MSDS 6372 Project 1

Infant Mortality Case Study

**Team members:**

Kottegoda Manjula

Messa Araya

Kumar Rajeev

Date:2016-09-13

**Introduction**

Our team decided 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 their tried look like as related with the response. Some of the explanatory variables such as birthdate have a strong linear association with the response.

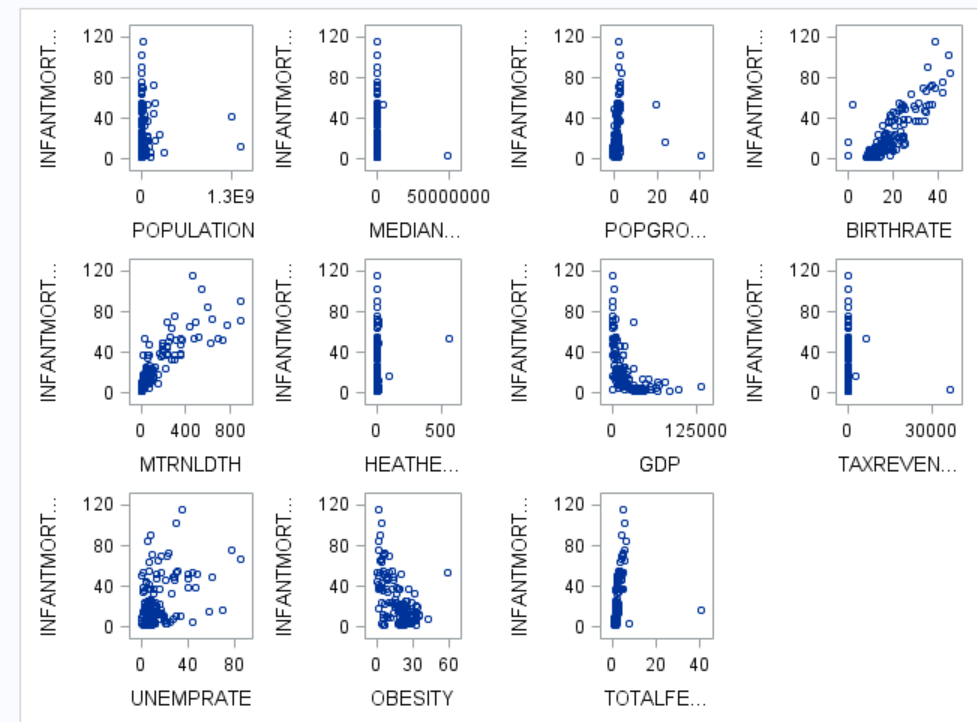
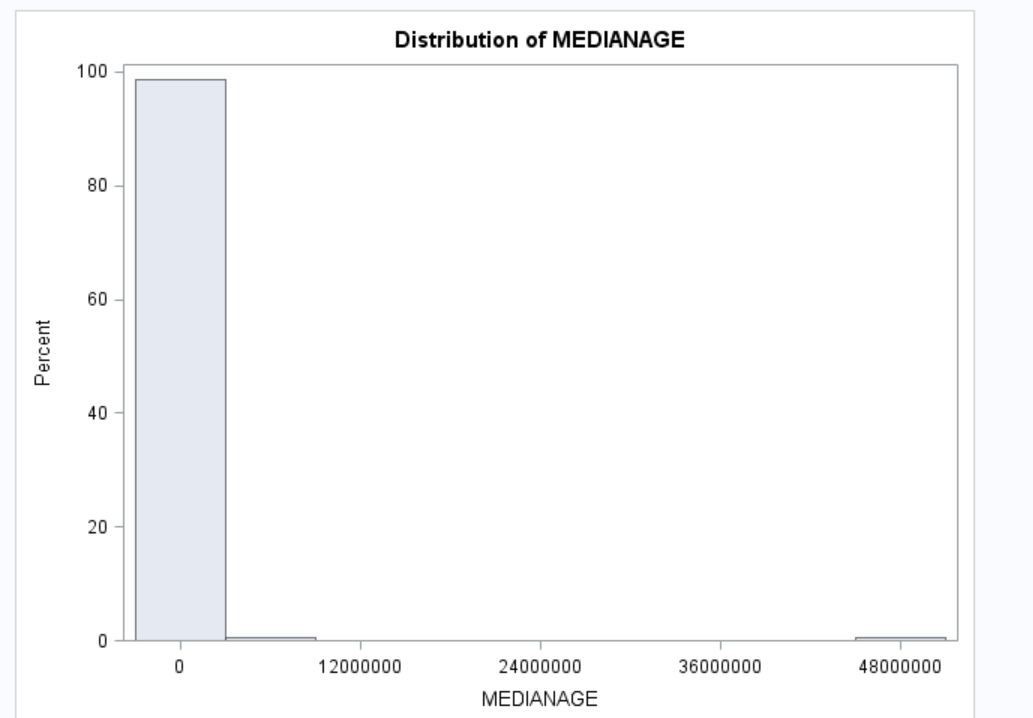
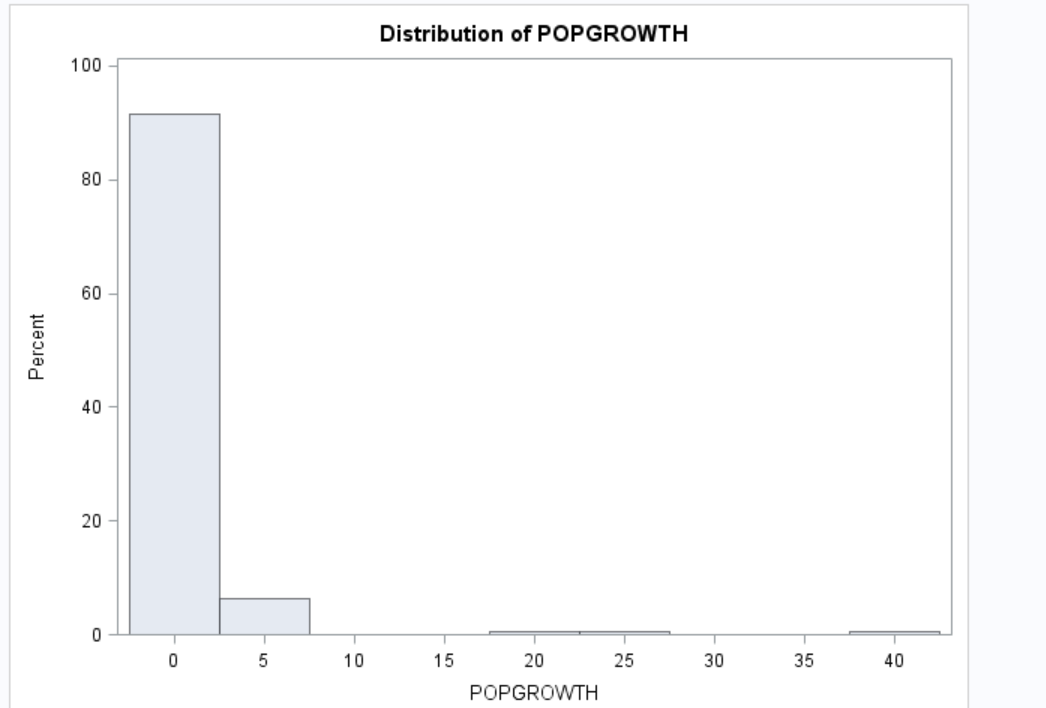
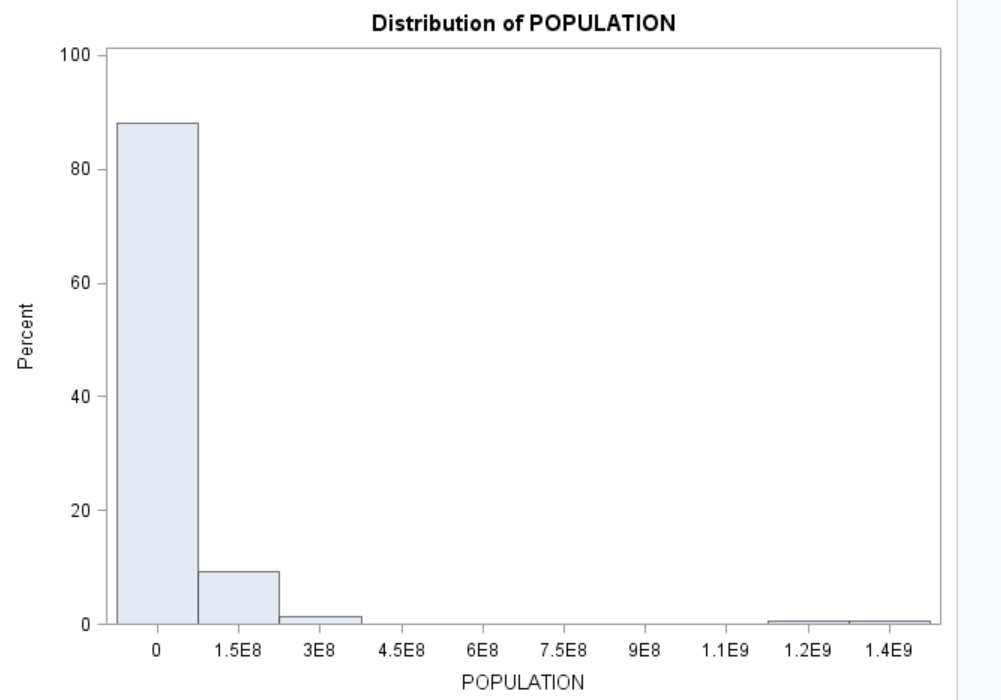
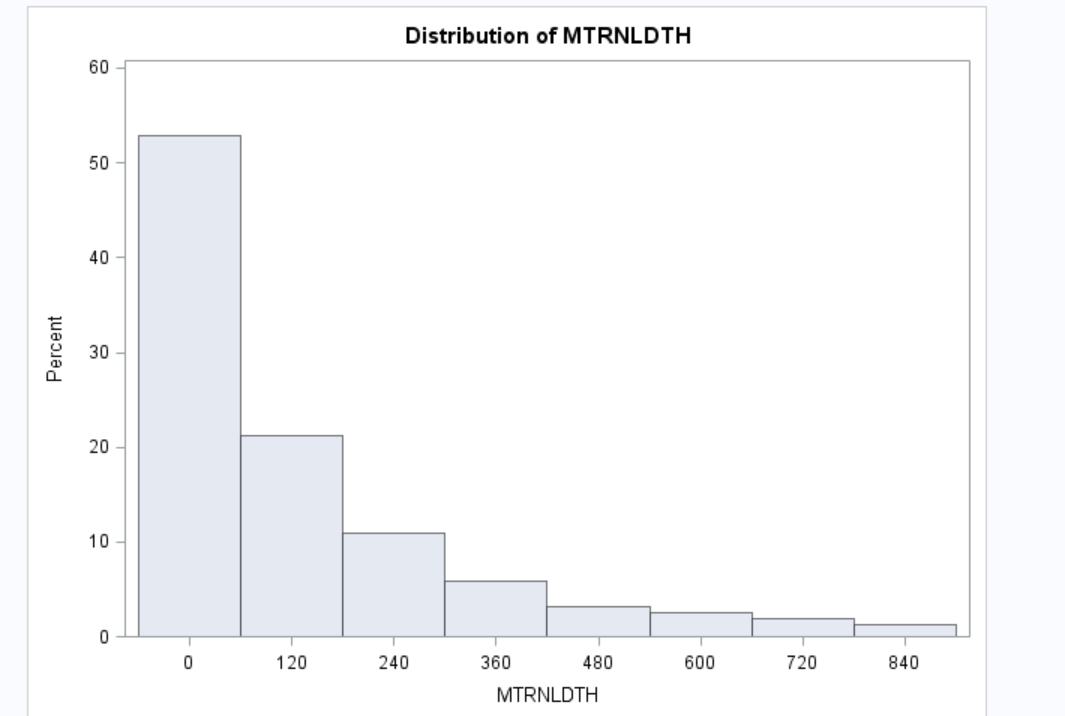
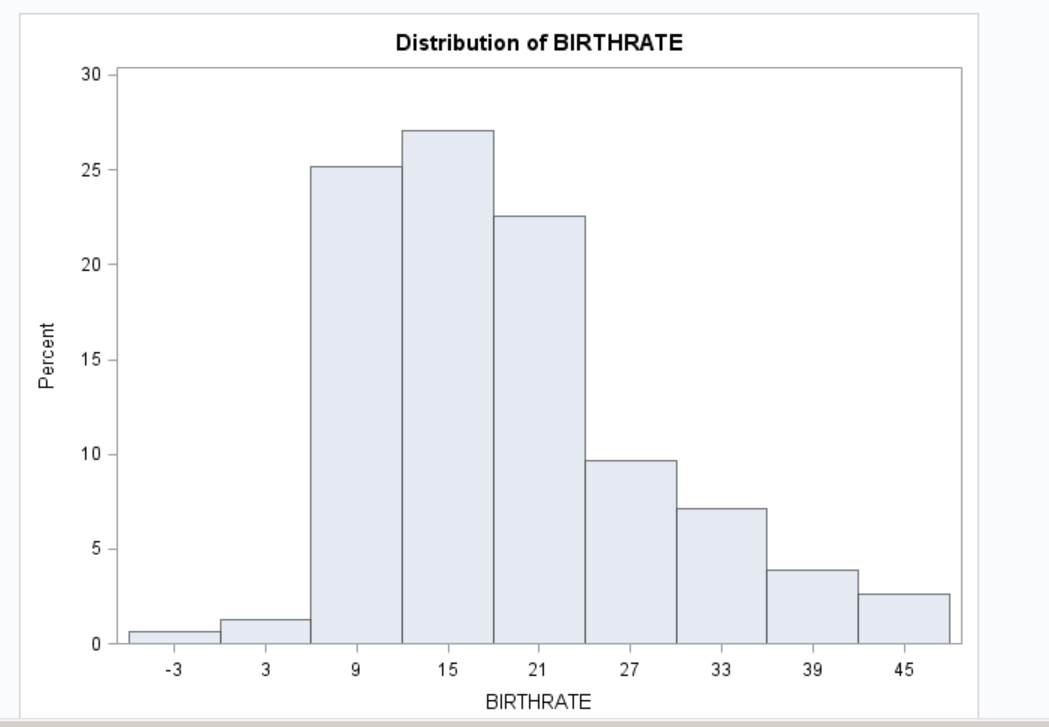
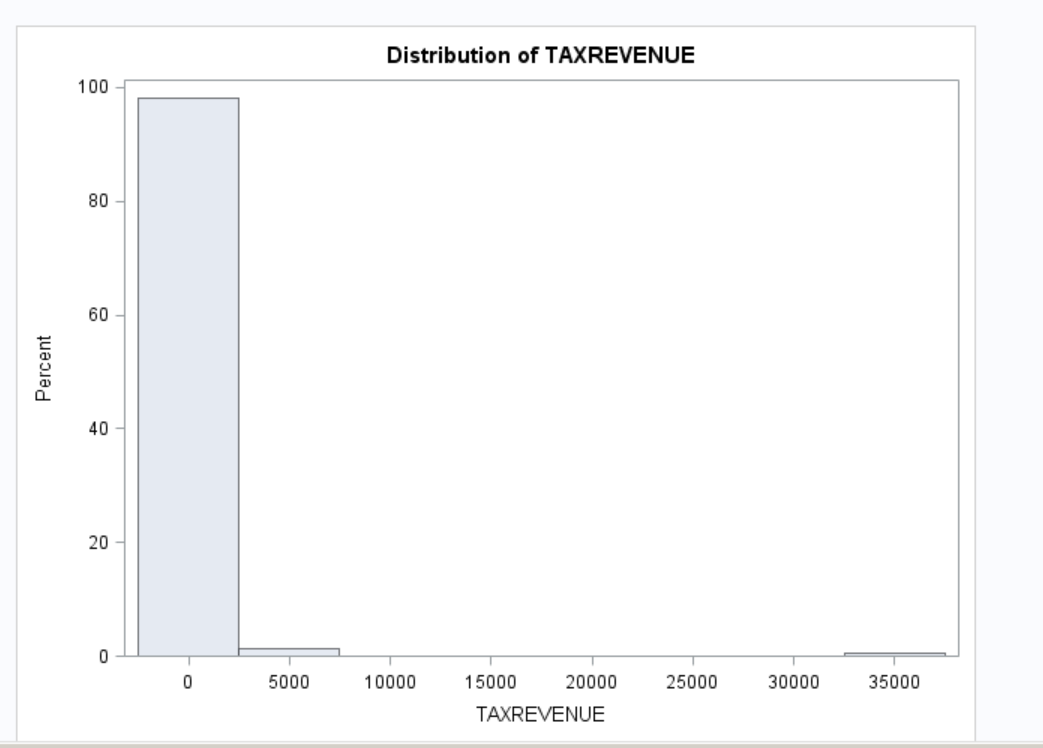
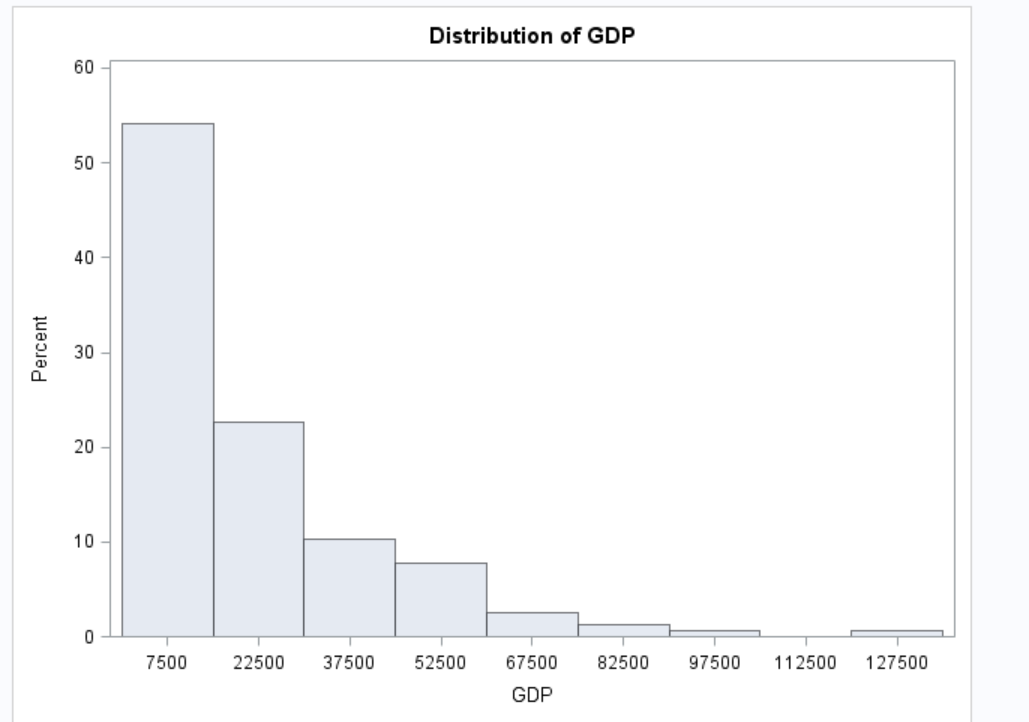
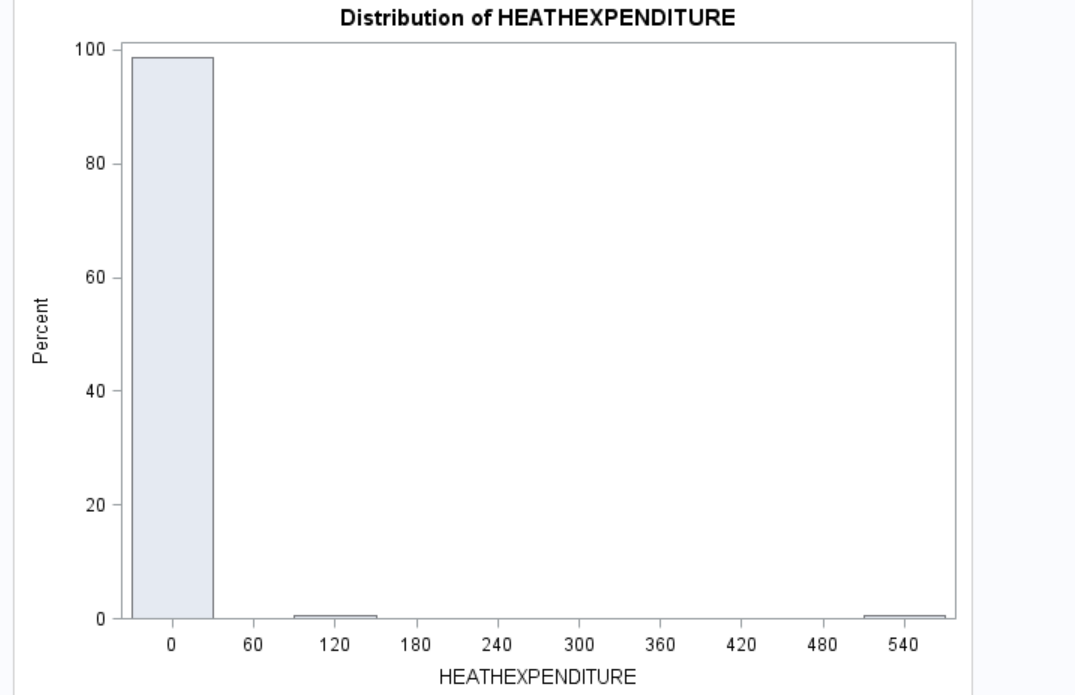


