**Econometrics – Group Project**

Tutor – Jeyhun Mammadov ph.d

Students – Jeyran Heydarova, Shahana Rasulova, Jahangir Huseynov and Mahammad Aliyev

Project Topic – Regression analysis to estimate what effects to predictive score for each country in taken sample of countries.

**Introduction**

Today there are a lot of estimators of countries, it might be Economical aspects (GDP, FDI), it can be educational estimators, military-force estimators, safety and crime level, bribery level, freedom in terms of political environment, ideological and religious points of views and level of tolerance and racism, level of generosity and spiritual-development, educational-level and so on. But we with our team-mates decided what if I don’t want to analyze each aspect separately and instead to analyze a “golden-middle” score to know better which country is more suitable in terms of above listed perspectives. We obtained different numbers for each category from statistical sources including “golden-middle” which is recently highly spread among immigrated people. We obtained this score from internet by making research. Basically, we will analyze which of above-mentioned categories effects this “golden-middle score” and describes it more. We took a sample of countries as observations and several of above-mentioned categories will be our variables.

**Data Description and Methodology**

A screenshot of a computer

Description automatically generated

As you see in this excell sheet, these are top 13 countries in overall analysis, we created an excel-sheet for our observations before making analysis in STATA and run Regression analysis. But we won’t analyze exactly the countries You see in this picture. They are too many. But we will choose random sample about 50+ countries out of 117.

Methodology of collection the data

Data description

**Our independent variables (Xs)**:

* GDP,
* Social Support
* Healthy Life-expectancy
* Freedom to make life choices
* Generosity
* Perception of corruption.

**Our dependent variable (Y):**

* Golden-Score (from 0 to 10, max rounded value is 8 -Finland)

Graphical user interface, table

Description automatically generated

It’s descriptive statistics for each variable(column). I got this picture using Python’s library (Pandas) it helps to calculate and quickly demonstrate descriptive statistics ignoring your missing values (if they are), and dummy values. But We don’t have dummy-categorical variables in our data, we are lucky from this point… :)

Before running the regression line, we should consider several points called Assumptions.

1. Linearity

To absorb the relations between independent variables and dependent one and checkup whether they are linear, we made scatter plot for each X for Y(Golden-Score).

**Golden-Score & GDP Golden-Score & Social Support**



**Golden-Score & Healthy Life-Expectancy Golden-Score & Freedom t.L.c. **

**Golden-Score & Generosity Golden-Score & Perceptions of corruption**

****

**As we see some variables are not linearly related to Golden-Score, thus we need to make transformations. We will use natural logarithmic transformation in order to fix linearity and reduce skewness of distribution. Variables such as Health-Life excpetancy, Social support, and Perception of corruption will be transformed. Below you can see the relation of Log(x) to the Golden Score.**

****

****

**We used Log for Social support but that was still non-linear so we squared it in order to suit parabola. No it’s linear.**

****

**Let’s run Regression according to this transformations**

Graphical user interface, text, table

Description automatically generated

**As You see for significance test we can claim that GDP, Freedom to make life choice, Log(Social\_Supporr)^2 has statistically significant impact on GoldenScore, while Log(perceptionofcorruption), Log(Healthy Life Expectancy) are statistically non-significant. We compared alpha to the p-value.**

**Let’s look at correlation between each variable and check up VIF**

A screenshot of a computer

Description automatically generated

We see that some independent variables are corelated which is not good.

Graphical user interface, text

Description automatically generated

The one predictor is linear function of the set of another function it’s tells us about multicollinearity. It can lead for our model to be overfit. So we well drop out the variable which has vif higher that 3.5. In this case we will drop out GDP and run regression once again. And below you can see new regression result.

Graphical user interface, text, table

Description automatically generated

We can see changes in significance of variables after dropping out the variable that caused multicoliniarity.

**Looking at residuals and checking assumption about residuals**

In stata we created new column which will show predicted Y, values fitted on regression line that we created. Graphical user interface, application, table, Excel

Description automatically generated

But know let’s create new variables which will show difference between real Y and predicted Y which is also called residuals (error) , and student’s t distribution for residuals.

Graphical user interface, application, table, Excel

Description automatically generated



After dropping out outliers ( 5 observations) the scatter plot of t-errors and fitted values looks so:



At first we made residuals t-distribution because it helps us to figure out some things clearly. For example we know that in normal distribution the data without extreme outliers is located between +-2 standard deviations around the mean and if they are even further 3 std from mean then they are extreme outliers. And we made the scatter plot, we see the main 95% of our data without potential outliers is within 2 std around mean..

First method

The [null hypothesis](https://www.statisticshowto.com/probability-and-statistics/null-hypothesis/)  is that the [variances](https://www.statisticshowto.com/probability-and-statistics/variance/) for the errors are equal. In math terms, that’s:  
H0 = σ2i = σ2.  
The [alternate hypothesis](https://www.statisticshowto.com/white-test/statisticshowto) (the one you’re testing), is that the variances are not equal:  
H1 = σ2i ≠ σ2.

Looking at scatter plot we obsorve that in lower level residuals has bigger variance but in upper level the variance gets less and all dots are close to the green line. So it’s heterogedastic.

## Second method Use Breusch-Pagan testing for homoscedasticity

H0-homogedasticity

H1-heterosgedasticity

Graphical user interface

Description automatically generated

p-value less that alpha means we reject null-hypothesis which claims homogedastisity. Eventually it’s heterogedastic.

**Last time running regression and interpreting it’s coefficients**

A screenshot of a computer

Description automatically generated

Our regression module is following:

**Golden-Score = 5.4 + 5.32\*Log(Social\_Support\_squared) + 0.27\* Log(Perception of Coorruption) + 1.5\*Log(HealthyLifeExpectancy) – 0.88\*Generosity + 2.59\*FreedomToMAkeLifeChoices**

**After some changes in data we run regression again. Now we see that except generosity all other coeffcients of variables has statistically significant impact on Golden-Score. The are also jointly significant too.**

**The variation of dependant variable is explained by model (R^2) is 89,46%. Which is pretty good!**

* Relation between Golden-Score and Log(Social\_support\_squared) is Level-Log regression. So 1% increase in Social Support will cause b1/100= 0.0532 unit increase in Goldes-Score.
* Relation between Golden-Score and Log(Perception of corruption) is also Level-Log.

Again 1% increase in perceptionTocorruption will cause 0.0027 unit increase in Golde-Score.

* Golden Score & Log(HealthLifeExpectancy) is also Level-Log. So 1% increase in Healthy Life expectancy will cause 0.015 unit increase in Golden\_Score
* 1 unit increase in Generosity will cause to decrease by -0.88 in Golden-Score.
* 1 unit increase in **FreedomToMakeLifeChoices will cause to increase by 2.59 in Gonder-Score.**