A study of depression influencing factors with a multiple linear regression model for some Countries in the World

Kevin Ferreira - 6182425
Department of Mathematics and Statistics
Florida International University

Advisor-Dr. B. M. Golam Kibria

Date: April 13, 2023

A study of depression influencing factors with a multiple linear regression model for some Countries in the World

Abstract

This study will determine the factors that affect depression of 40 countries from various regions around the world. The regression factors in question are internet users, alcohol consumption, country GDP, average annual working hours, unemployment rate, life expectancy, median age, global health security index, tertiary school enrollment, and average temperature of the 40 countries used in this study. All of this data has been collected via websites where most of the data was extracted from Ourworldindata. Afterwards, the data has been fitted in a regression model. Although the model meets the assumptions of normality, the p-value is not statistically significant which means that this model does not support that the y is not affected by at least one regressor variable. Since the model is not viable, I have performed a series of transformations to obtain a better model which were taking the square root of y, the log of y, the inverse of y, and the inverse of square root of y. After the tests, I have decided that the inverse of y transformation produced the best model and I went to perform a stepwise regression analysis to obtain the final model. The final model supports that depression is affected by alcohol consumption and country GDP.

1. Introduction

Depression is a mental disorder that causes a person to constantly feel sad and hopeless. Depression can have a huge impact on a person's life such as negatively hindering their ability to function in everyday activities and can even lead to self-harm or suicide. There are various factors that could cause a person to have depression and so this project uses ten factors that might cause depression. Some people may develop depression due to genetics, other people may develop depression due to a variety of psychological and environmental factors. According to the World Health Organization, approximately 280 million people in the world have depression which means depression is a serious mental disorder that affects many people around the world. The aim of this study is to determine if there is a relationship between depression and the 10 regressor variables which are listed below.

Y1: Prevalence rate of depression.

X1: The percentage of people that use the internet.

X2: The amount of alcohol consumed in liters.

X3: The country's GDP is trillions of USD.

X4: The average amount of annual working hours.

X5: The unemployment rate of each country.

X6: The life expectancy of each population which also accounts for both sexes.

X7: The median age of the country's population.

X8: The global health security of each country.

X9: The tertiary school enrollment of each country.

X10: The average temperature in each country is Celsius.

These are the 40 countries that will be used in this study and are used in this order.

1. USA	11. Japan	21. Portugal	31. Morocco
2. Canada	12. Australia	22. Ireland	32. Indonesia
3. Germany	13. Mexico	23. Netherlands	33. Thailand
4. France	14. Brazil	24. New Zealand	34. Vietnam
5. UK	15. Argentina	25. South Korea	35. Pakistan
6. Spain	16. Colombia	26. South Africa	36. Bangladesh
7. Norway	17. Chile	27. Nigeria	37. Iran
8. Russia	18. Cuba	28. Ethiopia	38. Israel
9. India	19. Sweden	29. Saudi Arabia	39. Turkey
10. China	20. Finland	30. Egypt	40. Iraq

There are many other factors that may cause depression, for example, the happiness index of a country. However, we have considered available variables that we can collect in a limited period of time. A multiple linear regression model will be constructed to show a relationship between the significant regressors and depression. For more research on depression, we refer our readers to Xu et al (2017) and fitting regression model, we refer to Montgomery (2013) and Guzman and Kibria (2019) among others.

The organization of the project is as follows: Different linear regression models are fitted along with diagnostics of the models given in section 2. The multicollinearity problem is discussed in section 3. Finding the best model is outlined in section 4. This project ends up with some concluding remarks in section 5.

2. Linear Regression Model

The linear regression model and its summary were generated from R Studio. We have fitted different types of regression models in this section.

