Quantitative Methods II - Statistics
Assignment

Group ID number:	10

Group members	Philip Bunford (3127438); Antonio Puzalkov (3122854); Nicolò Cobianchi (3100444)

SECTION 1

Descriptive statistics

Provide a brief description of your sample in terms of gender, age, education, country and eventually other variables that you consider as relevant (approximately from 1000 to 3000 characters for comments).

We have started by renaming every question asked in the section G file of the European Social survey into 7 different variables, to use them in STATA:

rename frprtpl G1
rename gvintcz G2
rename poltran G3
rename ifredu G4
rename ifrjob G5
rename evfredu G6
rename evfrjob G7

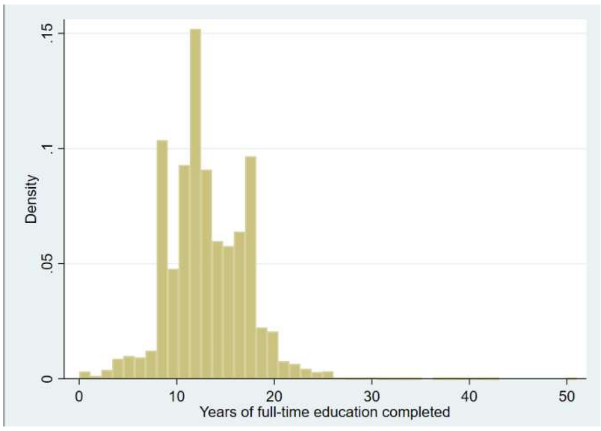
The variables concerning our research are delineated by: age(agea), gender(gndr), years of education(eduyrs), all the variables above and the country(country) of origin relating to the respondent (or unit).

Sample Size: 36.015

Gender(gndr): is a categorical nominal variable that can only take 2 values: male and female. It takes 1 for males and 2 for females. [1;2] no missing values reported.
1 is labelled as male. There were 16982 males in total that responded to the survey.
2 is labelled as female. There were 19033 females in total that responded to the survey.
We took the gender variable (gndr) and we turned it into the dummy variable (gndr_r), in which the two original numbers denoting whether the respondent was either male or female were turned into 0 and 1 [0,1]. 0 is now the female and 1 denotes the male respondent. This lets us determine the sample proportion for each gender.

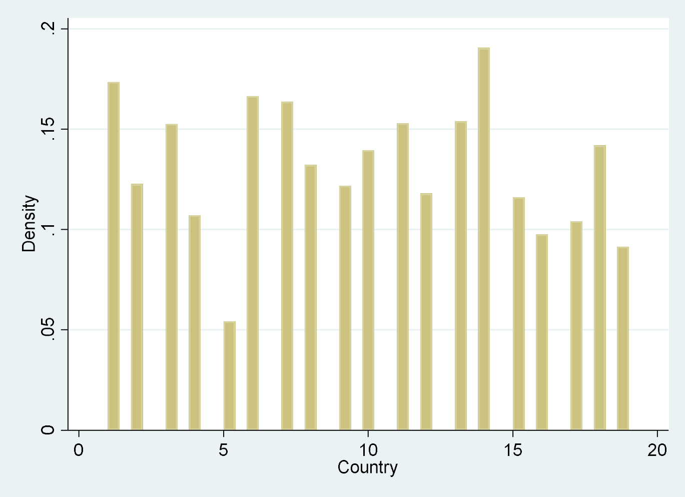
Years of Education(eudyrs): this is a numerical variable that takes values within the range [0;51] with 505 missing values.
This variable represents the years a given unit has spent into education.
It has the following distribution which can be defined to be slightly right skewed.

***hist** eduyrs*



Country (country): it is a categorical nominal variable:

For the purpose of regression, which we will later show we generated a new variable called (ducountry) that took values 0 and 1 in which 1 represents the value "Hungary" and 0 represents the other 18 EU countries. We created ducountry to highlight when a respondent (unit) answered either from Hungary or the other European countries, in order to calculate the proportion of units answering from Hungary in the sample.
These are the countries present in the original dataset in alphabetical order:
AT,BE,BG,CH,CY,CZ,DE,EE,FI,FR,GB,**HU**,IE,IT,NL,NO,PL,RS,SI.
The units are distributed as follows:

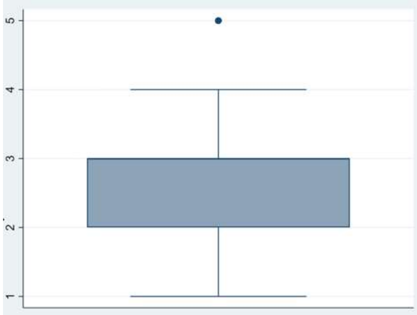


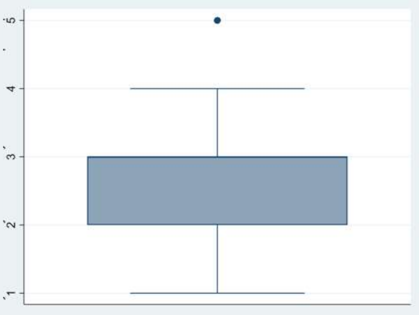
HU: Hungary is the value of interest for our regression.

G1: is a categorical ordinal variable that measures fairness of participation in politics within the respondent's country. The levels of fairness range from "Not at all" to "A great deal" on a likert scale ranging from [1;5].
The mean being: 2.749626

G2: is a categorical ordinal variable that measures the Government's interest regarding all citizens, within the respondent's country. The levels of interest range from "Not at all" to "A great deal" on a likert scale ranging from [1;5].
The mean being: 2.571665

G3: is a categorical ordinal variable that measures the Government's transparency withing a country's political framework, within the respondent's country. The levels of interest range from "Not at all" to "A great deal" on a likert scale ranging from [1;5].
The mean being: 2.433519





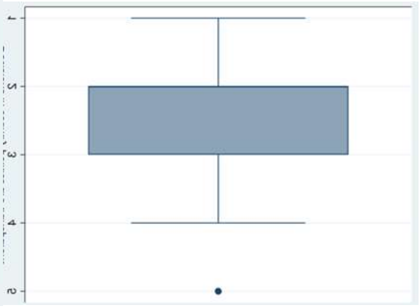
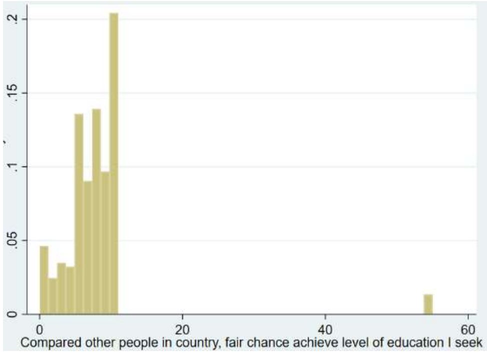


Figure 1: Box Plot for G1 *Figure 2: Box plot for G3* *Figure 3: Box plot for G3*