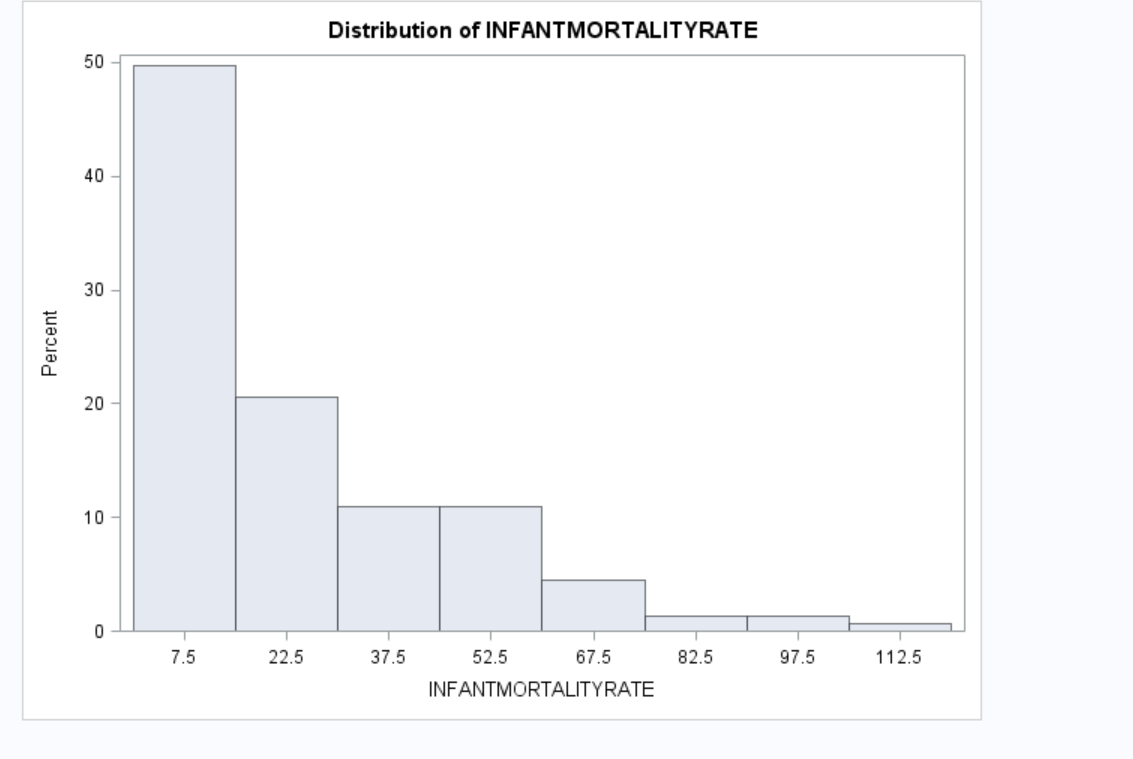
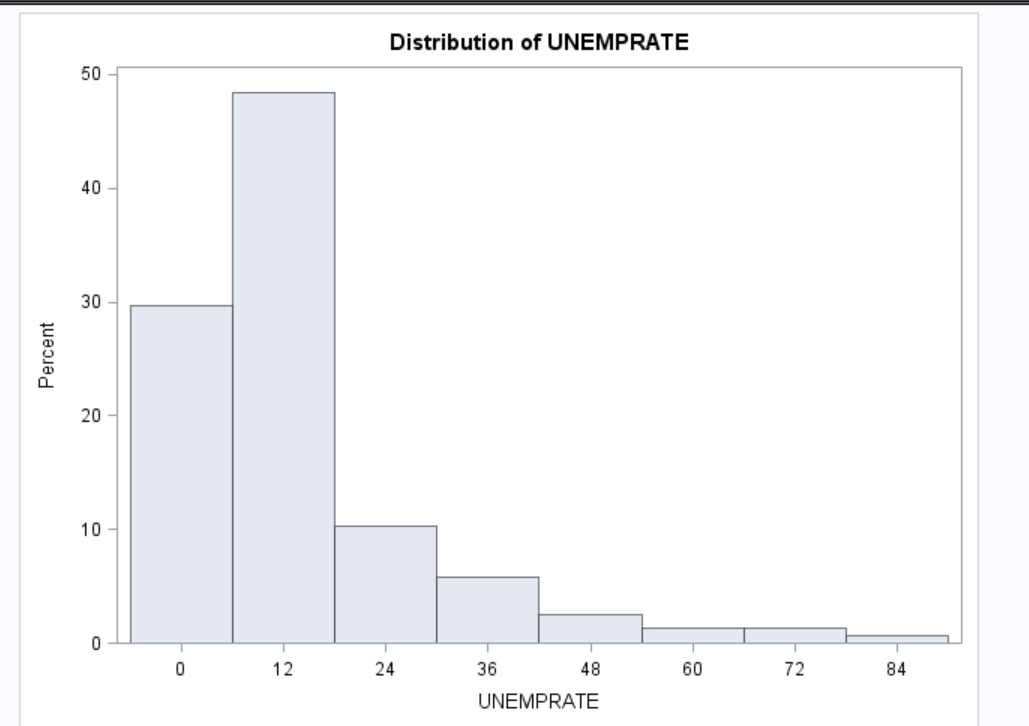
Fig. 2 Scatter plots for explanatory vs. response variables

Using a proc univariate individual data distributions were examined as shown using histograms (Fig. 3). Almost all distributions are not normal instead they all right skewed.









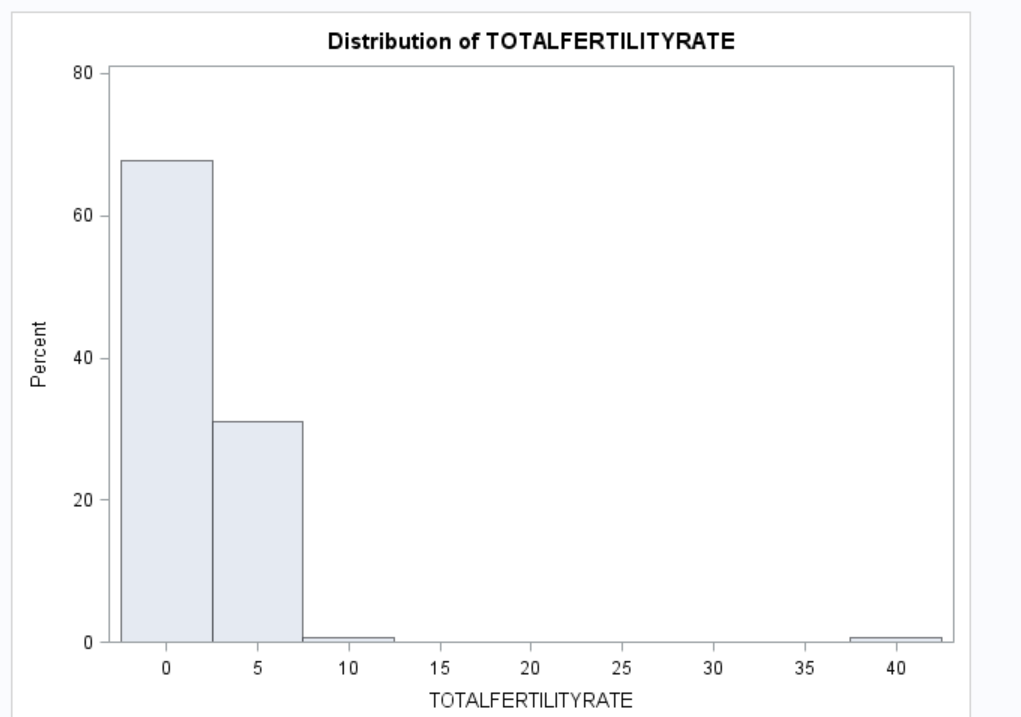
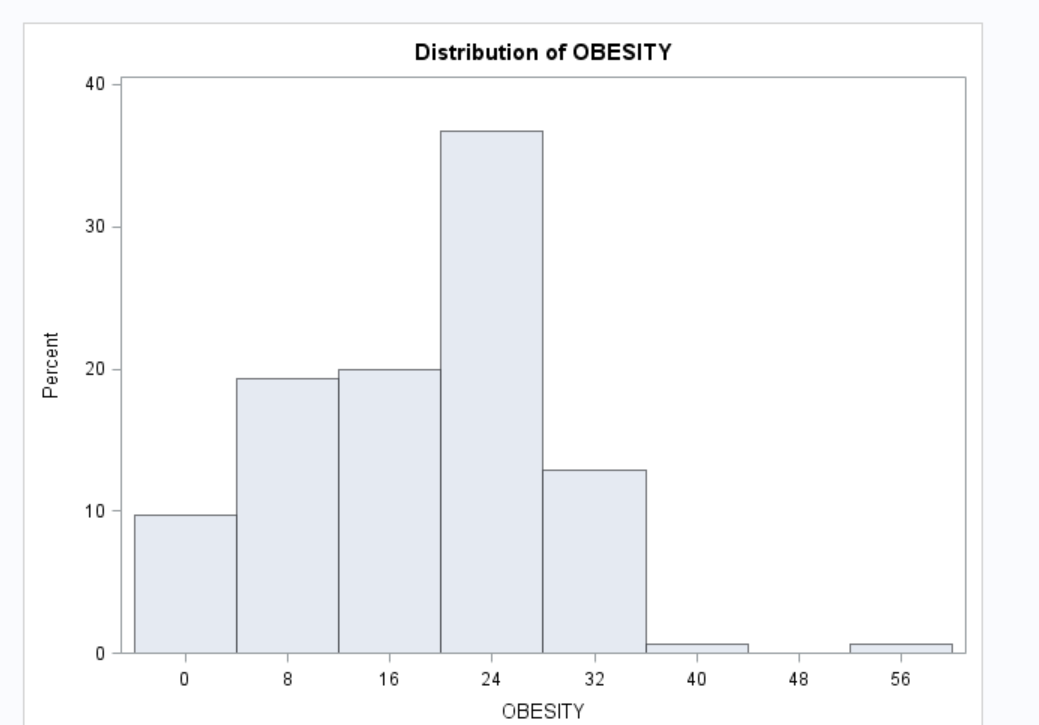


