Regression Analysis of Milton Friedman Dataset

Jacob McGraw
Lily Cai
Muhirwa Eric Kalisa
Van Ngo

May 24, 2021

Discussion

Milton Friedman's famous economic dataset is a landmark source that contains information on several economic indices. Friedman's dataset catalogs data from over two hundred countries from 1980 to 2006. Our group decided to use this dataset to model various economic data. Our group analyzed the prolific Milton Friedman dataset with a new sense of clarity after performing a series of data aggregation techniques with the Python programming language. Furthermore, our group performed a series of regression analyses to find a series of models that accurately determines economic activity on as many levels as possible. A brief period of exploratory analysis revealed several corollary relationships.

First, our group will test GDP per capita vs. the number of procedures to start a business for both developed and developing countries. According to the International Monetary Fund, countries are classified as developed or developing (2018, *International Monetary Fund*). The hypothesis is that the differences between developed and developing nations can be modeled with a non-interactive multiple additive regression model. Our group will use a scatterplot to speculate on a possible relationship, fit the model, and test practicality with QQ plots and residual plots. Second, our group will perform regression analysis on a possible relationship between GDP per capita and infant mortality. The hypothesis is that the relationship between GDP per capita and infant mortality can be illustrated with a simple linear regression model. To test the model's validity, our group will look at a scatterplot, fit the model, and test the residuals. Third, our group will model GDP per capita as a function of average life expectancy to determine a possible linear relationship. The hypothesis is that life expectancy will have a positive non-linear relationship with GPD (meaning that GDP will generally increase as life expectancy increases). Our group plans to observe a scatterplot to speculate on a possible connection and fit an appropriate regression model. Subsequently, our group will test the respective models with QQ plots on the residuals and residual plots. Finally, our group will perform a similar analysis on GDP per capita as a function of the average infant mortality rate and the average life expectancy. The hypothesis is that the relationship between GDP per capita and average mortality rate can be expressed with multiple additive regression models with interaction. Our group will test the model with a scatterplot, QQ plots, and residual plots.

Data Aggregation and Collection Methods

Our group had to aggregate the data before we could analyze it. Therefore, our group used the Python programming language to manipulate the original dataset in several ways. The Python code used to manage the data is in the project deliverable. First, our group subset the data to 2006 (because 2006 was the year with the least missing data). Then, the countries that had missing data in the inflation column were removed. However, several of the columns in the dataset had missing values. Average tariff, years of school between 15 to 25, average democracy, average life expectancy, number of procedures to start a business, and infant mortality rate all had missing values. The dataset may have left entries blank to avoid redundancies or simply had missing data. Regardless, our group took the average of each of the indices above from 1980 to 2006 and used those values as substitutes to account for missing values. After that, the columns were named accordingly. For example, the number of procedures to start a business was changed to an average number of procedures to start a business. Subsequently, our group performed additional research to determine which countries were considered developed or developing in 2006. Our group then provided labels (developed or undeveloped) on the countries developing status according to the International Monetary Fund. Then, the countries that still had missing entries were removed entirely from the dataset. The final variables are:

- Country-An identifying variable.
- Population-A variable indicating the population in the country in 2006
- GDP per capita-A variable indicating the purchasing power of the specified country in the year 2006.
- Inflation rate-A variable indicating the average percentage change in the cost of goods and services to the average consumer in the year 2006.
- Average highest tax rate-A variable indicating the average highest tax rate from the years 1980 to 2006.
- Average life expectancy-A variable indicating the average age of a person's life from the years 1980 to 2006
- Number of procedures to start a business-A variable indicating how many bureaucratic steps are needed to start a business in the year 2006.
- Average number of procedures to start a business-A variable indicating how many bureaucratic steps are needed to start a business in the years 1980 to 2006
- Average democracy-A variable indicating the average democracy from the years 1980 to 2006. It should be noted that democracy is a variable from 1 to ten (where one is least democratic and 10 is most democratic) that indicates the competitiveness and openness of a country's political institutions.
- Average infant mortality rate-A variable indicating the average rate of infant mortalities per 1000 births from the years 1980 to 2006.
- Developing status-A categorical variable indicating if the given country is developing or developed according to the International Monetary Fund.

The summary statistics for this data are:

Developing/Developed	Maximum	Minimum	Mean	Median
Population	1312000000/299000000	751175/4125376	69493060/39278195	13412920/11112985
GDP Per Capita	20890/38165	629.8/10939	6612.4/28643	5176/30237
Inflation	14.5/3.9	0.04/0.2	6.1/2.2	5.6/2
Highest Average Tax Rate	52.5/11.2	0.4/0.2	17/5.1	16.2/4.8
Average Life Expectancy	76.7/79.3	41.4/74.2	63.4/76.7	66.9/76.4
Number of	17/15	5/2	10.5/6.1	10/6

Procedures to Start a Business				
Average Number of Procedures to Start a Business	17/15	5/2	10.9/6.4	11/6
Average Democracy	10/10	0/2	4.9/9.4	5.1/10
Average Infant Mortality	176.3/17.9	6.9/4.5	53.2/7.5	46.8/6.9

Table 1: Summary statistics for developing/developed countries

Developed	Developing
23	67

Table 2: Number of developed and developing countries

Regression Modeling

Part i. Additive Multiple Regression for GDP and Procedures to Start a Business

Part a. Exploratory Analysis

One of the most promising relationships in the Friedman dataset is the GDP per capita as a function of the number of procedures to start a business for both developed and developing countries. The following scatterplot illustrates the relationship between the variables.