2.1 Model 1: The first model includes all 10 regression variables.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.2389300	1.8014029	0.688	0.4971	
X1	-0.0037008	0.0080466	-0.460	0.6490	
X2	0.0810320	0.0303666	2.668	0.0123	*
X3	-0.0036174	0.0024425	-1.481	0.1494	
X4	0.0002357	0.0001997	1.180	0.2475	
X5	0.0119126	0.0171092	0.696	0.4918	
X6	0.0369585	0.0297015	1.244	0.2233	
X7	-0.0028898	0.0202399	-0.143	0.8875	
X8	-0.0146333	0.0112161	-1.305	0.2023	
X9	0.0077249	0.0054566	1.416	0.1675	
X10	0.0086945	0.0126575	0.687	0.4976	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

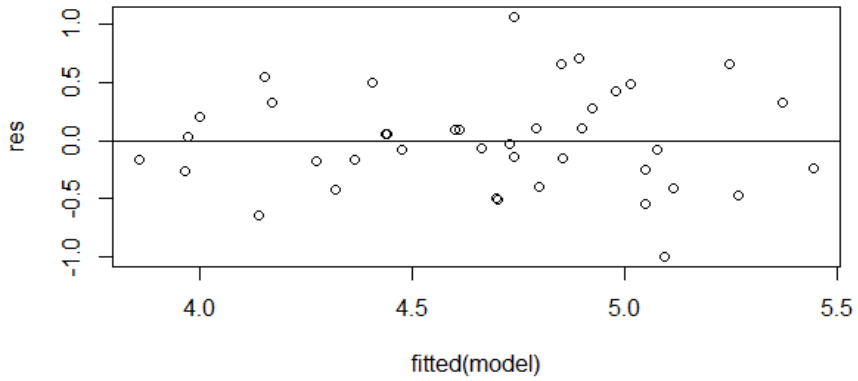
Residual standard error: 0.494 on 29 degrees of freedom
Multiple R-squared: 0.4812, Adjusted R-squared: 0.3024
F-statistic: 2.69 on 10 and 29 DF, p-value: 0.01825

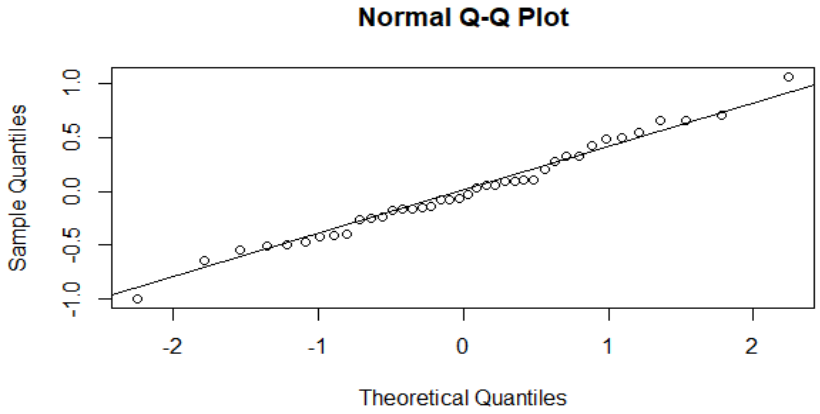
The fitted first model based on all regressors is given below:

$$\hat{y} = 1.4248549 - 0.0037008X_1 + 0.0810320X_2 - 0.0036174X_3 + 0.0002357X_4 + 0.0119126X_5 + 0.0369585X_6 - 0.0028898X_7 - 0.0146333X_8 + 0.0077249X_9 + 0.0086945X_{10}$$

From these summary statistics, we can see that the R-squared shows that 48% of the total variation is explained by the 10 regression variables. Although it is not as high as ideally it should, it still supports the model as a good fit for data. From the p-values, we can see that X_2 is only the significant regressor.

The F-value=2.69 with 10 and 29 DF and p-value=0.01825. That means overall the model is significant at 2% significance level. We can infer that at least one of the 10 regressors contributes significantly to the model.





Both the residual and qq-plot show that the distribution of the residuals is normal and the variance is constant. The residual plot is scattered, and the points stick to the line with no skew.

2.2 Model 2: Stepwise Regression of Data

Coefficients:

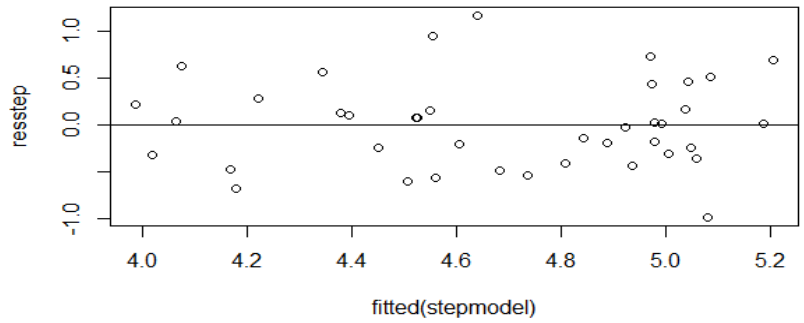
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.885842	0.191186	20.325	<2e-16	***
X2	0.049477	0.021008	2.355	0.0239	*
X9	0.007017	0.003082	2.276	0.0287	*

