NATIONAL ECONOMICS UNIVERSITY

-------*-------**

GROUP ASSIGNMENT
ECONOMETRICS

TOPIC

"Socio-Medical Factors Affecting Suicide Rates:
The Role of Mental Illness, Alcohol Abuse, and Health Expenditure"

Group 3

Nguyễn Danh Dũng - 11230523
Trần Thu Hiền - 11230534
Bùi Việt Huy - 11230542
Đỗ Công Huy - 11230543
Phạm Khánh Linh - 11230560
Nguyễn Thị Hà Phương - 11230585
Class: DSEB 65B

Lecturer: Ph.D Nguyễn Văn Quý

HANOI, APRIL 2025

TABLE OF CONTENTS

I. Introduction

Suicide is a significant global public health crisis, accounting for 1.1% of all deaths worldwide in 2021, meaning that 1 in every 100 deaths is by suicide (WHO). According to World Health Organization, each year, approximately 727,000 individuals take their own lives, and many more attempt suicide, leaving devastating impacts on families, communities, and entire nations. Particularly alarming is its prevalence among young people, ranking as the third leading cause of death among those aged 15–29 globally. These statistics reflect not only personal tragedies but also broader socio-medical issues that demand urgent attention and effective intervention.

Healthy People 2030 highlights the need for prevention and health promotion strategies to reduce suicide-related behaviors and improve mental health outcomes at the population level. Suicide, as a leading cause of death worldwide, is intricately linked to a range of mental health conditions that significantly elevate the risk of suicidal thoughts and behaviors. These conditions, which include depression, anxiety, bipolar disorder, schizophrenia, eating disorder and substance abuse (such as alcohol dependency), are often identified as primary contributors to suicide. These disorders often lead to feelings of hopelessness, despair, and isolation, all of which can create an overwhelming emotional burden on individuals. When mental health deteriorates to the point where individuals are unable to cope with these feelings, the risk of suicide can become tragically high. In addition, increases in economic uncertainty, measured by the World Uncertainty Index (WUI), are associated with higher suicide rates globally (Khan, 2023). Contributing economic factors such as unemployment, GDP fluctuations, and inflation may also influence suicide risk.

By applying econometric models and analyzing relevant data, this report aims to provide empirical evidence on how variations in the prevalence of mental illness, health expenditure and economics factors are statistically associated with changes in suicide rates. The findings of this research are expected to contribute valuable insights that can assist policymakers, mental health organisations, and healthcare providers in designing more effective prevention strategies and implementing supportive public health interventions. Understanding the relationship between mental health disorder and suicide is essential for developing effective prevention strategies and public health policies to reduce suicide risks and promote better mental health outcomes in society.

II. Literature review

Understanding the multifaceted determinants of suicide requires a close examination of mental health, substance abuse, and healthcare-related factors. Existing literature offers significant insights into how these variables affect suicide risk, yet many studies fall short of providing a comprehensive, cross-national, and quantitative foundation necessary for policy-oriented analysis. This section critically evaluates key sources on the relationship between suicide rates and the prevalence of mental disorders, alcohol abuse, and health expenditure, forming the empirical basis for this report's econometric modeling.

One notable report, *Mental Health and Suicide Statistics* by The Jed Foundation, reports that approximately 36.2% of young adults aged 18–25 in the U.S. experienced a mental, behavioral, or emotional issue in the past year up from 22.1% in 2016 (SAMHSA, 2023). Focused on depression, SAMHSA (2023) reports that 19.5% of adolescents aged 12 to 17 experienced a major depressive episode within the past year. Among college students, the Healthy Minds Study (2023) reveals that 36% have been diagnosed with anxiety, and 30% have been diagnosed with depression. In addition to mental health statistics, the aforementioned sources also highlight suicide rates. Furthermore, SAMHSA (2023) states that 13.6% of adults aged 18 to 25 reported having serious thoughts of suicide in the past year. These rising prevalence rates not only underscore a public health emergency but also provide measurable variables suitable for econometric analysis. These studies offers valuable, up-to-date statistical insights into the mental health challenges and suicide risks faced by young people; however, it primarily focuses on prevalence rates and lacks in-depth analysis of underlying causes or contributing sociocultural factors.

Secondly, the article *Suicide in Global Mental Health* by Lovero KL, Dos Santos PF, Come AX, Wainberg ML, and Oquendo MA introduces a broader perspective compared to the previous studies, by incorporating not only mental health and social factors, but also economic stressors as contributors to suicide risk. The authors emphasize that mental disorders such as depression, anxiety, bipolar disorder, and schizophrenia are significant risk factors for suicidal ideation and behaviors. A notable concern highlighted in the article is that many individuals in low-income and middle-income countries (LMICs) live with untreated mental health conditions due to stigma and limited access to care. Regarding the influence of social and economic conditions on suicide rates, the article identifies several contributing factors, including poverty, unemployment, gender inequality, domestic violence, and exposure to trauma. These conditions not only

deteriorate mental health but also elevate the risk of suicide. However, the article still lacks region-specific statistical breakdowns that could be used directly in empirical models.

Subsequently, the article *Mental Health Conditions Can Contribute to Suicide Risk*, authored by Farzana Akkas and Allison Corr (2022), offers a focused exploration of the relationship between mental health disorders, access to care, and suicide risk. Their article emphasizes that over 50% of individuals who die by suicide have no formal mental health diagnosis, pointing to diagnostic failures and underreporting. This finding suggests that a substantial proportion of at-risk individuals may remain undetected by healthcare systems, either due to underreporting, misdiagnosis, or barriers to care. Furthermore, the article highlights major concerns regarding limited access to mental health care, which remains a pervasive challenge in many societies. It is noted that fewer than half of those experiencing suicidal thoughts actually receive any form of mental health treatment. The article thus reinforces the notion that enhancing both the availability and quality of mental health services is essential in addressing suicide prevention effectively. This article presents a crucial and nuanced understanding of how systemic barriers in mental health care intersect with psychiatric disorders to influence suicide risk; however, its narrow focus limits the exploration of broader sociocultural or economic determinants that may also play a significant role.

Last but not least, a recent report published in Molecular Psychiatry (2025), titled *Risk of suicide and all-cause death in patients with mental disorders: a nationwide cohort study*, offers valuable insights into two critical dimensions of suicide risk: the elevated risk associated with mental disorders and the influence of illness duration on suicide vulnerability. The study presents robust empirical evidence indicating that individuals diagnosed with mental health conditions are at a significantly higher risk of suicide when compared to those without such diagnoses. This conclusion is based on a large-scale cohort analysis involving nearly four million adults in South Korea, tracked over an average follow-up period of 11.1 years, providing a strong statistical foundation for its findings. With respect to the second focus of the study the role of illness duration the results reveal that the risk of suicide increases as the duration of certain mental disorders extends. Specifically, individuals suffering from depression, insomnia, alcohol use disorder, and substance use disorders exhibit higher suicide risk over time. Notably, for patients with bipolar disorder, the study found that suicide risk peaked within the first year following diagnosis, emphasizing the critical nature of early-stage intervention. Additionally, the

findings underscore that individuals with conditions such as personality disorders, bipolar disorder, and schizophrenia face especially heightened risks, reinforcing the necessity of timely diagnosis, sustained monitoring, and targeted therapeutic strategies. Overall, the study highlights the urgency of strengthening mental health care systems to identify and support high-risk individuals at the earliest possible stage. This large-scale longitudinal study provides compelling empirical evidence linking both the presence and duration of mental disorders to elevated suicide risk, yet it remains limited by its geographic focus, potentially restricting the generalizability of its findings to other cultural or healthcare contexts.

III. Methodology

1. Data description

a. Data overview

The dataset employed in this study contains 900 cross-sectional observations, collected across various countries at five distinct time points spaced every four years between 2000 and 2016 (i.e., 2000, 2004, 2008, 2012, and 2016). Each observation represents a unique combination of country and year, although the dataset does not form a full panel due to the temporal spacing and potential missing data for some countries in specific years. The dataset comprises a total of 10 core variables used for econometric analysis, including 1 dependent variable (Suicide mortality rate) and 9 independent variables, which span across mental health prevalence, economic indicators, and health system factors. This structure allows for cross-sectional econometric techniques to examine potential associations between socio-economic and mental health variables and suicide outcomes across different countries and time periods.

b. Variables description

Names	***Description***
Country name	Indicates the official name of each country included in the dataset. It is useful for identifying the geographical and cultural context of each observation.

Country code	A standardized alphanumeric code that facilitates data merging and consistency across datasets, particularly when discrepancies in country naming exist.
Continent	A categorical variable denoting the continent each country belongs to. It is used to control for region-specific effects that may influence suicide rates or mental health systems.
Year	This variable records the year corresponding to each observation. Each country appears only five times, corresponding to the years 2000, 2004, 2008, 2012, and 2016. Each year-country pair is treated as an independent cross-sectional observation, rather than forming a balanced panel dataset.
Suicide mortality rate	This is the dependent variable, measuring the number of suicide deaths per 100,000 individuals in a given country and year. It is used to assess the severity of suicide as a public health issue across countries. The rate is scaled to enhance interpretability and comparability across countries and over time.
Bipolar disorder	These variables represent the estimated prevalence of respective mental health conditions within the population. The prevalence variables are measured as the number of cases per 100 individuals in the population.
Anxiety disorder	
Depression	
Eating disorders	

Alcohol disorder	Indicates the number of individuals affected by alcohol-related disorders per 100 population. This variable serves as a proxy for substance-related mental health risks.
Current health expenditure per capita	Indicates the number of individuals affected by alcohol-related disorders per 100 population. This variable serves as a proxy for substance-related mental health risks.
GDP per capita	Measures the gross domestic product divided by midyear population, also in current US dollars. This variable reflects the economic development level of each country.
Inflation	Represents the year-over-year percentage change in consumer prices, serving as an indicator of economic stability and purchasing power within each country.
Unemployment	Represents the share of the labor force that is unemployed.

c. Data sources

To investigate the impact of various socio-economic and health-related factors on suicide mortality rates, this study compiled data from a range of reputable international sources. The suicide mortality rate (per 100,000 population) was obtained from the World Bank database World Bank. Key economic indicators, including GDP per capita, inflation (annual %), and the unemployment rate, were also retrieved from the World Bank Open Data to capture macroeconomic conditions across countries. Additionally, current health expenditure per capita was used as a proxy for the strength of national health systems and access to healthcare services. To account for the influence of mental health conditions, we incorporated data on the prevalence of various disorders; such as depression, anxiety, bipolar disorder, eating disorders, and substance use disorders, which is from the Global

Health Data Analysis 1990–2019 dataset, publicly available on Kaggle (Kamau Munyoro, 2019). These datasets collectively enable a comprehensive cross-country analysis of the determinants of suicide mortality over time.

2. Model evaluation criteria and tests for assumptions

a. Model fit and comparison criteria

R-squared (R^2): The R-squared determines the proportion of variance in the dependent variable that can be explained by the independent variable. In other words, r-squared shows how well the data fit the regression model (the goodness of fit).

Adjusted R-squared: The adjusted R-squared is a modified version of R-squared that accounts for predictors that are not significant in a regression model. In other words, the adjusted R-squared shows whether adding additional predictors improves a regression model or not.

