The Environmental and Socio-Economical Factors of Depression

Baielli Alisia Sara, Cerretti Diego

January 2023 - Bocconi University

TABLE OF CONTENTS

Introduction and Research Question **1**

The Dataset **1**

Multivariate Linear Regression **2**

Model Selection **3**

Hypothesis Testing **4**

Conclusion **5**

Appendix **6**

Introduction and Research Question

Depression is the disorder affecting the most people worldwide. Despite the great efforts made by the medical community, this complex mental health disease is still considered incurable[1]. While there are a variety of effective treatments available for managing the symptoms of the illness, including medications, therapy, and lifestyle changes, there is no cure to the underlying disorder itself. Recent studies explored the trends in depression rates [Fig. 1], highlighting a concerning increase in the portion of the population affected by the disease in the last few years. There are ongoing debates among scientists about the potential drivers of this problematic growth. While the causes of depression are complex and varied, research has shown that both environmental and socio-economic factors can play a role in the development and severity of this illness. It is therefore of interest to analyse the statistical evidence surrounding the link between these factors and depression rates.

Our statistical analysis aims to investigate the correlation between socio-economical and environmental data of a country and its depression rate. We developed the same inquiry on a sample of countries worldwide and on the set of OECD countries. Ultimately, we tested whether OECD countries, considered by many as the most developed, display higher depression rates compared to other countries.

The Dataset

As previously mentioned, we developed our analysis on two separate parallel inquiries: the investigation of the drivers of depression on a sample of countries worldwide and on the set of OECD member states. The OECD, or the *Organisation for Economic Cooperation and Development*, is an intergovernmental organisation that works to promote economic growth, prosperity, and sustainable development in its member countries. It currently has 37 member countries, including some of the most developed ones. We built two different datasets for our study in order to individually focus on the two samples. There are multiple reasons why we chose to also analyse OECD countries on a stand alone basis. First of all, developed countries are arguably more inclined to provide unbiased, high-quality and up-to-date data. Secondly, the OECD is a sample of reasonable size from which to gather our data because its number of members is small but still sufficiently large to carry out a statistical study. In addition, the OECD publishes a great amount of open-source datasets on its website. Lastly, our global-scale inquiry lacks some of the data that is available for the most developed countries only.

In order to represent our dependent variable, i.e. depression, we used the latest available data regarding the percentage of population affected by the disease in several countries. Then, we proceeded building our datasets by looking for data relative to possible drivers of depression in a country's population. These drivers can be divided into three main macro-categories: "social", "economical" and "environmental". The factors taken into account are the following:

- Social: peacefulness, education level, internet usage.
- Economical: income inequality, wealth
- Environmental: weather, urbanisation

[1] Mental Health America, https://screening.mhanational.org/content/depression-curable/?layout=actions_b

In the inquiry concerning OECD countries, we were able to additionally include working hours as an independent variable.

Peacefulness is indicated in both datasets by the 2022 GPI index, or Global Peace Index. We represented the education level in countries worldwide using the 2019 Education Index. For OECD countries, on the other hand, we fulfilled the same purpose by using data relative to the percentage of the countries' populations which have not pursued upper secondary education[2]. Average internet usage is represented in both inquiries by the percentage of a country's population browsing the internet in 2017. Income inequality is described in the two datasets using the latest GINI indices (obtained by two different sources). We collected GDP per capita data for both inquiries in order to represent a population's average wealth. We considered the countries' average temperatures as a variable to evaluate the potential impact of weather. Urbanisation is captured by the percentage of the population living in urban areas. Lastly, average labour time was included in the OECD dataset to describe the amount of labour in each country's population.

For each country, we merged the information relative to these indices to the corresponding depression rate. Then, we cleaned our data by filtering out those countries which lacked information on at least one of the variables taken into consideration. The final global and OECD datasets had sizes of 140 and 37 countries, respectively.

To compare the depression rates between members of the OECD and those which are not members, we had to build a third dataset, which contains all non-OECD countries in the global dataset.

Multivariate Linear Regression

Multivariate linear regression is a statistical method that allows us to understand the relationship between a multidimensional dependent variable and independent variables, also known as predictors or explanatory variables. The multiple linear regression model for n dependent variables $Y_1,...,Y_n$ with corresponding p-dimensional predictor variables $(X_{1,1},...,X_{1,p}),...,(X_{n,1},...,X_{n,p})$ is described by the formula: $$Y_i = \sum_{j=1}^{p} b_j x_{i,j} + e_i \qquad i = 1,...,n$$

where $e_1,. . ., e_n$ are independent normally distributed random variables with expectation 0 and finite variance σ^2.

In this section we are going to interpret the multivariate linear regression associated with the correlation between socio-economical and environmental data of a country and its depression rate: respectively our independent and dependent variables. We repeat this operation considering both the Global dataset and OECD dataset.

