Predicting GDP from Demographic Information

Jaipal Sandhu, Lakshmi S. Menon, Mohammed Khawar Khan, Sahana Srinivasan
University of California San Diego

Introduction

Countries around the world are characterized by many relevant attributes that define their economic, social, and health standards. Among these attributes are population, infant mortality rates, climate, agricultural rates, and birth rates. Often, some of these attributes can be used to predict the outcome of other relevant qualities of each country. In other words, it is likely that there is some correlation between some of the variables that are used to represent the countries. Though some of these correlations may be intuitively hypothesized based off of the socioeconomic standing of a country, other correlations may be more ambiguous and hard to observe without a proper linear regression analysis. The 'Countries of the World' dataset that our team found on Kaggle contains 19 attributes for 220 different countries, and we decided to use the method of linear regression to observe any trends in data based off of the use of either single or multiple input variables and a single output variable. We decided to focus on GDP as our output variable because we hypothesized that GDP is highly dependent upon the social, cultural, land/population attributes, and literacy rates of a country.

Research Question: Can the Gross Domestic Product (GDP) of a country be predicted from its demographic and economic information?

Procedure

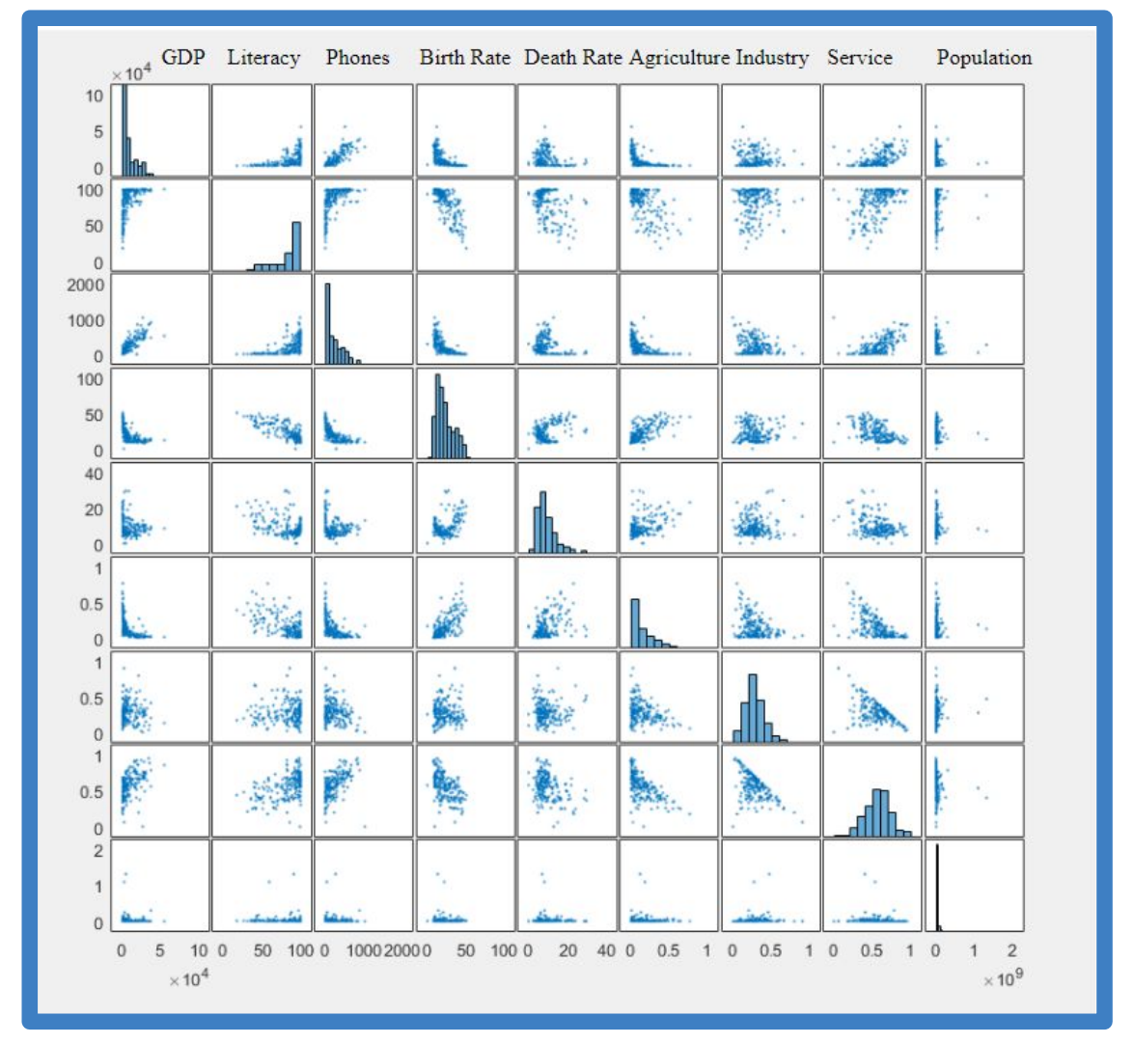
Data Cleaning and Preprocessing (Done in Excel):