Akaike Information Criterion (AIC): The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data it was generated from. AIC is used to compare different possible models and determine which one is the best fit for the data. AIC is calculated from the number of independent variables used to build the model and the maximum likelihood estimate of the model (how well the model reproduces the data). Lower AIC values suggest better model fit.

$$AIC = 2k - 2\ln(L),$$

in which k is the number of parameters,
and ln(L) is the log-likelihood of the model on the data.

Bayesian Information Criterion (BIC): The BIC is a criterion that evaluates the likelihood of a model given the data and penalizes the number of parameters in the model. It is useful for model comparison when selecting between competing models. Choosing the model with the lowest BIC value means that the model has a good fit to the data and a low complexity.

$$BIC = k.\ln(n) - 2\ln(L),$$

in which k is the number of parameters,
ln(n) is the log of the number of data points,
and ln(L) is the log-likelihood of the model on the data.

b. Tests for assumptions

Ramsey RESET Test: Ramsey's RESET test is a test of whether the functional form of the regression is appropriate. In other words, we test whether the relationship between the dependent variable and the independent variables really should be linear or whether a non-linear form would be more appropriate.

- Null Hypothesis: The model is correctly specified.
- Alternative Hypothesis: The model is not correctly specified.

Zero Mean Test: Ensures that the residuals have an expected value of zero, which is a fundamental assumption in OLS regression.

- Null Hypothesis: The expected value of the residuals is zero.
- Alternative Hypothesis: The expected value of the residuals is not zero.

Variance Inflation Factor (VIF): A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model. This can adversely affect the regression results. Thus, the variance inflation factor can estimate how much the variance of a regression coefficient is inflated due to multicollinearity.

- VIF < 5: indicates low multicollinearity
- VIF between 5 and 10: suggests moderate multicollinearity
- VIF > 10: signals a serious multicollinearity problem that may distort the coefficient estimates

Correlation Matrix: A correlation matrix is a statistical technique used to evaluate the relationship between two variables in a data set. In multiple linear regression, the correlation matrix determines the correlation coefficients between the independent variables of a model.

Breusch-Pagan Test: This test is used to assess the presence of heteroskedasticity in a regression model. One of the fundamental assumptions of linear regression is homoskedasticity, meaning that the residuals (errors) should have constant variance across all levels of the independent variables. If this assumption is violated, heteroskedasticity is said to be present. In such cases, the reliability of the regression results may be compromised, as it can lead to inefficient estimates and biased standard errors.

- Null Hypothesis: Homoscedasticity is present (the residuals are distributed with equal variance).
- Alternative Hypothesis: Heteroscedasticity is present (the residuals are not distributed with equal variance).

Durbin-Watson Test: The Durbin Watson (DW) statistic is used as a test for checking autocorrelation in the residuals of a statistical regression analysis. If autocorrelation exists, it undervalues the standard error and may cause us to believe that predictors are significant when in reality they are not.

- Null Hypothesis: There is no correlation among the residuals.
- Alternative Hypothesis: The residuals are autocorrelated.

IV. Regression analysis

1. Preprocess data

The data preprocessing process began with the removal of observations containing missing values. This step was undertaken to ensure the accuracy and validity of the model, minimize potential biases and noise in the analysis, and reduce the risk of technical errors during model estimation.

Next, the data normalization step was performed. Since the suicide rate is measured per 100,000 people and the indicators for the prevalence of several disorders, such as bipolar disorder, anxiety disorder, depression and eating disorders, are measured using standard percentage values., this step will help normalize the prevalence of mental health disorders into prevalence per 100,000 people.

```
[ ]  variables_to_normalize = [
         'prevalence_of_bipolar_disorder',
         'prevalence_of_anxiety_disorder',
         'prevalence_of_eating_disorders',
         'prevalence_of_depression',
         'alcohol_use_disorders'
     ]

     for var in variables_to_normalize:
         df[var] = df[var] * 1000
```

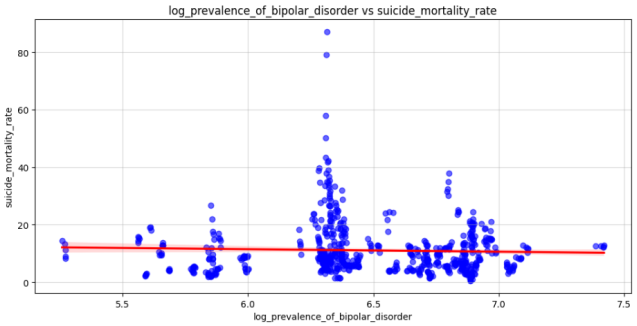
The following command provides summary statistics for all numerical columns, including count, mean, standard deviation, minimum, maximum, and the 25th, 50th, and 75th percentiles. This serves as an evaluative tool that enables further analysis of the variables.

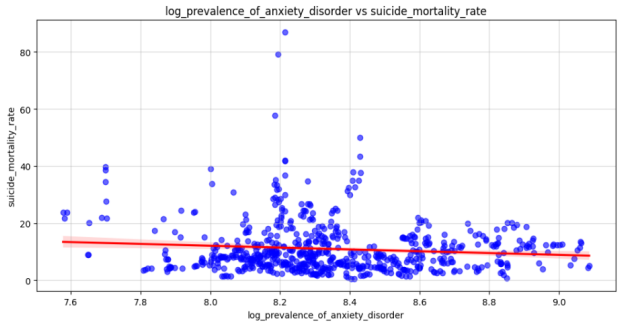
A copy of the original DataFrame is first created to ensure that the original data remains unchanged during the preprocessing steps. A list of excluded variables is then defined, including *suicide_mortality_rate*, *unemployment*, *inflation*, and *year*, as these variables are either not appropriate for log transformation (e.g. the dependent variable). Subsequently, all remaining numerical variables are log-transformed using the natural logarithm (np.log). This transformation is a common econometric technique aimed at reducing skewness and bringing variables closer to a normal distribution, thereby improving the accuracy and reliability of regression models. The transformed variables are then renamed with the prefix 'log_' for clarity and distinction from the original variables. Infinite values, which may arise from taking the logarithm of zero or negative numbers, are replaced with nan, and all rows containing missing values are removed to maintain data integrity. Finally, the dependent variable y is assigned as *suicide_mortality_rate*, while the independent variable matrix X includes *inflation*, *unemployment* and logarithm of the remaining predictors, excluding identifying columns such as *country_name*, *country_code*, *year*, and *continent*. Both X and y are explicitly converted to float type to ensure compatibility with subsequent econometric modeling procedure

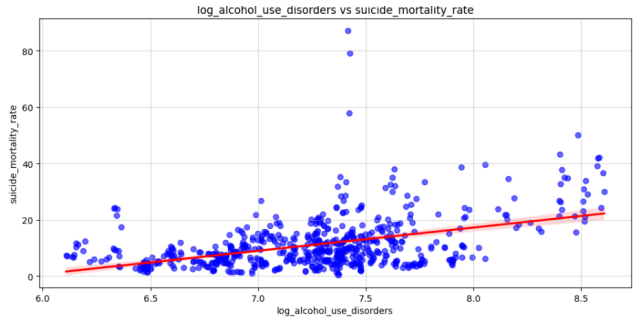
```
[ ]  excluded_vars = ['suicide_mortality_rate', 'unemployment', 'inflation', 'year']

     df_log = df.copy()
     log_vars = [col for col in df_log.select_dtypes(include=[np.number]).columns if col not in excluded_vars]

     for col in log_vars:
         df_log[col] = np.log(df_log[col])


     df_log.rename(columns={col: f"log_{col}" for col in log_vars}, inplace=True)


     df_log.replace([np.inf, -np.inf], np.nan, inplace=True)
     df_log.dropna(inplace=True)


     y = df_log['suicide_mortality_rate']
     X = df_log.drop(columns=['suicide_mortality_rate', 'country_name', 'country_code', 'year', 'continent'], errors='ignore')

     X = X.astype(float)
     y = y.astype(float)
```

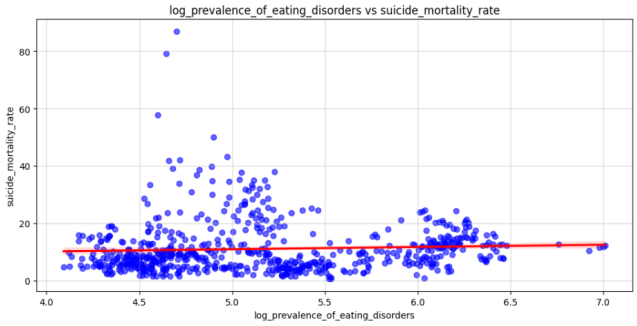
2. Descriptive statistics

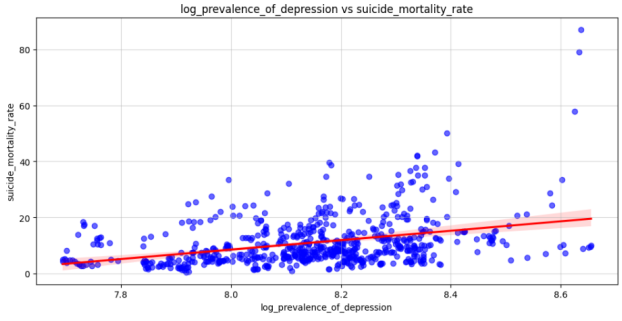
This analysis explores the relationship between independent variables and the suicide mortality rate. These variables include mental health disorders (bipolar, anxiety, depression), healthcare expenditure, GDP per capita, inflation, and unemployment rates. Scatter plots are used to visualize the correlation between each log-transformed independent variable and the suicide mortality rate, helping to identify potential relationships.