First off, we calculated the regression coefficients for each independent variable. These coefficients represent the strength and direction of the relationship between depression and the corresponding independent variables. Ideally, after calculating all regression coefficients, we could

[2] Keep in mind that we used two indicators of education with opposite values in the two datasets (high education indices are reasonably comparable to low percentages of populations which have not pursued upper secondary education). We therefore expect that their correlations to depression, if any, are respectively opposite.

predict the depression rate of a country on the basis of the x values only. In computing our model, we conventionally took as our initial hypothesis that the coefficients' value is 0, meaning that there is no relationship between the x variables and the y variable. Therefore, if our test yields low p-values, we witness significant correlations between the independent and dependent variables.

Interpreting the results relative to the Global dataset, we observe that the p-value of the F-statistics is *p-value: < 2.2e-16*, well below 0.05. This means that at least one of the predictor variables is significantly related to the dependent variable. By examining the coefficient table, which shows the estimate of regression beta coefficients and the associated p-values, we can gather that *Education* and *Internet usage* are statistically significant because their p-value is smaller than 0.05[3]. Then, to understand if the correlation is positive, we keep in mind that a positive coefficient estimate suggests that an increase in the predictor variable produces an increase in the dependent variable. In our case, both *Education* and *Internet usage* [Fig. 2, 4, 6] are positively correlated to depression. The overall quality[4] of the model can be assessed by examining the R-squared and Residual Standard Error, which represents the proportion of variance in the outcome variable y that may be predicted by knowing the value of all x variables. An R-squared value close to 1 reveals that the model is able to represent a large portion of the variance in the outcome variable. The Residual Standard Error estimate gives a measure of error of prediction. The lower the Residual Standard Error, the more accurate the model. In our model the Residual Standard Error is 0.4734 on 132 degrees of freedom. Therefore, we can deduce that our model is quite accurate[5].

The results found using the OECD dataset are consistent with the outcomes depicted above. However, we observe that *Education* [Fig. 5] is the only statistically significant variable. In contrast with the Global inquiry, *Education* is negatively correlated to depression, for the reasons mentioned in section "The Dataset". In the next section, using model selection methods, we will see how it is possible to find other statistically significant variables.

Overall, multivariate linear regression is a useful tool for understanding the relationships between variables and making predictions based on those relationships. However, by including all parameters we risk overfitting the model. As a result, the model may perform well on the training data, but not generalise to new data. In order to overcome this problem we will use model selection methods and check whether the assumptions required to state the statistical significance are met.

Model Selection

As previously stated, it is not always the optimal choice to include all measured predictor variables in the regression model. A way around this problem is to use model selection, a technique whose scope is to find the best-fitting model that is effective at making predictions or understanding relationships in the data. Model selection is important in statistical studies because it allows us to identify the most relevant variables and to avoid overfitting.

---

[3] Conventionally, in fact, we consider our significance level to be $\alpha = 0.05$.

[4] STHDA, http://www.sthda.com/english/articles/40-regression-analysis/168-multiple-linear-regression-in-r/

[5] To formally prove this, we can estimate the error rate by dividing the RSE by the mean outcome variable. In our case, the error rate is found to be 0.1055956 in the Global analysis, and 0.09473589 in the OECD analysis, which is fairly low.

In this section we will exploit two different approaches to model selection, the Step-up and Step-down methods. Then, we will compare their output to the initial model using AIC.

The step-up method takes an empty model including the intercept only, and no parameters. Firstly, it tests all $H_0 : \theta_i = 0$ separately and, if none of the hypotheses are rejected, we keep the empty model as our optimal. Otherwise, if any are rejected, we include in the model the parameter with the lowest p-value. Hence, we iterate the scheme enlarging our model until no parameters with p-value lower than our significance level are left.

The step-down model is specular to the step-up. We start off with the full model including all parameters and we test all $H_0 : \theta_i = 0$ separately. If all of them are rejected we choose the full model; otherwise we remove the θ_i with the highest p-value. We iterate the procedure keeping the smallest model, until all remaining parameters have p-values below the significance level.

In both methods, we find that the optimal model is dependent on *Education* and *Internet usage* only[6]. To check whether our selected models are statistically better than the one found via multivariate linear regression, we apply the Akaike's Information Criterion. AIC is a statistical measure that is used to evaluate the quality of a statistical model. It takes into account both the goodness-of-fit of the model and its complexity, with the goal of finding the model that strikes the right balance between these two factors. AIC allows you to compare different statistical models and select the ones that are most likely to be the best fit for your data. This tool can only be used by reasonably assuming that we are dealing with sufficiently large datasets. Comparing the three models with AIC we get a higher coefficient for the full model and an equal coefficient for the step-up and step-down models. Thus, we deduce that the step-up and step-down models, containing the only two explanatory variables (i.e. *Education* and *Internet usage*) are qualitatively best. This result is consistent in both inquiries. However, we surprisingly observe that *Internet usage* is directly proportional to depression in the Global regression, whereas it is inversely proportional in the OECD inquiry [Fig. 3, 7]. We will later discuss possible reasons why this might have happened.