- The original dataset used commas as decimal points, so we first had to replace those in order to enable the data to be read as coherent numeric data.
- We created a scatter plot matrix showing each variable against all the other variables. This gave us a preliminary idea about which variables may have a correlation and would be interesting to analyze.
- We then narrowed down our data to 9 variables.
- We removed any rows which were missing values for three or more variables. A total of 15 samples were removed. The resulting dataset contained 212 samples and 9 features.
- The 'literacy' variable appeared to have no entry for some samples, which we assumed was due to a lack of data. In order to prevent our analysis from being skewed by these outliers, we assigned the mean value of 'literacy' to each of these points. A total of 13 samples were assigned the mean value of 82.3.

Modelling:

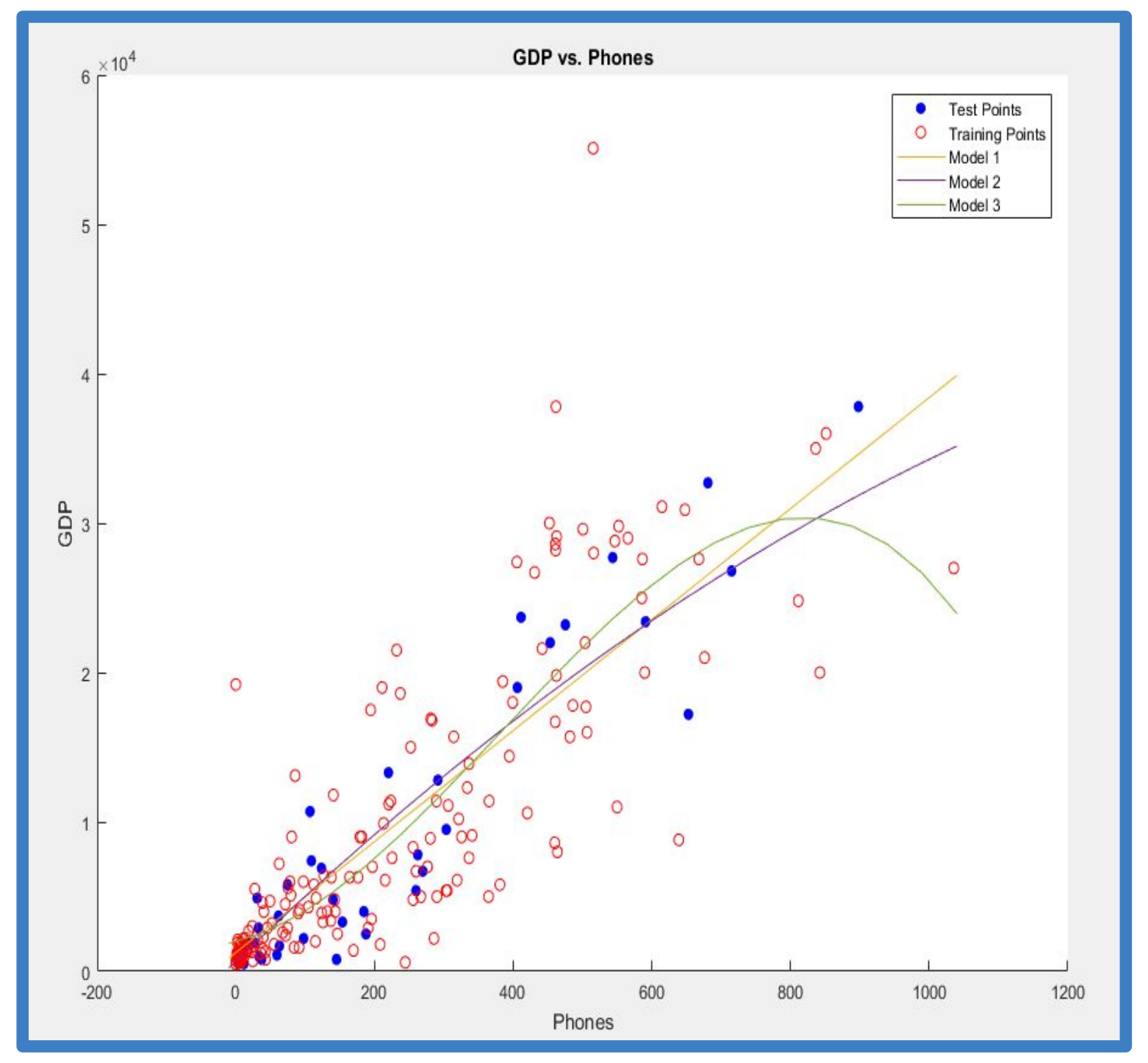
Since we were trying to predict GDP from other features, we considered the data to be labelled and decided to do linear regression. We chose two univariate correlations (GDP vs. Phones and GDP vs. Literacy) and two multivariate correlations (GDP vs. Agriculture, Industry, and Service and GDP vs. Birth Rate and Death Rate). For each correlation, we tried three models.