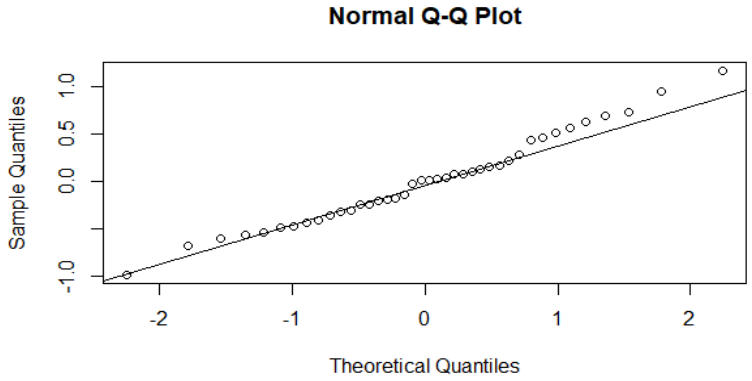
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4822 on 37 degrees of freedom
Multiple R-squared: 0.3693, Adjusted R-squared: 0.3353
F-statistic: 10.83 on 2 and 37 DF, p-value: 0.0001977

$\hat{y} = 3.885842 + 0.049477X2 + 0.007017X9$

After performing a stepwise backward regression, only x2 and x9 remain, which are alcohol consumption and life expectancy respectively. The R-squared is smaller than the data before transformation. The F-value=10.83 with 2 and 37 DF and p-value=0.0002. That means overall the model is significant at 1% significance level. We can infer that at least one of the 10 regressors contributes significantly to the model. In this case both X2 and X9 are significant.





The fitted vs residual data is scattered but the qq-plot has a bit of positive skew.

2.3 Model 3: Square Root Transformation

lm(formula = sqrt(Y1) ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.337e+00	4.124e-01	3.241	0.00298	**
X1	-8.280e-04	1.842e-03	-0.449	0.65643	
X2	1.864e-02	6.951e-03	2.681	0.01197	*
X3	-8.391e-04	5.591e-04	-1.501	0.14422	
X4	5.833e-05	4.572e-05	1.276	0.21217	
X5	2.906e-03	3.917e-03	0.742	0.46403	
X6	8.975e-03	6.799e-03	1.320	0.19717	
X7	-1.015e-03	4.633e-03	-0.219	0.82815	
X8	-3.309e-03	2.568e-03	-1.289	0.20768	

X9	1.719e-03	1.249e-03	1.376	0.17938
X10	1.749e-03	2.898e-03	0.603	0.55088

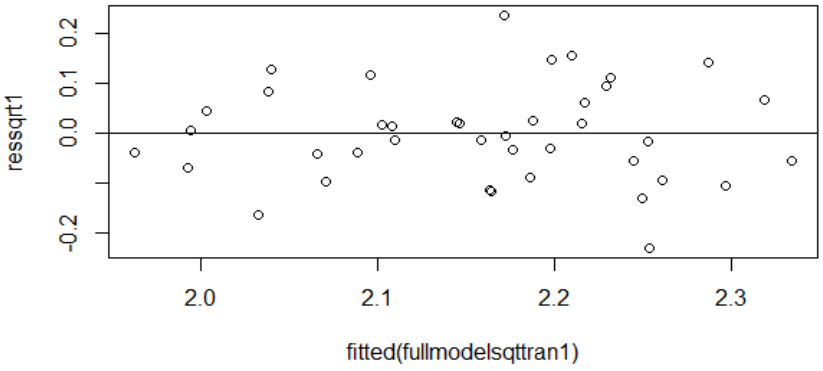
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