Lastly, we used the final model obtained with the step-up method to check the assumptions on the residuals, namely their normality, their homoscedasticity [Fig. 8, 9] and that their mean value is close to 0. In order to prove normality of the residuals, we visually observed their gaussian-shaped distribution with a histogram [Fig. 10, 11], verified the shape of their qq plot [Fig. 12, 13] and computed their corresponding Shapiro-Wilk tests. To check homoscedasticity, we graphically inspected whether the residuals' variance was close to being constant. Computing their mean value, we are able to evaluate whether it is reasonably close to 0. All assumptions were satisfied, meaning that we could assess the results of the regression with the new model.

Hypothesis Testing

After comparing our analyses, it is of interest to learn whether OECD countries are significantly more depressed than non-OECD countries. In order to do so, it is reasonable to perform a student's two-sample t-test. However, it is first essential to prove the prerequisites for t-testing: independence, homoscedasticity and normality of the dependent variables. Arguably, we can assume

[6] Note that this is not a trivial result: different outcomes of the two methods are often determined by the correlation between the independent variables.

that depression rates in OECD countries are not correlated in any way to those in non-OECD countries (thus implying independence). Furthermore, to attest homoscedasticity in the two models, it is enough to perform a variance test or visually observe the resemblances of their corresponding boxplots [Fig. 14]. Lastly, to check normality, we can assess a series of dedicated tests, such as the Kolmogorov-Smirnov test and the Shapiro-Wilk test, and visual approaches, for example by plotting their histograms [Fig. 15, 16] and qq plots [Fig. 17, 18].

Our null hypothesis is $H_0: \mu_x \leq \mu_y$, where μ_x is the mean of the depression rate in OECD countries and μ_y is the mean of the depression rate in non-OECD countries. Specularly, the alternative hypothesis is $H_1: \mu_x > \mu_y$. Performing the test, we conclude that we must reject the null hypothesis. As a matter of fact, we observe that the test statistic belongs to the confidence interval, and the resulting p-value is much smaller than the significance level of 0.05. Thus, we establish that OECD countries have a significantly higher depression rate than non-OECD countries [Fig.19].

Conclusion

To sum up, we found statistical evidence of direct significant correlation between the depression rate and average education level of a population. Moreover, we surprisingly ran into opposite conclusions when assessing the link between depression and the internet usage in a country. While higher internet usage is a driver of depression in our global inquiry, in the OECD analysis we observe that internet usage influences our dependent variable in the other way around. Lastly, as a result of our hypothesis testing, we manage to draw the strong conclusion that OECD members witness an averagely higher depression than non-OECD countries.

The greatest limitation in our study is arguably represented by the choice of handling data regarding OECD countries. As a matter of fact, with all the advantages following our strategy, comes a great issue. OECD members are not a generic sample of countries in the world. On the contrary, they represent the most developed societies. This implies that, generally speaking, OECD countries share some very similar characteristics. For this reason, when choosing to analyse this dataset, we run the risk of producing a biased analysis. When studying the correlations to depression rate, we must keep in mind that this subset of countries share, for example, very similar internet usage percentages. Therefore, since this variable ranges in a very small interval [figure 3], it becomes hard to distinguish its link with depression. This condition is the possible reason why our analysis yielded two different, opposite, conclusions when assessing the link between internet usage and depression.

Another possible weakness of our analysis is the choice of the indices representing our independent variables. For example, temperature isn't always linked to the typical weather of a country: it is not a direct implication of the amount of sunlight, fog, rain in a region. For instance, it could have been of interest to distinguish our countries by their climate region of belonging, rather than their temperature.

A fascinating future addition to our inquiry could be the focus on the effects of the recent COVID-19 pandemic on the depression rates of the involved nations. Moreover, if the relative data were available, it would be curious to repeat the same study by distinguishing the depression rates of female and male populations in a country.

Appendix

R code outputs:

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.12602 -0.26445 -0.00844  0.28016  1.44710

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.901e+00  4.742e-01   6.117 1.01e-08 ***
GINI            4.336e-03  6.414e-03   0.676   0.5002
GPI             1.024e-01  1.047e-01   0.978   0.3299
Avg_temp       -6.560e-03  7.218e-03  -0.909   0.3651
Education_idx   1.227e+00  5.397e-01   2.274   0.0246 *
GDP_pc         -2.740e-06  3.512e-06  -0.780   0.4367
Urbanisation    7.707e-04  3.236e-03   0.238   0.8121
Internet_usage  1.004e-02  3.983e-03   2.520   0.0129 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4734 on 132 degrees of freedom
Multiple R-squared:  0.4988,    Adjusted R-squared:  0.4723
F-statistic: 18.77 on 7 and 132 DF,  p-value: < 2.2e-16
```

Multivariate linear regression of Global inquiry

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.11554 -0.27903  0.00166  0.24263  1.58654

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.217941   0.184359  17.455  < 2e-16 ***
Internet_usage 0.008534   0.002913   2.930  0.00397 **
Education_idx  1.242748   0.448997   2.768  0.00642 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4706 on 137 degrees of freedom
Multiple R-squared:  0.4859,    Adjusted R-squared:  0.4784
F-statistic: 64.74 on 2 and 137 DF,  p-value: < 2.2e-16
```