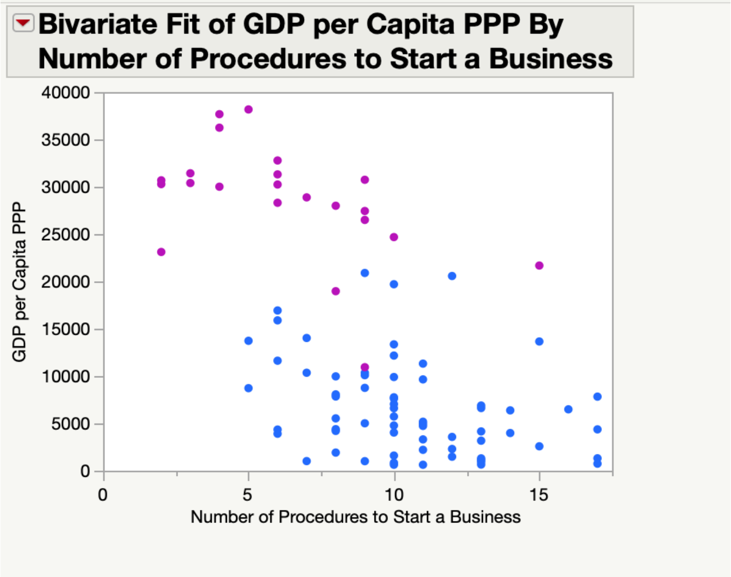
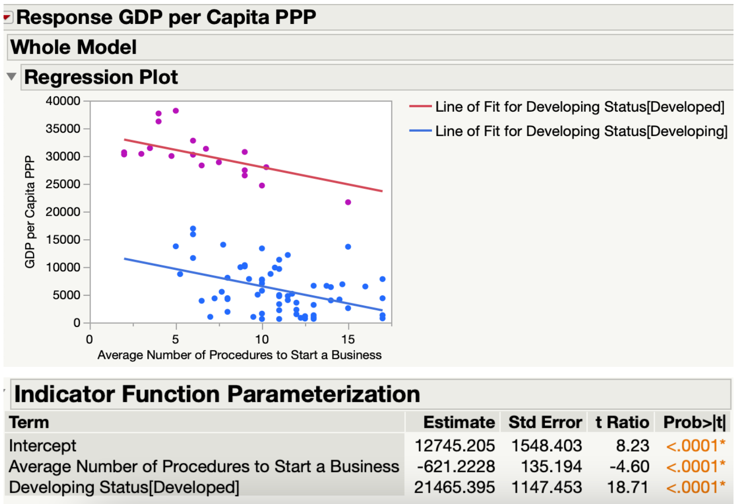


Figure 1: GDP per capita as a function of the number of procedures to start a business. Magenta signifies developed countries, and blue represents developing countries.

This scatterplot shows a possible relationship between GDP per capita and the number of procedures to start a business for developing and developed countries. Both types of countries have a negative linear relationship between the number of procedures to start a business and GDP per capita. Further analysis will be needed to determine a possible regression model for this data. Additionally, both groups appear to have several outliers that may affect the reliability of any hypothesized model. Upon additional analysis, the outliers New Zealand, Portugal, Chile, Slovenia, Czech, and Brazil were discovered. The preceding outliers have been dropped from this experiment to fit a more accurate model.

Part b. Model Fitting

The hypothesized model has one response variable (GDP per capita), one predictor variable (Number of Procedures to Start a Business), and one categorical variable (Developing Status; developing is coded as 1, developed is coded as 0). First, a full multiple additive regression model with interaction was fit through JMP. Multiple regression analysis with interaction reveals that fitted lines for this model produce nearly parallel slopes between the regression curves. Furthermore, JMP's numerical output is far more revealing. JMP reports that the R^2 value is 0.8916; this high value signifies that the additive regression model explains approximately 89 percent of the variation in GDP per capita is defined by the linear relationship with the number of procedures to start a business. Additionally, the p-value of the interaction term is 0.3982; this p-value is more prominent than any practical significance level. Therefore, the interaction term seems to explain minimal variation in our model. For this purpose, the interaction term was dropped, and the model was refit to an additive multiple regression model without interaction. When refit, the model has several interesting new behaviors. Subsequently, JMP provides the following output for the model.


Response GDP per Capita PPP

Whole Model

Regression Plot

40000
35000
30000
25000
20000
15000
10000
5000
0
GDP per Capita PPP
0
5
10
15
Average Number of Procedures to Start a Business
Line of Fit for Developing Status[Developed]
Line of Fit for Developing Status[Developing]

Indicator Function Parameterization

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	12745.205	1548.403	8.23	<.0001*
Average Number of Procedures to Start a Business	-621.2228	135.194	-4.60	<.0001*
Developing Status[Developed]	21465.395	1147.453	18.71	<.0001*

Figure 2: Multiple non-interactive additive regression model for GDP per capita as a function of the number of procedures to start a business. Magenta signifies developed countries, and blue signifies developing countries.

First, the R^2 value is slightly lower (at 0.8903), while the R^2 adjusted is marginally higher (0.000072 higher). While this new model causes slight inaccuracy, the loss of a superfluous term and a more straightforward structure makes it superior to the previous model.

Part c. Model Testing and Conclusions

The best way to test the model's practicality is to analyze the various residual plots.

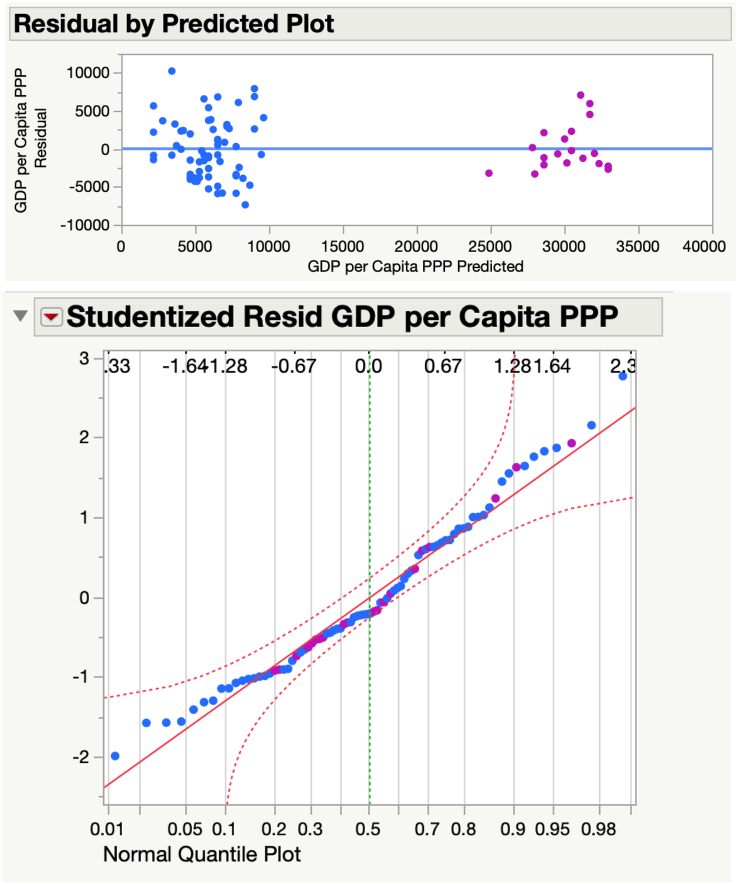


Figure 3: A plot of the standardized residuals as a function of GDP per capita (top) and a QQ plot of the standardized residuals (bottom). Magenta signifies developed countries, and blue represents developing countries.