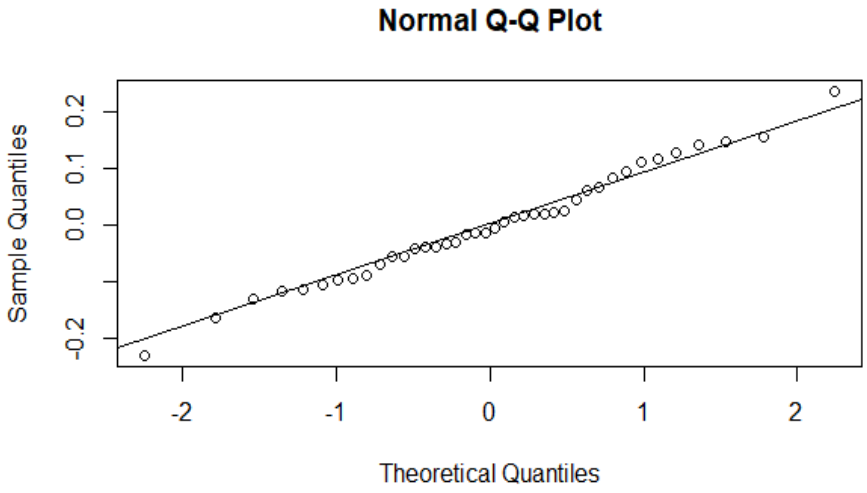
Residual standard error: 0.1131 on 29 degrees of freedom
Multiple R-squared: 0.4901, Adjusted R-squared: 0.3142
F-statistic: 2.787 on 10 and 29 DF, p-value: 0.01514

The fitted model 3 is given below:

$$\hat{y} = 38097.3714 - 99.3403X1 - 430.9478X2 + 24.5681X3 + 0.2247X4 - 18.8431X5 - 348.3754X6 + 67.5299X7 - 4.3799X8 + 44.8012X9 - 28.7977X10$$

From the summary, x2 which is alcohol consumption is significant. Compared to the data before transformation, the r-squared is slightly higher.





Both the residual and qq-plot show that the distribution of the residuals is normal and the variance is constant. The residual plot is scattered and the points stick to the line with no skew.

2.4 Model 4 Log Transformation

lm(formula = log(Y1) ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10)

Coefficients:

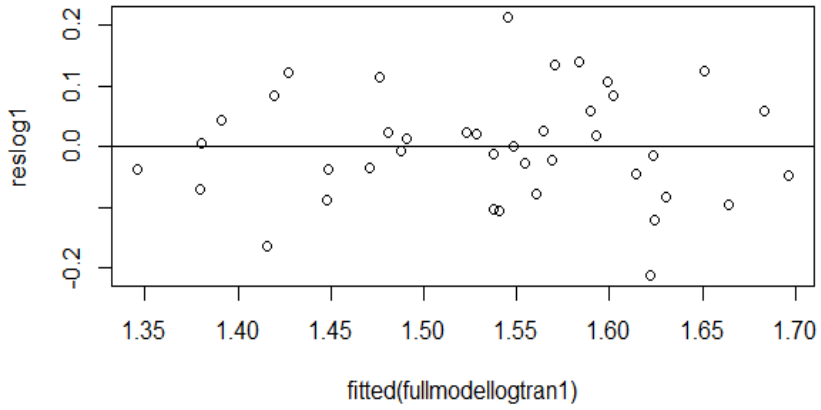
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	7.471e-01	3.796e-01	1.968	0.0587	.
X1	-7.473e-04	1.696e-03	-0.441	0.6627	
X2	1.720e-02	6.399e-03	2.687	0.0118	*
X3	-7.801e-04	5.147e-04	-1.516	0.1404	
X4	5.776e-05	4.209e-05	1.373	0.1804	
X5	2.840e-03	3.605e-03	0.788	0.4373	
X6	8.723e-03	6.259e-03	1.394	0.1740	
X7	-1.256e-03	4.265e-03	-0.295	0.7704	
X8	-2.994e-03	2.364e-03	-1.267	0.2153	
X9	1.534e-03	1.150e-03	1.334	0.1926	
X10	1.386e-03	2.667e-03	0.520	0.6072	

