01.020 Design Thinking Project III (MU)

Towards Zero Hunger:
A Cross-Country Analysis of Factors Affecting Food Insecurity

Cohort 08 Group 04

November 22, 2023

ID	Name	ID	Name
1006617	Michael Hoon Yong Hau	1006954	Joshua John Lee Shi Kai
1006651	Wong Qi Yuan Kenneth	1007085	Yaromir Viswanathan

Contents

1 Introduction

The world faces an unprecedented challenge — to feed a projected population of 9 billion by 2050 while ensuring safe, sustainable, and equitable food systems [**?**]. This imperative is not only a matter of meeting the demands of a growing population but is deeply rooted in the fundamental right of every individual to access safe, nutritious, and sufficient food, as underscored by the United Nations (UN) Sustainable Development Goal 2: Zero Hunger [**?**].

In light of this, our project aims to conduct a rigorous cross-country analysis, investigating the intricate factors that influence food insecurity on a global scale. Our problem statement is as follows:

> "How might we identify the **key factors of influence for food insecurity index** across countries, to facilitate informed and targeted governmental policy interventions?"

As such, our main target audience for this project is governmental organisations involved in policy-making for their country's food security. This ensures relevant insights to formulate effective policies and interventions, addressing food security issues based on specific socioeconomic indicators.

2 Data Collection and Wrangling

Most of our raw data is obtained from reputable sources such as the Food and Agricultural Organisation of the United Nations (FAO) and the World Bank Group. Since our data for each variable is sourced from multiple sources (for the year 2022), we have compiled everything into a single sheet as the input for our model.

2.1 Data Cleaning

To prepare the raw datasets into cleaned, usable information, we applied several data cleaning techniques, such as normalisation and imputation. For normalisation in Excel, we used the built-in `=STANDARDISE()` function, which applies z-score normalisation on each row value. As for data imputation, we applied an (Iterative) Multiple-Imputation (MI) technique [**?**] [**?**] to replace the missing data values in our dataset, instead of merely dropping the rows. As for the variables (Gini Index & Average Temperature) which are not available for the year 2022, we have imputed using the most recent data. The final dataset can be found in the Excel sheet.

2.2 Data Transformation

Furthermore, we have log-transformed GDPpc (Gross Domestic Product per capita) and Minimum Wage, as from the scatter plots obtained of all the predictor variables, they appear to follow a logarithmic trend, shown in Figure **??** and Figure **??**.

3 Multiple Linear Regression Model

Our Multiple Linear Regression Equation is as follows:

$$\begin{aligned}\textbf{FII} = {} & \beta_0 + \beta_1 \textbf{AgriculturalLand}(\%) + \beta_2 \textbf{CO}_\textbf{2}\textbf{Emissions} + \beta_3 \textbf{CPI} + \beta_4 \textbf{GDP} + \beta_5 \textbf{Population} \\ & + \beta_6 \textbf{InfantMortality} + \beta_7 \ln[\textbf{MinimumWage}] + \beta_8 \textbf{UnemploymentRate} \\ & + \beta_9 \textbf{LabourForceParticipation}(\%) + \beta_{10} \textbf{Temperature} + \beta_{11} \textbf{IdealTemperature} \\ & + \beta_{12} \textbf{PrecipitationDepth} + \beta_{13} \textbf{GiniIndex} + \beta_{14} \ln[\textbf{GDPpc}] + \beta_{15} \textbf{HDI}\end{aligned} \tag{1}$$

where **FII** is the Prevalence of Moderate and Severe Food Insecurity, **CPI** is the Consumer Price Index, and **HDI** is the Human Development Index.

3.1 OLS Parameter Estimates

Parameter estimates of our original model are given in Table **??**, Appendix **??**. Significance levels are indicated with stars below the table.

3.1.1 Individual Statistical Significance

From the *p-values* obtained in Table **??**, we conclude that **CO_2 Emissions**, **GDP**, **Population**, **Infant Mortality**, **Gini Index**, and **HDI** are all statistically significant at the 95% level.

