MSDS 6372 Project 1

Infant Mortality Case Study

**Team members:**

Kottegoda Manjula

Messa Araya

Kumar Rajeev

Date:2016-09-20

**Introduction**

Our team decided to analyze the world infant mortality rate, a widely used indicator of the level of health in a country, to see what factors influence this statistic. Infant mortality rate is the probability of a child born in a specific year or period dying before reaching the age of one, if subject to age-specific mortality rates of that period and is expressed as rate per 1000 live births (1). The data for the study was gathered from the CIA world fact book website. Our goal for the project was to find which characteristics of a nation contribute to the infant mortality rate of that nation. We analyzed eleven explanatory variables this study.

The explanatory variables looked at in the study are Population, Median Age of the country, Population growth rate, Birth rate, Maternal death rate, Health expenditure by government, Gross Domestic Product, Tax Revenue, Unemployment rate, Rate of Obesity and the Total fertility rate of the country. We wanted to see which factors were correlated and which factors had a direct effect on this measurement.

**Descriptive Statistics**

As explained on the introduction, we decided to work with the response variable of infant mortality rate. To begin our study, we need to determine what factors have an effect on infant mortality rate. We went through a process of cleaning through the data and making sure we looked at essential information and complete data for the most influential variables we wanted to use. We worked on all the explanatory variables to be quantifiable values. After exploring different groups of dataset as explanatory variables with complete data we ended up with a dataset of only 155 from the initial 238 observations.

We first look at the initial scatter plots of the original (non-transformed) data to determine the rough relationship of variables to each other. As it can be observed from the scatter plot matrix (Fig 1) some of the explanatory variables have linear relationship that indicates a correlation to each other.



Fig. 1 Scatter plot Matrix

Also we compared the scatter plots of each explanatory variables with response variable individually (Fig. 2) to look at how they are related with the response. Some of the explanatory variables such as birthrate have a strong linear association with the response.

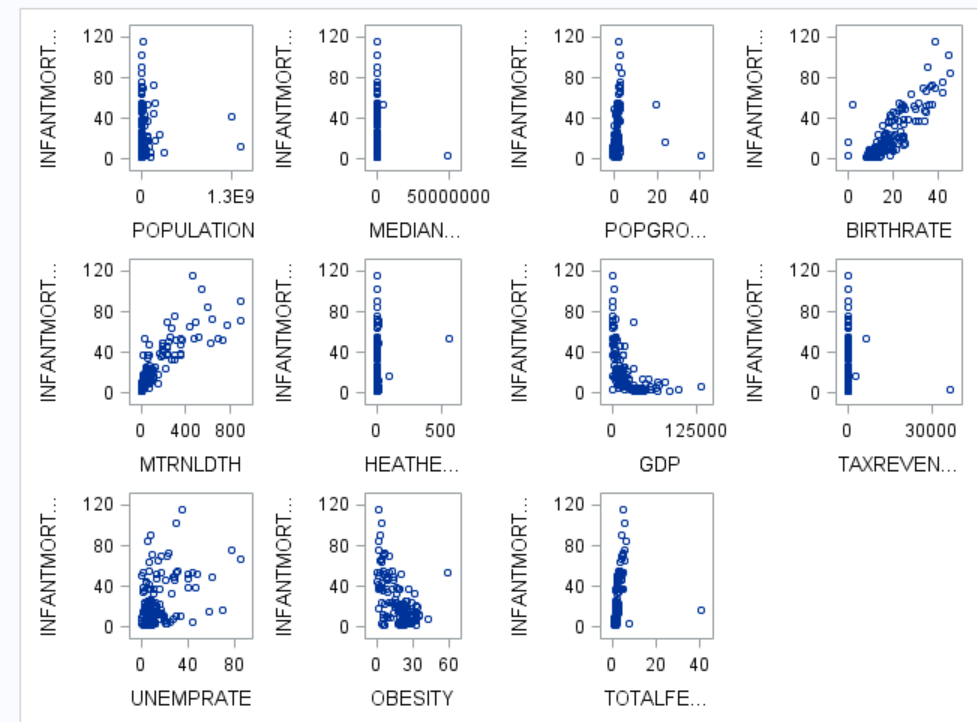


Fig. 2 Scatter plots for explanatory vs. response variables

The individual distributions are not necessarily important if the overall effect (residual plots) are satisfied with the linear regression assumptions. For that reason, we inspected the residual plots Cook’s D (Fig. 3) and Studentized residual (Fig. 4). The cook’s D shows all the plots are less than 0.25 which is not a problem. There leverages and outliers shown on the Studentized residual plot however they have low influence on the regression.