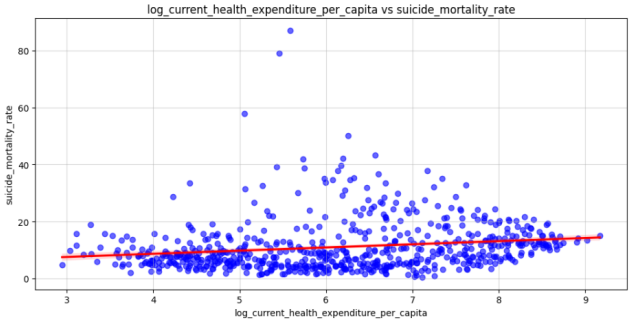


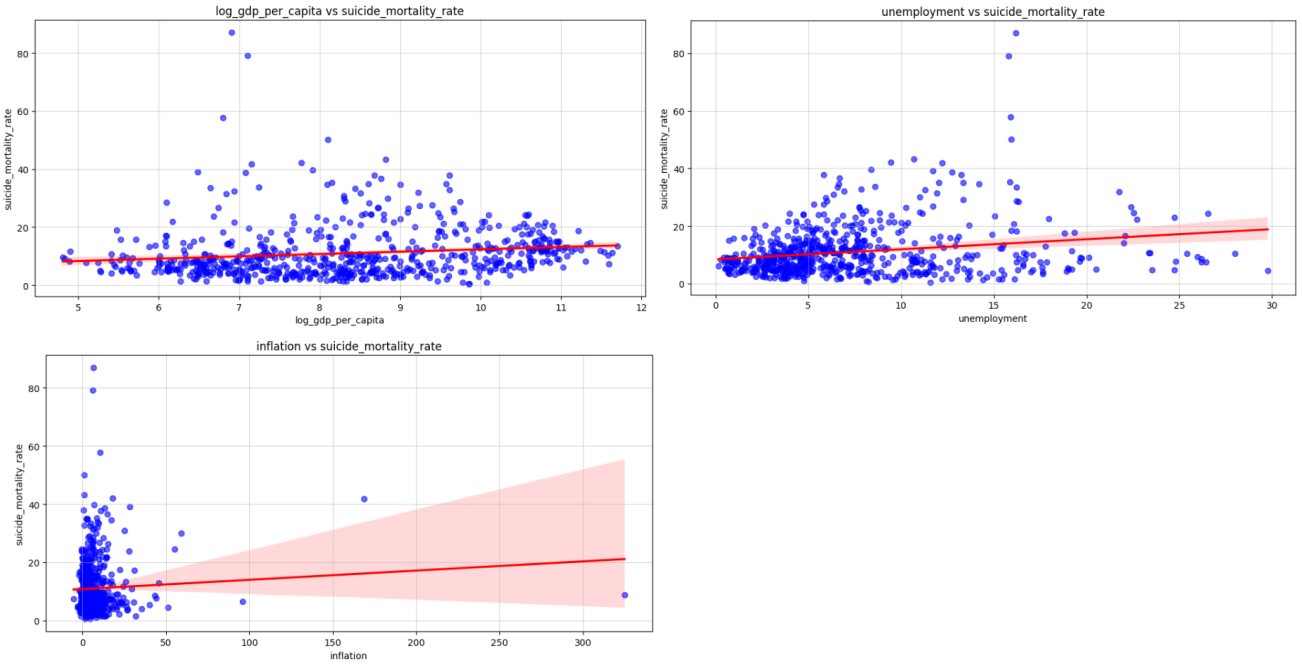












The analysis shows that the relationship between *log_prevalence_of_bipolar_disorder* and *log_prevalence_of_anxiety_disorder* with the *suicide_mortality_rate* is weak. The scatter plots with regression lines, which are nearly horizontal, indicate no clear upward or downward trend in the data. In contrast, both *log_prevalence_of_depression* and *log_alcohol_use_disorders* exhibit a strong positive linear relationship with the *suicide_mortality_rate*. The scatter plots show an increasing trend in suicide mortality rate as the prevalence of these disorders rises, indicating a clear and predictable relationship. On the other hand, the relationship between *log_prevalence_of_eating_disorders* and the *suicide_mortality_rate* is weak, with the regression line almost flat, indicating no significant change in the *suicide_mortality_rate* as the *log_prevalence_of_eating_disorders* changes.

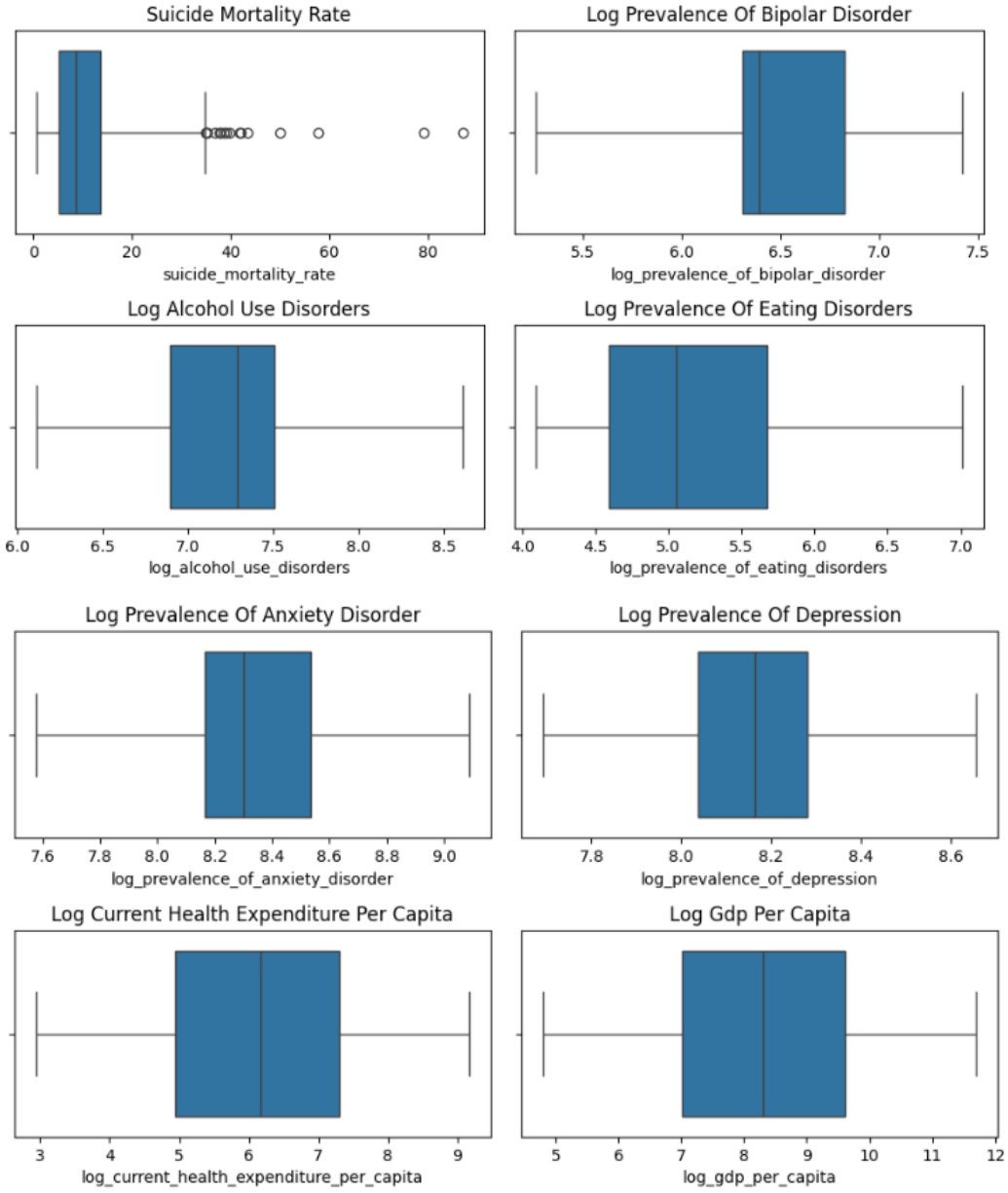
When it comes to economic factors, *log_current_health_expenditure_per_capita* and *log_gdp_per_capita* show a weak positive relationship with the *suicide_mortality_rate*, but there is no significant change in the data. *inflation* and *unemployment* exhibit weak and unclear relationships, with a large spread of data points, making it difficult to identify a definitive trend.

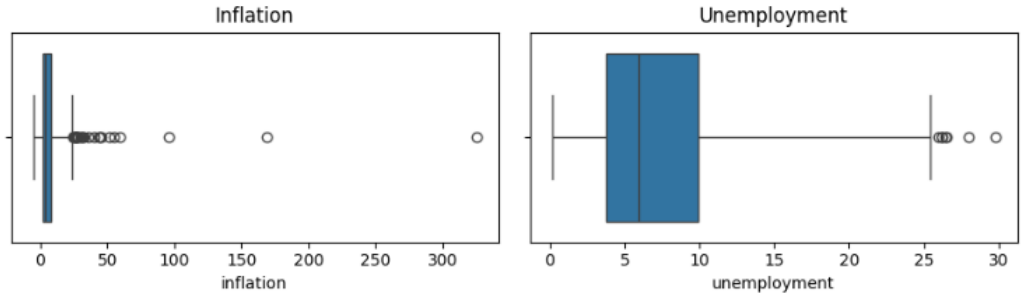
In conclusion, the analysis indicates that *log_prevalence_of_depression* and *log_alcohol_use_disorders* have a clear relationship with *suicide_mortality_rate*, while *log_prevalence_of_bipolar_disorder* and *log_prevalence_of_anxiety_disorder*, and

log_prevalence_of_eating_disorders have weak or no clear relationship. Economic factors require further study, as they currently show only weak relationships with the dependent variable.

3. Outliers

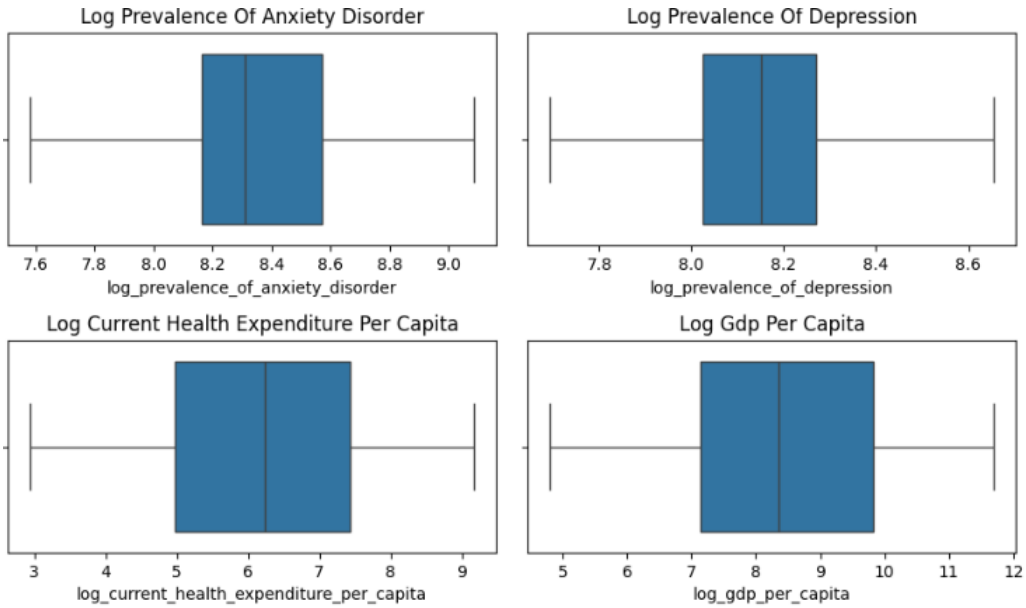
We will use boxplots to detect and handle outliers in the dataset. Detecting and addressing outliers is a crucial step in data cleaning, which helps improve the accuracy and reliability of subsequent analyses.