Fig. 3 Histograms for individual variable’s data distribution

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. 4) and Studentized residual (Fig. 5). 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.

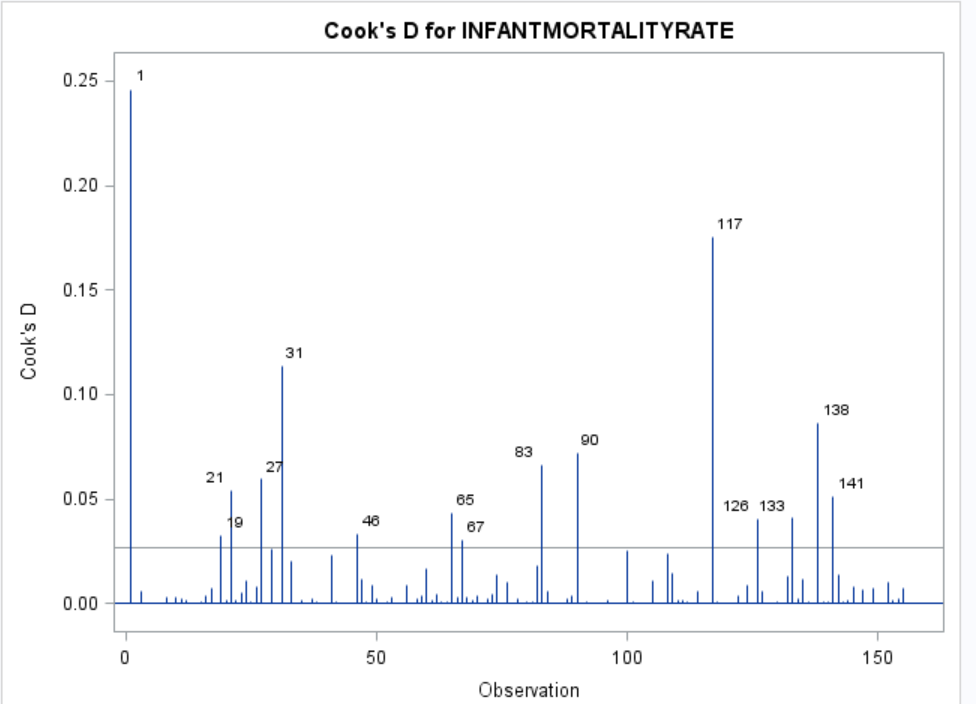


Fig. 4 Cook’s D

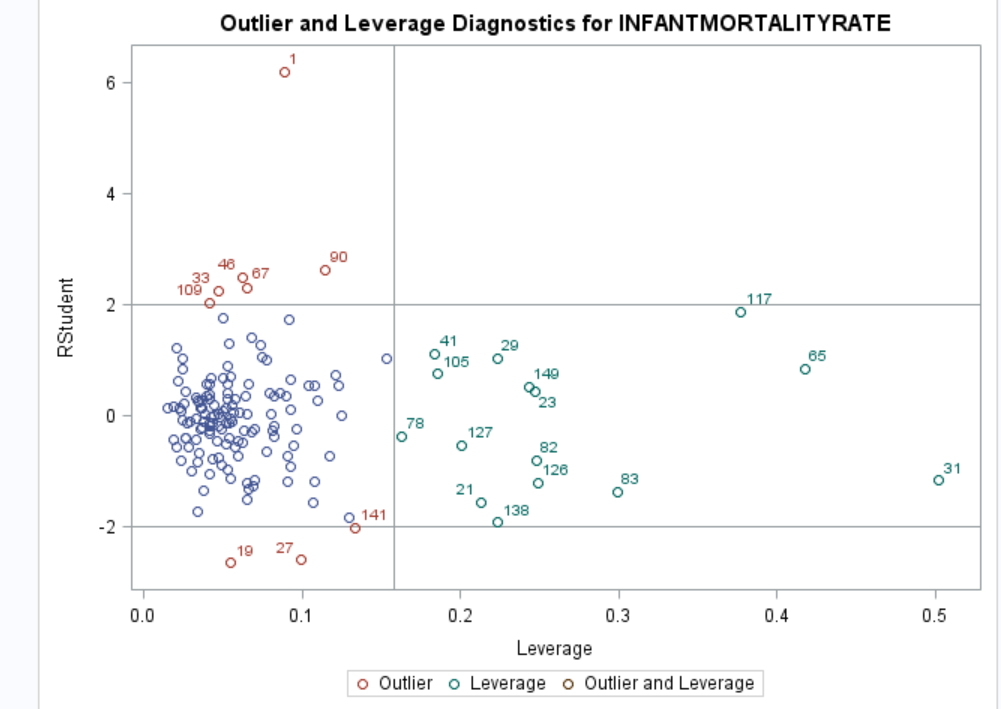


Fig. 5 Studentized residual distribution

In examining the residual plots (Fig. 6) 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.

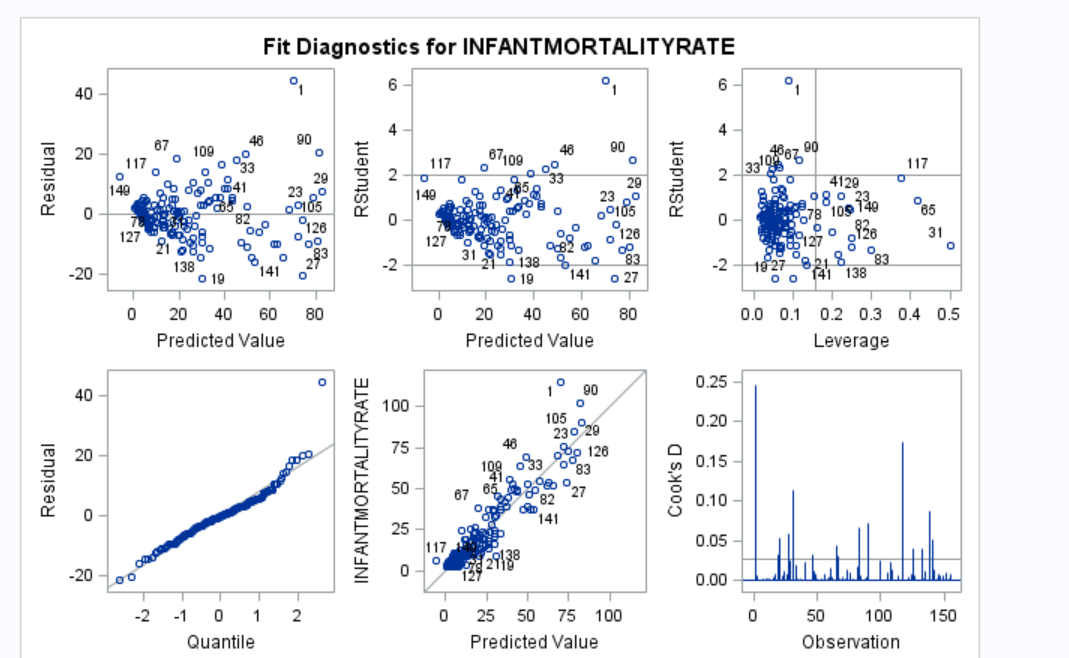
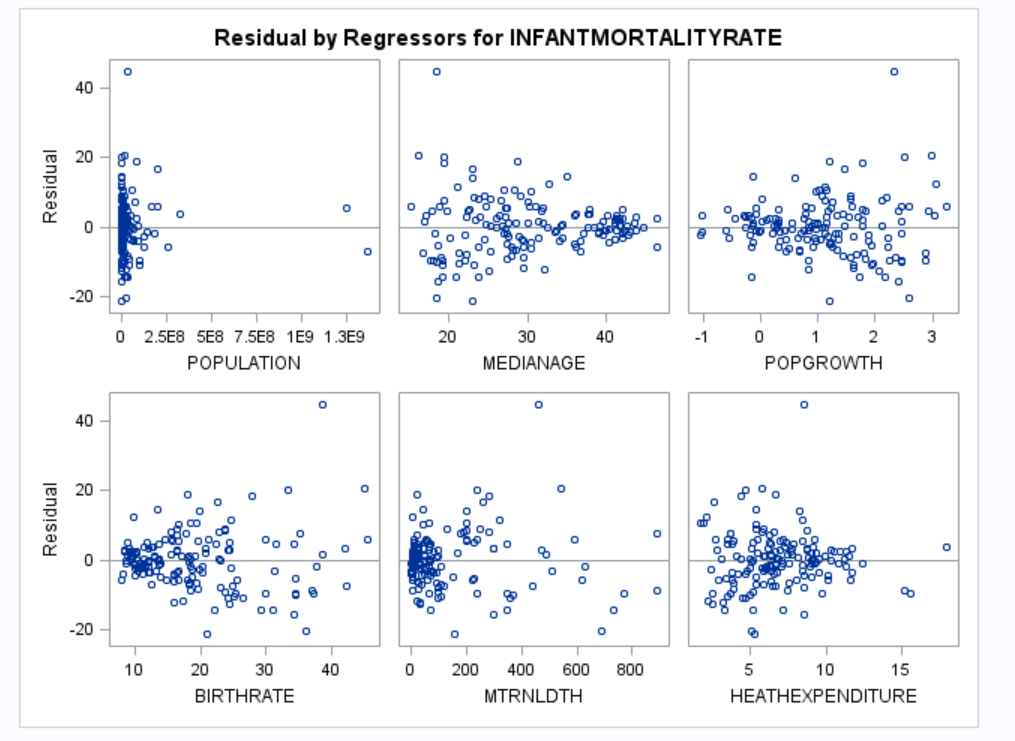


Fig. 6 Fit Diagnostic plots

There are significant differences in the distribution of the variables (repressors) in Fig 7. This is another information that we need to transform the data in order to fit the assumptions of a linear regression model.