Step-up and Step-down model of Global inquiry

	df	AIC
stepup_model_1	4	191.2278
stepdown_model_1	4	191.2278
fit1	9	197.6593

AIC of Step-up, Step-down and Multivariate Linear Regression models in Global inquiry

```
        Shapiro-Wilk normality test

data:  residuals(final_model_1)
W = 0.9895, p-value = 0.3751
```

Shapiro-Wilk Test for Residuals in Global inquiry

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.8406 -0.2787 -0.0516  0.2172  0.8530

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           7.518e+00  1.589e+00   4.732 5.77e-05 ***
GINI                  8.721e-01  2.258e+00   0.386   0.7023
GPI                  -1.600e-01  2.422e-01  -0.661   0.5142
Avg_temp              8.555e-03  1.752e-02   0.488   0.6292
Education_idx        -2.099e-02  8.169e-03  -2.570   0.0158 *
GDP_pc               -2.849e-07  7.172e-06  -0.040   0.9686
Urbanisation          1.875e-04  1.039e-02   0.018   0.9857
Internet_usage       -2.328e-02  1.704e-02  -1.366   0.1827
Avg_hrs_worked_yearly -1.391e-04  6.134e-04  -0.227   0.8222
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.475 on 28 degrees of freedom
Multiple R-squared:  0.289,     Adjusted R-squared:  0.08579
F-statistic: 1.422 on 8 and 28 DF,  p-value: 0.2307
```

Multivariate linear regression of OECD inquiry

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.85263 -0.31237 -0.03146  0.23281  0.87250

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.085913   0.779436   9.091 1.26e-10 ***
Education_idx -0.017456   0.005271  -3.312   0.0022 **
Internet_usage -0.020664   0.008712  -2.372   0.0235 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4388 on 34 degrees of freedom
Multiple R-squared:  0.2629,    Adjusted R-squared:  0.2196
F-statistic: 6.064 on 2 and 34 DF,  p-value: 0.005593
```

Step-up and Step-down models of OECD inquiry

	df	AIC
stepup_model_2	4	48.92391
stepdown_model_2	4	48.92391
fit2	9	57.66208

AIC of Step-up, Step-down and Multivariate Linear Regression models in OECD inquiry

```
        Shapiro-Wilk normality test

data:  residuals(new_model_2)
W = 0.97291, p-value = 0.4928
```

Shapiro-Wilk Test in OECD inquiry

```
       F test to compare two variances

data:  Depression_nonOECD and Depression_OECD
F = 1.4475, num df = 103, denom df = 36, p-value = 0.2077
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8110691 2.4047430
sample estimates:
ratio of variances
         1.447513
```

Test homoscedasticity of OECD and Non-OECD depression rate data

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Depression_OECD
D = 0.11043, p-value = 0.7576
alternative hypothesis: two-sided
```

Kolmogorov-Smirnov Test of OECD depression rate data

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  Depression_nonOECD
D = 0.095921, p-value = 0.2941
alternative hypothesis: two-sided
```

Kolmogorov-Smirnov Test of Non-OECD Depression rate data

```
        Shapiro-Wilk normality test

data:  Depression_OECD
W = 0.96355, p-value = 0.2612
```

Shapiro-Wilk test of OECD Depression rate data

```
        Shapiro-Wilk normality test

data:  Depression_nonOECD
W = 0.98391, p-value = 0.2405
```

Shapiro-Wilk test of Non-OECD Depression rate data

```
        Two Sample t-test

data:  Depression_OECD and Depression_nonOECD
t = 6.4938, df = 139, p-value = 6.888e-10
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.5308555       Inf
sample estimates:
mean of x mean of y
 5.013514  4.300962
```

Two-sample t-test

Plots:

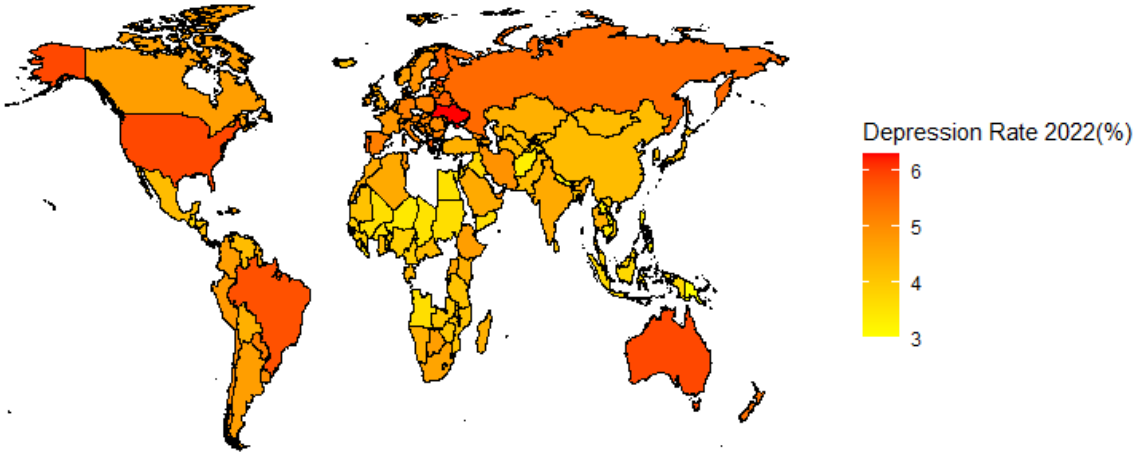
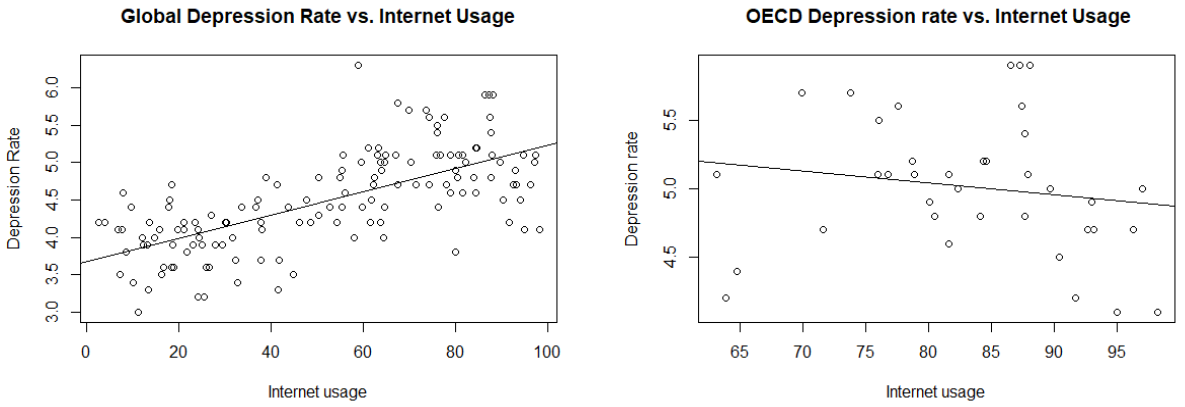
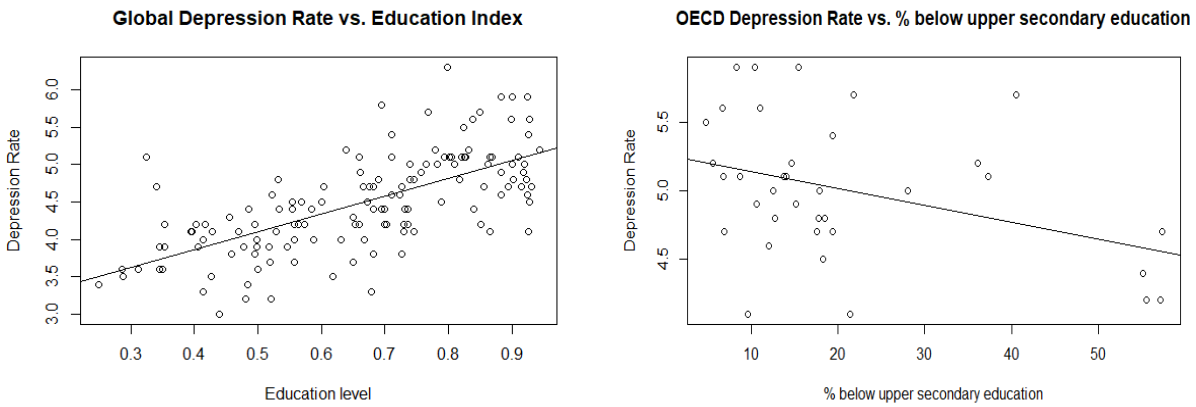