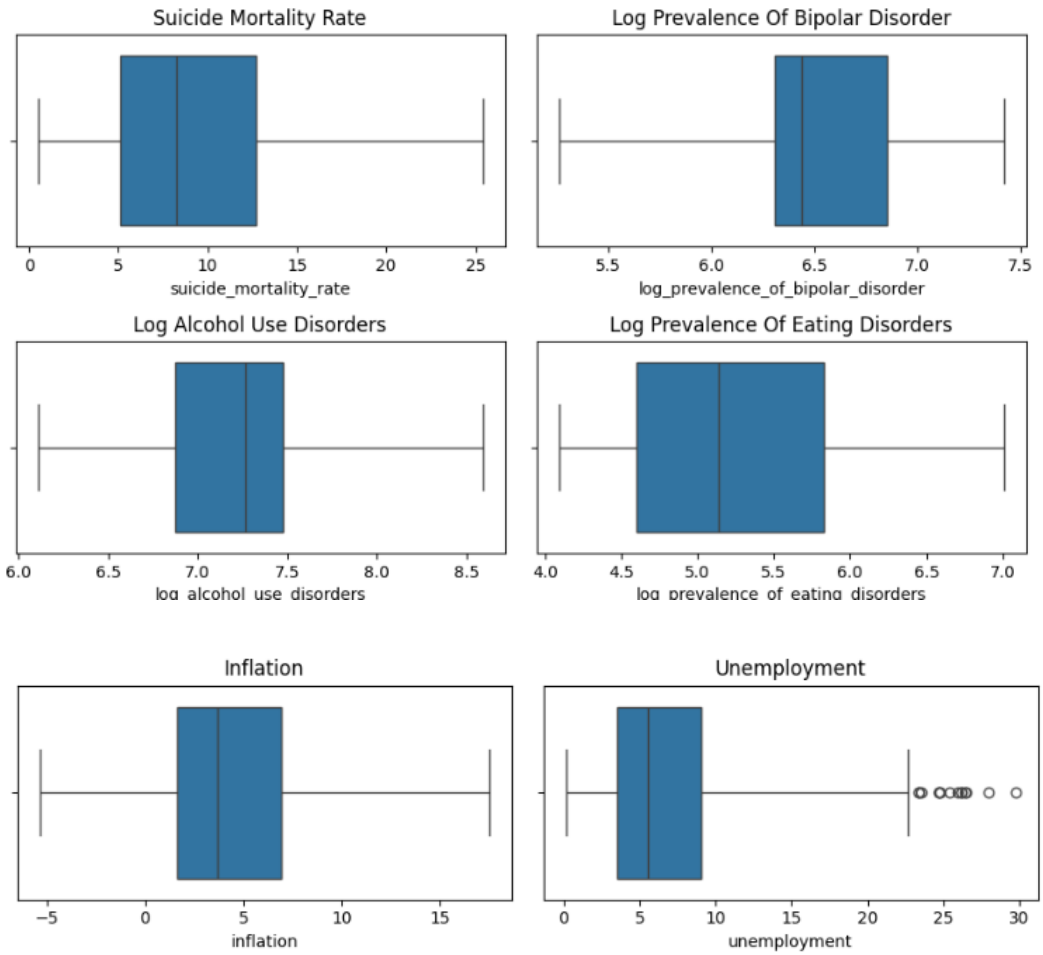




After using boxplots to detect outliers, the results reveal outliers in the *suicide_mortality_rate*, *inflation*, and *unemployment* variables. The outliers in *suicide_mortality_rate* and *inflation* could skew the analysis and affect the reliability of the model. Therefore, we will remove these outliers using the IQR (Interquartile Range) method to ensure the accuracy of our results and improve the robustness of the analysis.

However, for the *unemployment* variable, the outliers do not significantly affect the overall distribution and may reflect real economic conditions in certain regions. Therefore, we will retain these outliers to preserve important information about unemployment fluctuations in specific areas.





After removing the outliers, the boxplots display the distribution of the variables without any significant outlier points. The results show that most of the variables have an even distribution, with no noticeable outliers, except for the *unemployment*, where some outliers are still present.

```
<class 'pandas.core.frame.DataFrame'>
Index: 588 entries, 3 to 899
Data columns (total 14 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   country_name                          588 non-null    object
 1   country_code                          588 non-null    object
 2   year                                  588 non-null    int64
 3   suicide_mortality_rate                588 non-null    float64
 4   log_prevalence_of_bipolar_disorder    588 non-null    float64
 5   log_prevalence_of_anxiety_disorder    588 non-null    float64
 6   log_prevalence_of_depression          588 non-null    float64
 7   log_alcohol_use_disorders             588 non-null    float64
 8   log_prevalence_of_eating_disorders    588 non-null    float64
 9   continent                             588 non-null    object
 10  log_current_health_expenditure_per_capita  588 non-null  float64
 11  log_gdp_per_capita                    588 non-null    float64
 12  inflation                             588 non-null    float64
 13  unemployment                          588 non-null    float64
dtypes: float64(10), int64(1), object(3)
memory usage: 68.9+ KB
```

	year	suicide_mortality_rate	log_prevalence_of_bipolar_disorder	log_prevalence_of_anxiety_disorder	log_prevalence_of_depression	log_alcohol_use_disorders
count	588.000000	588.000000	588.000000	588.000000	588.000000	588.000000
mean	2008.578231	9.456803	6.488233	8.360441	8.139512	7.192856
std	5.619903	5.585093	0.406574	0.285906	0.185579	0.457026
min	2000.000000	0.500000	5.258540	7.582434	7.693523	6.109025
25%	2004.000000	5.075000	6.303451	8.164478	8.023982	6.873360
50%	2008.000000	8.250000	6.439094	8.309910	8.151592	7.269026
75%	2012.000000	12.700000	6.852073	8.571461	8.270801	7.473190
max	2016.000000	25.400000	7.419474	9.086528	8.655132	8.590631

log_prevalence_of_eating_disorders	log_current_health_expenditure_per_capita	log_gdp_per_capita	inflation	unemployment
588.000000	588.000000	588.000000	588.000000	588.000000
5.205924	6.213467	8.438998	4.691502	7.103080
0.678614	1.444915	1.598133	4.120504	5.320859
4.092023	2.944439	4.806225	-5.355400	0.150000
4.597677	4.978605	7.138408	1.665195	3.537250
5.134231	6.236404	8.349048	3.682269	5.534000
5.825012	7.438060	9.815937	6.936934	9.074000
7.009416	9.169507	11.698759	17.489449	29.770000

The analysis of the dataset reveals significant variation in the factors affecting suicide mortality rates across countries. The *suicide_mortality_rate* has an average value of 11.01, with a standard deviation of 8.96, and values ranging from 0.5 to 87.0. This shows substantial variation in suicide rates, reflecting differences across countries. The *log_prevalence_of_bipolar_disorder* also exhibits significant variation, with values ranging from 5.26 to 12.70, and a mean of 9.46.

The *log_prevalence_of_anxiety_disorder* has an average of 8.36, with a range from 7.58 to 9.09, and a standard deviation of 0.29, indicating relatively consistent patterns in mental health conditions related to anxiety. Similarly, *log_prevalence_of_depression* and

prevalence_of_eating_disorders	current_health_expenditure_per_capita	gdp_per_capita	inflation	unemployment
666.000000	666.000000	666.000000	666.000000	666.000000
219.626924	1107.799287	12481.886221	7.092488	7.378351
165.843149	1460.973061	18910.796118	16.056752	5.343756
59.860895	19.000000	122.269203	-5.355400	0.150000
99.085360	139.568929	1121.035085	2.002124	3.720000
156.225493	478.700905	3998.474565	4.064958	5.923500
291.954483	1470.479015	14924.741200	8.327062	9.929500
1107.007406	9599.891046	120422.137934	324.996872	29.770000

log_alcohol_use_disorders both show high means (8.14 and 7.19, respectively) and notable variability in their distribution, indicating that these conditions are prevalent in many countries, though their severity still varies.

Economic factors such as *log_current_health_expenditure_per_capita* and *log_gdp_per_capita* also demonstrate considerable variation. While *log_current_health_expenditure_per_capita* has a mean of 6.21 with a range from 2.94 to 9.17, *log_gdp_per_capita* has an average of 8.44 and ranges from 4.81 to 11.70, reflecting stark economic disparities across nations.

Finally, variables like *inflation* and *unemployment* show notable variation, with *inflation* ranging from -5.35 to 324.99, and a mean of 7.09, suggesting substantial economic instability in certain countries. *Unemployment* also varies widely, from 0.15 to 29.77, with an average of 7.37, highlighting significant differences in employment levels across the countries in the sample.

This detailed statistical breakdown emphasizes the wide disparities in both mental health conditions and economic factors, providing valuable insights into the potential drivers of suicide mortality across regions. These findings will inform future analyses and policy recommendations.

4. OLS modelling

$$\begin{aligned}\textbf{suicide\_mortality\_rate} = \beta0 &+ \beta1 * \text{log\_prevalence\_of\_bipolar\_disorder} \\ &+ \beta2 * \text{log\_prevalence\_of\_anxiety\_disorder} \\ &+ \beta3 * \text{log\_prevalence\_of\_depression} \\ &+ \beta4 * \text{log\_alcohol\_use\_disorders} \\ &+ \beta5 * \text{log\_prevalence\_of\_eating\_disorders} \\ &+ \beta6 * \text{log\_current\_health\_expenditure\_per\_capita} \\ &+ \beta7 * \text{log\_gdp\_per\_capita} \\ &+ \beta8 * \text{inflation} + \beta9 * \text{unemployment} + \varepsilon\end{aligned}$$

a. MODEL 1 - Model in case of full variables

Model summary and coefficient table interpretation

In this initial model, all the variables introduced in the previous section, including nine independent variables and one dependent variable, are included to build the regression

model. After running the model, the following output was obtained from the OLS regression.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     suicide_mortality_rate   R-squared:                       0.396
Model:                                OLS   Adj. R-squared:                  0.387
Method:                     Least Squares   F-statistic:                     42.09
Date:                    Sun, 13 Apr 2025   Prob (F-statistic):           7.81e-58
Time:                            16:24:47   Log-Likelihood:                -1697.1
No. Observations:                     588   AIC:                             3414.
Df Residuals:                         578   BIC:                             3458.
Df Model:                               9
Covariance Type:                nonrobust
=============================================================================================================
                                                coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
const                                       -51.6153     10.540     -4.897      0.000     -72.318     -30.913
log_prevalence_of_bipolar_disorder           -4.2680      0.892     -4.785      0.000      -6.020      -2.516
log_prevalence_of_anxiety_disorder           -5.8411      0.879     -6.646      0.000      -7.567      -4.115
log_prevalence_of_depression                 10.1844      1.061      9.603      0.000       8.101      12.267
log_alcohol_use_disorders                     4.9359      0.428     11.522      0.000       4.095       5.777
log_prevalence_of_eating_disorders            1.5204      1.123      1.354      0.176      -0.685       3.726
log_current_health_expenditure_per_capita     0.5409      0.499      1.083      0.279      -0.440       1.521
log_gdp_per_capita                            0.9337      0.473      1.974      0.049       0.005       1.863
inflation                                    -0.0468      0.047     -0.988      0.324      -0.140       0.046
unemployment                                  0.0371      0.036      1.034      0.302      -0.033       0.108
==============================================================================
Omnibus:                       49.330   Durbin-Watson:                   0.660
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               59.814
Skew:                           0.734   Prob(JB):                     1.03e-13
Kurtosis:                       3.534   Cond. No.                     1.24e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.24e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model 1 explains approximately 39.6% of the variation in suicide mortality rate, as indicated by the R-squared value of 0.396. The adjusted R-squared is slightly lower at 0.387, accounting for the number of predictors in the model. The F-statistic of 42.09, along with an extremely low p-value (7.81e10-58), indicates that the overall model is statistically significant. Additionally, the AIC and the BIC are 3414 and 3458, respectively.