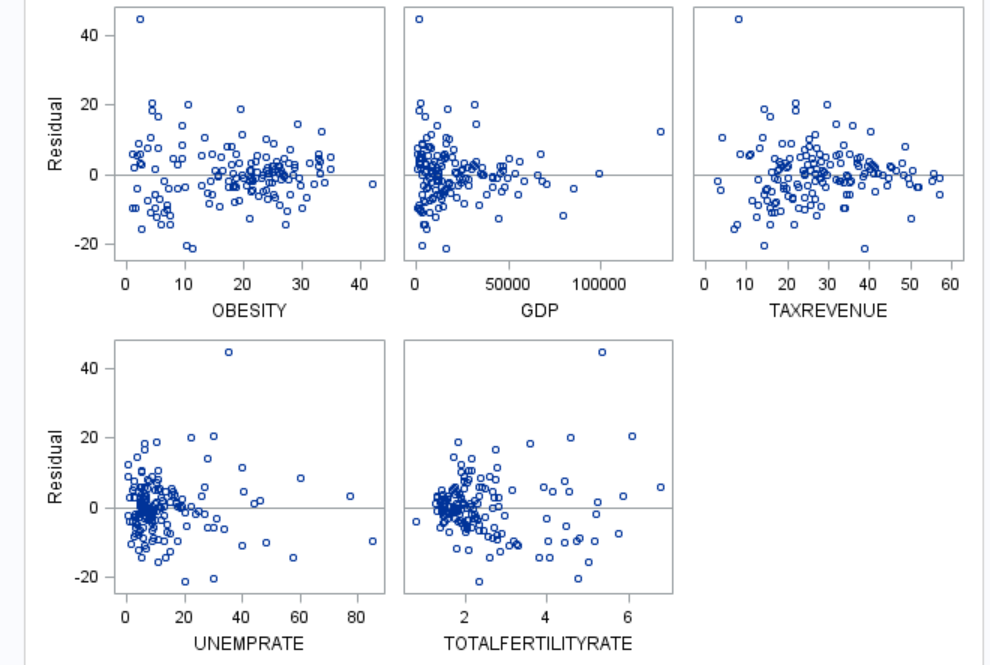


Fig. 7 Residual distribution for individual repressor variables

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** | | | | | | | | | | | | |
| **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 |
| **MEDIANAGE** | 0.0309 | 1.0000 | -0.7659 | -0.9063 | -0.6836 | 0.3810 | 0.5183 | 0.5446 | 0.5087 | -0.3330 | -0.8093 | -0.7960 |
| **POPGROWTH** | -0.0221 | -0.7659 | 1.0000 | 0.7850 | 0.5411 | -0.3055 | -0.4019 | -0.0663 | -0.3003 | 0.2173 | 0.7634 | 0.6363 |
| **BIRTHRATE** | -0.0371 | -0.9063 | 0.7850 | 1.0000 | 0.7841 | -0.2276 | -0.5653 | -0.5065 | -0.4002 | 0.3850 | 0.9728 | 0.8828 |
| **MTRNLDTH** | 0.0023 | -0.6836 | 0.5411 | 0.7841 | 1.0000 | -0.0153 | -0.5983 | -0.4702 | -0.3337 | 0.3791 | 0.7696 | 0.8504 |
| **HEATHEXPENDITURE** | -0.0615 | 0.3810 | -0.3055 | -0.2276 | -0.0153 | 1.0000 | 0.1831 | 0.1079 | 0.3431 | 0.1104 | -0.1493 | -0.1762 |
| **OBESITY** | -0.2040 | 0.5183 | -0.4019 | -0.5653 | -0.5983 | 0.1831 | 1.0000 | 0.4497 | 0.3663 | -0.2085 | -0.5326 | -0.6682 |
| **GDP** | -0.0638 | 0.5446 | -0.0663 | -0.5065 | -0.4702 | 0.1079 | 0.4497 | 1.0000 | 0.4558 | -0.3181 | -0.4227 | -0.5360 |
| **TAXREVENUE** | -0.1803 | 0.5087 | -0.3003 | -0.4002 | -0.3337 | 0.3431 | 0.3663 | 0.4558 | 1.0000 | -0.0447 | -0.3364 | -0.4090 |
| **UNEMPRATE** | -0.0980 | -0.3330 | 0.2173 | 0.3850 | 0.3791 | 0.1104 | -0.2085 | -0.3181 | -0.0447 | 1.0000 | 0.3661 | 0.4115 |
| **TOTALFERTILITYRATE** | -0.0369 | -0.8093 | 0.7634 | 0.9728 | 0.7696 | -0.1493 | -0.5326 | -0.4227 | -0.3364 | 0.3661 | 1.0000 | 0.8500 |
| **INFANTMORTALITYRATE** | 0.0301 | -0.7960 | 0.6363 | 0.8828 | 0.8504 | -0.1762 | -0.6682 | -0.5360 | -0.4090 | 0.4115 | 0.8500 | 1.0000 |

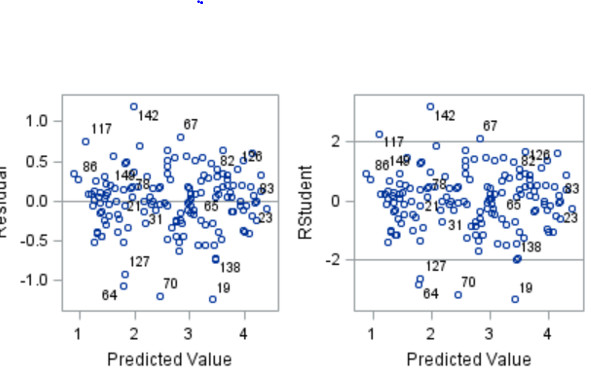
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Table 2 Parameter Estimates** | | | | | | |
| **Variable** | **DF** | **Parameter Estimate** | **Standard Error** | **t Value** | **Pr > |t|** | **Variance Inflation** |
| **Intercept** | **1** | -9.90132 | 14.65476 | -0.68 | 0.5004 | 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 reduce the variability and normalize the distribution. Our transformation was done only for the response variable.

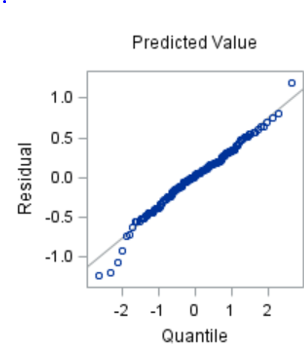
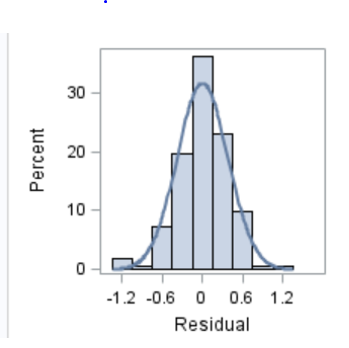
**Analysis**

*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. 8). 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 residuals are independent of each other. This is true as the residuals are uniformly distributed throughout the residual distributions shown above
* ***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.*

* ***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 8). 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 due to their high collinearity with **MEDIANAGE**. The remaining variables has a variation inflation factor (VIF) of 2.104 or less.

Before removing Collinearity

Correlation Table

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Correlation** | | | | | | | | | | | | | |
| **Variable** | **POPULATION** | **MEDIANAGE** | **POPGROWTH** | **BIRTHRATE** | **MTRNLDTH** | **HEATHEXPENDITURE** | **OBESITY** | **GDP** | **TAXREVENUE** | **UNEMPRATE** | **TOTALFERTILITYRATE** | **logINFANTMORATLITYRATE** |
| **POPULATION** | 1.0000 | 0.0342 | -0.0206 | -0.0402 | -0.0013 | -0.0619 | -0.2006 | -0.0606 | -0.1809 | -0.1004 | -0.0398 | 0.0590 |
| **MEDIANAGE** | 0.0342 | 1.0000 | -0.7542 | -0.9066 | -0.6857 | 0.3713 | 0.4903 | 0.5504 | 0.4511 | -0.3492 | -0.8113 | -0.8971 |
| **POPGROWTH** | -0.0206 | -0.7542 | 1.0000 | 0.7782 | 0.5436 | -0.3266 | -0.4107 | -0.0652 | -0.3062 | 0.2283 | 0.7588 | 0.5974 |
| **BIRTHRATE** | -0.0402 | -0.9066 | 0.7782 | 1.0000 | 0.7896 | -0.2346 | -0.5468 | -0.5103 | -0.3525 | 0.4088 | 0.9735 | 0.8424 |
| **MTRNLDTH** | -0.0013 | -0.6857 | 0.5436 | 0.7896 | 1.0000 | -0.0373 | -0.5903 | -0.4710 | -0.2967 | 0.4060 | 0.7763 | 0.7406 |
| **HEATHEXPENDITURE** | -0.0619 | 0.3713 | -0.3266 | -0.2346 | -0.0373 | 1.0000 | 0.2157 | 0.1002 | 0.3559 | 0.0828 | -0.1614 | -0.3326 |
| **OBESITY** | -0.2006 | 0.4903 | -0.4107 | -0.5468 | -0.5903 | 0.2157 | 1.0000 | 0.4237 | 0.3825 | -0.2147 | -0.5198 | -0.5749 |
| **GDP** | -0.0606 | 0.5504 | -0.0652 | -0.5103 | -0.4710 | 0.1002 | 0.4237 | 1.0000 | 0.4076 | -0.3242 | -0.4273 | -0.6703 |
| **TAXREVENUE** | -0.1809 | 0.4511 | -0.3062 | -0.3525 | -0.2967 | 0.3559 | 0.3825 | 0.4076 | 1.0000 | -0.0112 | -0.2966 | -0.4713 |
| **UNEMPRATE** | -0.1004 | -0.3492 | 0.2283 | 0.4088 | 0.4060 | 0.0828 | -0.2147 | -0.3242 | -0.0112 | 1.0000 | 0.3905 | 0.3778 |
| **TOTALFERTILITYRATE** | -0.0398 | -0.8113 | 0.7588 | 0.9735 | 0.7763 | -0.1614 | -0.5198 | -0.4273 | -0.2966 | 0.3905 | 1.0000 | 0.7480 |
| **logINFANTMORATLITYRATE** | 0.0590 | -0.8971 | 0.5974 | 0.8424 | 0.7406 | -0.3326 | -0.5749 | -0.6703 | -0.4713 | 0.3778 | 0.7480 | 1.0000 |