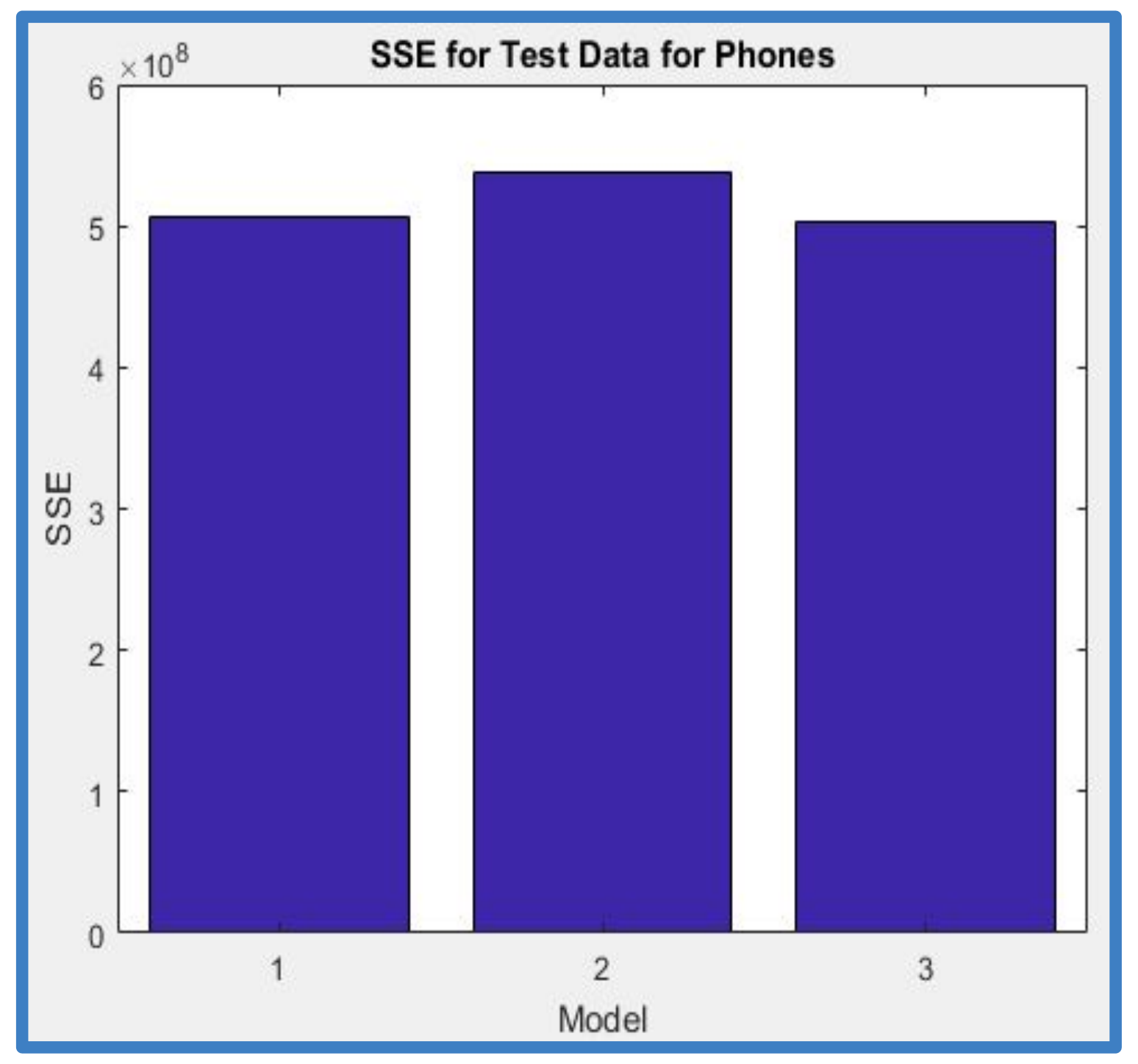
General Form of Models:
Model 1: GDP = w0 + w1*X
Model 2: GDP = w0 + w1*X + w2*X^2
Model 3: GDP = w0 + w1*X + w2*X^2 + w3*X^3