3.2 Significance Test for Equation

In order to test the significance of our model, we employ an F-test [**?**]. The null hypothesis of the F-test is given by:

$$H_0 : \boldsymbol{\beta_0} = \boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \boldsymbol{\beta_3} = \boldsymbol{\beta_4} = \cdots = \boldsymbol{\beta_{15}} = 0 \tag{2}$$

and the F-statistic is

$$F = \frac{\textbf{Mean Square Regression}}{\textbf{Mean Square Error}} = \frac{\sum_{i=1}^{N}(\hat{Y}_i - \hat{Y})^2/k}{\sum_{i=1}^{N}(Y_i - \hat{Y})^2/(n-k-1)} = 104.749 \tag{3}$$

The calculated F-statistic is $\boldsymbol{F = 104.749}$. At the 95% significance level, the critical value obtained from the F-distribution table on 15 and 179 degrees of freedom is $\boldsymbol{F = 2.04}$. Since the F-statistic $\boldsymbol{F = 104.749 > 2.04}$, then we can say that there is **sufficient evidence to reject the null hypothesis at the 95% significance level**, and **the equation is statistically significant.** However with this model, there is a high probability that over-fitting has occurred, and there are many variables which are not individually significant.

4 Model Improvements

With the above shortcomings of our model, we suggest several improvement techniques:

4.1 Feature Engineering

We have built a subset model via feature engineering, to separate the original model into an economic model in Equation **??** (with Economic factors) and a geographical model (with geographical-related factors) in Equation **??**. These variables are separated for a better comparison between the significance of economic and geographical factors affecting FII, allowing us to select the best variables for our final model. Parameter estimates of each model is given in Table **??** and Table **??**, Appendix **??**.

4.1.1 Economic Model

$$\begin{aligned}\textbf{FII} = \beta_0 + \beta_1\textbf{CPI} + \beta_2 \ln[\textbf{MinimumWage}] + \beta_3\textbf{Unemployment} + \beta_4\textbf{LabourForceParticipation} \\ + \beta_5\textbf{GiniIndex} + \beta_6\textbf{GDP}\end{aligned} \tag{4}$$

4.1.2 Geographic Model

$$\begin{aligned}\textbf{FII} = \beta_0 + \beta_1\textbf{AgriculturalLand}(\%) + \beta_2\textbf{CO}_\textbf{2}\textbf{Emissions} + \beta_3\textbf{Population} + \beta_4\textbf{InfantMortality} \\ + \beta_5\textbf{IdealTemperature} + \beta_6\textbf{PrecipitationDepth} + \beta_7\textbf{HDI}\end{aligned} \tag{5}$$

4.2 Statistical Significance

To determine the variables significant for our final model, we included variables which are statistically significant in both the original model and feature-engineered models (with reference to Tables **??**, **??**, **??**, Appendix **??**). Furthermore, we have made an exception for **ln [MinimumWage]**, as despite not being statistically significant in the original model, it has the highest t-statistic value (lowest p-value) in the economic model, which prompted us to include it in the final model.

4.3 Final Model

After accounting for all the factors in Sections **??** and **??**, our final model is as follows:

$$\begin{aligned}\textbf{FII} = 0.006\ \textbf{GDP} - 0.476\ \textbf{Population} + 0.560\ \textbf{InfantMortality} + 0.145\ \textbf{ln}\,[\text{MinimumWage}] \\ + 0.229\ \textbf{GiniIndex}\end{aligned} \tag{6}$$

where the coefficients have been included in the model, and $\beta_0 = 0$. Our model includes a mix of both geographical and economic factors, which provides a comprehensive analysis of the indicators affecting FII. From the coefficients, we can conclude that for most countries, **GDP**, **Population**, **Infant Mortality**, **Minimum Wage**, and **Gini Index** are significant factors in determining a country's **FII**.

5 Descriptive Statistics

5.1 Adjusted R^2

The adjusted $\boldsymbol{R^2}$ value for our original model is $\boldsymbol{R^2 = 0.889}$, which indicates a strong linear trend for the independent variables. For the Economic model, $\boldsymbol{R^2 = 0.647}$, while for Geographic Model, $\boldsymbol{R^2 = 0.861}$. Adjusted $\boldsymbol{R^2}$ accounts for the *number of independent variables* used for predicting the target variable, and hence is a better metric for our model with multiple predictors, compared to just $\boldsymbol{R^2}$ [**?**]. As for our final model, adjusted $\boldsymbol{R^2 = 0.862}$. Although it is lower than the original model, too high of an adjusted $\boldsymbol{R^2}$ value might indicate over-fitting. Adjusted $\boldsymbol{R^2}$ alone is also an insufficient (non-robust) indicator to model performance.