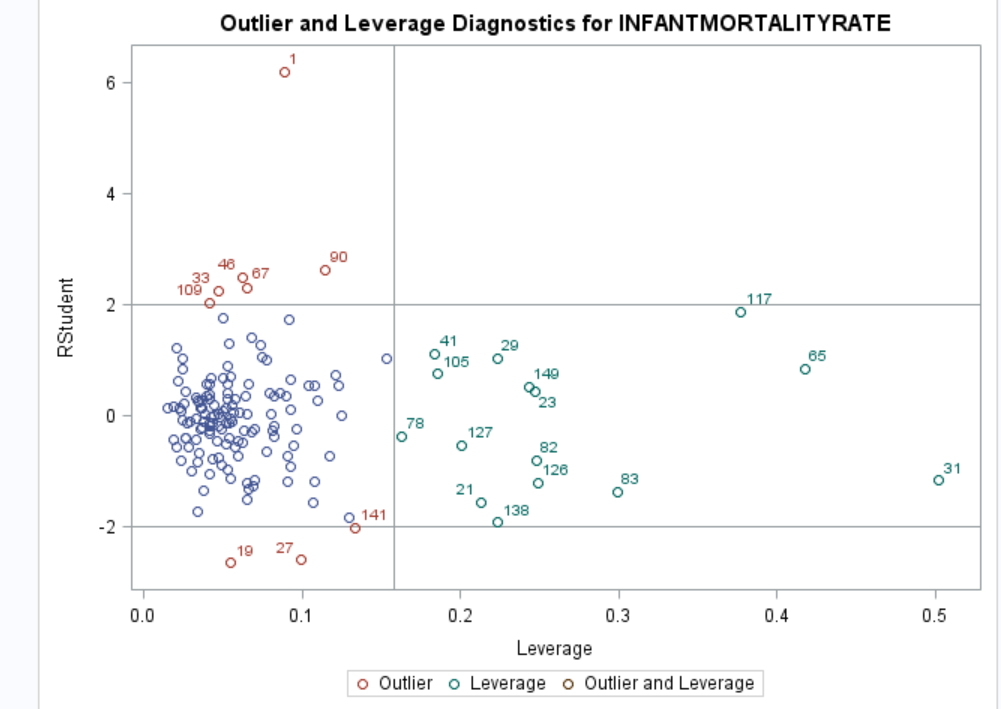
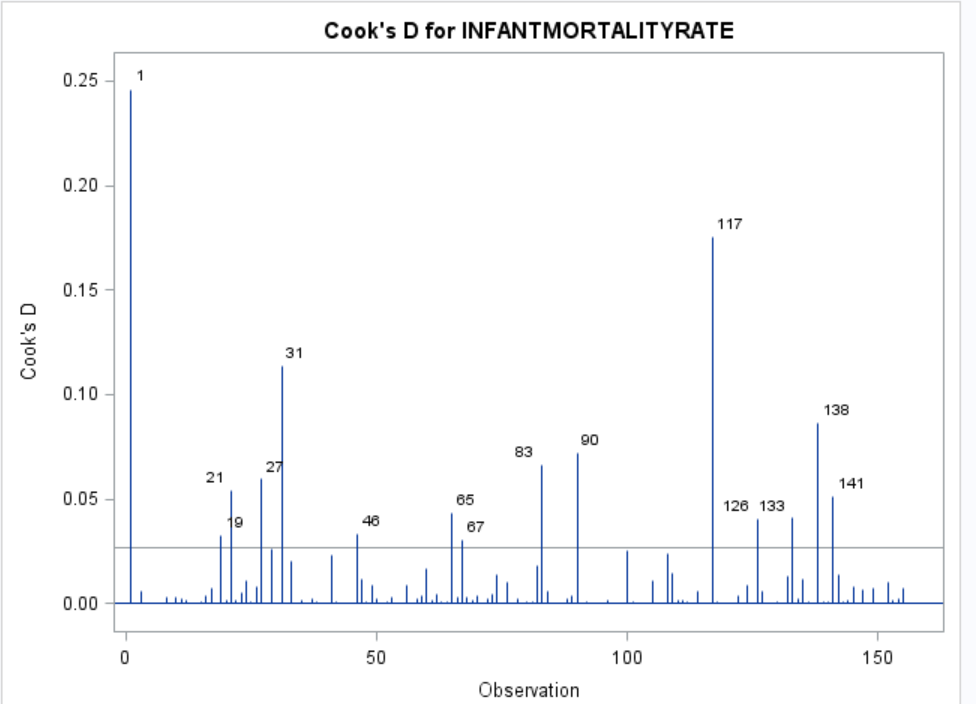