Table 1 below summarizes the estimated coefficients, corresponding p-values, and statistical significance of each independent variable included in model 1:

Variable	Coefficient	p-value	Significance
Intercept (const)	-51.6153	0.000	-
log_prevalence_of_bipolar_disorder	-4.2680	0.000	Significant
log_prevalence_of_anxiety_disorder	-5.8411	0.000	Significant
log_prevalence_of_depression	10.1844	0.000	Significant
log_alcohol_use_disorders	4.9359	0.000	Significant
log_prevalence_of_eating_disorder	1.5204	0.176	Not Significant
log_current_health_expenditure_per_capital	0.5409	0.279	Not Significant
log_gpd_per_capital	0.9337	0.049	Significant
inflation	-0.0468	0.324	Not Significant
unemployment	0.0371	0.302	Not Significant

table 1. OLS regression results of model 1

Ramsey's test for linear model

The Ramsey RESET test for model 1 yields an F-statistic of 0.7282 with a p-value of 0.3938. With a typical significance threshold of 5% ($\alpha = 0.05$), we fail to reject the null hypothesis that the model is correctly specified. This suggests that the current linear functional form of the model is adequate, and there is no strong evidence of misspecification due to omitted nonlinear relationships. Therefore, the linear specification appears appropriate for the data at hand.

t-test for zero mean

The t-statistic for the test of zero mean of the residuals in model 1 is 1.7065e-13, with a p-value of 0.999. Using a significance level of $\alpha = 5\%$, we observe that the p-value is significantly greater than 0.05. Therefore, we fail to reject the null hypothesis, meaning there is no statistical evidence suggesting that the mean of the residuals differs from zero. This confirms that the residuals in model 1 are centered around zero, satisfying the assumption of zero mean for the errors in OLS regression.

VIF to check multicollinear

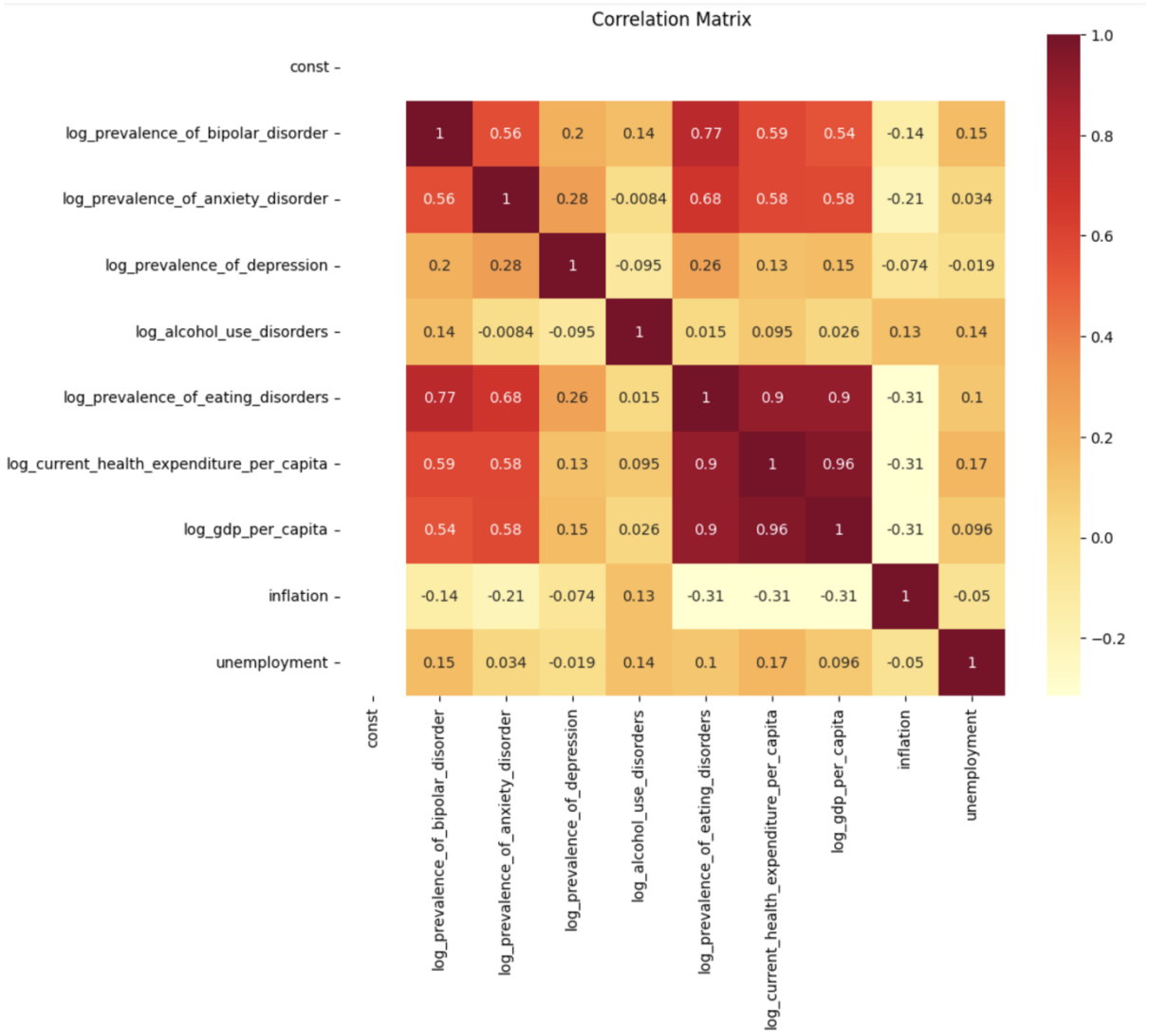
Variable	VIF
const	3413.808579
log_prevalence_of_bipolar_disorder	4.033381
log_prevalence_of_anxiety_disorder	1.936941
log_prevalence_of_depression	1.188261
log_alcohol_use_disorders	1.175772
log_prevalence_of_eating_disorder	17.818187
log_current_health_expenditure_per_capital	15.964839
log_gpd_per_capital	17.521044
inflation	1.167006
unemployment	1.119337

table 2. VIF results of model 1

In this model, most of the variables show low VIF values, indicating they do not pose multicollinearity concerns. However, there are three variables standing out, which include *log_prevalence_of_eating_disorders* (VIF = 17.82), *log_current_health_expenditure_per_capita* (VIF = 15.96) and *log_gdp_per_capita* (VIF = 17.52).

These three display severe multicollinearity, which may bias coefficient estimates and inflate standard errors. In subsequent models, we might consider removing these variables to address this issue.

To further investigate the cause of high multicollinearity, we construct a correlation matrix among all explanatory variables. A high correlation (typically above 0.8 or below -0.8) between two variables suggests they may be capturing the same underlying phenomenon. By visualizing the correlation matrix, we aim to identify which of the highly collinear variables such as *log_prevalence_of_eating_disorders*, *log_current_health_expenditure_per_capita*, and *log_gdp_per_capita* are closely correlated with each other. This analysis allows us to assess the degree of multicollinearity among these predictors. Based on the results, we can make informed decisions about whether to exclude one variable from each highly correlated pair.



Upon examining the correlation matrix, it becomes evident that *log_gdp_per_capita* is highly correlated with other key predictor variables, particularly *log_current_health_expenditure_per_capita* and *log_prevalence_of_eating_disorders*, with their correlation being 0.96 and 0.90, respectively. This strong correlation indicates that *log_gdp_per_capita* shares a substantial amount of information with these two variables. Including all of them in the model may introduce severe multicollinearity.

Moreover, the primary focus of our research is analyzing the socio-medical factors affecting suicide rates. In this context, *log_gdp_per_capita* is not a direct explanatory variable for suicide rates, whereas both *log_current_health_expenditure_per_capita* and *log_prevalence_of_eating_disorders* directly align with the chosen topic. These variables

provide more relevant insight into mental health and healthcare systems, which are central to the study.

For both statistical and thematic reasons, we therefore decide to remove *log_gdp_per_capita* from the model in order to reduce multicollinearity and better reflect the core objectives of our analysis.

b. MODEL 2 - Remove GDP per capita

Model summary and coefficient table interpretation

After addressing the issue of multicollinearity by removing the variable *log_gdp_per_capita*, we re-estimated the regression model. This revised model is referred to as model 2. Below are the OLS regression results for model 2:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     suicide_mortality_rate   R-squared:                       0.392
Model:                                OLS   Adj. R-squared:                  0.383
Method:                     Least Squares   F-statistic:                     46.64
Date:                    Sun, 13 Apr 2025   Prob (F-statistic):           7.44e-58
Time:                            16:24:49   Log-Likelihood:                -1699.0
No. Observations:                     588   AIC:                             3416.
Df Residuals:                         579   BIC:                             3455.
Df Model:                               8
Covariance Type:                nonrobust
=============================================================================================================
                                                coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------------------
const                                       -46.8013     10.280     -4.553      0.000     -66.992     -26.610
log_prevalence_of_bipolar_disorder           -4.9595      0.822     -6.031      0.000      -6.575      -3.344
log_prevalence_of_anxiety_disorder           -5.9346      0.880     -6.745      0.000      -7.663      -4.207
log_prevalence_of_depression                 10.0788      1.062      9.492      0.000       7.993      12.164
log_alcohol_use_disorders                     4.8824      0.429     11.392      0.000       4.041       5.724
log_prevalence_of_eating_disorders            2.5347      1.001      2.532      0.012       0.568       4.501
log_current_health_expenditure_per_capita     1.2402      0.353      3.516      0.000       0.547       1.933
inflation                                    -0.0413      0.047     -0.872      0.383      -0.134       0.052
unemployment                                  0.0268      0.036      0.752      0.453      -0.043       0.097
==============================================================================
Omnibus:                       50.846   Durbin-Watson:                   0.661
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               62.039
Skew:                           0.749   Prob(JB):                     3.38e-14
Kurtosis:                       3.539   Cond. No.                     1.11e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.11e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model 2 demonstrates a reasonable fit, explaining approximately 39.2% of the variation in suicide mortality rates, as indicated by the R-squared value. The adjusted R-squared is slightly lower at 0.383, accounting for the number of predictors, and still supports the model's explanatory strength. The F-statistic of 46.64, with a p-value less than 0.001, confirms that the overall model is statistically significant. Additionally, the AIC

(3416) and BIC (3455) values are slightly lower than those in model 1, suggesting a modest improvement in model parsimony and overall fit.

Table 3 below summarizes the estimated coefficients, corresponding p-values, and statistical significance of each independent variable included in model 2:

Variable	Coefficient	p-value	Significance
Intercept (const)	-46.8013	0.000	-
log_prevalence_of_bipolar_disorder	-4.9595	0.000	Significant
log_prevalence_of_anxiety_disorder	-5.9346	0.000	Significant
log_prevalence_of_depression	10.0788	0.000	Significant
log_alcohol_use_disorders	4.8824	0.000	Significant
log_prevalence_of_eating_disorder	2.5347	0.012	Significant
log_current_health_expenditure_per_capital	1.2402	0.000	Significant
inflation	-0.0413	0.383	Not Significant
unemployment	0.0268	0.453	Not Significant

table 3. OLS regression results of model 2

Model 2 witnesses a slight decline in R-squared, decreasing from 0.400 to 0.392, but offers several advantages. It achieves lower AIC and BIC values, indicating a better balance between model fit and complexity. Additionally, the lower condition number, from 1240 in model 1 to 1110 in this model, suggests improved multicollinearity, and the coefficient estimates are more stable and reliable as a result.