5.2 Significance Test for Final Model

Similar to Section **??**, we will conduct an F-test on our **final** model, where the null hypothesis is given by:

$$H_0 : \boldsymbol{\beta_0} = \boldsymbol{\beta_1} = \boldsymbol{\beta_2} = \boldsymbol{\beta_3} = \boldsymbol{\beta_4} = \boldsymbol{\beta_5} = 0 \tag{7}$$

and our calculated F-statistic is $\boldsymbol{F = 244.069}$. At the 95% significance level, the critical value obtained from the F-distribution table on 5 and 189 degrees of freedom is $\boldsymbol{F = 2.262}$. Since the F-statistic $\boldsymbol{F = 244.069 > 2.262}$, then we can say that there is **sufficient evidence to reject the null hypothesis at the 95% significance level**, and **the equation is statistically significant**. Furthermore, the F-stat for the final model is significantly larger than our original model, indicating a better fit.

5.3 Akaike Information Criterion (AIC)

The AIC value for our final model is $\boldsymbol{\text{AIC} = -380.784}$. For the other models, the values can be found in Table **??**. AIC evaluates how well a model fits the data it was generated from, and balances the goodness of fit and model complexity [**?**].

$$\boldsymbol{\text{AIC}} = 2k - 2\ln[\hat{L}] \tag{8}$$

where k is the number of estimated parameters in the model, and $\hat{L}$ is the maximum value of the likelihood function for our model. Comparatively for our original model, $\boldsymbol{\text{AIC} = -413.626}$ which is lower. This value is expected as AIC is more accepting of models with more complex predictors, if it gives a better goodness of fit. However for our case, it might be better to consider BIC as we are using a simpler model.

5.4 Bayesian Information Criterion (BIC)

The BIC value for our final model is $\boldsymbol{\text{BIC} = -361.146}$. For the other models, the values can be found in Table **??**. BIC also evaluates how well a model fits the data it was generated from, but it **balances the fit of the model with the number of parameters, heavily penalizing the models that are too complex** [**?**].

$$\boldsymbol{\text{BIC}} = k\ln[n] - 2\ln[\hat{L}] \tag{9}$$

where the '2' from AIC is replaced by '$\ln[n]$' in BIC, and n is the sample size of our dataset. Comparatively for our original model, $\boldsymbol{\text{BIC} = -361.258}$ which is roughly similar to the final model.

6 Gauss-Markov Assumptions

Since we have employed an Ordinary Least Squares (OLS)-based multiple linear regression model, we need to satisfy the Gauss-Markov assumptions [**?**] to ensure that our estimators are **unbiased and efficient** (Best Linear Unbiased Estimator - BLUE).

6.1 Normality of Residuals

To ensure normality of the residuals, we visualise using a Q-Q plot on Figure **??**, Appendix **??**. From the linear trend of the plot and the relatively high $\boldsymbol{R^2}$ value, we can deduce that the residuals are indeed normally distributed.

6.2 Independence of Errors

Since our data is cross-sectional as opposed to time-series, we will be using the Residuals vs. Fitted plot for analysis instead of a test for autocorrelation such as the Durbin-Watson Test. The scatter plot is in Figure **??**, Appendix **??**. From the trend of the plot, we can see that there is no clear trend of the errors and it is spread with constant variance around the x-axis. This indicates that the error terms are indeed independent of each other and the predicted variable.

6.3 Heteroskedasticity Analysis

To test for Heteroskedasticity, we can employ a White Test [**?**], in addition to analysing the Figures **??**, **??**, and **??**, Appendix **??**. From the figures, we can see that with the exception of GDP, the residuals are scattered evenly around the x-axis with constant variance, indicating that they are **homoskedastic**. As for GDP, the residuals follow a haphazard trend around the y-axis with **inconsistent variance**, indicating the presence of **heteroskedasticity**.