VIF Table

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | | | |
| **Variable** | **DF** | **Parameter**  **Estimate** | **Standard**  **Error** | **t Value** | **Pr > |t|** | **Variance**  **Inflation** |
| **Intercept** | 1 | 4.47684 | 0.61261 | 7.31 | <.0001 | 0 |
| **POPULATION** | 1 | 3.72676E-10 | 2.00704E-10 | 1.86 | 0.0654 | 1.11675 |
| **MEDIANAGE** | 1 | -0.05844 | 0.01368 | -4.27 | <.0001 | 15.62377 |
| **POPGROWTH** | 1 | -0.05529 | 0.07104 | -0.78 | 0.4376 | 5.45645 |
| **BIRTHRATE** | 1 | 0.06500 | 0.03201 | 2.03 | 0.0441 | 95.17585 |
| **MTRNLDTH** | 1 | 0.00111 | 0.00028127 | 3.94 | 0.0001 | 3.31826 |
| **HEATHEXPENDITURE** | 1 | -0.02409 | 0.01269 | -1.90 | 0.0596 | 1.48812 |
| **OBESITY** | 1 | -0.00492 | 0.00406 | -1.21 | 0.2279 | 1.90319 |
| **GDP** | 1 | -0.00000796 | 0.00000252 | -3.15 | 0.0020 | 3.34012 |
| **TAXREVENUE** | 1 | -0.00167 | 0.00295 | -0.57 | 0.5723 | 1.61532 |
| **UNEMPRATE** | 1 | 0.00174 | 0.00242 | 0.72 | 0.4728 | 1.33733 |
| **TOTALFERTILITYRATE** | 1 | -0.37576 | 0.17232 | -2.18 | 0.0308 | 46.83770 |

After removing the highly collinear variables from the model

VIF Table

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Parameter Estimates** | | | | | | | |
| **Variable** | **DF** | **Parameter**  **Estimate** | **Standard**  **Error** | **t Value** | **Pr > |t|** | **Variance**  **Inflation** |
| **Intercept** | 1 | 5.70201 | 0.14173 | 40.23 | <.0001 | 0 |
| **POPULATION** | 1 | 3.249E-10 | 2.13165E-10 | 1.52 | 0.1296 | 1.11128 |
| **MEDIANAGE** | 1 | -0.08479 | 0.00535 | -15.86 | <.0001 | 2.10416 |
| **HEATHEXPENDITURE** | 1 | -0.00990 | 0.01270 | -0.78 | 0.4369 | 1.31507 |
| **OBESITY** | 1 | -0.01192 | 0.00383 | -3.11 | 0.0022 | 1.49330 |
| **GDP** | 1 | -0.00001050 | 0.00000189 | -5.55 | <.0001 | 1.66007 |
| **TAXREVENUE** | 1 | -0.00063811 | 0.00306 | -0.21 | 0.8351 | 1.52821 |
| **UNEMPRATE** | 1 | 0.00355 | 0.00252 | 1.41 | 0.1617 | 1.28561 |

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 in countries (**logINFANTMORATLITYRATE**).

The result of the variable selection process is:

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

|  |  |
| --- | --- |
| **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 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **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 |  |  |

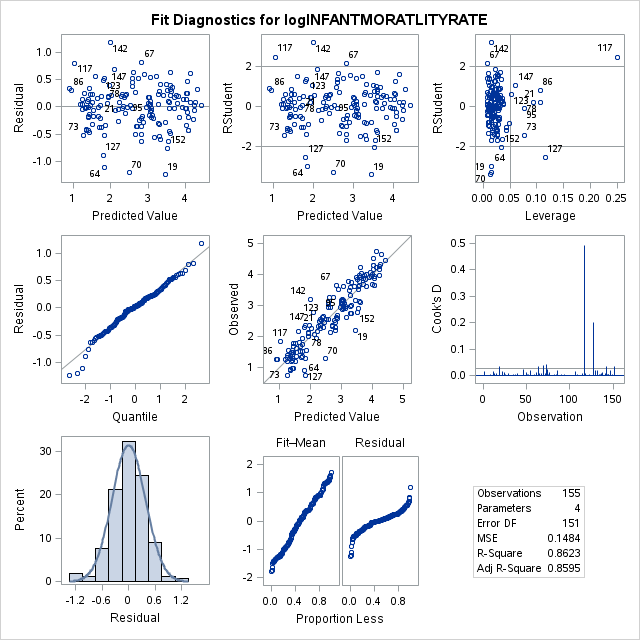
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Stop Details** | | | | |
| **Candidate**  **For** | **Effect** | **Candidate**  **AIC** |  | **Compare**  **AIC** |
| **Entry** | UNEMPRATE | -128.6751 | > | -130.1387 |

***Fit the model:***

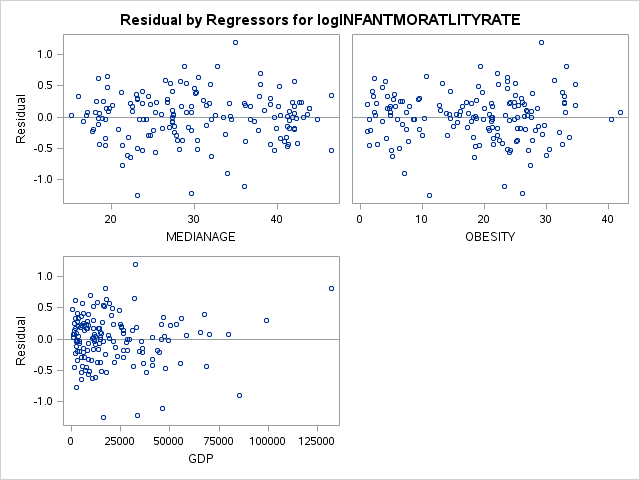
To make sure the variables above fit the model we ran a regression analysis. All the variables from the model appeared to be significant.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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.



The residual analysis is not the best but no distinct pattern in the residuals is present. The QQ-Plot and the histogram show the residuals are normally distributed. There is no significant leverage points. The Adjusted R2 is 0.8595.



The above residuals do not show a distinct pattern but there is some clustering, especially in GDP. There does not seem to be a significant reason to reject meeting the assumptions of linear regression.

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 |  |  |  |

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

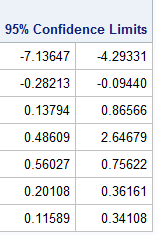
The model is a good fit (p-value < 0.0001) with an Adjusted R2 = 0.8595.

The final model equation is as follows:

|  |  |  |
| --- | --- | --- |
| Doubling of this variable | Multiplicative Change of the median of murders in a community | Decimal of multiplicative change |

**Interpretation & Conclusion**

There is significant evidence to suggest, in this sampling of 302 records, the doubling of percentage of white people in a city, the percentage of divorce people in a city, the mean number of people per family, the number of children born to parents who are not married, the land area in square miles and the number of police cars produces multiplicative changes in the median of murders in a city. Below are the 95% confidence intervals of the log of the variables.



The interpretation of these results can be misleading. In our opinion we think further research needs to be conducted to determine what each of these explanatory variables represents in a city. Taken out of context this information can be distorted and manipulated to fit an agenda or narrative but that is not in the scientific purpose of this report.

Do white people make a significant difference in a city to reduce murders? In the 1990’s white people, as a demographic group, may hold more disposable income. There may be merit in researching the amount of disposable income in a city as it relates to crime.

The social factors of divorce percentage, number of people per family and children born to parents who are not married may indicate the lack of mentorship from a respected person (parent or community leader) and an unstable social-economic environment. Further research will need to be conducted to determine what socio-economic factors contribute to the number of murders in a city.

Land area in square miles may indicate number of people and population density of a community. The larger the land area the more people are in the city. As the number of the people increases the crime rate and number of murders may increase. There may be merit in research if larger populations have more murders and higher murder rates.

The number of police cars may indicate similar factors as land area. Police cars may also indicate the operating budget and the need for more police because the crime rate is high thus the number of murders may be high. More law enforcement resources may be needed to react to the crime of a city. Further research needs to be conducted to determine the law enforcement indicators that underlie why the number of police cars would be an indicator for the number of murders in a city.

**Appendix**

SAS Code Explanatory Analysis:

Below is an example of the explanatory analysis looking for outliers and high leverage points. This was done several times in order to determine what was taken out. Then each variable was plotted to check the assumptions.