Fig. 4 Studentized residual distribution

Fig. 3 Cook’s D

In examining the residual plots (Fig. 5) the residuals of the variables are not randomly distributed. The variability is not uniformly distributed thought out the residual distribution points. This tells us that the data need some transformations.

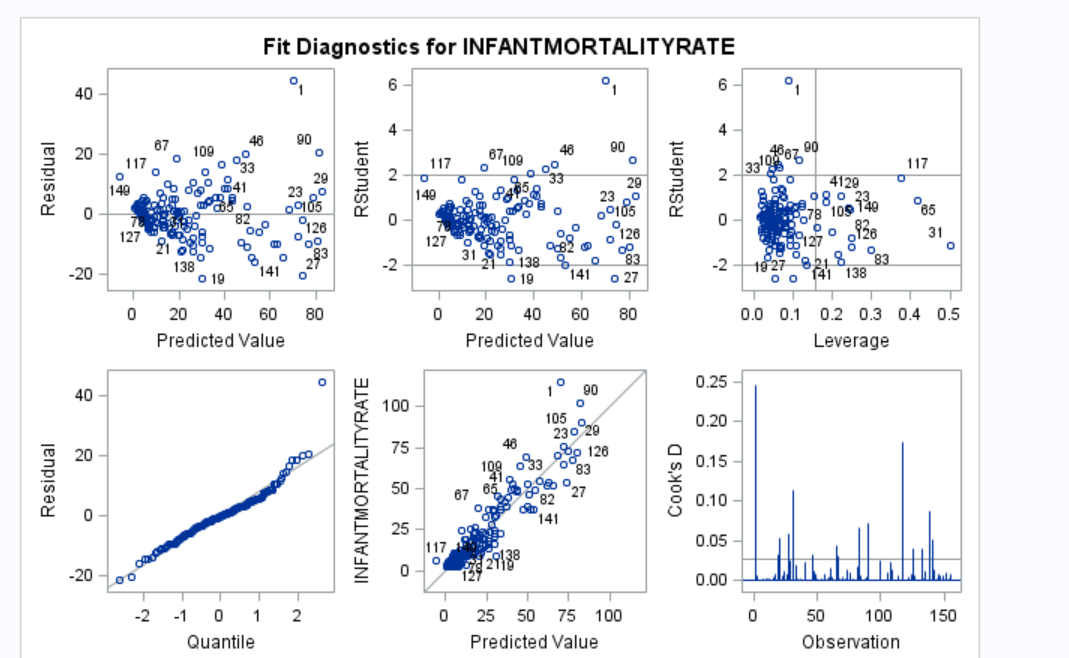


Fig. 5 Fit Diagnostic plots

There are significant differences in the distribution of the residuals for each regressor in Fig 6. This is another information that we need to transform the data in order to fit the assumptions of a linear regression model.