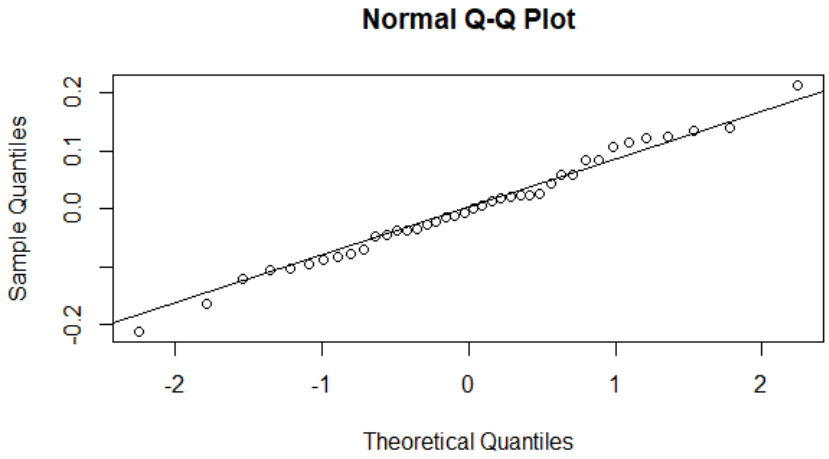
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1041 on 29 degrees of freedom
Multiple R-squared: 0.4983, Adjusted R-squared: 0.3252
F-statistic: 2.88 on 10 and 29 DF, p-value: 0.01268

$\hat{y} = 28.0942429 - 0.0252407X1 - 0.1528572X2 + 0.0098236X3 + 0.0005484X4 + 0.0083254X5 - 0.1407161X6 + 0.0514714X7 - 0.0350592X8 + 0.0200838X9 - 0.0058850X10$

From the summary, x2 which is alcohol consumption is significant. Compared to the data before transformation, the r-squared is slightly higher.





Both the residual and qq-plot show that the distribution of the residuals is normal and the variance is constant. The residual plot is scattered, and the points stick to the line with no skew.

Both transformations seem to produce similar results and summaries. Overall, the best transformation is taking the log of y because the R-squared is a bit higher and the p-value is smaller.

2.5 Model 5 Inverse Transformation

lm(formula = (1/Y1) ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.996e-01	8.181e-02	4.884	3.5e-05	***
X1	1.568e-04	3.654e-04	0.429	0.6711	
X2	-3.691e-03	1.379e-03	-2.676	0.0121	*
X3	1.695e-04	1.109e-04	1.528	0.1373	
X4	-1.420e-05	9.070e-06	-1.566	0.1283	
X5	-6.800e-04	7.770e-04	-0.875	0.3887	
X6	-2.065e-03	1.349e-03	-1.531	0.1366	
X7	4.037e-04	9.192e-04	0.439	0.6638	
X8	6.138e-04	5.094e-04	1.205	0.2379	
X9	-3.085e-04	2.478e-04	-1.245	0.2231	
X10	-2.034e-04	5.748e-04	-0.354	0.7260	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02244 on 29 degrees of freedom
Multiple R-squared: 0.512, Adjusted R-squared: 0.3437
F-statistic: 3.043 on 10 and 29 DF, p-value: 0.009306

From the summary, x2 which is alcohol consumption is significant. Compared to the data before transformation, the R-squared is slightly higher.

All four transformations seem to produce similar results and summaries. Overall, the best transformation is taking the inverse of y because the r-squared is higher and the p-value is much smaller.

2.6 Model 6 Stepwise Regression of Data After Transformation of Inverse of Y

Coefficients:

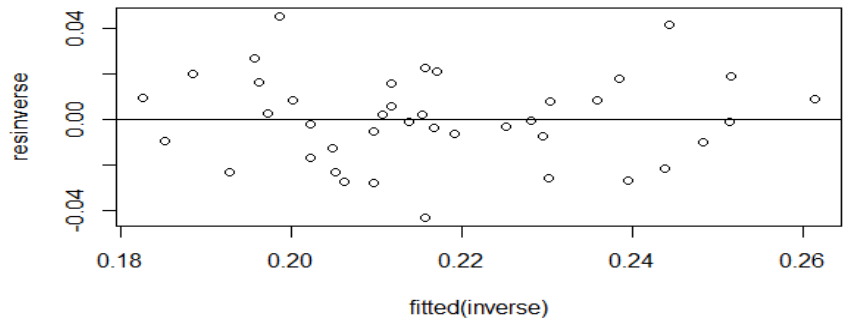
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.416e-01	6.073e-02	5.624	2.93e-06	***
X2	-3.424e-03	1.188e-03	-2.883	0.00689	**
X3	1.840e-04	1.022e-04	1.800	0.08107	.
X4	-1.412e-05	8.411e-06	-1.679	0.10256	
X6	-1.194e-03	8.902e-04	-1.341	0.18897	