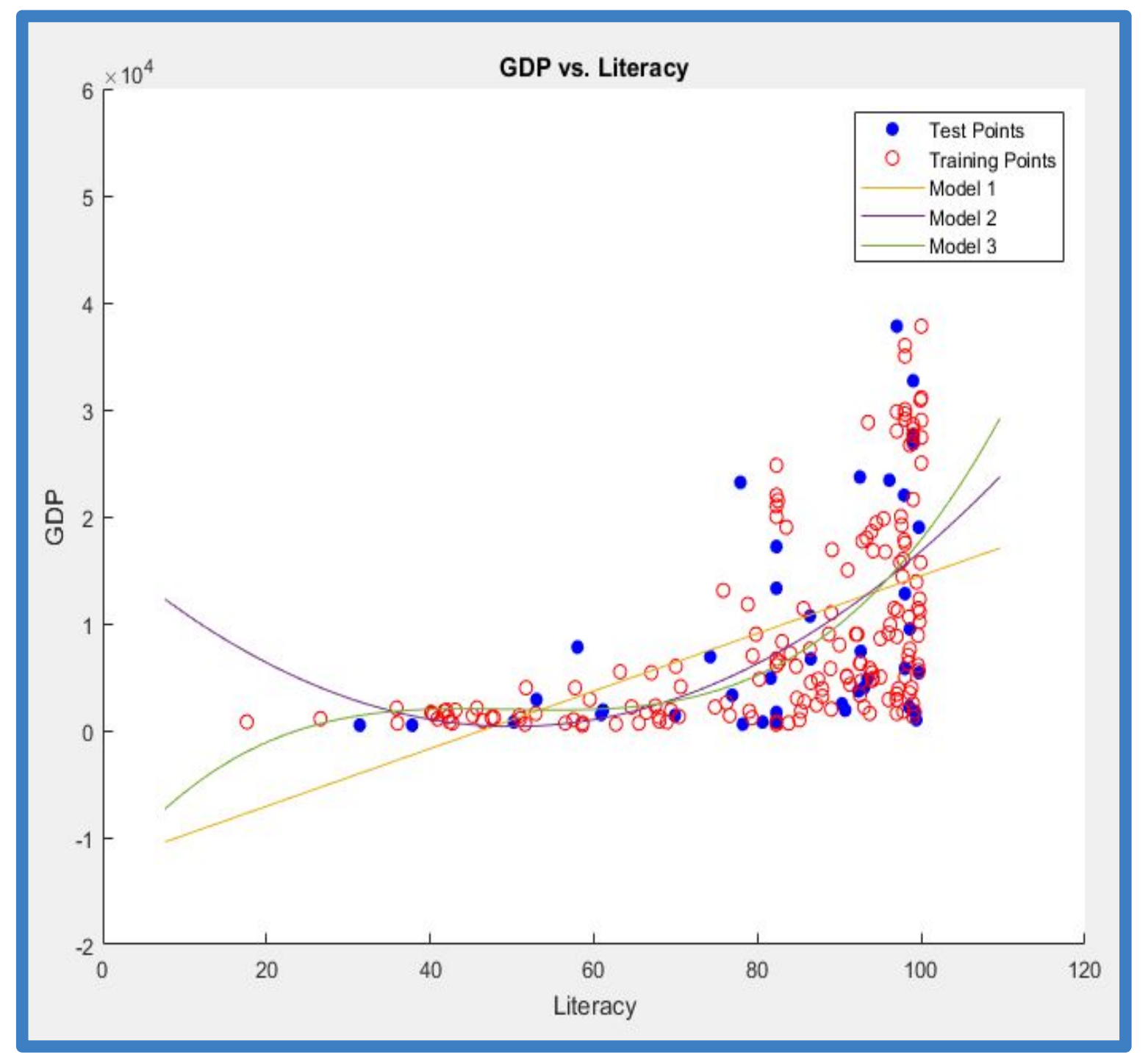
Results