JMP's residual analysis yields promising results. Looking at the residual plots, both of the different types of countries display significant and reasonably constant variance. Furthermore, the QQ plot for the residuals follows the pattern of a normal distribution. Because the residuals are normally distributed with a constant variance, the proposed model appears to meet the needed assumptions for a multiple linear regression model. Therefore, the additive non-interactive regression model is a good fit for the relationship between the average number of procedures to start a business and GDP per capita for both developing and developed countries. To summarize, developing and developed countries have a constant and significant difference between them according to this model. For example, this model indicates that the average difference between countries is 21465.395 GDP per capita when neither developing or developed countries have any procedures to start a business. Furthermore, both types of countries display a nearly identical trend; the more procedures to start a business, the less average GDP per capita. Another interesting observation is that developed countries universally have fewer procedures to start a business and larger GDP per capita; in contrast, developing countries generally have more procedures to start a business and lower GDP per capita.

Part ii. Simple Linear Regression for GDP and Infant Mortality Rate

Part a. Exploratory Analysis

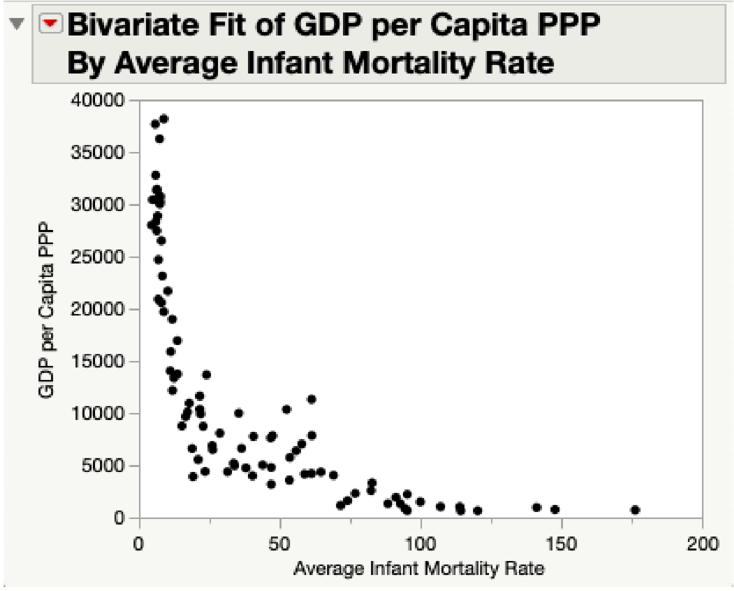


Figure 4: GDP per capita as a function of average infant mortality rate.

The scatterplot above shows a strong, negative, nonlinear association between infant mortality rate and GDP per capita. As the infant mortality rate increases, GDP per capita tends to decrease. There do not appear to be any outliers in the data.

Part b. Model Fitting

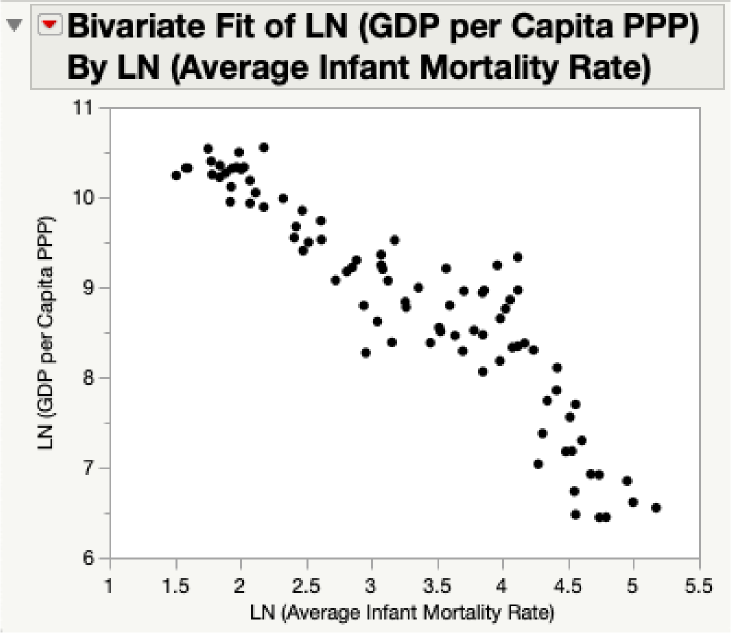
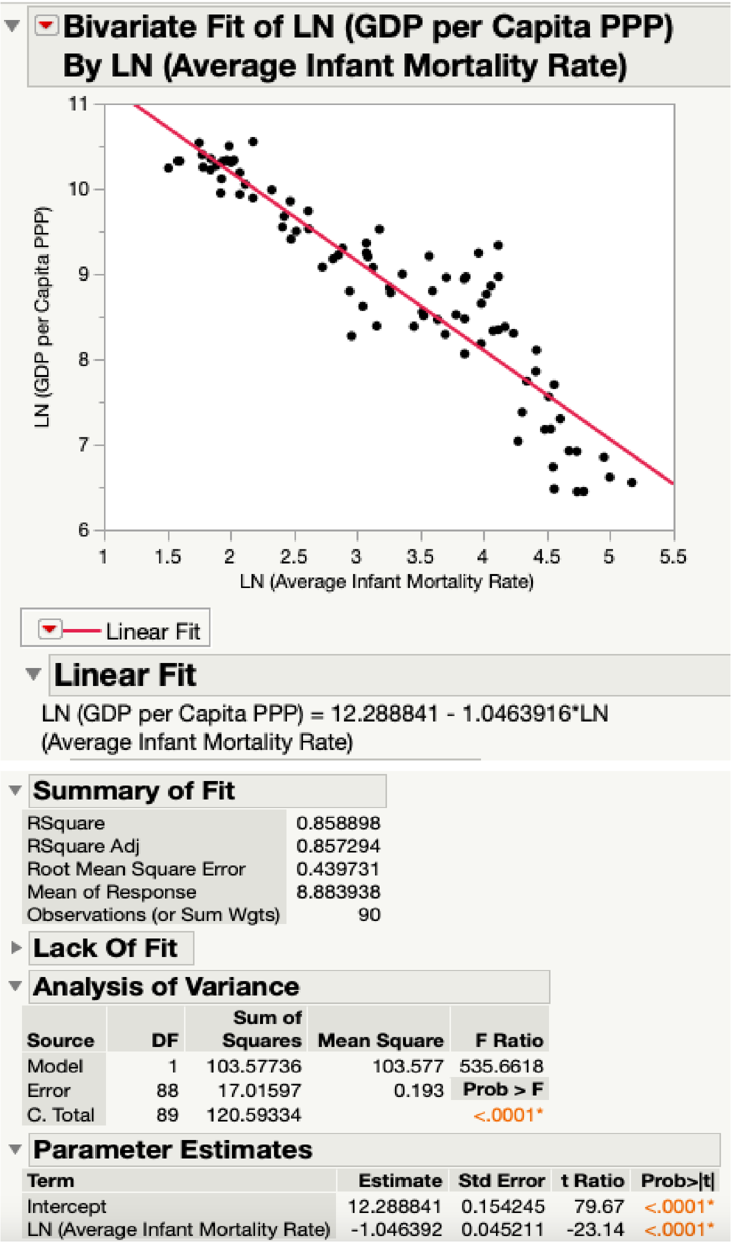