Ramsey's test for linear model

The Ramsey RESET test for model 2 returns an F-statistic of 1.1137 with a p-value of 0.2917. Similar to model 1, the p-value is above 0.05, indicating that there is no evidence of model misspecification. Thus, the functional form of model 2 appears to be correctly specified and does not violate the assumption of linearity.

t-test for zero mean

The t-statistic for the test of zero mean of the residuals in model 2 is 2.0519e-13, with a p-value of 0.999. Like model 1, the p-value is much greater than 0.05, meaning that we fail to reject the null hypothesis that the residuals have a mean of zero. Thus, the residuals in model 2 also meet the assumption of having a zero mean, which is a fundamental requirement in OLS regression.

VIF to check multicollinear

Variable	VIF
const	3231.134275
log_prevalence_of_bipolar_disorder	3.411409
log_prevalence_of_anxiety_disorder	1.931323
log_prevalence_of_depression	1.185238
log_alcohol_use_disorders	1.171068
log_prevalence_of_eating_disorder	14.089693
log_current_health_expenditure_per_capital	7.930084
inflation	1.163050
unemployment	1.095427

table 4. VIF results of model 2

From the VIF results of the variables in model 2, the VIF of *log_prevalence_of_eating_disorders* is 14.09, which exceeds the commonly accepted threshold of 10, indicating severe multicollinearity.

Although it was statistically significant in model 2, the multicollinearity may distort its estimated effect. To improve model stability and interpretability, we choose to remove this variable in the next model.

c. MODEL 3 - Remove Prevalence of eating disorders

Model summary and coefficient table interpretation

In Model 3, we excluded the variable *log_prevalence_of_eating_disorders* due to its high multicollinearity (VIF = 14.09 in model 2). The updated regression includes 7 predictors.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     suicide_mortality_rate   R-squared:                       0.385
Model:                                OLS   Adj. R-squared:                  0.378
Method:                     Least Squares   F-statistic:                     51.90
Date:                    Sun, 13 Apr 2025   Prob (F-statistic):           2.28e-57
Time:                            16:24:51   Log-Likelihood:                -1702.3
No. Observations:                     588   AIC:                             3421.
Df Residuals:                         580   BIC:                             3456.
Df Model:                               7
Covariance Type:                nonrobust
=========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                   -55.1756      9.779     -5.642      0.000     -74.382     -35.970
log_prevalence_of_bipolar_disorder       -3.5215      0.597     -5.895      0.000      -4.695      -2.348
log_prevalence_of_anxiety_disorder       -5.4350      0.861     -6.309      0.000      -7.127      -3.743
log_prevalence_of_depression             10.7547      1.033     10.416      0.000       8.727      12.783
log_alcohol_use_disorders                 4.5938      0.415     11.067      0.000       3.779       5.409
log_current_health_expenditure_per_capita 2.0155      0.176     11.459      0.000       1.670       2.361
inflation                                -0.0551      0.047     -1.166      0.244      -0.148       0.038
unemployment                              0.0104      0.035      0.294      0.769      -0.059       0.079
==============================================================================
Omnibus:                       51.217   Durbin-Watson:                   0.684
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               62.593
Skew:                           0.752   Prob(JB):                     2.56e-14
Kurtosis:                       3.542   Cond. No.                     1.01e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.01e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Model 3 shows a slight decrease in explanatory power compared to model 2. The R-squared value dropped from 0.392 in model 2 to 0.385 in model 3, indicating a marginal reduction in the proportion of variance explained. Similarly, the adjusted R-squared declined slightly from 0.383 to 0.378. However, these decreases are minimal and may be considered acceptable given the benefit of reducing multicollinearity. The AIC increased slightly from 3416 to 3421, and the BIC went from 3455 to 3456. Despite the minor increase in AIC and BIC, the F-statistic of model 3 improved to 51.90, suggesting stronger overall model significance. Overall, model 3 maintains good predictive power while offering improved stability and interpretability after the removal of a highly collinear variable.

Table 5 below summarizes the estimated coefficients, corresponding p-values, and statistical significance of each independent variable included in model 3:

Variable	Coefficient	p-value	Significance
Intercept (const)	-55.1756	0.000	-
log_prevalence_of_bipolar_disorder	-3.5315	0.000	Significant
log_prevalence_of_anxiety_disorder	-5.4350	0.000	Significant
log_prevalence_of_depression	10.7547	0.000	Significant
log_alcohol_use_disorders	4.5938	0.000	Significant
log_current_health_expenditure_per_capital	2.0155	0.000	Significant
inflation	-0.0551	0.244	Not Significant
unemployment	0.0104	0.769	Not Significant

table 5. OLS regression results of model 3

Model 2 had a slightly better AIC and R-squared values, indicating a marginally better fit; however, it suffered from severe multicollinearity due to the variable *log_prevalence_of_eating_disorders*. In contrast, model 3 resolves this issue by removing the problematic variable, resulting in a more stable and interpretable model. Despite a minor decline in model fit metrics, model 3 retains the significance and consistency of its key predictors, with effect directions remaining aligned with theoretical expectations. This reinforces the robustness and reliability of the model's core findings.

Ramsey's test for linear model

The Ramsey RESET test produced an F-statistic of 0.0152 with a p-value of 0.9018. Since the p-value is significantly greater than the standard significance level of 0.05, we fail to reject the null hypothesis that the model is correctly specified. This suggests there is

no evidence of omitted non-linear relationships among the predictors, and the linear functional form of the model appears appropriate.

Compared to previous models, this model demonstrates the strongest evidence of correct functional form, as the p-value is the highest among all three models. This reinforces that the linear specification is appropriate and that the model does not suffer from omitted variable bias due to non-linear relationships.

t-test for zero mean

The t-test for checking whether the residuals have a mean of zero resulted in a t-statistic of 2.052e-13 and a p-value of 0.999. Since the p-value is far above the 0.05 threshold, we do not reject the null hypothesis. This confirms that the residuals are centered around zero, satisfying a core assumption of the OLS regression.

VIF to check multicollinear

Variable	VIF
const	2896.598764
log_prevalence_of_bipolar_disorder	1.783510
log_prevalence_of_anxiety_disorder	1.834156
log_prevalence_of_depression	1.110304
log_alcohol_use_disorders	1.088232
log_current_health_expenditure_per_capital	1.953321
inflation	1.147681
unemployment	1.059134

table 6. VIF results of model 3

This model shows a substantial improvement in multicollinearity over both previous models. All VIF values are now below 2, with the highest being 1.953 for *log_current_health_expenditure_per_capita*. In comparison, model 1 and model 2 included variables such as *log_prevalence_of_eating_disorders* and *log_gdp_per_capita* that had VIFs exceeding 10, suggesting severe multicollinearity. By removing these problematic variables, the current model (model 3) achieves the lowest multicollinearity, making it more stable, interpretable, and statistically reliable.

BP-test for Heteroskedasticity

The Breusch-Pagan test for heteroskedasticity yields a Lagrange Multiplier (LM) statistic of 34.713 and an associated p-value of 1.27e-05. Additionally, the F-statistic is 5.198 with a corresponding p-value of 9.23e-06. Since both p-values are significantly less

than the common significance level of 0.05, we reject the null hypothesis of homoskedasticity. This indicates that the model exhibits heteroskedasticity, meaning the variance of the residuals is not constant across observations. This violation of the OLS assumption can lead to inefficient coefficient estimates and biased standard errors, which may affect statistical inference.

Durbin-Watson test for autocorrelation

The Durbin-Watson (DW) statistics for all independent variables in the model range from approximately 0.55 to 0.64, well below the threshold of 2, which signals the presence of positive autocorrelation in the residuals. Although the analysis treats the dataset as cross-sectional data, it is originally derived from a panel dataset that includes time-series components for each country. This temporal structure in the source data may carry over into the residuals, introducing serial correlation even in a cross-sectional framework. As a result, the DW test detects patterns typically associated with time-series data.

Therefore, the presence of autocorrelation in this context can be reasonably overlooked, as it reflects characteristics of the data's origin rather than a violation relevant to cross-sectional assumptions.

5. GLS regression and residual diagnostics and influential observations in the model

a. GLS regression

Because of the presence of heteroskedasticity and slight autocorrelation in preliminary OLS estimates, we employed a Generalized Least Squares (GLS) model to improve estimation efficiency and inference validity. Heteroskedasticity - evidenced by residual plots and significant Breusch-Pagan test - was an indication that the variance of the error term was not constant across observations. Further, a Durbin-Watson statistic well below 2 provided evidence of the existence of residual autocorrelation. Such misspecifications taint the reliability of OLS standard errors and hypothesis testing, so it becomes reasonable to utilize GLS to account for heteroskedasticity and correlation in the error structure.

For constructing the GLS model, we initially conducted an OLS regression and derived residuals. The residuals were then squared and modeled as a log-linear function of the fitted values and used to estimate variances specific to each observation. These variances were used to form a diagonal covariance matrix that was included in the GLS estimation, effectively weighting each observation in inverse proportion to its error

variance, which is predicted. This allows the model to down-weight the noisy or high-variance observations and lead to more efficient and unbiased estimates of coefficients. After estimating the GLS model with a variance-covariance matrix (σ) derived from the heteroskedasticity pattern of the residuals conditional on the fitted values, we apply the Newey-West HAC (Heteroskedasticity and Autocorrelation Consistent) estimator with a maximum lag length of 1 to adjust the standard errors. The application of HAC standard errors ensures the validity of hypothesis testing and p-values, even when the classical assumptions of independence and Homoskedasticity are violated. As a result, the statistical inference regarding the significance of explanatory variables becomes more robust and better reflects the actual characteristics of the time series data, which follow a four-year interval structure.

The model achieved an R-squared of 0.344, which means that approximately one-third of the cross-national differences in suicide mortality rates are explained by the predictors used. The AIC value of the above model was found to be 3391. Among the set of explanatory variables, the log prevalence of depression emerged as the strongest and statistically significant positive predictor, followed by alcohol use disorders. The results support existing empirical studies on mental health risk factors for suicide.

GLS + HAC:

GLS Regression Results

Dep. Variable:	suicide_mortality_rate	R-squared:	0.344
Model:	GLS	Adj. R-squared:	0.336
Method:	Least Squares	F-statistic:	29.46
Date:	Sun, 20 Apr 2025	Prob (F-statistic):	7.63e-35
Time:	15:44:37	Log-Likelihood:	-1687.6
No. Observations:	588	AIC:	3391.
Df Residuals:	580	BIC:	3426.
Df Model:	7		
Covariance Type:	HAC		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-65.8462	13.985	-4.708	0.000	-93.313	-38.380
log_prevalence_of_bipolar_disorder	-3.7441	0.796	-4.703	0.000	-5.308	-2.180
log_prevalence_of_anxiety_disorder	-4.5182	1.266	-3.569	0.000	-7.005	-2.032
log_prevalence_of_depression	11.3976	1.113	10.242	0.000	9.212	13.583
log_alcohol_use_disorders	4.7672	0.500	9.543	0.000	3.786	5.748
log_current_health_expenditure_per_capita	1.7376	0.257	6.772	0.000	1.234	2.241
inflation	-0.0859	0.047	-1.832	0.067	-0.178	0.006
unemployment	-0.0182	0.036	-0.506	0.613	-0.089	0.053

Omnibus:	65.530	Durbin-Watson:	0.673
Prob(Omnibus):	0.000	Jarque-Bera (JB):	85.635
Skew:	0.865	Prob(JB):	2.54e-19
Kurtosis:	3.710	Cond. No.	1.07e+03

In addition, the per capita health spending had a small but statistically significant positive coefficient, a somewhat surprising finding that could reflect the existence of confounding variables or reverse causality associated with health investment. Remaining variables, such as bipolar disorder prevalence and anxiety, showed smaller or statistically non-significant effects. It is interesting that the inflation rate came close to significance, while unemployment failed to demonstrate a strong association with suicide mortality in this baseline model.