(from the figures it is possible to visualize the Interquartile Range, the Median, the adjacents and eventually the outliers)

G4: is a categorical ordinal variable that measures the fairness in education seeking, within the respondent's country. The levels of interest range from "Does not apply at all" to "Applies completely" on a likert scale ranging from [0;11]. There is also a "I have not completed a level of education" as an additional value external to the scale.
The mean is: 7.730227



G5: is a categorical ordinal variable that measures the level of fairness to achieve the goal education when compared to the other people in the country. The levels of interest range from "Does not apply at all" to "Applies completely" on a likert scale ranging from [0;10]. There is also a "I have not completed a level of education" as an additional value external to the scale.
The mean is: 6.079154

G6: is a categorical ordinal variable, which can also be treater as a numerical variable since it gets many values ranging from [0;10].
The variable seeks to get the level of the fair chance of achieving the sought level of education in all the country.
The mean is: 6.217077

G7: is a categorical ordinal variable, which can also be treated as a numerical variable since it gets many values ranging from [0;10]. The variable seeks to get the level of Overall fairness to get the job the unit is seeking.
The mean is: 5.009731

Running the previous command in Stata SE we can visualize the following box plots:

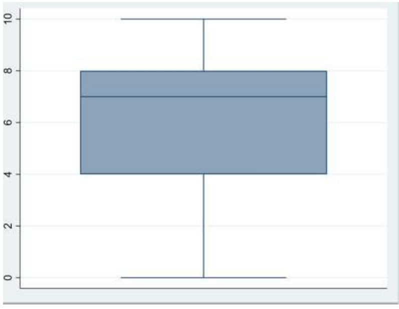


Figure 4: Box plot G5

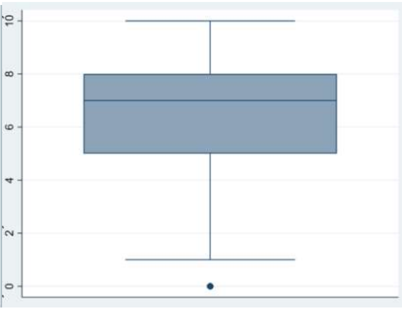


Figure 5: Box plot G6

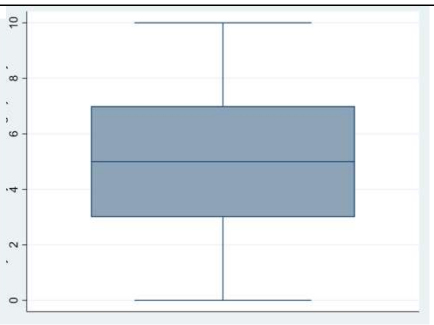


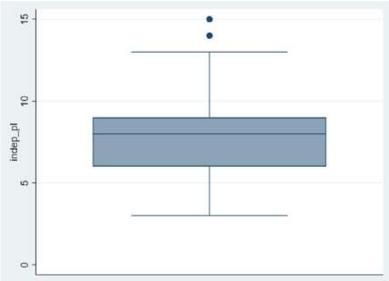
Figure 6: Box plot G7

Inference on the mean

Provide a brief explanation about the way you obtained the two additive indexes, provide some descriptive statistics about them and run inference on them, for example: workout confidence intervals, run a T test for hypothesis about relevant or interesting values, etc. (approximately from 1000 to 3000 characters for comments).

We generated two indexes, one being the dependent variable, first in the association analysis and then the regression analysis, and then independent variable.
Both **additive** indexes ought to be numerical:
The first generated index is the independent variable, called (indep_pl), and it is the sum of variables; G1, G2 and G3.
(Indep_pl) is the perceived level of functioning of the national political system:
given this index we can consider the fact that, the higher it gets, the more the poilitical system of the country is well-perceived as far as functioning is concerned, since is it made out of three variables which idenitifies 3 main characteristics concerning effectiveness (functioning).
The range of the variable is of [3;15]. We did consider the option of recoding it , but we decided that it would have been too manipulative with respect to the original dataset, and we thought it would be unnecessary when it comes to the regression, since the variables can have different intervals.
The variable has the following synthetic measurements:

- Its mean is calculated to be of 7.73 with respect to the sample
- Its St. Dev is 2.47, meaning that values spread around with respect to the mean on an average of 2.47
- The percentiles are the following:



10%	25%	50%	75%	90%
4	6	8	9	11

sum *indep_pl, d* & ***codebook*** *indep_pl*

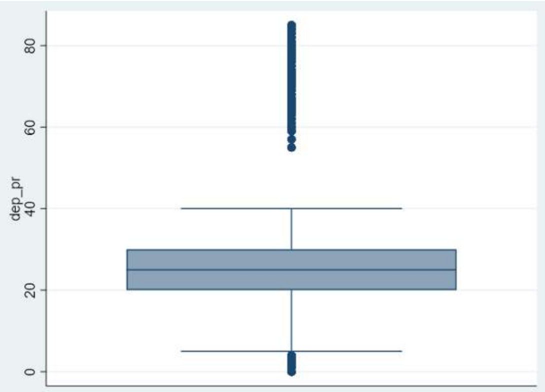
Graph 2-1: ***Graph*** *box indep_pl*

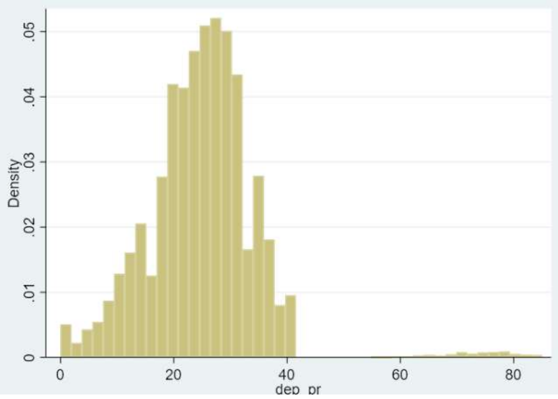
We do find outliers in our “Five Number Summary”, falling dinstantly from the whisker’s upper adjacent.

(dep_pr) is a numerical variable, derived as an additive index and measures the respondent’s perceived level of opportunities offered by the own country.
It is computed out of the variables (G4,G5,G6,G7), which are all summed, generating a new variable with a range [0;85].
This variable has he following synthetic measurement:

- Its Mean on the sample is 24.99
- Its St. Dev on the sample is 9.89
- Its percentiles are: ***sum*** dep_pr*, d* & ***codebook*** dep_pr

10%	25%	50%	75%	90%
13	20	25	30	35





3

Given the high amount of values the variable has we decided to include the the graphical rappresentation of the distribution using the ***hist*** command.

We can see that at the level of the sample the distribution is right-skewed, in fact the variable has the so called “tail” following on the right.

Inference:

We have computed the sample mean of both our variables, and we could predict that, increasing the sample size by 100 of both variables, drawn with respect to the same mean, this increase could hypothetically make it the real population mean, and we would end up with 90, 95 or 99 of these sample means falling within a specific interval, the so called confidence interval.