Figure 5: ln(GDP per capita) as a function of ln(average infant mortality rate).

Since the relationship between GDP per Capita and the average infant mortality rate was not linear, our group could not use a linear regression model to fit the data because the linearity assumption was violated. Our group decided to try different transformations on each variable and found that the variables ln(GDP per capita) and ln(average infant mortality rate) better satisfies the assumptions of the linear regression model. In the scatterplot above, there appears to be a negative linear relationship between these two transformed variables.



Bivariate Fit of LN (GDP per Capita PPP) By LN (Average Infant Mortality Rate)

Linear Fit

LN (GDP per Capita PPP) = 12.288841 - 1.0463916*LN (Average Infant Mortality Rate)

Summary of Fit

RSquare	0.858898
RSquare Adj	0.857294
Root Mean Square Error	0.439731
Mean of Response	8.883938
Observations (or Sum Wgts)	90

Lack Of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	103.57736	103.577	535.6618
Error	88	17.01597	0.193	Prob > F
C. Total	89	120.59334		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	12.288841	0.154245	79.67	<.0001*
LN (Average Infant Mortality Rate)	-1.046392	0.045211	-23.14	<.0001*

Figure 6: JMP output of a simple linear regression model for GDP per capita as a function of average infant mortality rate.

Using JMP, our group fitted a simple linear regression model to the transformed data. The model predicts that for every one unit increase in the ln(average Infant Mortality Rate), ln(GDP per capita) is expected to decrease by approximately -1.04639. The R^2 value is 0.8589, which indicates that 85.89% of the variation in ln(GDP per capita) is accounted for by the linear model.

The standard deviation of the residuals is 0.4397, which means that when the linear model is used to predict ln(GDP per capita), the error will typically be about 0.4397.

Part c. Model Testing and Conclusions

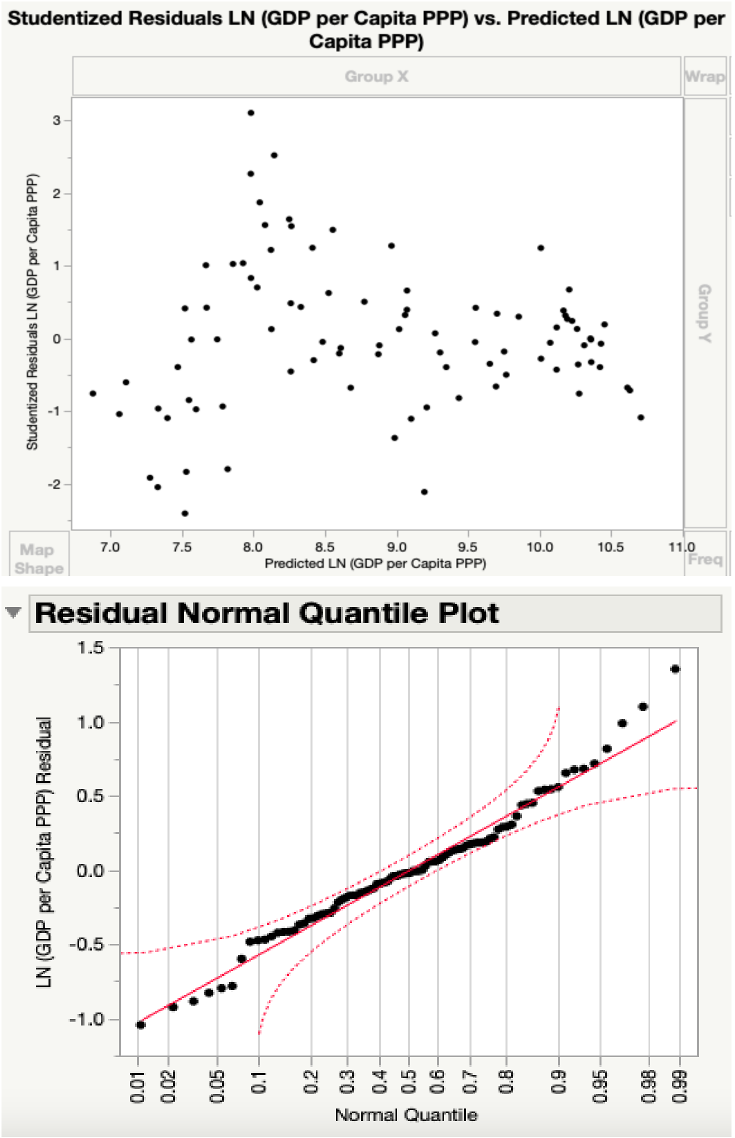


Figure 7: Residual plot of the studentized residuals as a function of ln(GDP per capita) (top). QQ plot of the studentized residuals (bottom).