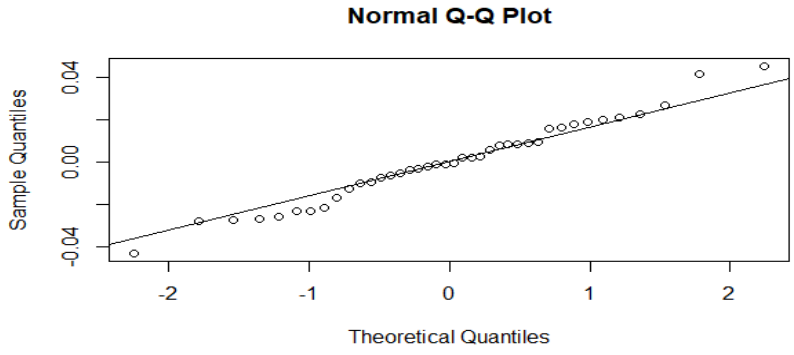
X8 6.943e-04 4.804e-04 1.445 0.15778
X9 -2.865e-04 2.085e-04 -1.374 0.17863

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0215 on 33 degrees of freedom
Multiple R-squared: 0.4902, Adjusted R-squared: 0.3976
F-statistic: 5.289 on 6 and 33 DF, p-value: 0.00065

The fitted vs residual and qq plot of residuals are given below





From the summary statistics we can see that both x2 and x3 which are alcohol consumption and country GDP respectively are important factors for the model. R-squared and p-value are still around the same. Both the residual and qq-plot show that the distribution of the residuals is normal and the variance is constant. The residual plot is scattered and the points stick to the line with no skew. Therefore, the inverse transformed final fitted model is

$\hat{y}$ = 3.416e-01 − 3.424e-03X2 + 1.840e-04X3 - 1.412e-05X4 - 1.194e-03X6 + 6.943e-04X8 + 2.865e-04X9

3 Checking Multicollinearity

One of the important assumptions for the linear regression model is that the regressors (independent variables) should be independent. Therefore, it is necessary to check this assumption. There are several ways available to check the multicollinearity problem. We will consider the popular correlation matrix, variance inflation factor and condition number.

3.1 The Correlation Matrix

The correlation matrix for all regressors is given below.

Correlation Matrix

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	1.00	0.49	-0.04	-0.20	-0.00	0.78	0.71	0.64	0.79	-0.58
x2	0.49	1.00	0.08	0.07	-0.04	0.39	0.56	0.71	0.50	-0.53
x3	-0.04	0.08	1.00	0.14	-0.17	-0.04	0.02	-0.08	-0.14	0.14
x4	-0.20	0.07	0.14	1.00	-0.17	-0.10	-0.05	0.11	-0.12	0.08
x5	-0.00	-0.04	-0.17	-0.17	1.00	-0.31	-0.23	-0.21	-0.11	0.08
x6	0.78	0.39	-0.04	-0.10	-0.31	1.00	0.78	0.65	0.76	-0.52
x7	0.71	0.56	-0.02	-0.05	-0.23	0.78	1.00	0.64	0.63	-0.58
x8	0.64	0.71	-0.08	0.11	-0.21	0.65	0.64	1.00	0.71	-0.58
x9	0.79	0.50	-0.14	-0.12	-0.11	0.76	0.63	0.71	1.00	-0.63
x10	-0.58	-0.53	0.14	0.08	0.08	-0.52	0.57	-0.58	-0.63	1.00

High Correlated Pairs have a correlation coefficient of 0.65: (x1,x6), (x1,x7), (x1,x9), (x2,x8), (x6, x7), (x6, x9), (x8,x9)

3.2 Variance Inflation Factor (VIF)

Now to see the severity of multicollinearity, we provide the variance inflation factor as follows

VIF

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
4.478144	2.672330	1.189062	1.210390	1.468705	5.583068	3.700056	3.681198	4.008472	1.986225

No VIF has a value over 10 . Therefore, we may assume that the data are not suffering from the problem of multicollinearity (see Momtgomery et al. 2013).