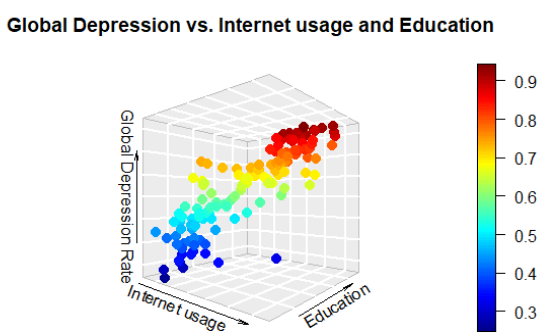
Figure 1: Worldwide Depression Rates

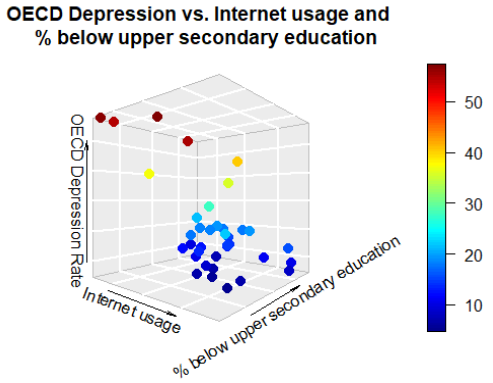


Figures 2,3: Correlation between Depression Rate and Internet Usage in Global and OECD countries

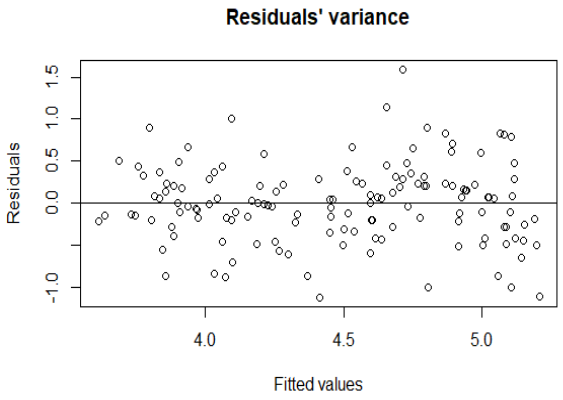


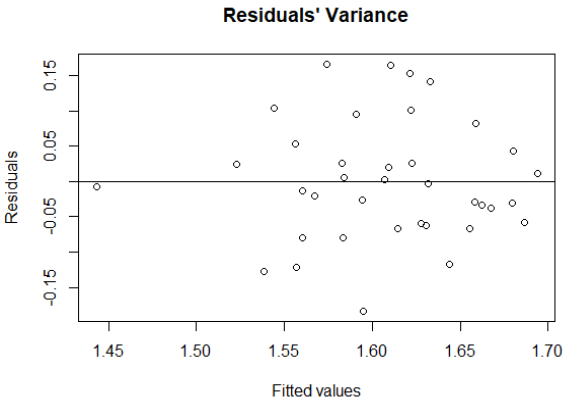
Figures 4,5: Correlation between Depression Rate and Education in Global and OECD countries



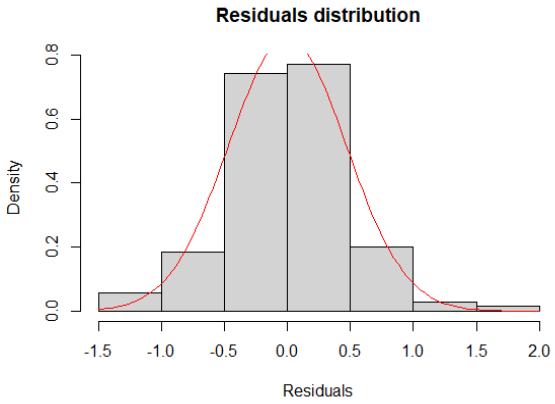


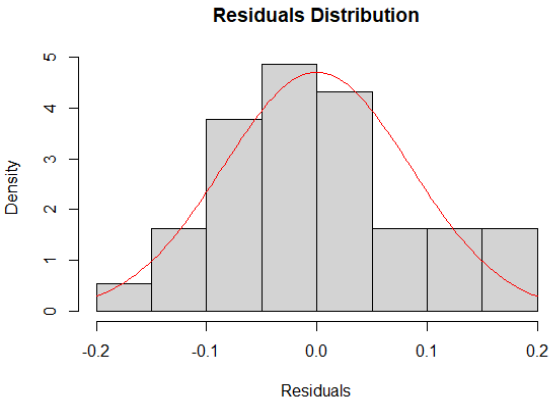
Figures 6,7: Correlation between Depression Rates and Education and Internet Usage in Global and OECD countries