The linear regression model assumes that the errors are independent and normally distributed with mean 0 and constant variance σ^2. To ensure that these assumptions are met, our group created a residual plot and a QQ-plot. In the QQ-plot, the plots all lie close to the y=x line, indicating that the residuals' distribution is normal; thus, the errors are normally distributed. In the residual plot, the points are randomly scattered, indicating that the variance is constant. Thus, the simple linear regression model is appropriate for modelling the relationship between ln(GDP per capita) and ln(average infant mortality rate). According to this model, our group can conclude that as ln(average infant mortality rate) increases, ln(GDP per capita) decreases.

Part iii. Simple Linear Regression for GDP and Life Expectancy

Part a. Exploratory Analysis

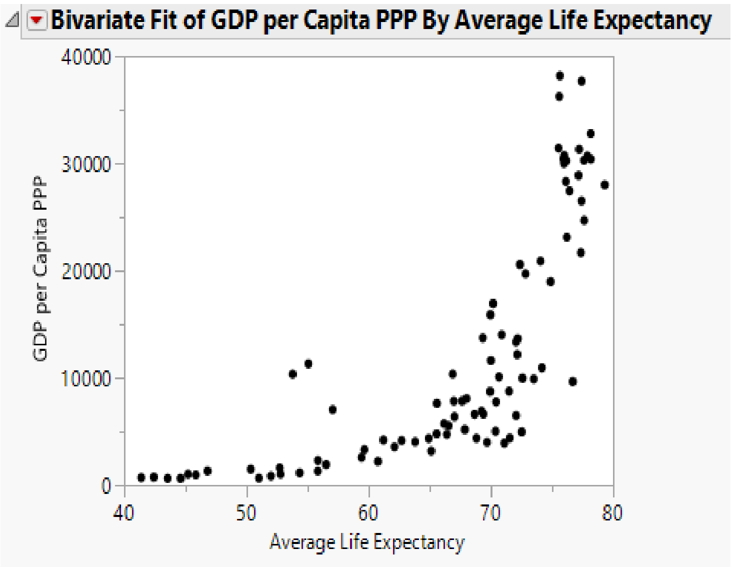


Figure 8: GDP per capita as a function of average life expectancy.

The scatterplot shows a strong, positive, and nonlinear relationship between GDP per capita and average life expectancy. Therefore, the GDP per capita will also increase when the average life expectancy increases. Note that there are several outliers in the plot. The outliers were not removed because they had no noticeable effect on the outcome of the model.

Part b. Model Fitting

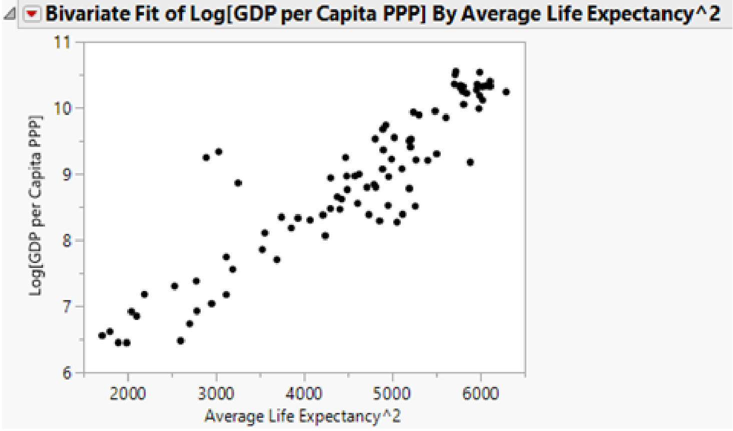
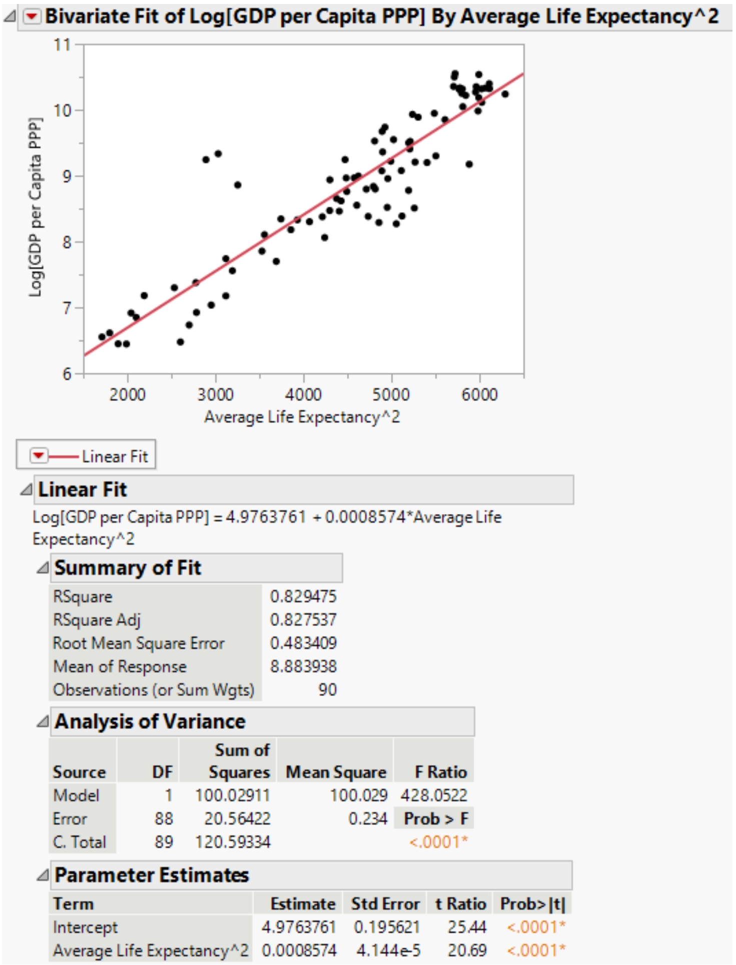


Figure 9: log(GDP per capita) as a function of (average life expectancy)^2.

The relationship between GDP per capita and the average life expectancy was not linear. It is inappropriate to use a linear regression model to fit the data while violating the linearity assumptions. After trying different transformations on each variable, our group concluded that the variables log(GDP per capita) and (average life expectancy)^2 satisfy the linear regression model's assumptions. Additionally, the scatter plot above indicates a positive linear relationship between the two transformed variables.