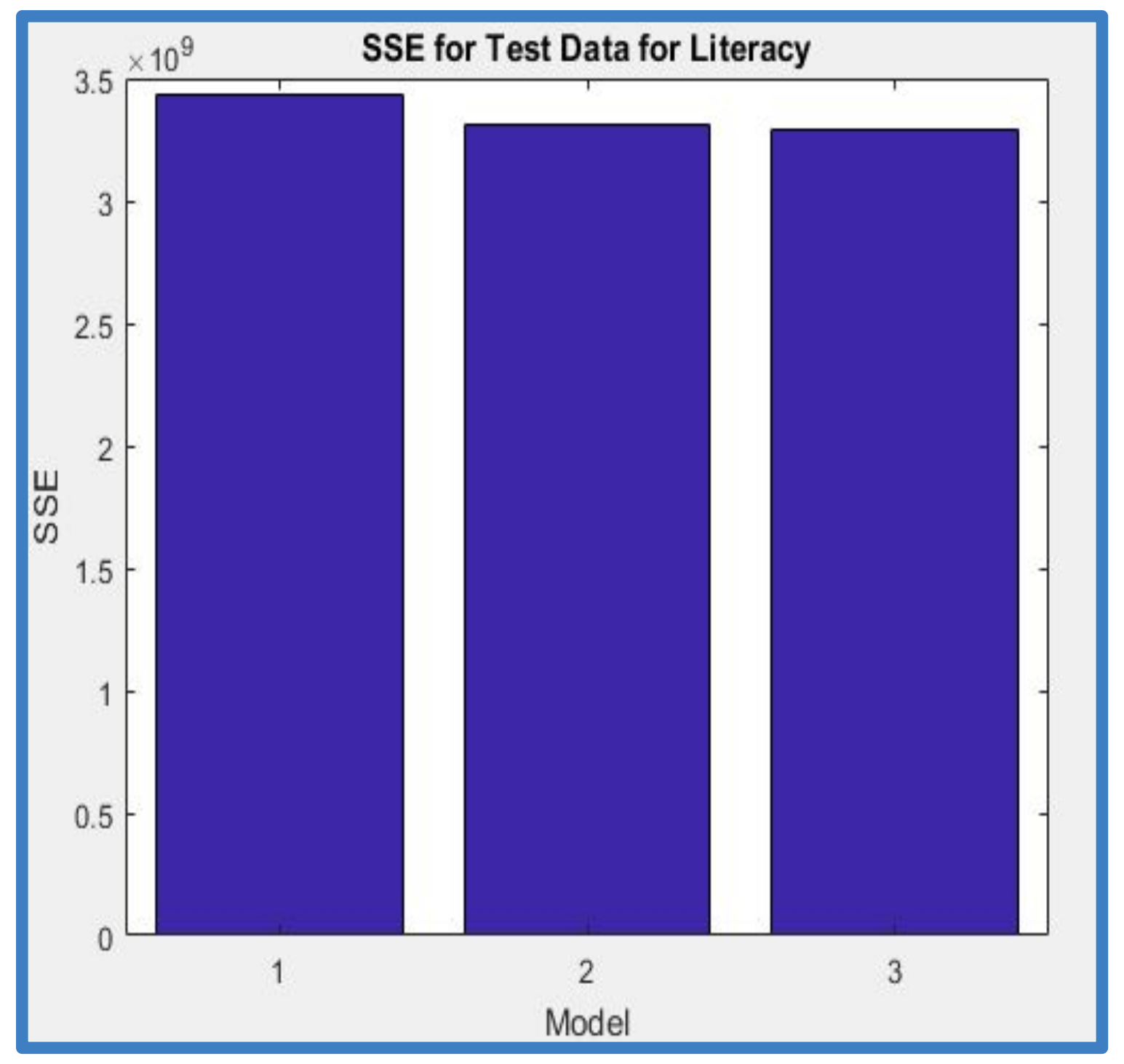
Univariate Models:





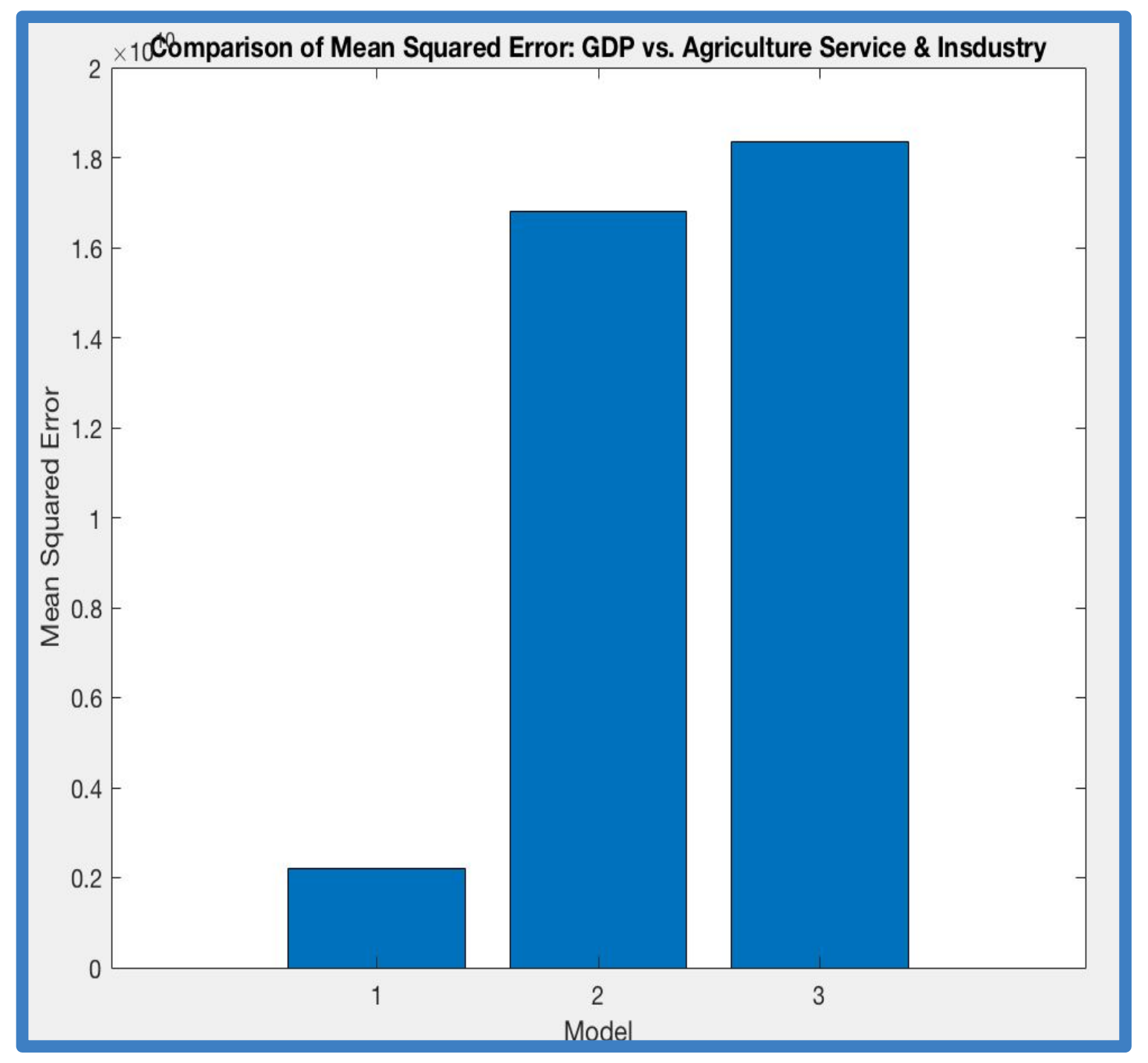
Model 1: GDP = 1257.3 + 37.1*Phones
Model 2: GDP = 667.4 + 44.6*Phones - 0.011*Phones^2
Model 3: GDP = 1977.7 + 11.5*Phones + 0.1*Phones^2 - 8.6e-5*Phones^3