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Begin data exploration for murders

group 02

explanatory variables agePct16t24 - pctWWage;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**proc** **sgscatter** data = stats01.zstats02proj01;

matrix murders /\*population houseHoldSize racePctBlack racePctWhite racePctAsian racePctHisp\*/ agePct12t21 agePct12t29 agePct16t24 agePct65Up numbUrban pctUrban /\*medIncome pctWWage pctWFarmSelf pctWInvInc pctWSocSec pctWPubAsst pctWRetire medFamInc perCapInc whitePerCap blackPerCap indianPerCap asianPerCap otherPerCap hispPerCap numUnderPov pctPopUnderPov pctLess9thGrade pctNoHsGrad catPctHsGrad highPctHsGrad medpctHsGrad cnfPopHsGrad pctBSorMore pctUnemployed pctEmploy pctEmplManu pctEmplProfServ pctOccupManu pctOccupMgmtProf malePctDivorce malePctNevMarr femalePctDivorce totalPctDivorce persPerFam pctFam2Par pctKids2Par pctYoungKids2Par pctTeen2Par pctWorkMomYoungKids pctWorkMom numKidsBornNeverMar pctKidsBornNeverMar numImmig pctImmigRecent pctImmigRec5 pctImmigRec8 pctImmigRec10 pctRecentImmig pctRecimmig5 pctRecImmit8 pctRecImmig10 pctSpeakEnglOnly pctNotSpeakEngWell pctLargeHouseGam pctLargeHouseOccup persPerOccupHouse persPerOwnOccHouse persPerRentOccHouse pctPersOwnOccup pctPersDenseHouse pctHouseLess3Br mednumBr HouseVacant pctHouseOccup pctHouseOwnOccup pctVacantBoarded pctVacmore6Mos medYrHouseBuilt pctHouseNoPhone pctWoFullPlumb ownOccLowQuart ownOccMedVal ownOccHiQuart ownOccQRange rentLowQ rentMedian rentHighQ rentQRange medRent medRentPctHouseInc medOwnCostPctInc medOwnCostPctIncNoMtg numInShelters numStreet pctForeignBorn pctBornSameState pctSameHouse85 pctSameCity85 lemasSwornFt lemasSwFtPerPop lemasSwFtFieldOps lemasSwFtFieldPerPop lemasTotalReg lemasTotRegPerPop policeRegPerOff policePerPop recialmatchCommPol pctPoliceWhite pctPoliceBlack pctPoliceHisp pctPoliceAsian

pctPoliceMinor offAssgnDrugUnits numKidsDrugsSeiz policeAveOtWorked landArea popDens pctUsePubTrans policeCars policeOperBudg lemasPctPoliceOnPatr lemasGangUnitDEploy lemasPctOffDrunUn policeBudgePerPop\*/ / diagonal = (histogram);

**run**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

looking for significant outliers & leverage

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**proc** **univariate** data = stats01.zstats02proj01;

var numbUrban pctUrban;

histogram;

**run**;

**proc** **reg** data = stats01.zstats02proj01 corr plots(label) = (rstudentleverage cooksd);

model murders = agePct16t24 agePct65Up numbUrban pctUrban medIncome pctWWage;

**run**;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Results of exploratory analysis

group 02

significant outliers of the population are:

obs 80 -> Los Angeles City; outlier and leverage

obs 59 -> New York City; leverage

obs 126 -> Washington DC; significant outlier

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

SAS Code Variable Transformation:

Most of the variables, only two did not need to be transformed, used a log() or logit() transformation. This transformation meet the assumptions of the linear regression assumtions.

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

This is the data that was transformed;

There were only several variables that were not transformed;

Most were transformed using a log() or logit() transformation;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**data** stats01.zstats02proj01\_otlv\_trans;

set stats01.zstats02proj01\_otlv;

/\* response variable \*/

logMurder = log(murders);

/\* explanatory variables \*/

logPop = log(population);

logHouseHoldSize = log(houseHoldSize);

logitPctBl = log(racePctBlack / (**100** - racePctBlack));

logitPctWh = log(racePctWhite / (**100** - racePctWhite));

logitPctAs = log(racePctAsian / (**100** - racePctAsian));

logitPctHsp = log(racePctHisp / (**100** - racePctHisp));

logitPct12t21 = log(agePct12t21 / (**100** - agePct12t21));

logitPct12t29 = log(agePct12t29 / (**100** - agePct12t29));

logitPct16t24 = log(agePct16t24 / (**100** - agePct16t24));

logitPct65Up = log(agePct65Up / (**100** - agePct65Up));

lognumbUrban = log(numbUrban);

logPctUrb = log(pctUrban);

logMedInc = log(medIncome);

logitPctWWage = log(pctWWage / (**100** - pctWWage));

logitPctWFarmSelf = log(pctWFarmSelf / (**100** - pctWFarmSelf));

logitPctWInvInc = log(pctWInvInc / (**100** - pctWInvInc));

logitPctWSocSec = log(pctWSocSec / (**100** - pctWSocSec));

logitPctPubAsst = log(pctWPubAsst / (**100** - pctWPubAsst));

logitPctWRetire = log(pctWRetire / (**100** - pctWRetire));

logMedFamInc = log(medFamInc);

logPerCapInc = log(perCapInc);

logWhitePereCap = log(whitePerCap);

logBlackPerCap = log(blackPerCap);

logIndPerCap = log(indianPerCap);

logMurd = log(murders);

logAsianPerCap = log(asianPerCap);

logOtherPerCap = log(otherPerCap);

logHispPerCap = log(hispPerCap);

logNumUndPov = log(numUnderPov);

logitPctUndPov = log(pctPopUnderPov / (**100** - pctPopUnderPov));

logitPctLess9Grade = log(pctLess9thGrade / (**100** - pctLess9thGrade));

logitNoHsGrad = log(pctNoHsGrad / (**100** - pctNoHsGrad));

logCnfPopHsGrad = log(cnfPopHsGrad);

logitPctBSorMore = log(pctBSorMore / (**100** - pctBSorMore));

logitPctUnempl = log(pctUnemployed / (**100** - pctUnemployed));

logitPctEmpl = log(pctEmploy / (**100** - pctEmploy));

logitPctEmplMan = log(pctEmplManu / (**100** - pctEmplManu));

logitPctEmplProfServ = log(pctEmplProfServ / (**100** - pctEmplProfServ));

logitPctEmplOccupMan = log(pctOccupManu / (**100** - pctOccupManu));

logitPctEmplOccupProf = log(pctOccupMgmtProf / (**100** - pctOccupMgmtProf));

logitPctMaleDiv = log(malePctDivorce / (**100** - malePctDivorce));

logitPctMaleNevMarr = log(malePctNevMarr / (**100** - malePctNevMarr));

logitPctFemaleDiv = log(femalePctDivorce / (**100** - femalePctDivorce));

logitPctTotDiv = log(totalPctDivorce / (**100** - totalPctDivorce));

logPersPerFam = log(persPerFam);

logitPctFam2Par = log(pctFam2Par / (**100** - pctFam2Par));

logitPctKids2Par = log(pctKids2Par / (**100** - pctKids2Par));

logitPctYoungKids2Par = log(pctYoungKids2Par / (**100** - pctYoungKids2Par));

logitPctTeen2Par = log(pctTeen2Par / (**100** - pctTeen2Par));

logitPctWorkMomYoungKids = log(pctWorkMomYoungKids / (**100** - pctWorkMomYoungKids));

logitPctWorkMom = log(pctWorkMom / (**100** - pctWorkMom));

logNumKidsBornNeverMar = log(numKidsBornNeverMar);

logitPctKidsBornNeverMar = log(pctKidsBornNeverMar / (**100** - pctKidsBornNeverMar));

logNumImmig = log(numImmig);

logPctImmigRecent = log(pctImmigRecent / (**100** - pctImmigRecent));

logPctImmigRec5 = log(pctImmigRec5 / (**100** - pctImmigRec5));

logPctImmigRec8 = log(pctImmigRec8 / (**100** - pctImmigRec8));

logPctImmigRec10 = log(pctImmigRec10 / (**100** - pctImmigRec10));

logitPctRecentImmig = log(pctRecentImmig / (**100** - pctRecentImmig));

logitPctRecimmig5 = log(pctRecimmig5 / (**100** - pctRecimmig5));

logitPctRecImmit8 = log(pctRecImmit8 / (**100** - pctRecImmit8));

logitPctRecImmig10 = log(pctRecImmig10 / (**100** - pctRecImmig10));

logitPctSpeakEnglOnly = log(pctSpeakEnglOnly / (**100** - pctSpeakEnglOnly));

logitPctNotSpeakEngWell = log(pctNotSpeakEngWell / (**100** - pctNotSpeakEngWell));

logitPctLargeHouseGam = log(pctLargeHouseGam / (**100** - pctLargeHouseGam));

logitPctLargeHouseOccup = log(pctLargeHouseOccup / (**100** - pctLargeHouseOccup));

logPersPerOccupHouse = log(persPerOccupHouse);

logPersPerOwnOccHouse = log(persPerOwnOccHouse);

logPersPerRentOccHouse = log(persPerRentOccHouse);

logitPctPersOwnOccup = log(pctPersOwnOccup / (**100** - pctPersOwnOccup));

logitPctPersDenseHouse = log(pctPersDenseHouse / (**100** - pctPersDenseHouse));

logitPctHouseLess3Br = log(pctHouseLess3Br / (**100** - pctHouseLess3Br));

logHouseVacant = log(HouseVacant);

logitPctHouseOccup = log(pctHouseOccup / (**100** - pctHouseOccup));

logitPctHouseOwnOccup = log(pctHouseOwnOccup / (**100** - pctHouseOwnOccup));

logitPctVacantBoarded = log(pctVacantBoarded / (**100** - pctVacantBoarded));

logitPctVacmore6Mos = log(pctVacmore6Mos / (**100** - pctVacmore6Mos));

logMedYrHouseBuilt = log(medYrHouseBuilt);

logitPctHouseNoPhone = log(pctHouseNoPhone / (**100** - pctHouseNoPhone));

logitPctWoFullPlumb = log(pctWoFullPlumb / (**100** - pctWoFullPlumb));

logOwnOccLowQuart = log(ownOccLowQuart);

logOwnOccMedVal = log(ownOccMedVal);

logOwnOccHiQuart = log(ownOccHiQuart);

logOwnOccQRange = log(ownOccQRange);

logRentLowQ = log(rentLowQ);

logRentMedian = log(rentMedian);

logRentHighQ = log(rentHighQ);

logRentQRange = log(rentQRange);

logMedRent = log(medRent);

logitMedRentPctHouseInc = log(medRentPctHouseInc / (**100** - medRentPctHouseInc));