6.3.1 White Test

The null hypothesis for the White Test is that the variances for the error terms in the equation are equal:

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2 \tag{10}$$

and the White test statistic is calculated: $\mathbf{LM = 195.000}$ (where LM is the Lagrange Multiplier test statistic: $\mathbf{LM} = nR^2$, with p-value = 1.00e−4. The critical value for this is $\boldsymbol{\chi^2 = 40.113}$ at 27 degrees of freedom. Since $\mathbf{LM = 195.000 > 40.113}$, following the $\boldsymbol{\chi^2}$ distribution, we have sufficient evidence to **reject the null hypothesis at the 95% significance level**, and thus the error terms exhibit heteroskedasticity. This might be attributed to the errors of GDP being highly heteroskedastic.

6.3.2 Consequences of Heteroskedastic Errors

If the errors are heteroskedastic, the standard errors of the estimated coefficients tend to be **biased and inconsistent**. The standard errors may be underestimated for observations with larger variances and overestimated for observations with smaller variances.

Furthermore, the t-statistic is calculated as the ratio of the estimated coefficient to its standard error (similar to the F-statistic in Equation **??**). When the standard errors are biased due to heteroskedasticity, the t-statistics (and F-statistic) can be distorted, leading to incorrect conclusions about the statistical significance of the coefficients, where we fail to reject the null hypothesis H_0.

6.4 Multicollinearity Considerations

To test for multicollinearity, we can use the Variance Inflation Factor (VIF). VIF is a measure to analyse the magnitude of multicollinearity of our model's parameters [**?**]. High VIF values (typically above 10) indicate untolerable multicollinearity. Since none of the predictors exhibit high VIF values, then we can conclude that there is little to no multicollinearity in our model.

	GDP	**Population**	**InfantMortality**	**ln [MinWage]**	**GiniIndex**	**FII**
VIF	1.766	3.399	7.166	5.624	1.684	7.457

Table 1: VIF values for **Final** Model

We can also use the correlation matrix in Table **??**, and the values highlighted in **red** signify relatively high correlation.

6.5 Endogeneity

When Endogeneity is present in a model, one or more independent variables are correlated with the error term, resulting in biased and inefficient parameter estimates [**?**]. Sources of Endogeneity in a model usually occur from Omitted Variable Bias (where the variable is correlated with both the independent variable in the model and with the error term) [**?**]. To test for Endogeneity in our model, we can again refer to Figures **??**, **??**, and **??**, Appendix **??**. From the scatter plots, we can see that there is no explicit trend (quadratic, logarithmic or otherwise) in the error terms, indicating that the variables are likely exogenous. Furthermore, a rigorous

analysis of endogeneity depends on the Data Generating Process [**?**], which we are unaware of as our dataset was obtained from external sources.

7 Limitations of Analysis

There are certain limitations of this project which we must acknowledge, for a comprehensive analysis of the suggested outcomes for determining FII.

7.1 Presence of Heteroskedasticity

As mentioned in Section **??**, the presence of Heteroskedasticity is a key limitation in our analysis. In future studies, to correct for heteroskedasticity, we may employ the method of Generalised Least Squares (GLS) [**?**] or any Heteroskedasticity and Autocorrelation Consistent Estimator (e.g. Newey-West Estimator) [**?**]. GLS explicitly models the variance-covariance structure of the error terms, allowing for the estimation of regression coefficients in the presence of heteroskedasticity. The specific variance-covariance structure used to calculate the GLS estimator is the power of a variance covariate. With this, we can then say that the estimators for our Multiple Linear Regression Model is the Best Linear Unbiased Estimator (BLUE). Conducting GLS estimation for our analysis is outside the scope of this project, however.

7.2 Generalisation of Countries

Another key limitation of our model involves the generalisation of countries sampled. In our dataset, we sampled data for all available countries, which have vastly differing socioeconomic indicators such as GDP per capita, especially for developing and developed countries. As such, the model is unable to identify certain intricacies specific to regions such as specific countries in Africa, which may be significant in explaining changes in the food insecurity index for those countries. As a further improvement for future localised models, a subset of the data could be used to train the model using only countries from a certain region, to obtain more accurate results of the predicted FII values.

7.3 Endogeneity Concerns

As mentioned in Section **??**, we are unsure of the existence of endogeneity as it usually requires rigorous analysis on the Data Generating Process. However, to potentially mitigate this issue, further studies can be conducted using Instrumental Variable (IV) Regression [**?**] or Two-Stage Least Squares (2SLS) Regression [**?**], which corrects for bias in the OLS model. We will not conduct IV or 2SLS regression in the scope of this project however, as determining suitable Instrumental Variables usually require months to years of research.