Bivariate Fit of Log[GDP per Capita PPP] By Average Life Expectancy^2

Linear Fit

Log[GDP per Capita PPP] = 4.9763761 + 0.0008574*Average Life Expectancy^2

Summary of Fit

RSquare	0.829475
RSquare Adj	0.827537
Root Mean Square Error	0.483409
Mean of Response	8.883938
Observations (or Sum Wgts)	90

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	100.02911	100.029	428.0522
Error	88	20.56422	0.234	Prob > F
C. Total	89	120.59334		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	4.9763761	0.195621	25.44	<.0001*
Average Life Expectancy^2	0.0008574	4.144e-5	20.69	<.0001*

Figure 10: JMP output of a simple linear regression model of GDP per capita as a function of average life expectancy.

The model predicts that for every one unit increase in (average life expectancy)^2, there is a corresponding 0.0008574 unit increase in the log (GDP per capita). The R^2 value is 0.8295 means 82.95% of the variation in log (GDP per capita) is explained through the regression on (average life expectancy)^2. Our group found that the p-value in the analysis of the variance table is less than 0.05, which indicates that the regression model statistically significantly predicts the outcome variable. Moreover, the p-value in the parameter estimates tables is also less than 0.05, which means the (average life expectancy)^2 has a statistically significant effect on log (GDP per capita ppp).

Part c. Model Testing and Conclusions

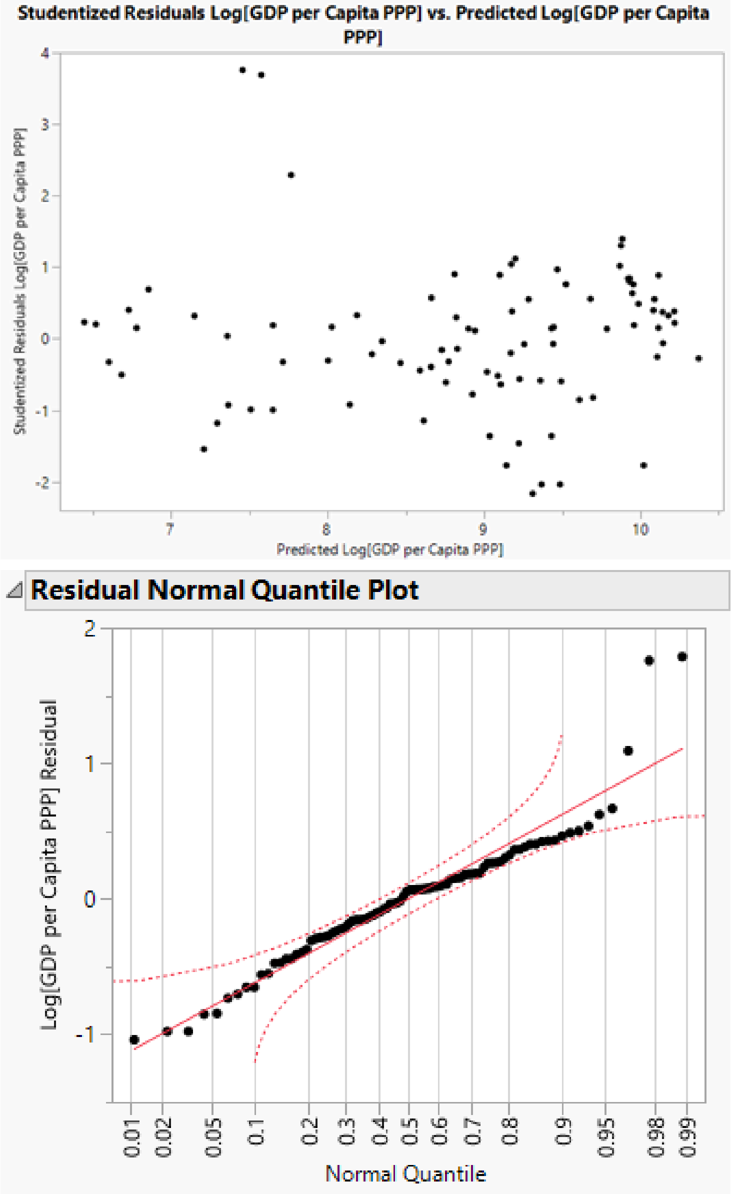


Figure 11: Residual plot of studentized residual as a function of log (GDP per capita) (top). QQ plot of the studentized residuals (bottom).

The linear regression model assumes that the error terms are independent and normally distributed with a mean 0 and a constant variance σ^2. By using the residual plot and the QQ plot, our group can verify the assumptions. The points are randomly spread out in the residual plot without a discernible pattern, indicating that the variance is constant. In the QQ plot, the points fall close to the y=x line (except the outliers). Our group concludes that the residuals have an approximately normal distribution. Therefore, the simple linear regression model is appropriate for this data. Thus, as (average life expectancy)^2 increases, the log(GDP per capita) increases.

Part iv. Multiple Linear Regression for GDP, Life Expectancy and Infant Mortality Rate

Part a. Exploratory Analysis

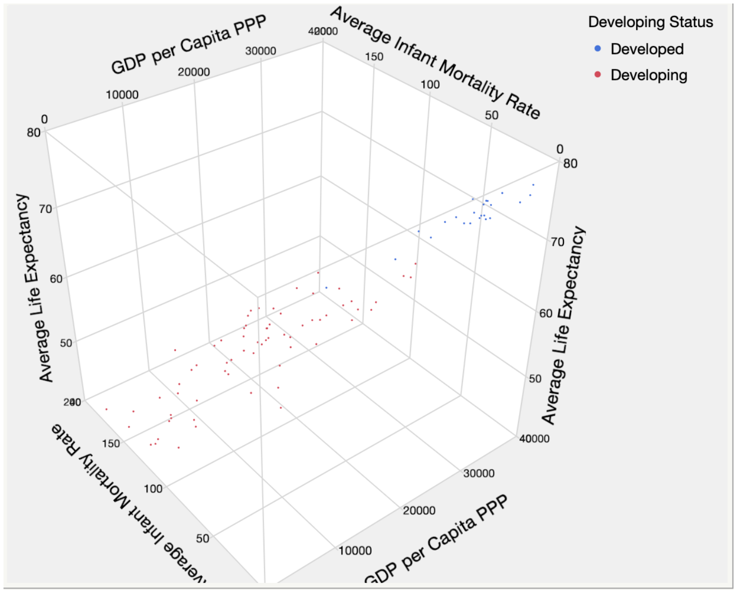


Figure 12: GDP per capita as a function of average life expectancy and average infant mortality rate.