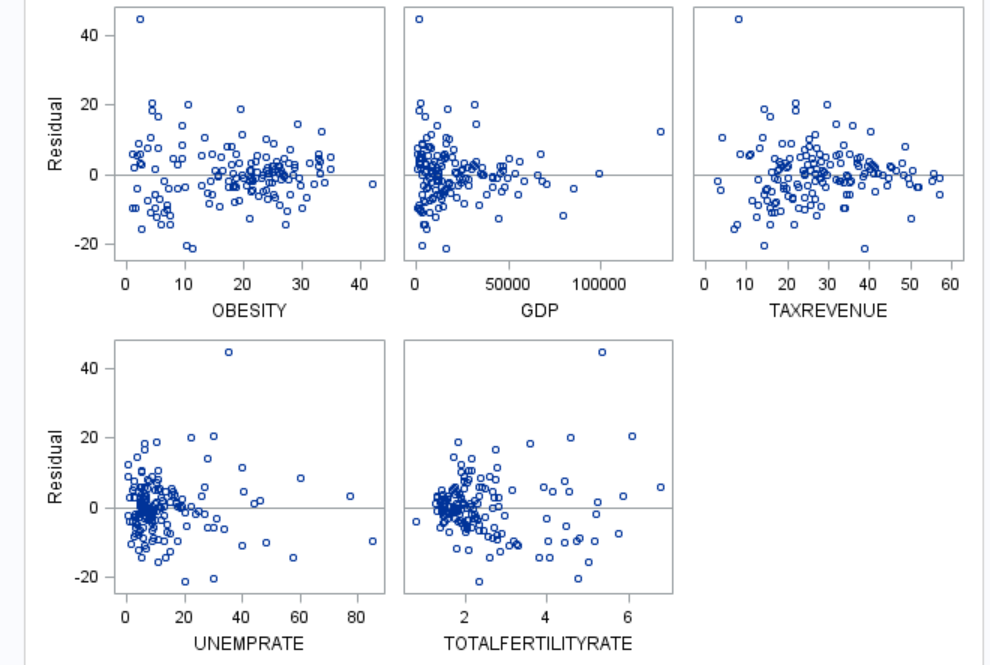
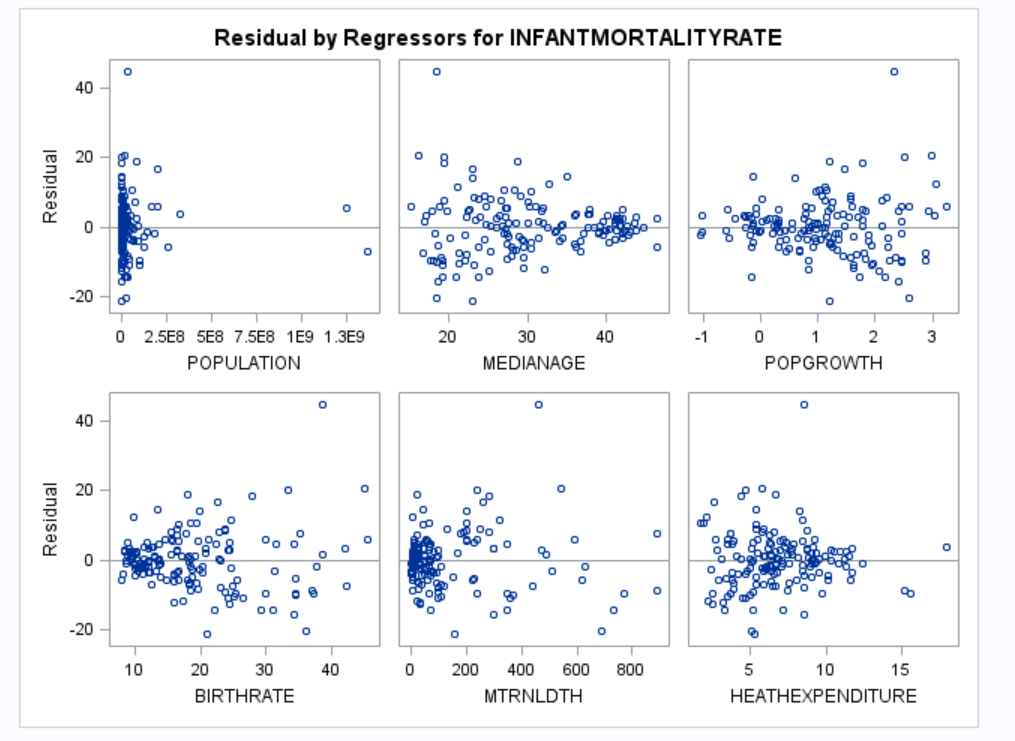


Fig. 6 Residual distribution for individual repressor variables

Before transformation, however, those that are highly correlated variables are filtered using correlation matrix (Table 1) and the variance inflation factor, VIF (Table 2). Based on that criteria we removed four variables (birthdate, population growth, maternity death, and total fertility rate).