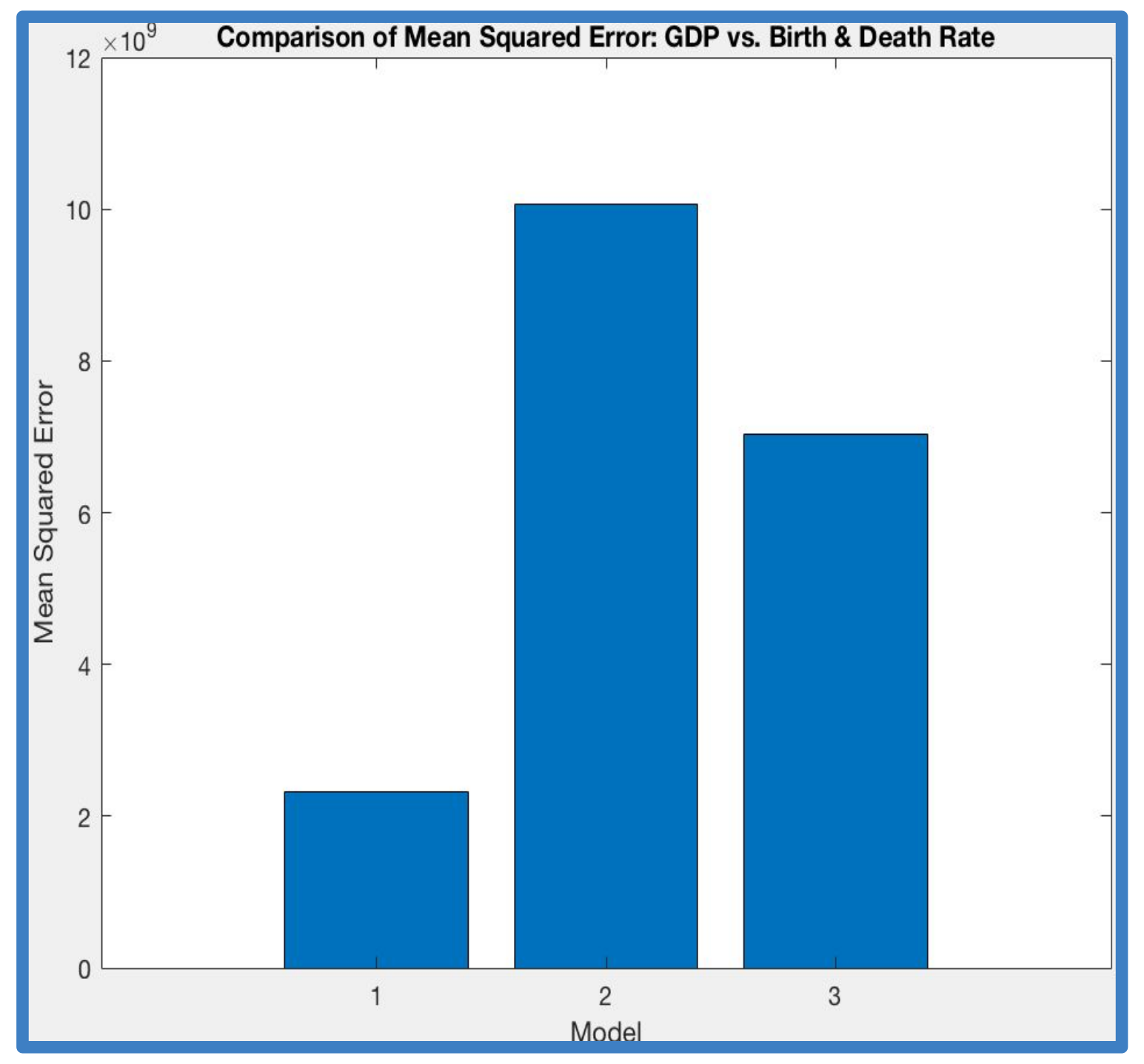




Model 1: -1.2467e+04 + 269.38*Literacy
Model 2: 1.6916e+04 - 660.3319*Literacy + 6.5910*Literacy^2
Model 3: -1.3593e+04 + 964.1021*Literacy - 19,6129*Literacy^2 + 0.1312*Literacy^3

Multivariate Models:





Best Models: GDP = 2.1146e+04 - 0.0598e+04*Birth Rate + 0.0192e+04*Death Rate
GDP = 3.6344e+04 -6.0250e+04*Agriculture -3.2875e+04*Industry -1.5323e+04*Service

Discussion

Both of our univariate models demonstrate that GDP can be predicted as an output of the amount of phones and the literacy rates that are observed in each country. The final model of the 'Phones v. GDP' graph yielded the lowest SSE_{test} value out of the three models we created. Generally, the trend appears to be that, for a greater number of phones, a country's GDP increases proportionally. Though some outliers skew the data, the trend is consistent with the expectation for the most part. The 'Literacy v. GDP' graph exhibits a similar trend, but the SSE_{test} are slightly higher than the values given by the 'Literacy v. Phones' model. Again, the third model yields the lowest SSE_{test} values, which informs us that it is best fit to represent the dataset.. The parabolic curve tells us that a greater literacy rate is mostly correlated with a higher GDP, with some exceptions. Neither graph is linear and, as a consequence, none of the trends are necessarily directly proportional.