logitMedOwnCostPctInc = log(medOwnCostPctInc / (**100** - medOwnCostPctInc));

logitMedOwnCostPctIncNoMtg = log(medOwnCostPctIncNoMtg / (**100** - medOwnCostPctIncNoMtg));

logNumInShelters = log(numInShelters);

logNumStreet = log(numStreet);

logitPctForeignBorn = log(pctForeignBorn / (**100** - pctForeignBorn));

logitPctBornSameState = log(pctBornSameState / (**100** - pctBornSameState));

logitPctSameHouse85 = log(pctSameHouse85 / (**100** - pctSameHouse85));

logitPctSameCity85 = log(pctSameCity85 / (**100** - pctSameCity85));

logLemasSwornFt = log(lemasSwornFt);

logLemasSwFtPerPop = log(lemasSwFtPerPop);

logLemasSwFtFieldOps = log(lemasSwFtFieldOps);

logLemasSwFtFieldPerPop = log(lemasSwFtFieldPerPop);

logLemasTotalReg = log(lemasTotalReg);

logLemasSwFtPerPop = log(lemasSwFtPerPop);

logPoliceRegPerOff = log(policeRegPerOff);

logPolicePerPop = log(policePerPop);

logitRecialmatchCommPol = log(recialmatchCommPol / (**100** - recialmatchCommPol));

logitPctPoliceWhite = log(pctPoliceWhite / (**100** - pctPoliceWhite));

logitPctPoliceBlack = log(pctPoliceBlack / (**100** - pctPoliceBlack));

logitPctPoliceHisp = log(pctPoliceHisp / (**100** - pctPoliceHisp));

logitPctPoliceAsian = log(pctPoliceAsian / (**100** - pctPoliceAsian));

logitPctPoliceMinor = log(pctPoliceMinor / (**100** - pctPoliceMinor));

logOffAssgnDrugUnits = log(offAssgnDrugUnits);

logNumKidsDrugsSeiz = log(numKidsDrugsSeiz);

logPoliceAveOtWorked = log(policeAveOtWorked);

logLandArea = log(landArea);

logPopDens = log(popDens);

logitPctUsePubTrans = log(pctUsePubTrans / (**100** - pctUsePubTrans));

logPoliceCars = log(policeCars);

logPoliceOperBudg = log(policeOperBudg);

logitLemasPctPoliceOnPatr = log(lemasPctPoliceOnPatr / (**100** - lemasPctPoliceOnPatr));

logitLemasPctOffDrunUn = log(lemasPctOffDrunUn / (**100** - lemasPctOffDrunUn));

logPoliceBudgePerPop = log(policeBudgePerPop);

**run**;

SAS Code Variable Selection:

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

This is the analysis portion of the test;

- Model name: TransAllGlmselectAfterColinCheck

- linear regression model (included the transformed variables and the few that are not transformed

- variable selection

- cross validation

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**proc** **glmselect** data = stats01.zstats02proj01\_otlv\_trans;

title "Stats 02 Project 01; GLMSELECT; Model: TransAllGlmselectAfterColinCheck";

/\* model \*/

model logMurd = /\*logPop\*/ logHouseHoldSize /\*logitPctBl\*/ logitPctWh /\*logitPctAs logitPctHsp logitPct12t21 logitPct12t29\*/ logitPct16t24 /\*logitPct65Up lognumbUrban logPctUrb\*/ logMedInc logitPctWWage /\*logitPctWFarmSelf logitPctWInvInc logitPctWSocSec\*/ logitPctPubAsst /\*logitPctWRetire logMedFamInc logPerCapInc\*/ logWhitePereCap /\*logBlackPerCap logIndPerCap logAsianPerCap logOtherPerCap logHispPerCap logNumUndPov logitPctUndPov logitPctLess9Grade\*/ logitNoHsGrad /\*logCnfPopHsGrad\*/ logitPctBSorMore /\*logitPctUnempl\*/ /\*logitPctEmpl logitPctEmplMan logitPctEmplProfServ logitPctEmplOccupMan logitPctEmplOccupProf logitPctMaleDiv\*/ logitPctMaleNevMarr /\*logitPctFemaleDiv\*/ logitPctTotDiv logPersPerFam logitPctFam2Par

/\*logitPctKids2Par logitPctYoungKids2Par logitPctTeen2Par logitPctWorkMomYoungKids logitPctWorkMom\*/ logNumKidsBornNeverMar /\*logitPctKidsBornNeverMar logNumImmig logPctImmigRecent logPctImmigRec5

logPctImmigRec8 logPctImmigRec10 logitPctRecentImmig logitPctRecimmig5 logitPctRecImmit8 logitPctRecImmig10\*/ logitPctSpeakEnglOnly /\*logitPctNotSpeakEngWell logitPctLargeHouseGam logitPctLargeHouseOccup

logPersPerOccupHouse logPersPerOwnOccHouse logPersPerRentOccHouse logitPctPersOwnOccup logitPctPersDenseHouse logitPctHouseLess3Br mednumBr\*/ logHouseVacant /\*logitPctHouseOccup logitPctHouseOwnOccup\*/ logitPctVacantBoarded logitPctVacmore6Mos /\*logMedYrHouseBuilt logitPctHouseNoPhone logitPctWoFullPlumb\*/ logOwnOccLowQuart /\*logOwnOccMedVal logOwnOccHiQuart logOwnOccQRange logRentLowQ\*/ logRentMedian /\*logRentHighQ logRentQRange logMedRent logitMedRentPctHouseInc logitMedOwnCostPctInc logitMedOwnCostPctIncNoMtg\*/ logNumInShelters logNumStreet /\*logitPctForeignBorn logitPctBornSameState logitPctSameHouse85 logitPctSameCity85\*/ logLemasSwornFt /\*logLemasSwFtPerPop logLemasSwFtFieldOps logLemasSwFtFieldPerPop logLemasTotalReg logPoliceRegPerOff logPolicePerPop\*/ logitRecialmatchCommPol

/\*logitPctPoliceWhite logitPctPoliceBlack logitPctPoliceHisp logitPctPoliceAsian logitPctPoliceMinor logOffAssgnDrugUnits\*/ logNumKidsDrugsSeiz logPoliceAveOtWorked logLandArea /\*logPopDens\*/ logitPctUsePubTrans logPoliceCars /\*logPoliceOperBudg logitLemasPctPoliceOnPatr lemasGangUnitDEploy\*/ logitLemasPctOffDrunUn

/\*logPoliceBudgePerPop\*/ /

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

OPTIONS

- variable selection criteria = LASSO

- variable stop criteria = AIC

- break up data = cross validation, method random in 5 parts

- additional statistics:

- adjusted R^2 -> linear correlation

- cp -> Mallows C(p)

- bic -> sawa bayesian information criterion

- sbc -> schwarz bayesian information criterion

- sl -> significance level of the F-Stat for entering and departing effects

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

selection = lasso (choose = cv stop = aic) cvmethod = random(**5**) stats = (adjrsq cp bic sbc sl);

**run**;

title ;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Results of the model TransAllGlmselectAfterColinCheck:

- seven variables show as selected to the AIC criterial using cross validation

- logitPctWh logitPctTotDiv logPersPerFam logNumKidsBornNeverMar logNumStreet logLandArea logPoliceCars

- the next step is to run the model in proc reg to look at the vif's to check the collinearity of the remaining variables

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

SAS Code Model Validation:

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Collinearity check of model TransAllGlmselectAfterColinCheck:

- will use proc reg to check the model fit and the vif's of the seven variables

- variables: logitPctWh logitPctTotDiv logPersPerFam logNumKidsBornNeverMar logNumStreet logLandArea logPoliceCars

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**proc** **reg** data = stats01.zstats02proj01\_otlv\_trans;

title "Stats 02 Project 01; REG; Model: TransAllGlmselectAfterColinCheck";

model logMurd = logitPctWh logitPctTotDiv logPersPerFam logNumKidsBornNeverMar logNumStreet logLandArea logPoliceCars /

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

OPTIONS

- lack of fit test

- variance inflation factors to check for collinearity different variables are measuring the same thing

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

lackfit vif ;

**run**;

title ;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Results of collinearity check of model TransAllGlmselectAfterColinCheck:

Lack of Fit Test:

- pass; looks like a good fit

Variable (VIF) P-value

- logitPctWh 2.37945 < 0.0001

- logitPctTotDiv 1.59913 0.002

- logPersPerFam 1.66595 0.0022

- logNumKidsBornNeverMar 3.31191 < 0.0001

- logNumStreet 1.35758 0.1351

- logLandArea 2.04826 < 0.0001

- logPoliceCars 2.86895 0.0375

Stats:

- the adjusted R^2 = 0.8658

Interesting findings:

- logNumStreet does not appear significant in the model

- re-run proc reg without logNumStreet in model

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Collinearity check of model TransAllGlmselectAfterColinCheck:

- will use proc reg to check the model fit and the vif's of the six variables

- took out logNumStreet because the analysis showed it was significantly greater than p-value of 0.05

- variables: logitPctWh logitPctTotDiv logPersPerFam logNumKidsBornNeverMar logLandArea logPoliceCars

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

**proc** **reg** data = stats01.zstats02proj01\_otlv\_trans;

title "Stats 02 Project 01; REG; Model: TransAllGlmselectAfterColinCheck";

model logMurd = logitPctWh logitPctTotDiv logPersPerFam logNumKidsBornNeverMar logLandArea logPoliceCars /

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

OPTIONS

- lack of fit test

- variance inflation factors to check for collinearity different variables are measuring the same thing

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/

lackfit vif ;

**run**;

title ;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

Results of collinearity check of model TransAllGlmselectAfterColinCheck:

Lack of Fit Test:

- pass; looks like a good fit

Variable (VIF) P-value

- logitPctWh 2.31540 0.0001

- logitPctTotDiv 1.60870 0.0071

- logPersPerFam 1.60605 0.0046

- logNumKidsBornNeverMar 3.39043 < 0.0001

- logLandArea 1.89040 < 0.0001

- logPoliceCars 2.36108 < 0.0001

Stats:

- the adjusted R^2 = 0.8575

Interesting findings:

- analysis complete

- the next step is the write-up and conclusions

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*/