8 Conclusion

Given the statistically significant variables found in Section **??**, we conclude that the key factors of influence on Food Insecurity are **GDP**, **Population**, **Infant Mortality**, **Minimum Wage**, and **Gini Index**. As such, governments should aim to employ relevant macroeconomic policy interventions in those particular domains to ensure that the food insecurity index is sufficiently low.

Appendix A Data Visualisation

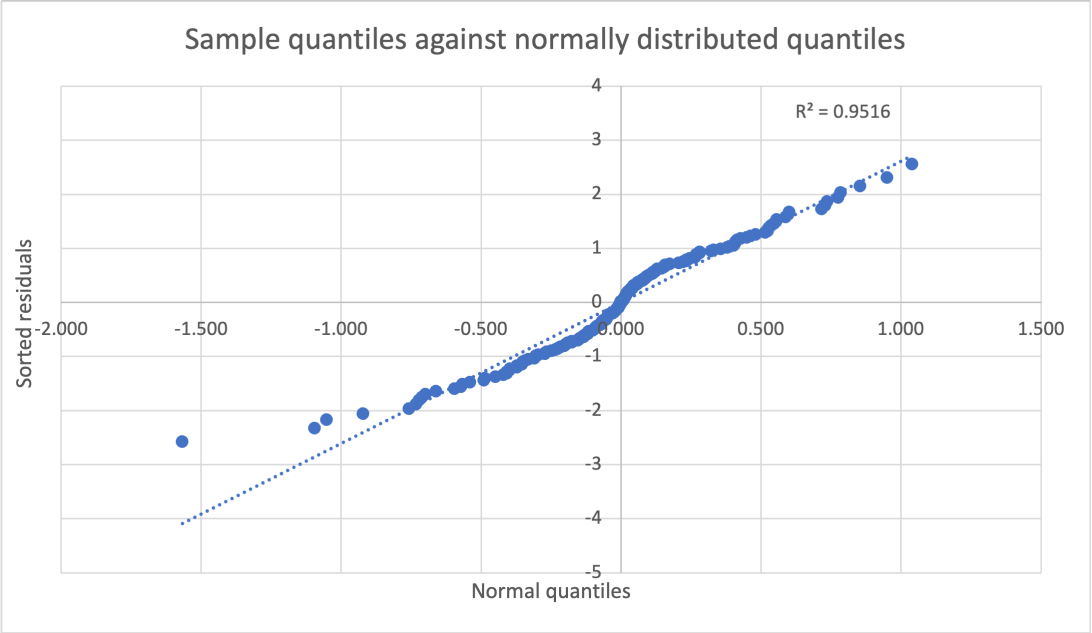
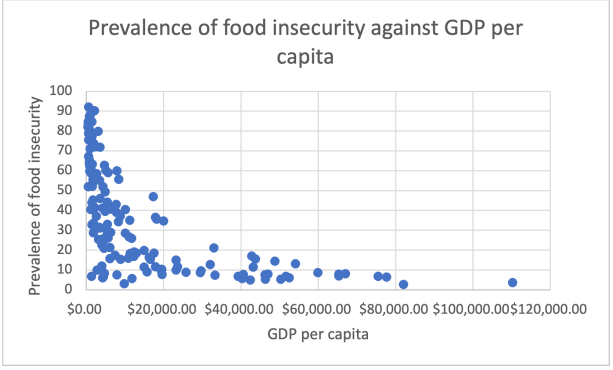
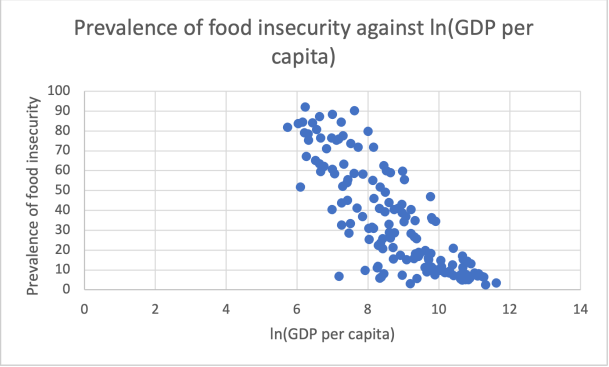