Yet we are not talking about a probabilistic framework, we are observing the values and we will observe them with a certain “confidence level” in a “confidence interval”

For the first variable (indep_pl) the estimated mean with respect to the sample, and the one we would infere onto the population is 7.73.

We want to assess, with a specific confidence, thus with a given level of sureness, non-statistically speaking, the specific interval in which we will see the population mean, given our sample and given ours as a first hypothesis:

mean *indep_pl, level(##)*

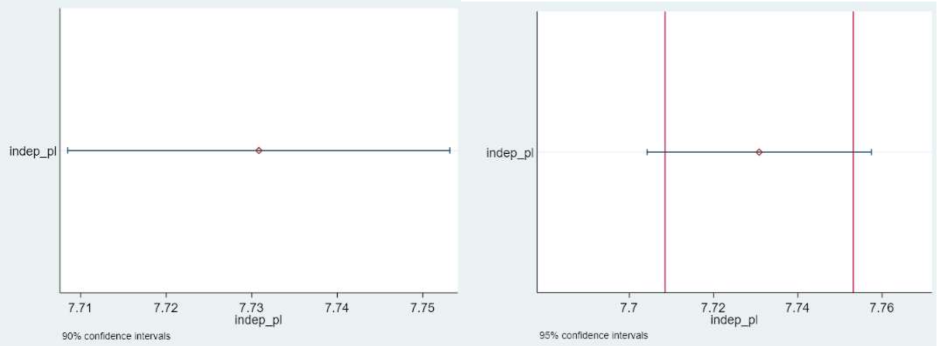
- If we are 99% confident, the interval of confidence is: [7.695863;7.765806] (6 significant figures following the decimal point), while the mean is always the same for the sample.
- If we want to be 95% confident, the interval of confidence is: [7.704225;7.757444] (6 significant figures following the decimal point).
- If we want to be 90% confident, the interval of confidence is: [7.708503;7.753165] (6 significant figures following the decimal point).

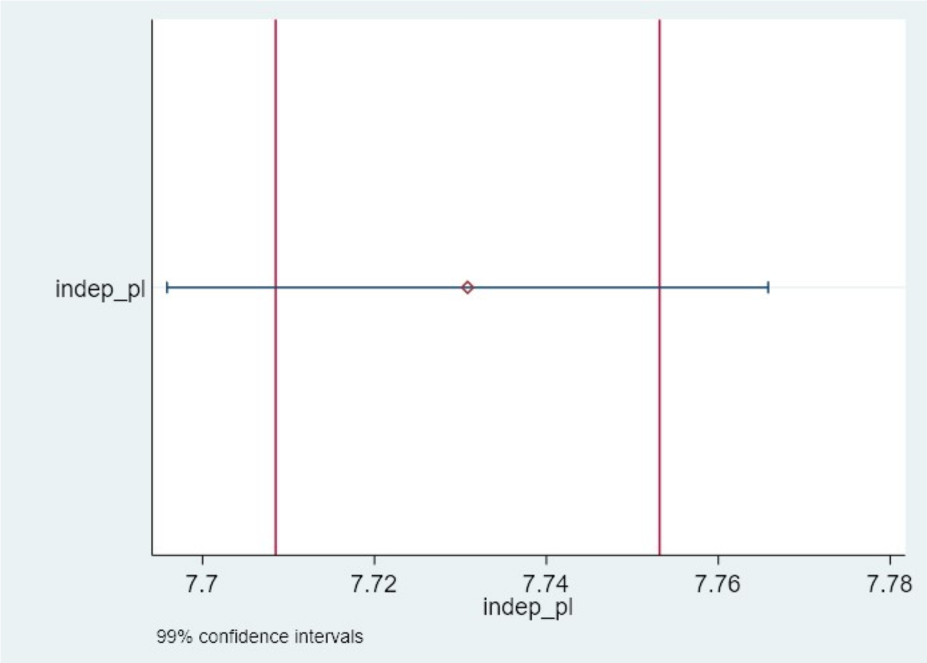
We do see a trend in these numbers: the lower the level of confidence the shorter the interval.

We can visualize it with the additional package for Stata SE called ciplot.

ssc install *ciplot*

ciplot *indep_pl, level(##) horizontal* → we then use the Stata graph editor to add reference line corresponding to the 90% confidence interval. (in the “objects” form on the right, we clicked on the x-axis option and then we added reference lines)





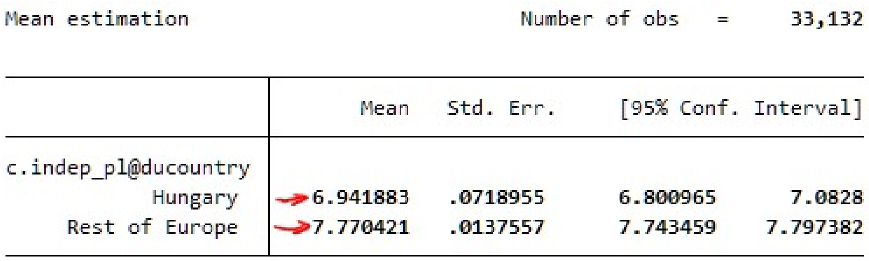
As we can see graphically the confidence intervals change and they change greatly: the red lines are the lines corresponding to the values of the extremes of the confidence interval of the 90% confidence level.
The same procedure is then applied to the (dep_pr) variable and we obtain the following intervals:

- For 99% confidence and mean 24.99199 the confidence interval is [24.85244 ; 25.13154]
- For 95% confidence and mean 24.99199 the confidence interval is [24.88581 ; 25.09817]
- For 90% confidence and mean 24.99199 the confidence interval is [24.90288 ; 25.0811]

Since we have established the confidence levels and intervals we can state that the population mean will be within the extremes of a certain interval, considering a given level of confidence (i.e. if we are 90% confident the population mean will be within the [24.90288 ; 25.0811] interval .

