0.1355

a) What will be the difference in consumption between a woman and a man with the same income?

b) Test at the 5 % level, the following statements:

1) There are no differences in beer consumption across gender
2) There are no differences in the marginal propensity to consume beer across gender

6. There are situations where we can observe data before and after an exogenous policy change. In the case where policy only affects a subgroup of the population we can interpret the outcomes as a natural experiment which allows us to analyze the policy's effect by considering agents'behavior. Put differently, even though data is not experimental we can consider it as such if the policy only affects a subgroup (treatment group) but not the rest of the population (control group). The idea behind a natural experiment is that the assignment of individuals to the treatment or control group is exogenous and does not depend on their actions. Put differently, by accident or exogenously some individuals end up in the control group and

others in the treatment group.

In the most simple case there are two periods and two groups. One control group A that includes those not affected by the policy change and one **treatment** (or experimental) group B, including those individuals affected by the change
Let $\overline{Y}_{A1}$ and $\overline{Y}_{A2}$ be the mean of Y for the control group in the periods 1 and 2 respectively, and $\overline{Y}_{B1}$ and $\overline{Y}_{B2}$ the means for the treatment group.

If data could be generated from a true experiment we could measure the treatment effect ignoring the first period (before the policy change) and just compare individuals from treatment and control group after the policy change. Put differently, we would evaluate the effect of the policy change by calculating the difference in means of treatment and control group after the policy change,

$$\left(\overline{Y}_{B2} - \overline{Y}_{A2}\right)$$

In practice, data is not truly experimental and thus the problem of this estimator is that part of the difference in means of control and treatment group after the policy change could be due to systematic and unobservable differences between both groups that are unrelated to the change in policy.
An interesting measure could be the effect of the change in policy on the treatment group.

$$\left(\overline{Y}_{B2} - \overline{Y}_{B1}\right).$$

The problem of this measure is that the mean of the treatment group could change along time (from period 1 to 2) for reasons other than the exogenous policy change.
An appropriate measure to capture the treatment effect (i.e. the change in economic policy) is to compare the changes in both groups treatment and control respectively. In this sense we can control for the ex ante differences between both groups as well as within group changes that are independent of the policy change:

$$\left(\overline{Y}_{B2} - \overline{Y}_{B1}\right) - \left(\overline{Y}_{A2} - \overline{Y}_{A1}\right)$$

Consider the binary variables *dB*, that takes the value of 1 if the individual belongs to the treatment group and 0 otherwise and variable *d2*, that takes the value of 1 for the second period (after the policy change) and 0 for the first period (before the policy change). The most simple equation to analyze the impact of the policy change can be written as follows

$$Y = \beta_0 + \delta_0\ d2 + \beta_1\ dB + \delta_1\ \ (d2 \times dB) + u, \tag{3}$$

where Y is the variable of interest, and:
d2 is a binary variables that captures aggregate factors that affect Y over time and in the same fashion for both treatment and control group;
dB captures possible ex ante differences between control and treatment group (before the policy change).

Note that if we did not take into account the differences between treatment and control groups before the policy change we would possibly be incorrectly attributing these differences

to treatment effects.

The coefficient of interest, δ_1, is associated to the interaction between both binary variables, *d2* and *dB* (their product being equal to a new binary variable that takes the value of 1 for individuals in the treatment group in the second period)
Let $\overline{Y}_{A1}$ and $\overline{Y}_{A2}$ be the mean of Y for the control group in periods 1 and 2 respectively, and $\overline{Y}_{B1}$ and $\overline{Y}_{B2}$ the means for the treatment group. Thus, it is easy to show that the OLS estimator of δ_1, d_1, can be written as

$$d_1 = \left(\overline{Y}_{B2} - \overline{Y}_{B1}\right) - \left(\overline{Y}_{A2} - \overline{Y}_{A1}\right) \tag{4}$$

This estimator is called **difference-in-difference** estimator (DED).
Certainly, for the DED estimator to be able to consistently evaluate the effects of a policy change it is necessary that there is no systematic relation between this policy change and other unobserved factors (contained in u) that affect Y.

In July of 1980 the state of Kentucky (EE.UU.) increased the ceiling on subsidies for job-related accidents or illnesses. Those subsidies are equal to a percentage of individualsíncome with an upper limit (ceiling). Thus, the increase in the upper limit only affected high-income workers. This policy change reduced the opportunity cost of a sick leave for high-income workers. The policy change allows us to evaluate if a more generous public system of subsidies for job-related accidents or illnesses leads to longer sick leaves.
The file `KENTUCKY.GDT`includes data for the state of Kentucky of workers who have experienced some type of job-related accident or illness. The variable *d2* equals 1 for observations after the policy change on the ceiling of the subsidy and 0 otherwise, and *dB* is a binary variable that takes the value of 1 for high-income workers affected by the policy change and 0 otherwise.

a) Evaluate the effect of the policy change on the natural logarithm of the duration of sick leave (in days) *ldur* using the DED estimator proposed before. What is the percentage increase (or decrease) in the mean duration of sick leave after the policy change?

b) In most applications, the equation (3) includes observable factors affecting Y. Thus, this allows for the possibility that there are systematic differences in theses factors in each group and thus one can isolate in d_1 the pure effect of the policy change. (In this case, d_1 does not have such a simple representation as in (4), even though conceptually the idea remains the same).
Reestimate this effect controlling also for workers'gender (*sexo*), marital status (*casado*), as well as binary variables for the type of accident or illness (*cabeza*, *cuello*, *brazos*, *tronco*, *lumbares*, *piernas*, *enfocup* –this last one refers to pains due to the job itself) and the logarithm of age (*ledad*).
How do results change? Which estimation of the effect of the policy change do you think is better and why?

c) Given the value of R^2, can we deduce that the results are of little relevance?