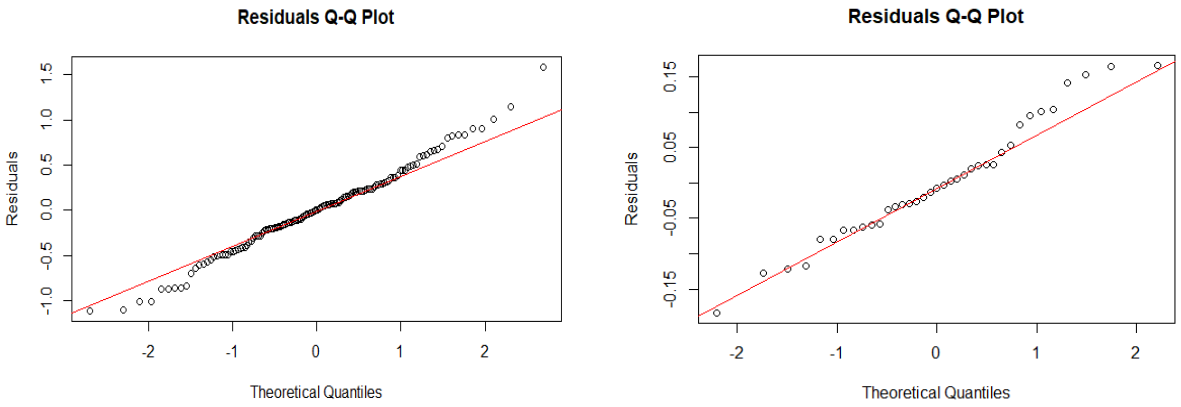


Figures 8,9: Variance of residuals in global and OECD countries





Figures 10,11: Histograms of residuals' distributions in Global and OECD inquiries respectively



Figures 12, 13: Residuals Q-Q Plots in Global and OECD countries

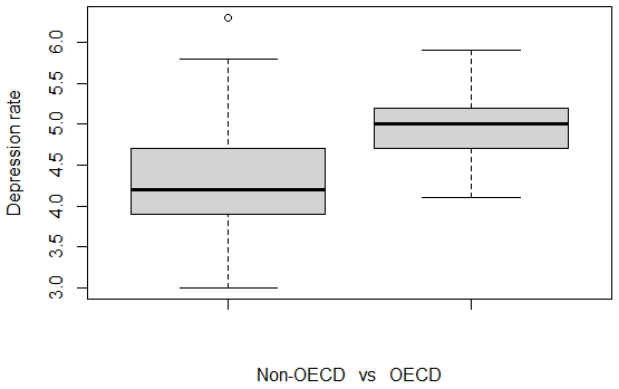
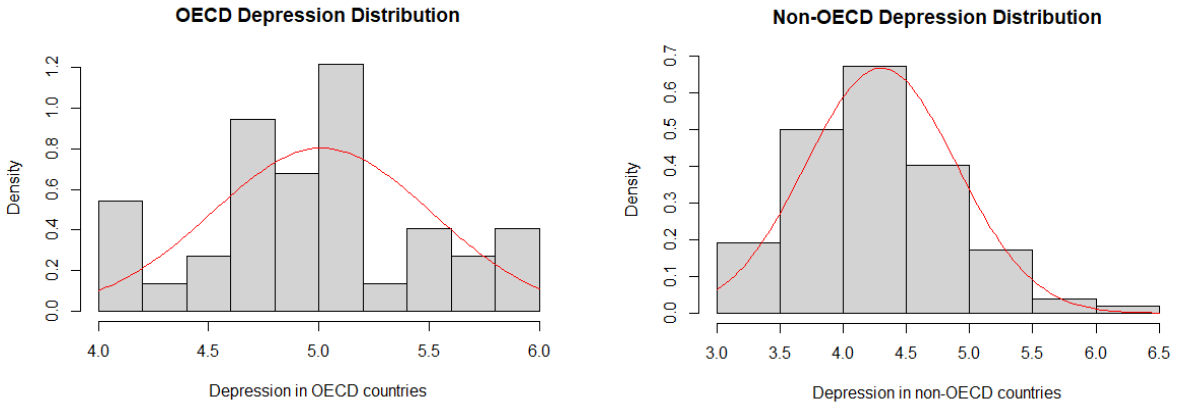
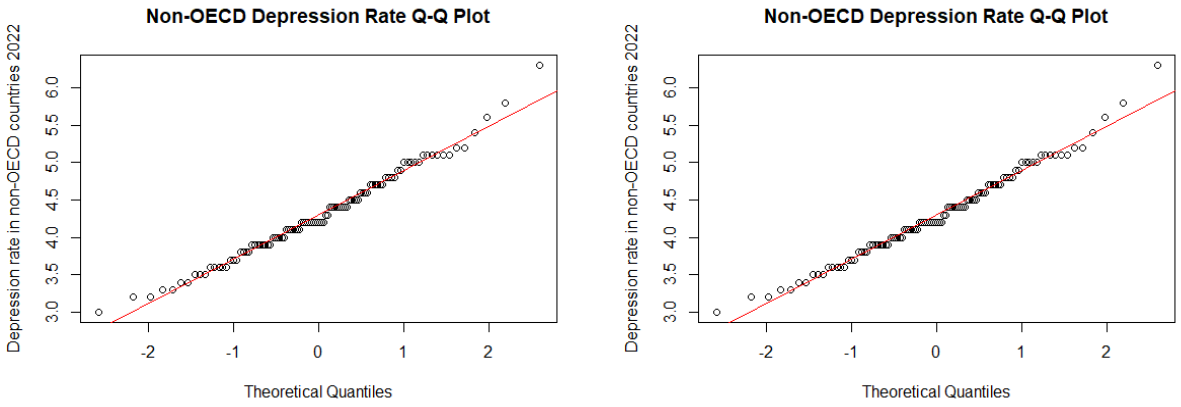