Now the goal is to compare the estimated mean of our country's population with respect to the parameter estimated from the entire sample for both variables.
Since we have already declared the existance of our dummy variable representing the proportion of the sample coming from Hungary we will use it to compare the means.
We run the following commands:
For the variable (indep_pl):

***mean** indep_pl, **over**(ducountry)*



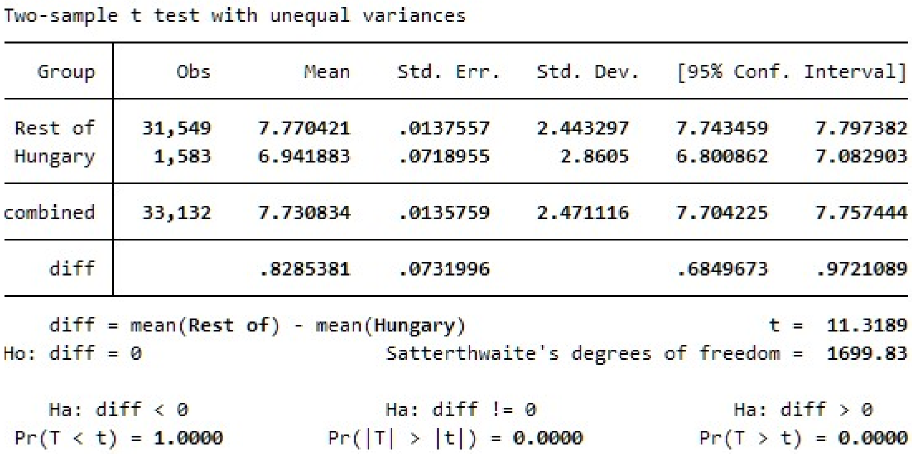
```
Mean estimation                          Number of obs   =     33,132

--------------------------------------------------------------
                  |       Mean   Std. Err.     [95% Conf. Interval]
------------------+-------------------------------------------
c.indep_pl@ducountry |
          Hungary |   6.941883   .0718955      6.800965      7.0828
   Rest of Europe |   7.770421   .0137557      7.743459    7.797382
--------------------------------------------------------------
```

Let’s comment on the results:

Since the indep_pl variable is an index measuring the perceived level of functioning of the national political system, we can confidently state that, at the level of the sample, there is a difference in the average percieved level of functioning of the political system in Hungary when compared to the rest of Europe and, given as premise this sample, the confidence intervals, if 95% confident, will not overlap on any interval or range of values.

Now we shall verify if there is a significant mean difference:

***ttest** indep_pl, by(ducountry) unequal*



```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 Rest of |  31,549    7.770421    .0137557    2.443297    7.743459    7.797382
 Hungary |   1,583    6.941883    .0718955      2.8605    6.800862    7.082903
---------+--------------------------------------------------------------------
combined |  33,132    7.730834    .0135759    2.471116    7.704225    7.757444
---------+--------------------------------------------------------------------
    diff |            .8285381    .0731996                .6849673    .9721089
------------------------------------------------------------------------------
    diff = mean(Rest of) - mean(Hungary)                          t =  11.3189
Ho: diff = 0                     Satterthwaite's degrees of freedom =  1699.83

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

The results are clear: at any level of significance, our P-value, or probability to encounter an even more distant result from ‘0’ is impossibile, thus the null-hypothesis, for which there is no difference in the means, can be rejected, hence we accept the alternative hypothesis, which states that the difference is not equal to ‘0’.

There is a significant association between the indep_pl and the aspect of living in Hungary.

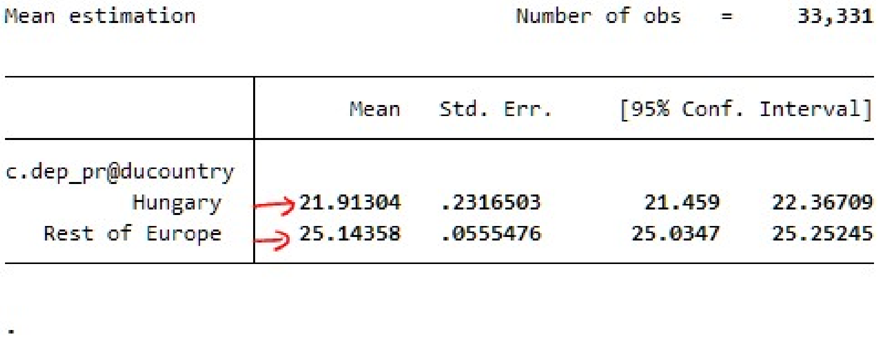
We can say that Hungarian people have a significant lower average of confidence and trust in the functioning of their political system when compared to the rest of Europe.

The next variable is the dep_pr, which, as a reminder for the following analysis represents the respondent’s perceived level of opportunities offered by the own country.

Our goal is to assess significantly if there is a difference in the opportunities a citizen finds in Hungary when compared to the rest of Europe, and determine, thanks to the Mean difference, by how much do they differ anyways.

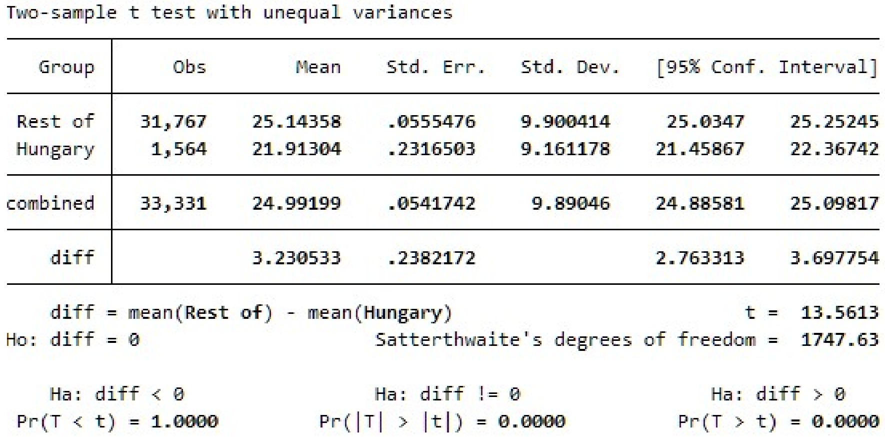
We run the same command we used before for the previous variable:

***mean** dep_pr, **over**(ducountry)*



```
Mean estimation                        Number of obs   =     33,331

--------------------------------------------------------------
                     |       Mean   Std. Err.     [95% Conf. Interval]
---------------------+----------------------------------------
c.dep_pr@ducountry   |
             Hungary |   21.91304    .2316503        21.459    22.36709
      Rest of Europe |   25.14358    .0555476       25.0347    25.25245
--------------------------------------------------------------

.
```

There is difference when the mean is compared through out the groups, and the confidence interval, when assessing for a 95% confidence level are not overlapping, yet we have to verify if there is a significant difference and whether there might be a real association between living in "Hungary" (thus being a unit from Hungary) and stating that there is a relative perceived level of opportunity.
"Do units from Hungary feel significantly less positively towards the level of opportunity in their country when compared to the rest of Europe?"



```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 Rest of |  31,767    25.14358    .0555476    9.900414     25.0347    25.25245
 Hungary |   1,564    21.91304    .2316503    9.161178    21.45867    22.36742
---------+--------------------------------------------------------------------
combined |  33,331    24.99199    .0541742     9.89046    24.88581    25.09817
---------+--------------------------------------------------------------------
    diff |            3.230533    .2382172                2.763313    3.697754
------------------------------------------------------------------------------
    diff = mean(Rest of) - mean(Hungary)                          t =  13.5613
Ho: diff = 0                     Satterthwaite's degrees of freedom =  1747.63

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

The answer is in fact affirmative, there is a significant, at any level of significance, association between living in Hungary and having a lower perception of opportunity when compared to the rest of Europe. The interpreation in fact is that given the null-hypothesis being an absence of association, thus a situation in which the difference is equal to zero, we test for the opposite and the result is significant since the P-value, value which we already defined, is the lowest the possible.