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1 Correlation Table (Partial Snapshot)** | | | | | | | | | | | | |
| **Variable** | **POPULATION** | **MEDIANAGE** | **POPGROWTH** | **BIRTHRATE** | **MTRNLDTH** | **HEATHEXPENDITURE** | **OBESITY** | **GDP** | **TAXREVENUE** | **UNEMPRATE** | **TOTALFERTILITYRATE** | **INFANTMORTALITYRATE** |
| **POPULATION** | 1.0000 | 0.0309 | -0.0221 | -0.0371 | 0.0023 | -0.0615 | -0.2040 | -0.0638 | -0.1803 | -0.0980 | -0.0369 | 0.0301 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 2 Parameter Estimates** | | | | | | |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** |
| **Intercept** | **1** | -9.90132 | 14.65476 | -0.68 | 0.5004 | 0 |
| **POPULATION** | **1** | 2.221289E-9 | 4.759207E-9 | 0.47 | 0.6414 | 1.13294 |
| **MEDIANAGE** | **1** | 0.27147 | 0.33530 | 0.81 | 0.4195 | 16.54200 |
| **POPGROWTH** | **1** | -1.08842 | 1.72270 | -0.63 | 0.5285 | 5.62747 |
| **BIRTHRATE** | **1** | 2.08309 | 0.75644 | 2.75 | 0.0067 | 92.80646 |
| **MTRNLDTH** | **1** | 0.04092 | 0.00666 | 6.14 | <.0001 | 3.23993 |
| **HEATHEXPENDITURE** | **1** | -0.29763 | 0.30795 | -0.97 | 0.3355 | 1.51907 |
| **OBESITY** | **1** | -0.35050 | 0.09795 | -3.58 | 0.0005 | 1.89215 |
| **GDP** | **1** | -0.00000467 | 0.00005956 | -0.08 | 0.9376 | 3.32547 |
| **TAXREVENUE** | **1** | -0.05494 | 0.07406 | -0.74 | 0.4594 | 1.67995 |
| **UNEMPRATE** | **1** | 0.09894 | 0.05788 | 1.71 | 0.0896 | 1.30952 |
| **TOTALFERTILITYRATE** | **1** | -3.95380 | 4.07130 | -0.97 | 0.3332 | 45.74967 |

Looking at the residual and scatter plots the most appropriate options will be the Log() transformation. This transformation will uniform the variability and normalize the distribution. Our transformation was done only for the response variable.

**Analysis**

*Linear regression is appropriate. Because the following major simple linear regression assumptions are satisfied.*

* ***Linearity between variables****-The linearity assumption can be assessed from the predicted vs. residual/Studentized residual scatter plots (Fig. 7). Although there is widespread scatter in the residual plot, there is no apparent pattern or relationship between the residuals and predicted values. The residuals appear to be almost evenly spread above and below 0 for the differing values of predicted values. We can conclude that the linear model is appropriate for the given data*
* ***Independence between observations****-* The observations are independent of each other.
* ***Normality of populations or errors****- We can see this from quantile vs residuals (QQ) plot that the data do not appear to depart substantially from a normal distribution. The robustness of regression analysis with modest departures from normality enables us to conclude that we should not be overly concerned about departures from this normality assumption in the given data. Also the histogram shows almost symmetric or normal density plot. This is shown in Fig.9 below.*