There is a linear relationship between GDP per capita vs. average life expectancy and GDP per capita vs. average infant mortality rate as shown earlier. This 3D scatterplot shows that the relationship between GDP per capita and average mortality rate grows faster in developing countries than the relationship between GDP per capita and average life expectancy. Thus, our group will analyze this possible correlational relationship, grouping the two predictors into developed and developing countries.

Part b. Model Fitting

Diagnostic plots suggest that there will be a need for variable transformation to avoid the variance depending on the predictors; therefore, our group will use a natural log transformation.

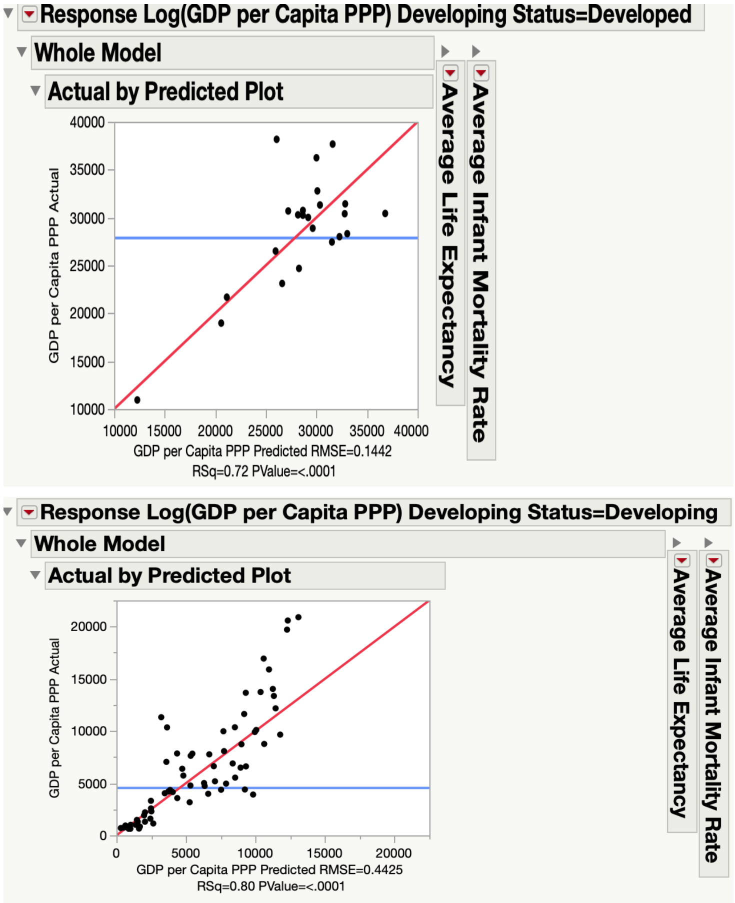
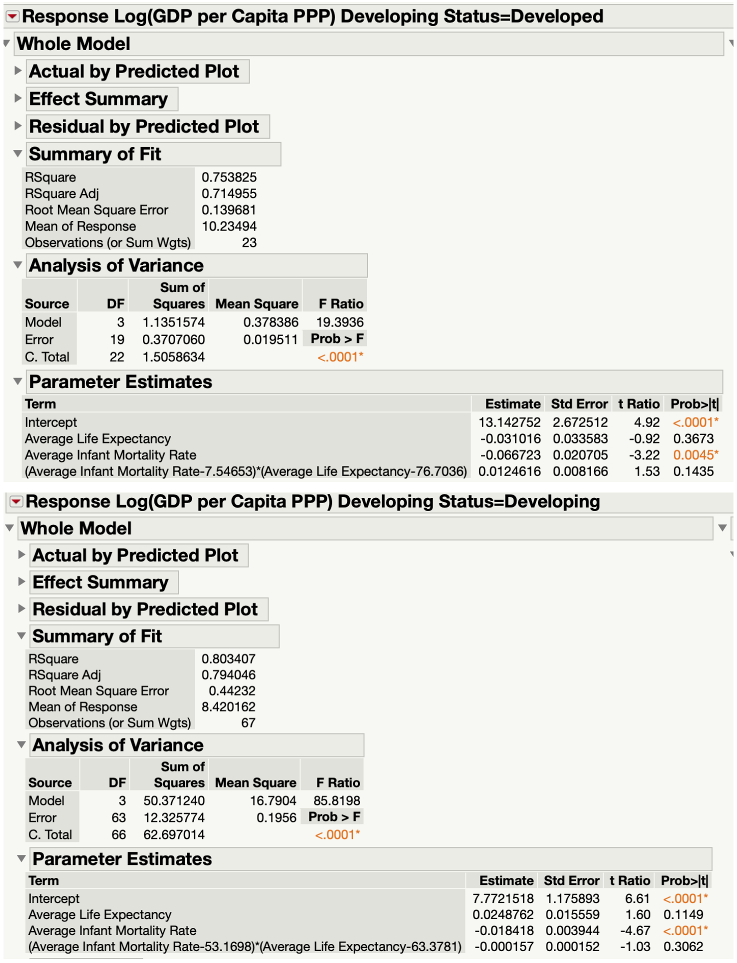


Figure 13: log (GDP per capita) as a function of (average life expectancy) and (average Infant Mortality Rate) for developed countries and developing countries.

The correlation term in both instances shows that more than 70% of the variation in GDP per capita can be explained through both of these variables. However, the p-value on the life expectancy slope parameter is more than 0.1 for both developed and developing countries. In this instance, the null hypothesis cannot be rejected; thus, the life expectancy term does not significantly predict the infant mortality rate in this model. Our group will investigate the relationship by adding a cross term in the predictors.