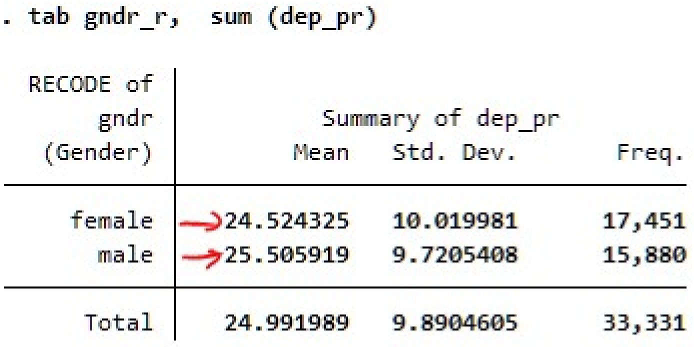
Inference on association

Analyze the association between the dependent variable and each of the other variables (that you are going to use in the regression model as independent variables), using the appropriate technique based on the type of the variables and pointing out whether the association is significant or not (approximately from 2000 to 4000 characters for comments). If you want and if you have time, you can run further analysis (e.g.: correlation between the items of each additive index, association between age, gender, country, education, etc.).

We then proceeded by running an analysis on association both at the level of the sample and of the population, the association between the dependent variable (**dep_pr**) and the variables that will be used in the multiple regression model (**gndr_r**, **agea, eduyrs**, **ducountry**), and finally between the two indexes (alone and controlling for variables **gndr_r** and **ducountry**).

Association between the dependent index at the level of the sample (made up by G4, G5, G6, G7) and the dummy variable accounting for the gender **(gndr_r**); since the dependent variable is numerical, and the independent variable is categorical nominal the command we use is:

*tab **gndr_r**, sum (**dep_pr**)*

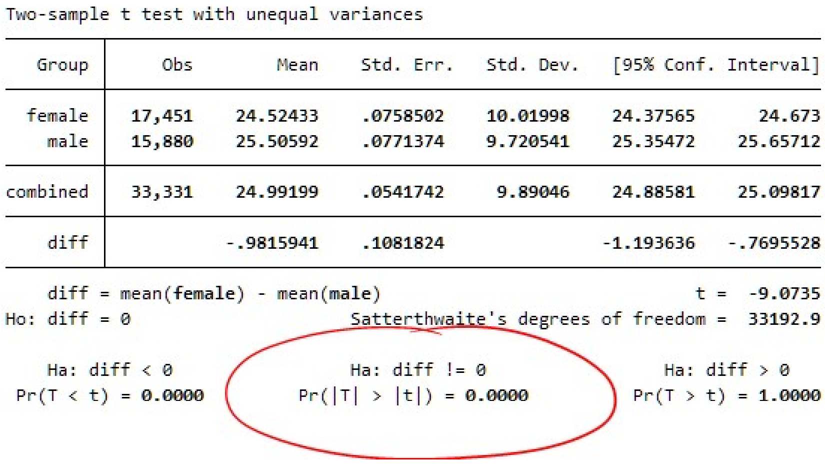


```
. tab gndr_r,  sum (dep_pr)

  RECODE of  |
       gndr  |       Summary of dep_pr
   (Gender)  |        Mean   Std. Dev.       Freq.
-------------+------------------------------------
     female  |   24.524325   10.019981      17,451
       male  |   25.505919   9.7205408      15,880
-------------+------------------------------------
      Total  |   24.991989   9.8904605      33,331
```

Since the two means at the level of the sample, the female one and the male one are different , our mean comparison leads us to infer that at the level of the sample there is association between the variables **gndr_r** and **dep_pr.**

The possible association between **gndr_r** and **dep_pr** is now analyzed at the level of the population. Knowing that we are dealing with a categorical independent variable and a numerical dependent, we run a test on the mean difference, in order to see whether it can be concluded that the association between the two variables is significant. We run a hypothesis testing test assuming as null hypothesis that the difference between the population means of the dependent variable in the categories specified by the independent one is equal to 0. The command we use is:

*ttest **dep_pr**, by (**gndr_r**) unequal*

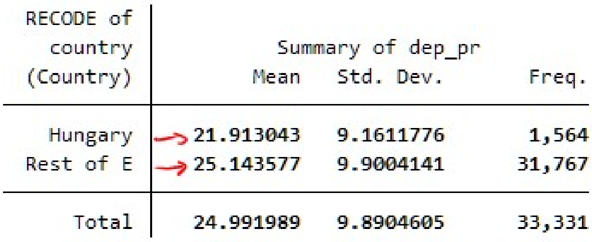


```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  female |  17,451    24.52433    .0758502    10.01998    24.37565      24.673
    male |  15,880    25.50592    .0771374    9.720541    25.35472    25.65712
---------+--------------------------------------------------------------------
combined |  33,331    24.99199    .0541742     9.89046    24.88581    25.09817
---------+--------------------------------------------------------------------
    diff |           -.9815941    .1081824               -1.193636   -.7695528
------------------------------------------------------------------------------
    diff = mean(female) - mean(male)                              t =  -9.0735
Ho: diff = 0                     Satterthwaite's degrees of freedom =  33192.9

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

We see that the p-value for the two-sided test is smaller than every significance level, thus we can say that the difference between the means of **dep_pr** in the two groups specified by **gndr_r** is significant, there is enough evidence to conclude that there is significant association between the two variables.

- The association is now investigated for **ducountry** independent variable, the procedure is exactly the same as before, since we have a categorical independent variable and a numerical dependent (**dep_pr**) **;**
  below the output of the analysis at the level of the sample and of the population:

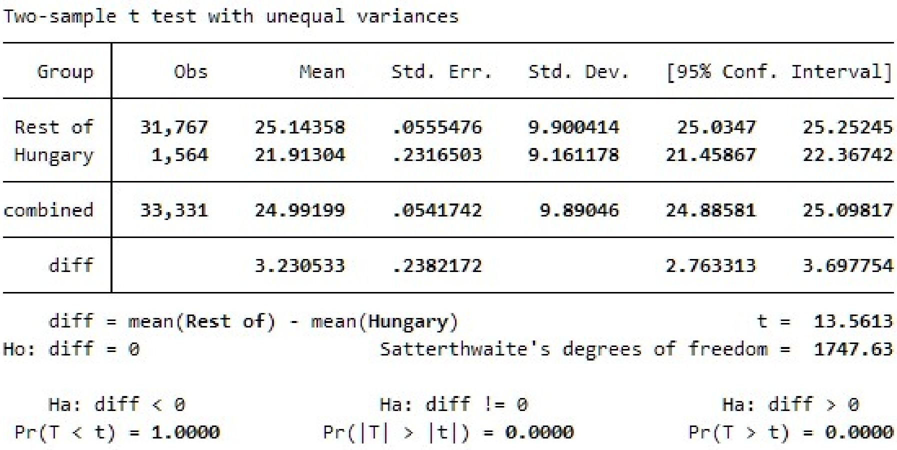
tab ***ducountry*** *, sum (****dep_pr****)*