3.3 Condition Number and Condition Indices

The ordered eigenvalues of the X'X matrix are,

4.8354 1.4048 0.9939 0.8990 0.5843 0.4490 0.3733 0.1857 0.1590 0.1156

Condition Number: K = 4.8354/0.1156 = 41.8287
The condition indices are

K1: 1
K2: 3.4421
K3: 4.8651
K4: 5.3786
K5: 8.2755
K6: 10.7693
K7: 12.9531
K8: 26.0388
K9: 30.4113
K10: 41.8287

From the above correlation matrix, VIFs, condition indices, and condition Number, we can see that there is no substantial Multicollinearity that must be addressed amongst the 10 regressors. This means that performing a Ridge Regression analysis on the models will not improve the models. More on multicollinearity problems, see Kibria (2003), Kibria and Banik (2016) among others.

4. Finding the best fitted model

There are several ways one can compare models or to find the best fitted model. One of the very simple way is as follows:

Model	F-value (p-value)	R-square	Residual SE	Number of significant regressors

Model 1	2.69 (0.018)	0.48	0.494	X2
Model 2	10.83 (0.0002)	0.369	0.482	X2, X9
Model 3	2.787 (0.015)	0.49	0.113	X2
Model 4	2.88 (0.012)	0.50	0.104	X2
Model 5	3.043 (0.01)	0.51	0.022	X2
Model 6	5.289 (0.001)	0.49	.0215	X2, X3, X4 (at 10% significance level)

If we review the above table and consider all model characteristics, model 6 performs the best. Therefore, the model 6 will be our final model. The final fitted model is

$\hat{y}$ = 3.416e-01 − 3.424e-03X2 + 1.840e-04X3 - 1.412e-05X4 - 1.194e-03X6 + 6.943e-04X8 + 2.865e-04X9

Now, if you want to give better ways to find the best fitted model, you may follow

4. Some Concluding Remarks

After doing some analysis of the data and some transformation of the data as well, I was able to find that there is a model of a relationship between depression and 10 regressor factors. The final model showed there is a relationship between depression and the two regressor factors which are alcohol consumption and country gdp. The final model was obtained by taking the inverse of y and then performing a stepwise function of the transformed data. The model has met the assumptions of normality and the r-squared is at 49% which is decent. There were also plenty of tests to show that multicollinearity is not a problem in this model. If one wishes to do more in this study, they could add more countries and do tests to detect any outliers and then omit them in designing more models.

References

Guzman, C. I and Kibria, B. M. G. (2019). Developing Multiple Linear Regression Models for

the Number of Citations: A Case Study of Florida International University Professors. *International Journal of Statistics and Reliability Engineering,* 6(2), 75-81.

Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computation*, 32, 419-435.

Kibria, B. M. G. and S. Banik (2016). Some Ridge Regression Estimators and Their Performances. *Journal of Modern Applied Statistical Methods. 15 (1), 206-238.*

Montgomery, Douglas C., Elizabeth A. Peck, & G. Geoffrey Vining. *Introduction to Linear Regression Analysis (Fifth Edition)*. John Wiley & Sons, 2013.

Xu T, Zhu G, Han S. (2017) Study of depression influencing factors with zero-inflated regression models in a large-scale population survey. BMJ Open 2017;7:e016471. doi:10.1136/bmjopen-2017-016471

Depression Rates by Country 2023, https://worldpopulationreview.com/country-rankings/depression-rates-by-country.

Team, Our World in Data, et al. "Our World in Data." *Our World in Data*, https://ourworldindata.org/.

"Unemployment, Total (% of Total Labor Force) (Modeled ILO Estimate)." *Data*, https://data.worldbank.org/indicator/SL.UEM.TOTL.ZS?end=2021&most_recent_year_desc=false&start=2021&view=bar.

"Life Expectancy of the World Population." *Worldometer*, https://www.worldometers.info/demographics/life-expectancy/.

"The 2021 Global Health Security Index." *GHS Index*, 8 Dec. 2021, https://www.ghsindex.org/#:~:text=The%20Data%3A,is%20essentially%20unchanged%20from%202019.

Average Temperature by Country, https://tradingeconomics.com/country-list/temperature.