Figure 1: QQ Plot of Model Residuals



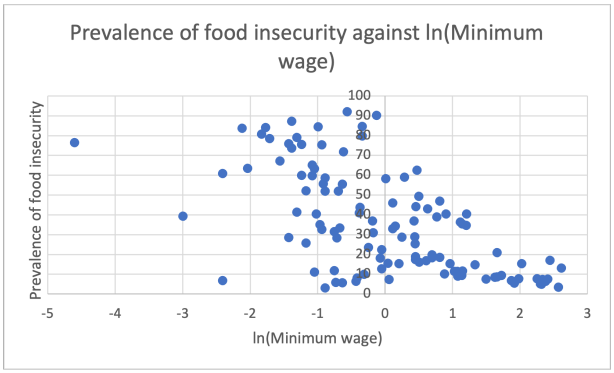
(a) FII vs. GDPpc



(b) FII vs. ln [GDPpc]

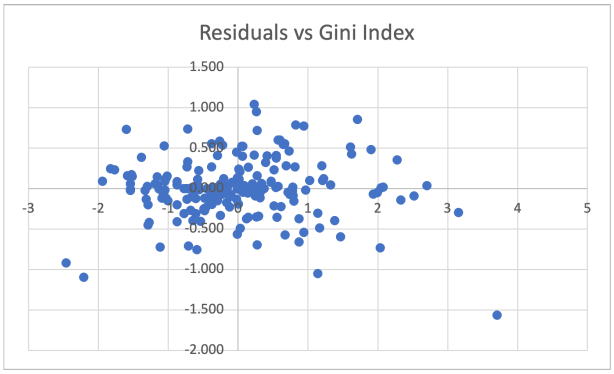
Figure 2: Scatter Plots of FII against GDPpc

(a) FII vs. MinimumWage

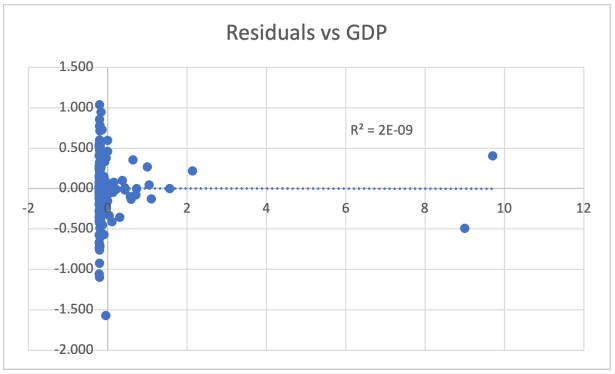


(b) FII vs. ln [MinimumWage]