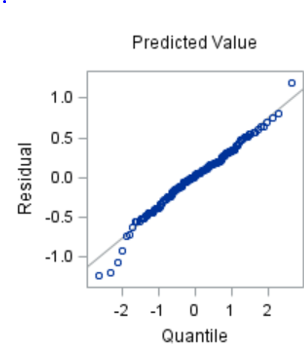
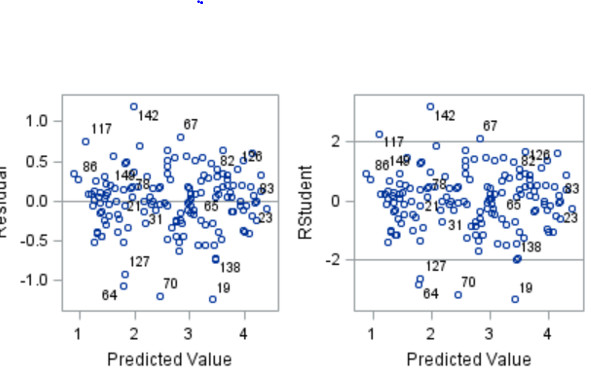
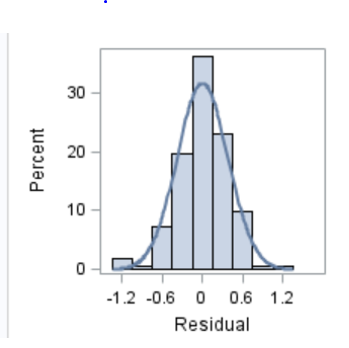
 

Fig. 7 QQ Plot-Residual distribution for individual regressor variables

* ***Equal Variance-****we can evaluate the assumption of equal variance from a plot of the residuals with predicted values again. The scatter plots do not appear to be major differences in the variability of the residuals for different predicted values (Fig 7). Thus, we can conclude that there is no apparent violation in the assumption of equal variance at each level of predicted values.*
* Observations must be quantified: This is true as each observation is quantifiable by decimal

The variance inflation factor of the variables is greatly improved after transformation as shown in Table 3 below. This shows the collinearity between explanatory variables is eliminated.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 3 Parameter Estimates** | | | | | | |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** |
| **Intercept** | **1** | 5.69923 | 0.14281 | 39.91 | <.0001 | 0 |
| **POPULATION** | **1** | 3.09531E-10 | 2.1613E-10 | 1.43 | 0.1543 | 1.12529 |
| **MEDIANAGE** | **1** | -0.08383 | 0.00563 | -14.88 | <.0001 | 2.24749 |
| **HEATHEXPENDITURE** | **1** | -0.00983 | 0.01308 | -0.75 | 0.4535 | 1.31972 |
| **OBESITY** | **1** | -0.01263 | 0.00403 | -3.13 | 0.0021 | 1.54158 |
| **GDP** | **1** | -0.00001036 | 0.00000193 | -5.36 | <.0001 | 1.68672 |
| **TAXREVENUE** | **1** | -0.00113 | 0.00331 | -0.34 | 0.7334 | 1.61279 |
| **UNEMPRATE** | **1** | 0.00350 | 0.00261 | 1.34 | 0.1814 | 1.27864 |

***Build a Model:***

Considering the availability of the data, to simplify the model we reduced the number variables to 11. Multiple collinearity tests were conducted to determine if the remaining 11 variables were measuring similar aspects of the model. Variables such as **TOTALFERTILITYRATE, POPGROWTH, BIRTHRATE** were taken out of the model due to their high collinearity with **MEDIANAGE**. The remaining variables have a variation inflation factor (VIF) of 2.104 or less.

We conducted a variable selection procedure using the LARS algorithm with the cross validation splitting the data randomly into five groups (4 used to test the variables and 1 to validate) and the stop criteria is AIC. The process selected variables which affect the log of infant mortality rate of countries (**logINFANTMORATLITYRATE**).

The result of the variable selection process is:

**Table 7 Parameter Estimate Table**

|  |  |
| --- | --- |
| **Root MSE** | 0.39102 |
| **Dependent Mean** | 2.68713 |
| **R-Square** | 0.8581 |
| **Adj R-Sq** | 0.8553 |
| **AIC** | -130.13871 |
| **AICC** | -129.73602 |
| **BIC** | -285.19296 |
| **C(p)** | 9.05654 |
| **SBC** | -274.96501 |
| **CV PRESS** | 23.85201 |

|  |  |  |
| --- | --- | --- |
| **Parameter Estimates** | | |
| **Parameter** | **DF** | **Estimate** |
| **Intercept** | 1 | 5.608757 |
| **MEDIANAGE** | 1 | -0.083869 |
| **OBESITY** | 1 | -0.010689 |
| **GDP** | 1 | -0.000009589 |

**Table 8 ANOVA Table**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | |
| **Source** | **DF** | **Sum of**  **Squares** | **Mean**  **Square** | **F Value** |
| **Model** | 3 | 139.62285 | 46.54095 | 304.39 |
| **Error** | 151 | 23.08780 | 0.15290 |  |
| **Corrected Total** | 154 | 162.71065 |  |  |

***Fit the model:***

To make sure the variables above fit the model we ran a regression analysis. All the variables from the model appeared to be significant.