Response Log(GDP per Capita PPP) Developing Status=Developed

Whole Model

Actual by Predicted Plot

Effect Summary

Residual by Predicted Plot

Summary of Fit

RSquare	0.753825
RSquare Adj	0.714955
Root Mean Square Error	0.139681
Mean of Response	10.23494
Observations (or Sum Wgts)	23

Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	3	1.1351574	0.378386	19.3936
Error	19	0.3707060	0.019511	Prob > F
C. Total	22	1.5058634		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	13.142752	2.672512	4.92	<.0001*
Average Life Expectancy	-0.031016	0.033583	-0.92	0.3673
Average Infant Mortality Rate	-0.066723	0.020705	-3.22	0.0045*
(Average Infant Mortality Rate-7.54653)*(Average Life Expectancy-76.7036)	0.0124616	0.008166	1.53	0.1435

Response Log(GDP per Capita PPP) Developing Status=Developing

Whole Model

Actual by Predicted Plot

Effect Summary

Residual by Predicted Plot

Summary of Fit

RSquare	0.803407
RSquare Adj	0.794046
Root Mean Square Error	0.44232
Mean of Response	8.420162
Observations (or Sum Wgts)	67

Analysis of Variance

		Sum of		
Source	DF	Squares	Mean Square	F Ratio
Model	3	50.371240	16.7904	85.8198
Error	63	12.325774	0.1956	Prob > F
C. Total	66	62.697014		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.7721518	1.175893	6.61	<.0001*
Average Life Expectancy	0.0248762	0.015559	1.60	0.1149
Average Infant Mortality Rate	-0.018418	0.003944	-4.67	<.0001*
(Average Infant Mortality Rate-53.1698)*(Average Life Expectancy-63.3781)	-0.000157	0.000152	-1.03	0.3062

Figure 14: JMP output of multiple linear regression of log (GDP per capita) as a function of (average life expectancy) and (average Infant mortality rate) for developed countries (top). JMP output of multiple linear regression of log (GDP per capita) as a function of (average life expectancy) and (average Infant mortality rate) for developing countries (bottom).

Even when the cross term is included, the p-value on the average life expectancy renders it insignificant in both instances. Furthermore, the p-value also renders the cross term between the average infant mortality rate and the average life expectancy insignificant. This effect confirms our group's suspicion; due to the nature of these predictors, there is no need to include both. Therefore, our group would instead include the average infant mortality rate. Hence, for every 1 unit of increase in the infant mortality rate, there is a negative increase in the natural log of GDP per capita by 0.018 while other predictors remain fixed.

Part c. Model Testing and Conclusions

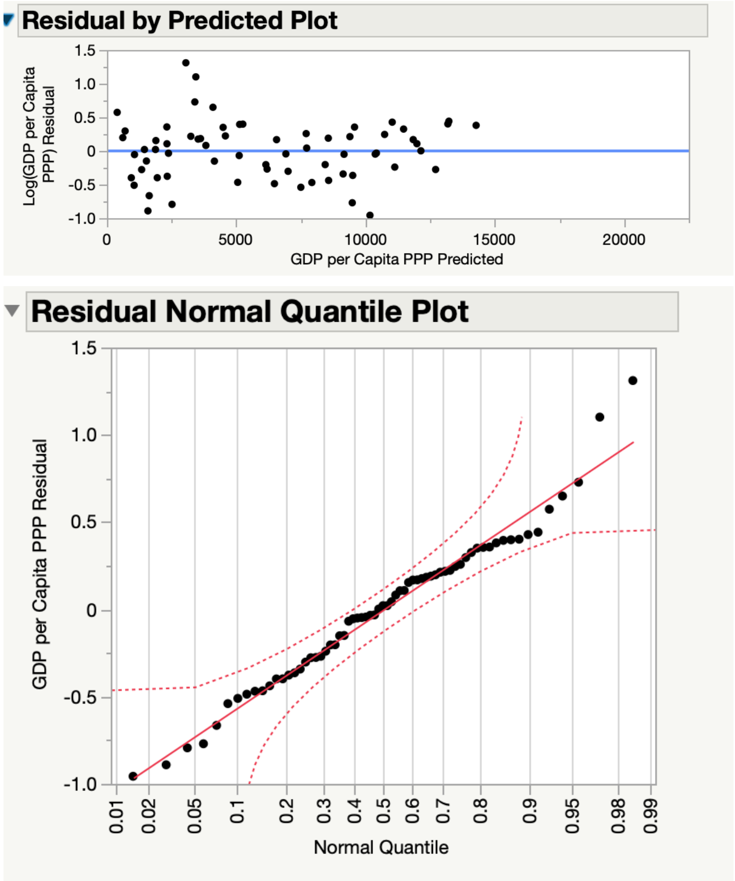


Figure 16: Residual plot of studentized residual as a function of log (GDP per capita) (top). QQ plot of the studentized residuals (bottom).

The linear regression model assumes that the error terms are independent and normally distributed with a mean 0 and a constant variance σ^2. Using the residual plot and the QQ plot, we can confirm these assumptions. The points are randomly spread out with no discernible pattern in the residual plot, indicating that the variance is independent. In the QQ plot, the points fall close to the y=x line other than the tails of the line; hence, our group feels confident in asserting the assumptions of the normal distribution. Therefore, the simple linear regression model is appropriate for this data.

Conclusions and Recommendations

The experiment yielded several interesting results on multiple indices. First, both developed and developing countries show a negative linear relationship between the average number of procedures to start a business and GDP per capita. According to the proposed model, the difference in the average GDP per capita between developed and developing countries is 21465.395 GDP for any specific number of procedures to start a business. Next, countries that have high GDP per capita tend to have lower average infant mortality rates. Fitting a linear regression model to the transformed variables ln(GDP per capita) and ln(Average infant mortality rate) shows a negative linear relationship between these two variables. Moreover, countries with high GDP per capita have a higher average life expectancy compared to others. Next, The results of the fitted regression model of the transformed variables log(GDP per capita) and (average life expectancy)^2 indicate a positive linear relationship between those variables. Therefore, as a country experiences higher life expectancy, it will experience a drastically improved GDP per capita. Finally, our group performed a series of analyses to determine a possible multiple additive regression model for GDP per capita as a function of both average infant mortality and average life expectancy for both developing and developed countries. Ultimately, life expectancy did not significantly impact the infant mortality predictor, so the interaction term was dropped, and the model was refit. The refit model was GDP per capita as a function of log(average infant mortality). Once refit, the model passed all residual tests and predicted the model behavior accurately. While the study results were promising, several additional analyses could have been performed on this dataset. For example, both forwards and backward selection could be used to determine a more comprehensive model. Furthermore, the differing world bank regions from the original dataset could have factored into the analysis (for example, an ANOVA analysis of GDP per capita for the differing areas). In conclusion, the Friedman dataset continues to be a fascinating source for analyzing economic activity more than a decade after its initial creation.

References

(October 2018). Database—WEO Groups and Aggregates Information. *International Monetary Fund*. https://www.imf.org/external/pubs/ft/weo/2018/02/weodata/groups.htm