Our approach to analyzing the multivariate data was slightly different. In our 'GDP v. Birth and Death Rate,' our first model has a fairly low SSE_{test} value but our second and third models have much higher values, so we know that a linear graph is the best representation of the dataset. From the weights, we know that the correlation of birthrate to GDP is slightly negative and the correlation of death rate to GDP is slightly positive. The next multivariate model was based off Agriculture, Industry, and Service as inputs and GDP as the output. The weights show that each variable has a negative correlation to GDP. Again, the first model is the most powerful representation of the given values, thus a linear fit is the most appropriate.

Conclusion

Our goal in creating the models was to be able to extrapolate further output values (GDP values) based off of various future possible values for our input variables. The effectiveness in each of the models we created in allowing us to extrapolate future data varies, with some graphs exhibiting a better fit to the data and subsequently allowing for a more appropriate analysis of future potential data points.

For further research, it would be helpful to analyze more models with different combinations of variables, because although our models show a general trend, they may not be as strong for precise prediction. A limitation of our study is that we focused on only some of the variables provided in the dataset, as well as the fact that we restricted our models to a limited number of combinations which we thought may be useful.

References

The dataset for this project was chosen from the collection of datasets made available on kaggle.com.

Link to dataset: https://www.kaggle.com/fernandol/countries-of-the-world