Figure 3: Scatter Plots of FII against Minimum Wage

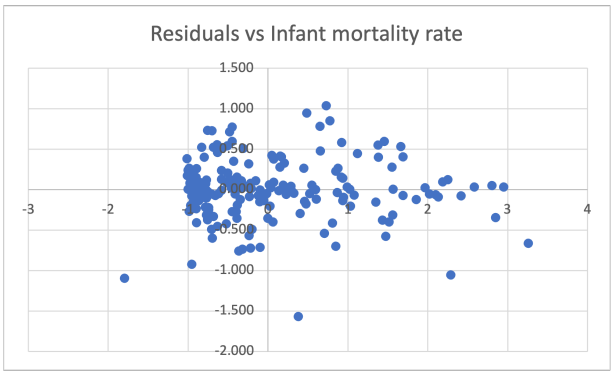


(a) Residuals vs. GiniIndex

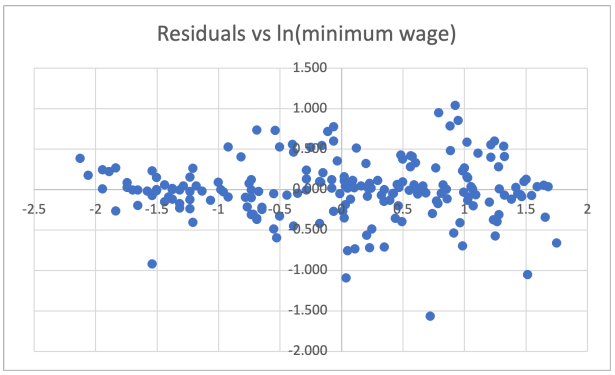


(b) Residuals vs. GDP

Figure 4: Residuals vs. GDP



(a) Residuals vs. InfantMortality



(b) Residuals vs. MinimumWage

Figure 5: Scatter Plots of Residuals against Predictors

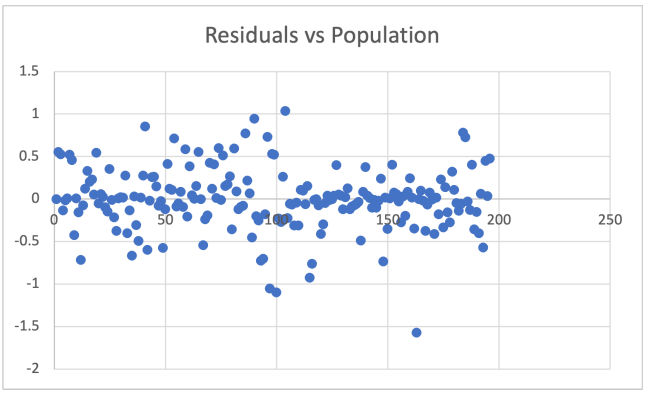


Figure 6: Scatter Plots of Residuals against Population

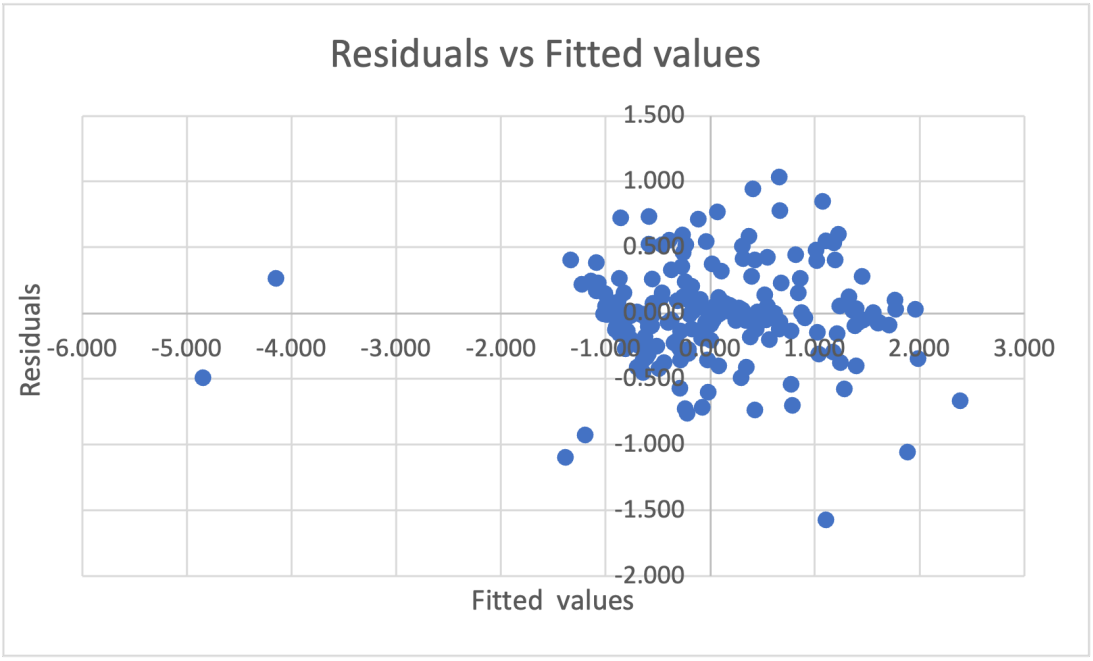


Figure 7: Scatter Plots of Residuals against Fitted Values of Model

Appendix B OLS Parameter Estimates for Each Model

	Estimate	**Std. Error**	**t-statistic**	**p-value**
β_0	0.000	0.024	0.000	1.000
β_1 **AgriculturalLand(%)**	-0.017	0.027	-0.620	0.536
β_2 $\boldsymbol{CO_2}$**Emissions**	-0.193	0.097	-1.996	0.047 **
β_3 **CPI**	0.004	0.026	0.150	0.881
β_4 **GDP**	0.190	0.071	2.662	0.008 ***
β_5 **Population**	-0.431	0.049	-8.818	1.00e−4 ***
β_6 **InfantMortality**	0.418	0.059	7.118	1.00e−4 ***
β_7 **ln [MinimumWage]**	-0.079	0.067	-1.177	0.241
β_8 **Unemployment(%)**	-0.029	0.032	-0.896	0.371
β_9 **LabourForceParticipation**	-0.045	0.030	-1.516	0.131
β_{10} **Temperature**	0.069	0.044	1.571	0.118
β_{11} **IdealTemperature**	0.051	0.033	1.546	0.124
β_{12} **PrecipitationDepth**	-0.013	0.027	-0.490	0.625
β_{13} **GiniIndex**	0.193	0.034	5.601	1.00e−4 ***
β_{14} **ln [GDPpc]**	-0.017	0.044	-0.381	0.704
β_{15} **HDI**	-0.351	0.078	-4.493	1.00e−4 ***

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 2: **Original** Multiple Linear Regression Model Parameter Estimates

	Estimate	Std. Error	t-statistic	p-value
β_0	0.000	0.043	0.000	1.000
β_1 **CPI**	0.044	0.044	1.009	0.314
β_2 **ln [MinimumWage]**	0.572	0.049	11.593	1.00e−4 ***
β_3 **Unemployment(%)**	-0.007	0.054	-0.125	0.901
β_4 **LabourForceParticipation**	0.017	0.052	0.338	0.736
β_5 **GiniIndex**	0.222	0.054	4.107	1.00e−4 ***
β_6 **GDP**	-0.301	0.043	-6.940	1.00e−4 ***