Although the GLS model improved OLS estimation by correcting for heteroskedasticity, residual diagnostics continued to exhibit some issues. Residuals continued to be non-normally distributed, as indicated by large Jarque-Bera and Omnibus test statistics, and a Durbin-Watson statistic value of 0.673, which showed that autocorrelation, albeit smaller, was still present. These findings motivated additional investigation of outlier and influence diagnostics.

b. Residual diagnostics and influential observations in the GLS model

We examined the standardized residuals (Z-scores) derived from the GLS model residuals. Observations with absolute Z-scores greater than 3 were flagged as statistical outliers. Simultaneously, a GLS-specific influence score was computed using a leave-one-out (LOO) approach. For each observation, the model was re-estimated without that data point, and the Euclidean norm of the change in the estimated coefficients was recorded.

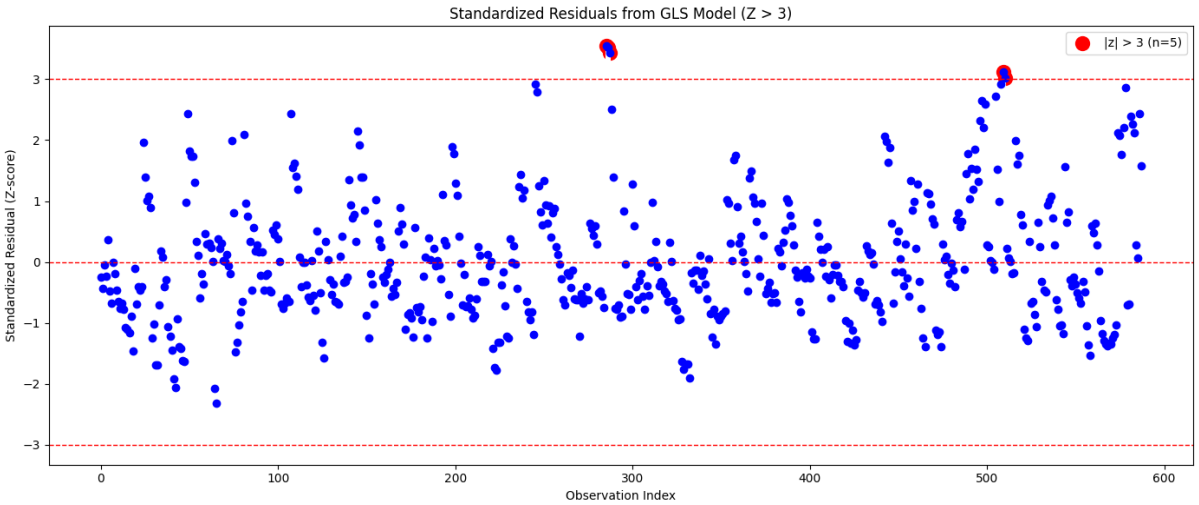
Outlier Detection by Z-score

```
from scipy.stats import zscore
residuals_gls = gls_results.resid
standardized_gls_resid = zscore(residuals_gls)
df_log["gls_outlier_z>3"] = np.abs(standardized_gls_resid) > 3
```

Influence Score by Leave-One-Out GLS

```
original_params = gls_results.params.values
influential_scores = []

# Get the actual indices of df
df_indices = df_log.index.to_list()

# Loop through the valid indices to avoid out-of-bounds errors
for idx in tqdm(df_indices):
    try:
        # Remove the row based on the index
        X_loo = X.drop(index=idx)
        y_loo = y.drop(index=idx)
        sigma_loo = np.delete(np.delete(sigma_matrix, df_log.index.get_loc(idx), axis=0), df_log.index.get_loc(idx), axis=1)

        # Fit GLS without that observation
        model_loo = sm.GLS(y_loo, X_loo, sigma=sigma_loo).fit()

        # Measure the deviation between the new and original coefficients
        score = np.linalg.norm(model_loo.params.values - original_params)
        influential_scores.append(score)
    except:
        influential_scores.append(np.nan)  # fallback in case of error

# Add the influence_score column to df_log
df_log["gls_influence_score"] = influential_scores
```

A scatterplot of standardized residuals revealed several observations exceeding the ±3 threshold. Japan, particularly in the year 2000, showed a suicide mortality rate of 23.9 per 100,000 people with a Z-score well above 3. This suggests an extreme deviation from model predictions and flags the data point as an outlier.



	country_name	year	suicide_mortality_rate	gls_outlier_z>3	gls_influence_score
410	Japan	2000	23.9	True	3.679987
411	Japan	2004	24.1	True	3.357421
565	Myanmar	2000	4.7	False	3.172833
412	Japan	2008	24.4	True	2.945951
155	Central African Republic	2000	19.0	False	2.568986
566	Myanmar	2004	4.2	False	1.873788
413	Japan	2012	21.6	False	1.777083
157	Central African Republic	2008	15.1	False	1.609349
111	Brazil	2004	4.4	False	1.493545
156	Central African Republic	2004	15.7	False	1.483687

A few instances of high Z-scores and influence statistics such as Japan in 2000 represent sociocultural and contextual circumstances that lie outside the parameters of the mental health variables covered in this analysis. The historically significant suicide rate in Japan is directly related to cultural beliefs about honor, duty, and failure. While there is no longer the extreme practice of seppuku, vestiges of these cultural beliefs persist, affecting

societal perceptions of self-injury. Furthermore, Japan is characterized by high levels of work-related stress and a longstanding stigma surrounding mental health care factors that exacerbate the effect of psychiatric illness but are poorly captured by standard regression controls. Consequently, Japan stands out as a cultural outlier, in which the causes of suicide can be played out indirectly through social institutions instead of through clinically defined indicators.

In contrast, Myanmar represents a case typified by institutional weakness. Since the early 2000s, and more so after waves of political unrest since the 2010s, Myanmar has struggled with civil wars, public health collapses, and untrustworthy health statistics. In these circumstances, suicide rates tend to mirror a combination of exposure to trauma, economic volatility, and limited access to mental health care. Consequently, the Myanmar data point exhibits a type of omitted variable bias, with political decline acting as an unmeasured proxy for psychiatric distress.

5 observations were identified as outliers or influential cases, which originated primarily from nations with cultural distinctiveness (Japan), political instability (Myanmar). These cases do not represent random aberrations; instead, they reflect the model's existing limitations in representing cross-national variability in suicide risk factors sufficiently. Eliminating these observations will enhance estimation stability, while also focusing more on the topic of the paper.

c. Model fit before vs. after removing anomalies

To assess model robustness, we compared GLS regression results before and after removing extreme outliers and influential observations.

In the original model, R-squared was 0.344, AIC was 3391, after removing problematic observations (those with $|z| > 3$ and high influence scores), the cleaned model showed better fit:

```
GLS + HAC:
                            GLS Regression Results
==============================================================================
Dep. Variable:     suicide_mortality_rate   R-squared:                       0.363
Model:                                GLS   Adj. R-squared:                  0.356
Method:                     Least Squares   F-statistic:                     39.26
Date:                    Sun, 20 Apr 2025   Prob (F-statistic):           4.44e-45
Time:                            15:45:10   Log-Likelihood:                -1642.0
No. Observations:                     582   AIC:                             3300.
Df Residuals:                         574   BIC:                             3335.
Df Model:                               7
Covariance Type:                      HAC
=========================================================================================================
                                            coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------------
const                                   -77.5773     11.765     -6.594      0.000    -100.686     -54.469
log_prevalence_of_bipolar_disorder       -3.9368      0.777     -5.067      0.000      -5.463      -2.411
log_prevalence_of_anxiety_disorder       -3.6492      1.047     -3.486      0.001      -5.706      -1.593
log_prevalence_of_depression             11.8557      0.985     12.031      0.000       9.920      13.791
log_alcohol_use_disorders                 5.0860      0.454     11.204      0.000       4.194       5.978
log_current_health_expenditure_per_capita 1.6412      0.224      7.336      0.000       1.202       2.081
inflation                                -0.0691      0.044     -1.562      0.119      -0.156       0.018
unemployment                             -0.0044      0.035     -0.128      0.898      -0.072       0.064
==============================================================================
Omnibus:                       41.966   Durbin-Watson:                   0.706
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               49.203
Skew:                           0.701   Prob(JB):                     2.07e-11
Kurtosis:                       3.248   Cond. No.                     1.09e+03
==============================================================================
```

R-squared increased to 0.363, AIC dropped to 3300, Residuals became more normal (Omnibus: 65.5 → 41.97; JB: 85.6 → 49.2), and Durbin-Watson improved to 0.706.

Key coefficient directions and significance remained stable, especially for depression and alcohol use. However, the inflation variable became non-significant, and health expenditure slightly decreased in effect.

In short, the cleaned model fits better, is statistically more stable, and supports more trustworthy inferences about global suicide mortality.

d. Check assumption with GLS_HAC

After estimating the GLS model, it is essential to revisit key assumptions that may no longer hold. This step ensures the reliability of statistical inference and the robustness of the model under more complex error structures.

First, we apply the Ramsey RESET test to assess potential misspecification of the functional form. While the OLS version of this test typically relies on the F-test or LM-test using the residual sum of squares or the R^2 statistic, these approaches become invalid in the context of GLS due to the presence of heteroskedasticity and possible autocorrelation in the error terms. Therefore, we employ a GLS-specific version of the RESET test, which utilizes the Likelihood Ratio (LR) test. This version compares the log-likelihood of the original model with that of an expanded model that includes higher-order terms of the fitted

values. This adjustment ensures that the RESET test properly accounts for the structure of the GLS model and maintains the validity of conclusions regarding functional form misspecification.

```
GLS_LR_statistic: 3.4279297443490577
df: 2
p_value: 0.18015010261590234
conclusion: Right function
robust_standard_errors: [4.09023539e+01 2.28777274e+00 1.75283713e+00 5.62366846e+00
 2.63291125e+00 8.32942852e-01 5.78682008e-02 3.43369434e-02
 5.55845823e-02 1.82120338e-03]
```

The test result yields a p-value of approximately 0.18, which exceeds the 5% significance level. As a result, we fail to reject the null hypothesis of correct model specification. This suggests that the current GLS model is acceptable in terms of functional form.

We also examine the zero-mean residual assumption. Unlike OLS, in which residuals are guaranteed to have a mean of zero (given an intercept), GLS residuals do not inherently satisfy this property due to the presence of weights or a non-identity covariance structure. We apply a technically adjusted test to evaluate this condition and find that the zero-mean assumption is not violated, indicating no systematic bias in the residuals.

```
Mean of residuals (GLS): 0.16183716521616884
T-statistic: 0.888, p-value: 0.375
```

V. Results

The Generalized Least Squares (GLS) regression was conducted to examine the impact of various mental health-related factors on the suicide mortality rate across a sample of 582 observations. After control for Heteroskedasticity, the model demonstrates a moderate explanatory power, with an R-squared of 0.363 and an adjusted R-squared of 0.356. For a study on suicide rates, an R-squared value of this magnitude is considered reasonable, as nearly 40% of the variance in the dependent variable is explained by the independent variables. A social issue is subject to numerous influences from various fields and different conditions, so our model can be considered quite effective. The F-statistic (39.26, $p < 0.001$) confirms the joint statistical significance of the explanatory variables, indicating that the model is appropriate for explaining variations in suicide rates.