**Table 9 Parameter Estimate Table**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | | | |
| **Variable** | **DF** | **Parameter**  **Estimate** | **Standard**  **Error** | **t Value** | **Pr > |t|** | **Variance**  **Inflation** |
| **Intercept** | 1 | 5.77668 | 0.12180 | 47.43 | <.0001 | 0 |
| **MEDIANAGE** | 1 | -0.08666 | 0.00470 | -18.43 | <.0001 | 1.62217 |
| **OBESITY** | 1 | -0.01377 | 0.00369 | -3.73 | 0.0003 | 1.37802 |
| **GDP** | 1 | -0.00001092 | 0.00000180 | -6.05 | <.0001 | 1.50180 |

The above information shows the significant variables, the p-values and the VIF of those variables.

***Check the Model Fit (lack of fit test):***

|  |  |  |  |
| --- | --- | --- | --- |
| **Root MSE** | 0.38523 | **R-Square** | 0.8623 |
| **Dependent Mean** | 2.68713 | **Adj R-Sq** | 0.8595 |
| **Coeff Var** | 14.33622 |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Analysis of Variance** | | | | | | |
| **Source** | **DF** | **Sum of**  **Squares** | **Mean**  **Square** | **F Value** | **Pr > F** |
| **Model** | 3 | 140.30151 | 46.76717 | 315.13 | <.0001 |
| **Error** | 151 | 22.40914 | 0.14840 |  |  |
| **Lack of Fit** | 151 | 22.40914 | 0.14840 | . | . |
| **Pure Error** | 0 | 0 | . |  |  |
| **Corrected Total** | 154 | 162.71065 |  |  |  |

The model is a good fit (p-value < 0.0001) with an Adjusted R2 = 0.8595.

The final model equation is as follows:

**or**

|  |  |
| --- | --- |
| Unit Increment of this variable | Multiplicative Change of the median of Infant Mortality Rate |
| MEDIANAGE | 0.917 |
| OBESITY | 0.986 |
| GDP | 1.00001092 |

**Interpretation & Conclusion**

With the data that was analyzed in the study we see a clear relationship between the infant mortality rate and the median age of a nation’s population, obesity rate of that particular nation as well as the GDP. All of the indicators mentioned above show a negative correlation to the response which in this case is the infant mortality rate. A lower infant mortality rate is the desired goal of all nations and as such these variables can be used as a reliable measurement of a country’s health.

The interpretation of the resulting model yields some interesting findings as well as general expected ones. For example, the high level of Collinearity seen amongst the fertility rate, population growth, birthrate and the median age are somewhat expected. As in younger populations being more fertile and having more children resulting in higher birthrates resulting in a higher population growth rates.

While the model’s selected variables are somewhat expected, some of the variables that were left out are equally intriguing. The most striking observation being that high healthcare expenditure did not necessarily translate to a statistically significant lower infant mortality rate. While one can argue that this is counter intuitive further study needs to be done on the subject to see where the actual spend ends up. For example, countries with aging, obese and other disease prone populations will get burdened by high costs of treating diseases rather than preventative care there by skewing the measurement against common sense expectations.

While looking at the model the Median age looks as if it has the biggest coefficient and thus effect, in fact it’s the GDP that has the highest influence. The GDP coefficient is quite small simply because the GDP value is significantly large when comprised to the other two explanatory variables in the model. Looking at the maps one can also argue that the GDP plays the biggest role in decreasing the infant mortality rate than the other two variables.