Figure 14: Boxplot of Depression distribution in OECD and non-OECD countries



Figures 15,16: Histograms of Depression distribution in OECD and Non-OECD countries



Figures 17,18: Q-Q Plots of OECD and Non-OECD distributions

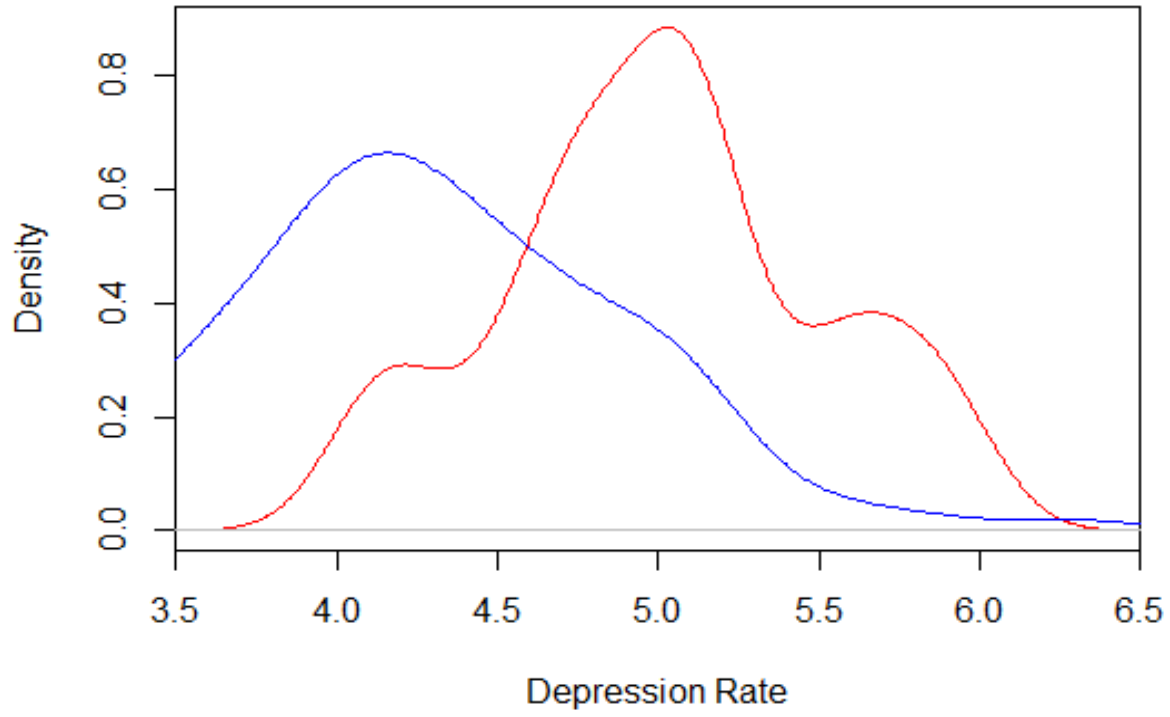


Figure 19: Comparison of density of Depression rates in OECD (red) and Non-OECD (blue) countries

Sitography:

Depression rate by country: https://worldpopulationreview.com/country-rankings/depression-rates-by-country

GINI coefficient by country: https://worldpopulationreview.com/country-rankings/gini-coefficient-by-country

GPI by country: https://worldpopulationreview.com/country-rankings/most-peaceful-countries

Average temperature by country:
https://worldpopulationreview.com/country-rankings/hottest-countries-in-the-world

Education index by country: https://en.wikipedia.org/wiki/Education_Index

GDP by country: https://worldpopulationreview.com/country-rankings/gdp-per-capita-by-country

Urbanisation by country: https://en.wikipedia.org/wiki/Urbanization_by_country

Internet usage by country: https://data.worldbank.org/indicator/IT.NET.USER.ZS

OECD Adult education level by country: https://data.oecd.org/eduatt/adult-education-level.htm

OECD Average labour time by country: https://data.oecd.org/emp/hours-worked.htm

OECD GINI: https://data.oecd.org/inequality/income-inequality.htm

Bibliography:

Saveanu, R.V. and Nemeroff, C.B., 2012. Etiology of depression: genetic and environmental factors. *Psychiatric clinics*, *35*(1), pp.51-71.

Renaud-Charest, Olivier, et al. "Onset and frequency of depression in post-COVID-19 syndrome: A systematic review." *Journal of psychiatric research* 144 (2021): 129-137.

Goodwin, Renee D., et al. "Trends in US Depression Prevalence From 2015 to 2020: The Widening Treatment Gap." *American Journal of Preventive Medicine* 63.5 (2022): 726-733.

Bijma, F. et al. “An Introduction to Mathematical Statistics” (2016)