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 3: **Economic** Multiple Linear Regression Model Parameter Estimates

	Estimate	Std. Error	t-statistic	p-value
β_0	0.000	0.027	0.000	1.000
β_1 **AgriculturalLand(%)**	0.000	0.029	0.011	0.991
β_2 **CO_2Emissions**	0.047	0.047	1.004	0.317
β_3 **Population**	-0.517	0.047	-11.012	1.00e−4 ***
β_4 **InfantMortalityRate**	0.394	0.057	6.974	1.00e−4 ***
β_5 **IdealTemperature**	0.120	0.029	4.1032	1.00e−4 ***
β_6 **PrecipitationDepth**	-0.020	0.029	0.692	0.490
β_7 **HDI**	-0.409	0.058	-7.064	1.00e−4 ***

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 4: **Geographic** Multiple Linear Regression Model Parameter Estimates

	Estimate	Std. Error	t-statistic	p-value
β_0	0.000	0.027	0.000	1.000
β_1 **GDP**	0.006	0.035	0.176	0.860
β_2 **Population**	-0.476	0.035	-13.676	1.00e−4 ***
β_3 **InfantMortality**	0.560	0.059	9.576	1.00e−4 ***
β_4 **ln [MinimumWage]**	0.145	0.062	2.333	0.021 **
β_5 **GiniIndex**	0.229	0.030	7.582	1.00e−4 ***

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

Table 5: **Final** Multiple Linear Regression Model Parameter Estimates

Appendix C Inferential Statistics of Models

Model Type	Adjusted R^2	F-Statistic	AIC	BIC
Original Model	0.889	104.749	-413.626	-361.258
Economic Model	0.647	60.272	-196.199	-173.288
Geographical Model	0.861	172.132	-376.433	-350.249
Final Model	0.862	244.069	-380.784	-361.146

Table 6: Inferential Statistics of All Models

Appendix D Correlation Matrix of Final Model

	GDP	**Population**	**InfantMortality**	**ln [MinWage]**	**GiniIndex**	**FII**
GDP	**1**	0.625	-0.150	-0.170	-0.011	-0.403
Population	**0.625**	**1**	0.006	0.027	0.008	-0.463
InfantMortality	-0.150	0.006	**1**	0.888	0.352	0.766
ln [MinWage]	-0.170	0.027	**0.888**	**1**	0.453	0.733
GiniIndex	-0.011	0.008	**0.352**	**0.453**	**1**	0.489
FII	-0.403	-0.463	0.766	0.733	0.489	**1**

Table 7: **Final** Multiple Linear Regression Model Correlation Matrix **(Mirrored along Diagonal)**