**Appendix**

SAS Code

\*1) Importing the data set\*;

**data** Project1;

infile "<[\\Client\C$\Users\ARAYA ABADI\Desktop\EXPERMENTALSTAT2\_6372\PROJECT1DataSet.csv](%5C%5C%5C%5CClient%5C%5CC$%5C%5CUsers%5C%5CARAYA%20ABADI%5C%5CDesktop%5C%5CEXPERMENTALSTAT2_6372%5C%5CPROJECT1DataSet.csv)>"dlm="," firstobs=**2**;

input COUNTRY $ POPULATION MEDIANAGE POPGROWTH BIRTHRATE MTRNLDTH HEATHEXPENDITURE GDP TAXREVENUE

UNEMPRATE INFANTMORTALITYRATE OBESITY TOTALFERTILITYRATE;

**run**;

**proc** **print** noobs data=Project1;

**run**;

\*2) a closer look on the variable scatterplots\*;

**proc** **sgscatter** data = Project1;

matrix POPULATION MEDIANAGE POPGROWTH BIRTHRATE MTRNLDTH HEATHEXPENDITURE GDP TAXREVENUE

UNEMPRATE INFANTMORTALITYRATE OBESITY TOTALFERTILITYRATE/diagonal=(histogram kernel)

ellipse=(type=mean);

**run**;

\*3) scatterplot as compared the explanatory variables with the response(y)\*;

**proc** **sgscatter** data=Project1;

plot INFANTMORTALITYRATE\*POPULATION INFANTMORTALITYRATE\*MEDIANAGE INFANTMORTALITYRATE\*POPGROWTH

INFANTMORTALITYRATE\*BIRTHRATE INFANTMORTALITYRATE\*MTRNLDTH INFANTMORTALITYRATE\*HEATHEXPENDITURE

INFANTMORTALITYRATE\*GDP INFANTMORTALITYRATE\*TAXREVENUE INFANTMORTALITYRATE\*UNEMPRATE INFANTMORTALITYRATE\*OBESITY

INFANTMORTALITYRATE\*TOTALFERTILITYRATE;

**run**;

\*4) Looking the variables using univariate\*;

**proc** **univariate** data=PROJECT1;

histogram ;

**run**;

\*5) Looking the regression before transformation\*;

**Proc** **Reg** Data=PROJECT1 corr plots(label)=(rstudentleverage cooksd);

\*plots(unpack label);

model INFANTMORTALITYRATE = POPULATION MEDIANAGE POPGROWTH BIRTHRATE MTRNLDTH HEATHEXPENDITURE OBESITY GDP TAXREVENUE UNEMPRATE TOTALFERTILITYRATE / partial VIF ;

**run**;

\*6) Transformation of the response\*;

**data** Project2;

set Project1;

logINFANTMORTALITYRATE=log(INFANTMORTALITYRATE);

**run**;

**proc** **print** data=Project2;

**run**;

\*7) After Removing highly corrolated variables and transforming the response\*;

**Proc** **Reg** Data=Project2 corr plots(label)=(rstudentleverage cooksd);

\*plots(unpack label);

model logINFANTMORTALITYRATE = POPULATION MEDIANAGE HEATHEXPENDITURE OBESITY GDP TAXREVENUE UNEMPRATE /\*BIRTHRATE POPGROWTH MTRNLDTH TOTALFERTILITYRATE\*/ / partial VIF ;

**run**;

\*8)/\*Model Building - Variable Selection Process Using LARS Algorithm\*/

Proc GLMSELECT data=Poject2;

model logINFANTMORTALITYRATE = POPULATION MEDIANAGE HEATHEXPENDITURE OBESITY GDP TAXREVENUE UNEMPRATE / selection=LARS(choose = cv stop = aic) cvmethod = random(5) stats = (adjrsq cp bic sbc sl);

run;

\*9)/\* Model Building - Fit Linear Regression Model\*/

Proc Reg Data=Project2 corr plots(label)=(rstudentleverage cooksd);

\*plots(unpack label);

model logINFANTMORTALITYRATE = MEDIANAGE OBESITY GDP /partial VIF R ;

run;

\*10)/\* Model Building - LACK of FIT Test\*/

Proc Reg Data=Project2 corr plots(label)=(rstudentleverage cooksd);

\*plots(unpack label);

model logINFANTMORTALITYRATE = MEDIANAGE OBESITY GDP /lackfit VIF ;

run;