```
  RECODE of  |
    country  |         Summary of dep_pr
  (Country)  |        Mean   Std. Dev.       Freq.
-------------+------------------------------------
    Hungary  |   21.913043   9.1611776       1,564
  Rest of E  |   25.143577   9.9004141      31,767
-------------+------------------------------------
      Total  |   24.991989   9.8904605      33,331
```

Since the two means at the level of the sample, the one for Hungary and the one for "Rest of Europe", are different, our mean comparison leads us to say that at the level of the sample there is association between the variables **ducountry** and **dep_pr.**

ttest ***dep_pr****, by (****ducountry****) unequal*



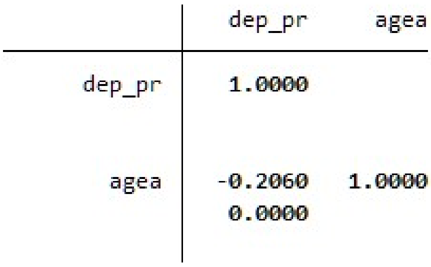
```
Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 Rest of |  31,767    25.14358    .0555476    9.900414     25.0347    25.25245
 Hungary |   1,564    21.91304    .2316503    9.161178    21.45867    22.36742
---------+--------------------------------------------------------------------
combined |  33,331    24.99199    .0541742     9.89046    24.88581    25.09817
---------+--------------------------------------------------------------------
    diff |            3.230533    .2382172                2.763313    3.697754
------------------------------------------------------------------------------
    diff = mean(Rest of) - mean(Hungary)                          t =  13.5613
Ho: diff = 0                     Satterthwaite's degrees of freedom =  1747.63

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

As a reminder of what we previously said:
Even for this analysis the p-value of the two-sided test is smaller than every significance level, thus we can say that the difference between the means of ***dep_pr*** *in the two groups specified by* ***ducountry*** *is significant, there is enough evidence to conclude that there is significant association between the two variables.*

Now the association is investigated between the variables **agea** and **dep_pr**, we are dealing with two numerical variables, therefore we will use the method of correlation between two interval variables, and by means of the Pearson coefficient we will see if the two variables are correlated at the level of both sample and population, we can run a single analysis to assess it thanks to the command " *pwcorr*

***dep_pr agea**, sig*" , in fact it let us discover the Pearson coefficient (R) at the level of the sample, and it will be done a hypothesis testing test on the estimator **ρ** (population Pearson coeff.) , assuming as null hypothesis that there is no correlation at the level of the population, thereby : correlation coefficient equal to 0.

*pwcorr **dep_pr agea**, sig*

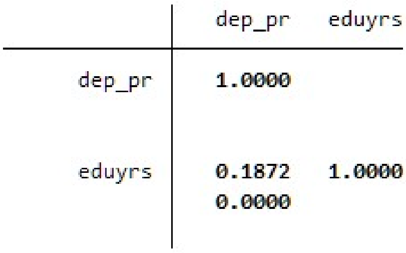


```
             |   dep_pr     agea
-------------+------------------
      dep_pr |   1.0000
             |
             |
        agea |  -0.2060   1.0000
             |   0.0000
             |
```

The correlation coefficient is –0.2060, meaning that there is a moderate-low negative correlation between the two variables (at the level of the sample), the p-value is smaller than any level of significance, we can reject the null hypothesis that **ρ=0,** claiming, then, that we have enough empirical evidence to say that the parameter (population Pearson index) is significant and that there is significant correlation at the level of the population between **agea** and **dep_pr**.

- Now we run analysis on the association between **eduyrs** and **dep_pr**, we are in the same situation as before (two interval variables), so it will be run a correlation test to see the value of the Pearson coefficient and if we can reject the null hypothesis that at the level of the population it is equal to 0.

*pwcorr **dep_pr eduyrs**, sig*



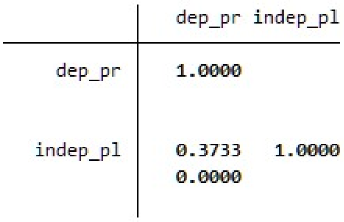
```
             |   dep_pr   eduyrs
-------------+------------------
      dep_pr |   1.0000
             |
             |
      eduyrs |   0.1872   1.0000
             |   0.0000
             |
```

The correlation coefficient is now 0.1872, we have so a moderate-low positive correlation (at the level of the sample), again the p-value is smaller than every significance level, thus we have enough evidence to claim that the parameter is significant and that there is significant correlation between the variables **eduyrs** and **dep_pr.**

Finally, we run analysis on the association between the two additive indexes **dep_pr**and **indep_pl**, both variables are numerical, so the procedure is analogous as the last two analyses: correlation test and hypothesis testing test assuming as null hypothesis that the population correlation coefficient is equal to

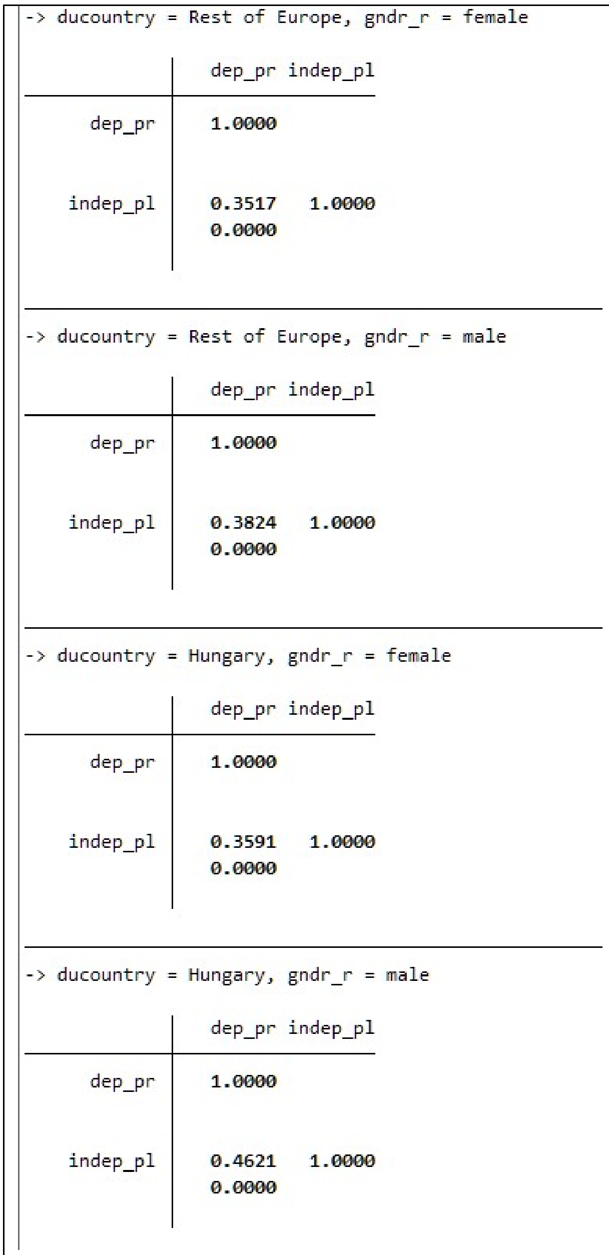
0. We wanted even to analyze the correlation between the two variables controlling for **ducountry** and **gndr_r**, seeing if the association become stronger/weaker or disappears with those two variables. Below all the commands and outputs:

pwcorr ***dep_pr indep_pl,*** *sig*



```
             |   dep_pr indep_pl
-------------+------------------
      dep_pr |   1.0000
             |
             |
    indep_pl |   0.3733   1.0000
             |   0.0000
             |
```

bysort ***ducountry gndr_r****: pwcorr* ***dep_pr indep_pl****, sig*



```
-> ducountry = Rest of Europe, gndr_r = female

             |   dep_pr indep_pl
-------------+------------------
      dep_pr |   1.0000
             |
             |
    indep_pl |   0.3517   1.0000
             |   0.0000
             |

------------------------------------------------------
-> ducountry = Rest of Europe, gndr_r = male

             |   dep_pr indep_pl
-------------+------------------
      dep_pr |   1.0000
             |
             |
    indep_pl |   0.3824   1.0000
             |   0.0000
             |

------------------------------------------------------
-> ducountry = Hungary, gndr_r = female

             |   dep_pr indep_pl
-------------+------------------
      dep_pr |   1.0000
             |
             |
    indep_pl |   0.3591   1.0000
             |   0.0000
             |

------------------------------------------------------
-> ducountry = Hungary, gndr_r = male

             |   dep_pr indep_pl
-------------+------------------
      dep_pr |   1.0000
             |
             |
    indep_pl |   0.4621   1.0000
             |   0.0000
             |
```

What emerges from this analysis is that there is a moderate positive correlation at the level of the sample between the two additive indexes (R = 0.3733), since the p-value is smaller than every significance level we can reject the null hypothesis that **ρ = 0**, the parameter is therefore significant and the correlation between **indep_pl** and **dep_pr** is significant too.

The other analysis tells us that there is a discrepancy for what concern the gender male in controlling for Hungary, in fact the correlation appears to be stronger when by-sorting for **ducountry** = Rest Of Europe and **gndr_r** = male than the same gender in the rest of Europe, the opposite gender in both rest of Europe and Hungary and the total European population, we do not observe other prominent variations in the correlation index when looking at the other outputs of our analysis. Indeed, the Pearson coefficient between **dep_pr** and **indep_pl** will be higher in Hungary (driven by males) than in the rest of Europe.

SECTION 2

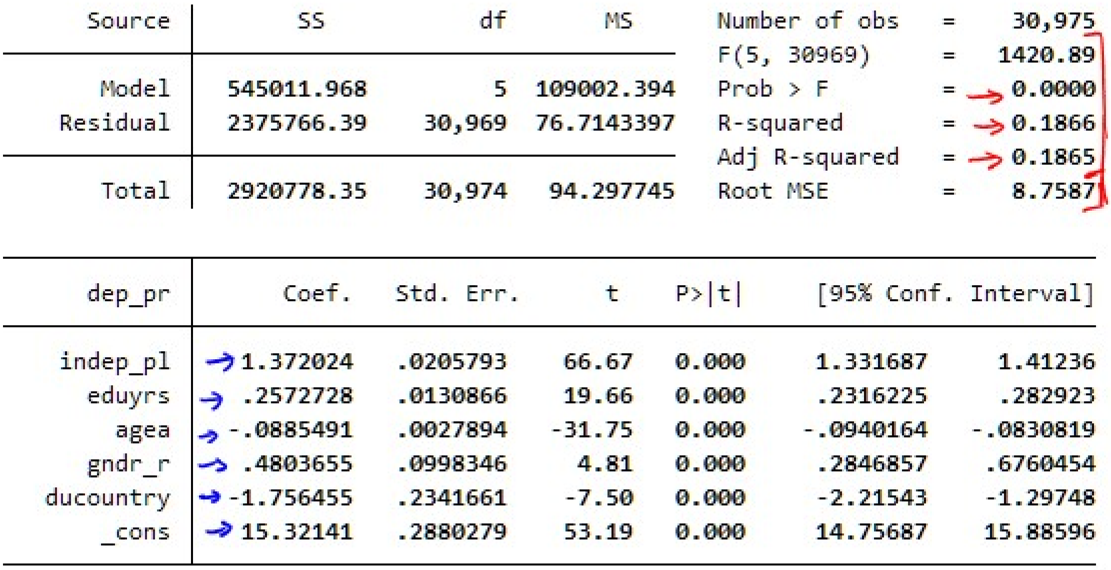
Regression model

Write a brief explanation of the model (e.g.: number and name of the independent variables used, way of coding categorical covariates in dummy variables, etc.) and report the results of your analysis, evaluating the goodness of fit and the significance of the model, interpreting the estimated coefficients and their significance, drawing for some conclusions (approximately from 4000 to 10000 characters for comments).

Our regression model is ought to be multiple, thus controlling the dependency of one variable and the other with several different, in order to evaluate their strength, influence and whether the independent variable we are checking if it has a spurious or not spurious with respect to the dependent variable.
The variable we are considering as our dependent for the analysis is the **dep_pr**, which on the scatterplot and on the plot in which the regression line is computed, represents the y-axis, since of course is dependent on the value of the estimated coefficients.
The so-called controlling variable, also called explanatory variables are the following:

- (**eduyrs**), labelled as educational years. As previously described, it is an interval variable, also called numerical, which takes values going starting from 0 and going to 51. Range [0;51].
  Considered the fact that it is a numerical variable, the consideration we had of it was as a relative index, the higher it is the more the person has studied, thus approximatively has more academic qualifications.
  We did not perform a recoding since the interval, or the numbers could be used as they are.
  We could have made the values collapse into categories of years of studying considering specific ranges, but there is no real application in our specific case, thus we used the variable as already present in the database.
- (**agea**), labelled as age of the unit responding, it takes values going from 15 to 90. Range [15;90]
  We did consider a recodification into a categorical ordinal variable, but we decided not to, since the concept of "age" expressed with numerical values is more precise and easier to handle in the regression modelling.
- **(gndr_r**), the recoding of the **(gndr**) variable, now containing the values 0 and 1 for "Female" and "Male" respectively, rather than the previously coded 1 and 2 for "Male" and "Female".
  The variable is in range [0;1] and we recoded it from its previous form, in order to calculate the proportion of the sample of a given gender, in our case the proportion of "Males" and the sample minus the proportion gives the "female" units within the sample.
  The goal of this is to simplify the creation of the regression model.
  The part of the sample presenting a certain characteristic is the part which contributes positively or negatively to the regression.
  If the characteristic of being "male" is inserted in the regression, hence considering the **gndr_r** proportion of "male" then we can assess if this characteristic is influent, positively or negatively, to the dependent variable, in our case to the fact that the unit perceives a high or low level of opportunity in his country.
- (**ducountry**) this variable again is related to the country and considers the proportion of the sample which completed the answer from "Hungary" or from the "rest of Europe", it takes values 0 and 1, where 0 stands for "Rest of Europe" while 1 stands for "Hungary".
  The idea behind the recoding and generation of this dummy variable is the same as the one gender one, our goal is to assess the proportion of the sample which presents the characteristic of the "Country: Hungary".
  Furthermore, our objective is to verify if this characteristic contributes positively or negatively to the regression, and verify the strength of the association, which, we will later see, is quite strong.