Variable	Coefficient	p-value	Significance
log_prevalence_of_bipolar_disorder	-3.9368	0.000	Significant
log_prevalence_of_anxiety_disorder	-3.6492	0.001	Significant
log_prevalence_of_depression	11.8557	0.000	Significant
log_alcohol_use_disorders	5.086	0.000	Significant
log_current_health_expenditure_per_capita	1.6412	0.000	Significant
inflation	-0.0691	0.119	Not Significant
unemployment	-0.0044	0.898	Not Significant

table 7. GLS model estimation results

The regression results reveal that the *log_prevalence_of_depression* is the most influential predictor, exhibiting a strong positive relationship with suicide mortality. Specifically, a one-unit increase in the log-transformed prevalence of depression is associated with an 11.86-unit increase in the suicide mortality rate, ceteris paribus. This means that at a 95% confidence level, considering a group of 100,000 people, with an increase of 1,000 cases, there will be an additional 11 suicides. Medical research identifies depression as one of the primary causes of suicide. According to Depression and Bipolar Suppor Alliance, depression is the cause of over two-thirds of the 30,000 reported suicides in the U.S. each year. So that, the result from the GLS model is consistent with the extensive body of literature identifying depression as a leading risk factor for suicide. Another notable outcome, the *log_prevalence_of_alcohol_use_disorders* also shows a positive and statistically significant association with suicide rates ($\beta = 5.086$, $p < 0.01$), highlighting the detrimental role of substance abuse in exacerbating suicidal behavior. The abuse of stimulants such as alcohol and beer not only affects the economy, disrupts social order, increases the rate of traffic accidents... but also diminishes public health, raises the suicide rate, and imposes invisible pressures on the global health system.

Conversely, both the *log_prevalence_of_bipolar_disorder* ($\beta = -3.937$, $p < 0.01$) and *log_prevalence_of_anxiety_disorder* ($\beta = -3.649$, $p < 0.01$) are negatively associated with the suicide mortality rate. This result is unexpected; unlike depression, the rates of bipolar disorder and anxiety disorders exhibit an inverse relationship with the suicide rate. It would be illogical to conclude that a higher prevalence of these conditions would lead to a reduction in suicides. This noteworthy finding may arise from diagnosis rates, availability of treatment, or higher levels of mental health support provided to these patient groups. The accurate awareness of the dangers of these diseases within society has enabled

countries to implement timely actions. To further elucidate this pertinent point in the model, additional in-depth studies may be necessary.

Furthermore, the *log_of_current_health_expenditure_per_capita* is found to have a positive and significant effect on suicide rates ($\beta = 1.641$, $p < 0.01$). While counterintuitive, this result could imply that countries that allocate greater resources to healthcare tend to be wealthier and possess more resources. In developed and prosperous nations, the complexities of life can increase significantly, as individuals encounter numerous pressures stemming from high societal standards, financial concerns, and achievement-related stressors. Additionally, these countries experience higher rates of internet connectivity, which can lead to challenges such as verbal aggression and a pervasive sense of insecurity in cyberspace. From a different point of view, countries with higher healthcare spending possess more comprehensive systems for recording and reporting suicide cases, or it may point to deeper socio-economic factors not captured directly by the model. Research by Wolters Kluwer (2015) also indicates that a significant proportion of individuals who attempt suicide have had recent interactions with healthcare providers. These frequent interactions may contribute to more accurate recording and recognition of suicide cases in healthcare systems with higher utilization

In contrast, the macroeconomic variables included in the model: inflation and unemployment do not exhibit statistically significant effects on suicide mortality ($p = 0.119$ and $p = 0.898$, respectively). These findings suggest that, within the context of this model, mental health-related factors play a more prominent role than economic conditions in influencing suicide outcomes. Although inflation's coefficient is negative and close to marginal significance, it remains statistically insignificant at the 5% level. Meanwhile, the unemployment rate displays an almost negligible effect, with both the coefficient and statistical significance indicating no discernible relationship. These results suggest that short-term fluctuations in economic conditions may not directly or uniformly translate into changes in suicide mortality, at least not to a statistically detectable extent in the presence of mental health-related controls.

Several diagnostic tests were conducted to assess the validity of the model assumptions. The Durbin-Watson statistic (0.706) suggests the presence of positive autocorrelation in the residuals. Moreover, the Omnibus test and Jarque-Bera test both return significant results ($p < 0.001$), indicating that the residuals deviate from normality.

These issues may affect the efficiency of the coefficient estimates and should be addressed in further research.

VI. Conclusion and recommendations

Based on data compiled over 20 years, using a simple linear regression model, we have gained a more objective and specific perspective on the relationship between factors related to healthcare quality and mental health with the issue of suicide in countries around the world. Accordingly, depression is considered one of the main causes of suicide incidents, as the increase in the rate of depression will accompany an increase in the overall suicide rate within a sample group of 100,000 people. Nations must accord greater attention to this pressing public health issue. There is a critical need for expanded medical research and advancements in therapeutic approaches. Simultaneously, it is essential to implement integrated measures aimed at fostering cultural and societal development, enhancing the spiritual well-being of the populace, reinforcing psychological education within schools, and alleviating the mounting pressures faced by young people. Another salient factor potentially influencing suicide rates is the widespread abuse of stimulants, particularly alcohol and beer. Excessive consumption of such substances exerts detrimental effects on both physical and mental health. It is, therefore, imperative that nations around the world adopt more stringent regulations on alcohol consumption. Governments must grasp the severity of this issue and have the courage to prioritize public health and safety over the relentless pursuit of economic growth metrics.

Furthermore, the study raises several critical questions for policymakers. According to the findings derived from the linear model, an increase in national healthcare expenditures appears to correlate with a rise in suicide rates. This paradoxical relationship warrants deeper investigation: does it reflect enhanced capabilities in suicide detection and reporting, as healthcare systems become more adept at identifying and documenting such cases? Or does it reveal a misallocation of resources and an overarching neglect of suicide prevention an urgent challenge in the 21st century?

Another dimension that warrants deeper investigation one that appears to be a principal contributor to suicide is the prevalence of mental disorders, including anxiety disorders and bipolar disorder, which paradoxically show an inverse correlation with suicide rates. The number of diagnosed cases is inversely proportional to the number of suicides. This presents a particularly thought-provoking inquiry: has the progression of

medical science led to higher recovery rates from these conditions, thereby mitigating their fatal consequences? There is a clear need for further research in this area. Of course, the objective is not to increase the incidence of such disorders, but rather to drive both metrics disease prevalence and suicide rates down simultaneously.

As previously mentioned, suicide is a multifaceted societal issue that, at a macro level, is shaped by a wide array of influences. It cannot be fully understood or addressed by relying solely on a limited set of variables related predominantly to economic and healthcare factors. Nonetheless, this study offers a more objective and comprehensive perspective on global suicide trends over the past two decades. It also substantiates and underscores the direct harm caused by depression and the abuse of stimulants in relation to this urgent issue. Furthermore, the findings pose critical questions functioning both as a challenge and a cautionary signal to governments regarding whether they have genuinely given due consideration to the current state of societal instability, and whether they are charting an appropriate course toward collective well-being and national happiness.

The statistical model used in this research only accounts for suicides that have been officially recorded merely the visible tip of the iceberg. In reality, countless individuals hover on the precipice between life and death, contemplating self-destruction; this hidden figure is undoubtedly far greater than the documented cases. Therefore, it is high time we devote serious attention to this issue and to the mental health of our communities, particularly among the youth. The rapid acceleration of the global economy is producing far-reaching repercussions: society is growing increasingly complex, and individuals are burdened with mounting pressures. We must profoundly reconsider this trajectory since human beings innovate to attain comfort in life, not to create comfort on the path to death.

VII. References

Healthy People 2030 (U.S. Department of Health and Human Services). (n.d.). *Reduce the suicide rate - MHMD-01.* Retrieved April 2025, from Healthy People 2030. U.S. Department of Health and Human Services: https://health.gov/healthypeople/objectives-and-data/browse-objectives/mental-health-and-mental-disorders/reduce-suicide-rate-mhmd-01

Khan, A. R. (2023). *Cultural Perspectives of Suicide in Bangladesh.* Retrieved from Springer Nature Switzerland AG: https://doi.org/10.1007/978-981-99-0289-7_4

Er, T. S., Demir, E., & Sari, E. (2023, June). *Suicide and economic uncertainty: New findings in a global setting.* Retrieved from SSM Population Health: https://doi.org/10.1016/j.ssmph.2023.101387

The Jed Foundation. (n.d.). *Mental health and suicide statistics.* Retrieved April 2025, from The Jed Foundation: https://jedfoundation.org/mental-health-and-suicide-statistics/

Centers for Disease Control and Prevention. (2024). *Mental health and suicide risk among high school students and protective factors - Youth Risk Behavior Survey, United States, 2023.* Retrieved from CDC: https://www.cdc.gov/mmwr/volumes/73/su/su7304a9.htm

Lovero, K. L., Santos, P. F., Come, A. X., Wainberg, M. L., & Oquendo, M. A. (2023). *Suicide in global mental health.* Retrieved from Springer Nature Link: https://doi.org/10.1007/s11920-023-01423-x

Akkas, F., & Corr, A. (2022, May 2). *Mental health conditions can contribute to suicide risk.* Retrieved from The Pew Charitable Trusts: https://www.pewtrusts.org/en/research-and-analysis/articles/2022/05/02/mental-health-conditions-can-contribute-to-suicide-risk

Kim, H., Jung , J., Han, K., & Jeon, H. (2025). *Risk of suicide and all-cause death in patients with mental disorders: A nationwide cohort study.* Retrieved from National Library of Medicine: https://pubmed.ncbi.nlm.nih.gov/39843548/

Ahmedani, B. K. (2015, April 15). *High rate of healthcare visits before suicide attempts.* Retrieved from Wolters Kluwer: https://www.wolterskluwer.com/en/news/high-rate-of-healthcare-visits-before-suicide-attempts

World Health Organization. (2025, March 25). *Suicide.* Retrieved from World Health Organization: https://www.who.int/news-room/fact-sheets/detail/suicide

Lee, S. (2025, March 12). *A Guide to Bayesian Information Criterion for Model Selection.* Retrieved from Number Analytics: https://www.numberanalytics.com/blog/bayesian-information-criterion-guide

Bevans, R. (2020, March 26). *Akaike Information Criterion | When & How to Use It (Example).* Retrieved from Scribbr: https://www.scribbr.com/statistics/akaike-information-criterion/

Bobbitt, Z. (2020, December 31). *The Breusch-Pagan Test: Definition & Example.* Retrieved from Statology: https://www.statology.org/breusch-pagan-test/

Bobbitt, Z. (2021, January 21). *The Durbin-Watson Test: Definition & Example.* Retrieved from Statology: https://www.statology.org/durbin-watson-test/

timeseriesreasoning. (n.d.). *A Tutorial On Generalized Least Squares Estimation Using Python And Statsmodels.* Retrieved from Statistical Modeling and Forecasting: https://timeseriesreasoning.com/contents/generalized-least-squares-tutorial/