The command used in order to run a multiple linear regression analysis is the following:
***regress** dep_pr indep_pl eduyrs agea gndr_r ducountry*



```
      Source |       SS           df       MS      Number of obs   =    30,975
-------------+----------------------------------   F(5, 30969)     =   1420.89
       Model |  545011.968         5  109002.394   Prob > F        =    0.0000
    Residual |  2375766.39    30,969  76.7143397   R-squared       =    0.1866
-------------+----------------------------------   Adj R-squared   =    0.1865
       Total |  2920778.35    30,974   94.297745   Root MSE        =    8.7587

------------------------------------------------------------------------------
      dep_pr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    indep_pl |   1.372024   .0205793    66.67   0.000     1.331687     1.41236
      eduyrs |   .2572728   .0130866    19.66   0.000     .2316225     .282923
        agea |  -.0885491   .0027894   -31.75   0.000    -.0940164   -.0830819
      gndr_r |   .4803655   .0998346     4.81   0.000     .2846857    .6760454
   ducountry |  -1.756455   .2341661    -7.50   0.000     -2.21543    -1.29748
       _cons |   15.32141   .2880279    53.19   0.000     14.75687    15.88596
------------------------------------------------------------------------------
```

Starting now to analyze the estimator of the coefficients (β1 , β2, ... β5) of the explanatory variables we have that:

- When a unit increase occurs in **indep_pl** , the average change of **dep_pr** is 1.372024, while keeping fixed and constant all the other explanatory variables or controlling for them, the test statistic of the first coefficient tells us that the its estimator falls 66.67 standard errors above zero, the p-value is smaller than every significance level, thus we can reject the null hypothesis which is : β1=0 , rejecting the null hypothesis let us to say that the coefficient is significant , thereby **indep_pl**, provides a significant contribution to the explanation of the dependent variable.
  This, statistically translated means that there is a significant correlation between the perceived level of functioning of the national political system and the perceived level of opportunities in the country, thus leading to the affirmation that the more a person (respondent or unit) has a positive consideration of the national political system than they will also consider the country to be offering a greater number of opportunities.
  controlling for the other explanatory variables:
- **indep_pl** is significant; in addition, we are 95% confident that the value of the parameter β1 is between 1.331687 and 1.41236.

- Moving now to B2 we have that the dependent variable faces an average change of 0.2572728 when the variable **eduyrs** increase by 1 unit, controlling for all the other ind. variables; the estimator falls 19.66 standard errors above 0, the p-value is smaller than every significance level, meaning , once again, that the null hypothesis, which is β2 = 0 , can be rejected, stating therefore that the coefficient is significant and that the variable **eduyrs**, gives a significant contribution to the explanation of **dep_pr** , keeping fixed and constant all the other variables; here we are 95% confident that our parameter β2 falls in the interval between 0.2316225 and 0.282923.
  This leads to the consideration that, despite having a low influence on the association, the more a person has studied, or better said, the more a person has had years of education, the more, significantly speaking, this aspect influences the perceived level of opportunities within a country.
- For what concern B3 we see that , when the explanatory variable **agea** increase by 1 unit we see an average change of the dependent variable that is negative this time (-0.0885491) , leading us to say that an increase in **agea** will take **dep_pr** to a slow decrease; the test statistic this time is -31.75 and again we can reject the null hypothesis at every level of significance , by looking at the p-value, **agea** is thus significant (like the two variables before), here we are 95% confident that our parameter value will fall in the interval [-.0940164 -.0830819].
  This means that the elder a unit is the less they feel positive about the opportunities in the given country.
- The estimator of the coefficient β4 , is the one of the dummy variable **gndr_r** , and its value (0.4803655) tell us the estimate of the average difference in the level of **dep_pr** , in the two groups of male and female (we have to remember than when acting with a dummy binary variable βx is the difference between the two models generated by the dummy variable: the one of those who show a characteristic – the one of those who do not show it). Once again, we reject the null hypothesis β4 =0, and we state the significance of our variable.
  This leads us to state that the model that presents the characteristic (thus includes the value of the presence of the gender characteristic), hence the fact that the unit is "male", the positive direction between being a "male" and feeling positively with respect to the opportunities in the country is significant.
- With respect to the last explanatory variable of our multiple linear regression model, we have the two groups of the dummy variable **ducountry** : Hungary and Rest of Europe, the estimator B5 tells us the estimate of the average difference in the level of **dep_pr** in the groups of those who show the feature "Hungary" and those who do not show the feature "Hungary". Once again, we reject the null hypothesis, and we conclude that also this variable is significant.

Of Course, we conclude that all the five explanatory variables contribute to the explanation of our dependent variable, either positively or negatively.

Our country specifically contributes negatively to the regression, since the estimated coefficient which links the dependent variable to the **ducountry**, dummy representing the characteristic of the unit being from "Hungary" is negative.

Now we have to assess the goodness of the model , and we start doing it by seeing what proportion of the total variability of the model comes from the independent variables, this indication is provided by the R-squared coefficient , which is the ratio between the model sum of squares (variability accounted by the explanatory variables with their linear association) and the total sum of squares (variability explained by the independent variables + variability that the independent variables does not explain). R-squared coefficient is 0.1866, therefore 18.66% of the total variability is explained by the independent variables. Another important coefficient is the adjusted R-squared which is essential when comparing different models, since it is a coefficient that does not depend on the size of the sample or on the number of explanatory variables.

We move now to analyze the F-test, a hypothesis testing procedure which tells us if the model is significant, which means that at least one explanatory variable explain the dependent variable **dep_pr**, the F test verifies the null hypothesis that each coefficient (β1 +β2...) is equal to 0, against the alternative hypothesis, which claims that at least one coefficient is!=0; as always, the decision whether rejecting or not the null hypothesis is taken by looking at the p-value. In our model the F-test is 1420.89, the p-value represents the probability of observing a value of the F-test higher than the one observed, since the p-value is 0.000, it is smaller than every level of significance and we reject the null hypothesis, we have enough evidence to conclude that at least one explanatory variable is significant, and that all the variables together contribute significantly to the explanation of the **dep_pr**, we can state , finally, that our model is significant.