

### Data cleaning Methodology:

I used Python to clean my data as follows:

```
# importing libraries
```

```
import numpy as np
```

```
import pandas as pd
```

```
#import data file
```

```
data = pd.read_csv('C:\\\\Users\\\\rimo\\\\Desktop\\\\Term 2 Winter 2019\\\\DSC 423 Data Analysis and Regression\\\\Group Project\\\\suicide-rates-overview-1985-to-2016\\\\master.csv')
```

```
#Checking data
```

```
data.head()
```

```
data.columns.values
```

```
#Changing column names to appropriate names
```

```
data.columns = ['country', 'year', 'sex', 'age', 'suicides_no', 'population',
```

```
    'suicidesper100kpop', 'country-year', 'HDI for year',
```

```
    'gdp_for_year_dollars', 'gdp_per_capita_dollars', 'generation']
```

```
data.columns.values
```

```
#Remove commas within the gdp_for_year_dollars column observations
```

```
data['gdp_for_year_dollars'] = data['gdp_for_year_dollars'].str.replace(',', '')
```

```
#Check data
```

```
data.info()
```

```
len(data.country.unique()) #number of the countries 101
```

```
#Subset the data to years 2006 to 2010 for the individual analysis
```

```
suicide = data[(data['year'] >= 2006) & (data['year'] <= 2010)]
```

```
#check the info and years in the new data set
```

```
suicide.info()      #5196 observations
```

```
suicide.head()
```

```
suicide.describe()
```

```
suicide.year.unique()
```

```
#check number of countries and generations used
```

```
len(suicide.country.unique())  #93 countries
```

```
#check for null values
```

```
suicide.isnull().sum().sort_values(ascending=False)
```

```
#since there is 4188 "HDI for year" nulls out of 5196 which is 80.6%, I would drop the column.
```

```
#Also, we don't need country-year
```

```
suicide = suicide.drop(['HDI for year', 'country-year'], axis=1)
```

```
#Import continent data set to create a continent column
```

```
#Source: https://www.kaggle.com/statchaitya/country-to-continent#countryContinent.csv
```

```

country = pd.read_csv('C:\\Users\\rimo\\Desktop\\Term 2 Winter 2019\\DSC 423 Data Analysis and Regression\\Group Project\\suicide-rates-overview-1985-to-2016\\countryContinent.csv', encoding = "ISO-8859-1")

#check data
country.info()
country.continent.unique()

#Drop unnecessary columns
country = country.drop(['code_2', 'code_3', 'country_code', 'iso_3166_2', 'sub_region', 'region_code', 'sub_region_code'], axis=1)

#Check data and country names
country.info()      # 249 observations
country.country.unique()
suicide.country.unique()

#Compare country names per country in both data sets and check for different standardizations
#After careful checking, there are 4 countries have different names in the data set
#Unify country names to implement the aggregation
replacements = {
    'Saint Vincent and the Grenadines': 'Saint Vincent and Grenadines',
    'United States of America': 'United States',
    'United Kingdom of Great Britain and Northern Ireland': 'United Kingdom',
    'Korea (Republic of)': 'Republic of Korea'
}
country['country'].replace(replacements, inplace=True)

#Aggregate the two data sets using left join to create a continent column using country name as a key
result = pd.merge(suicide, country[['country', 'continent']], on='country', how='left')

#check the resulted data set
result.info()      #5196 observations
result.head()
result.describe()

#Check how to limit the observations so they are from 2k to 3k obs
#Count the number of observations for each continent
result.groupby('continent').count()           # Europe has 2124 obs

#I would exclude Europe to reduce the amount of observations to around 3072 obs.
my_dataframe = result[result.continent != 'Europe']

#I would exclude all the zero suicide rates in case of log transformation and to limit the observations as well
my_dataframe = my_dataframe[my_dataframe.suicidesper100kpop != 0]

#Check my new and final data frame information

```

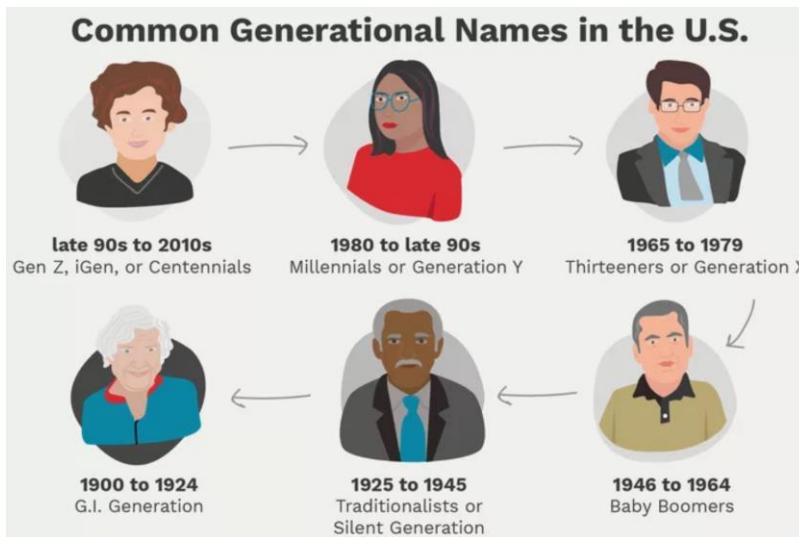
```

my_dataframe.info()           #2139 observations
my_dataframe.head()

#Check the unique values in the qualitative independant variables to create dummy variables
my_dataframe.continent.unique()
#[‘Americas’, ‘Asia’, ‘Oceania’, ‘Africa’]
my_dataframe.year.unique()
#[2007, 2006, 2008, 2009, 2010]
my_dataframe.sex.unique()
#[‘male’, ‘female’]
my_dataframe.generation.unique()
#[‘Boomers’, ‘Silent’, ‘Millenials’, ‘Generation X’, ‘Generation Z’]

#Finally, import the data frame into a csv file
my_dataframe.to_csv(r‘C:\\Users\\rimo\\Desktop\\Term 2 Winter 2019\\DSC 423 Data Analysis and Regression\\Group Project\\suicide-rates-overview-1985-to-2016\\finalsuicide.csv’, header=True, index=False)

```



Number of generations = 5  
 Number of dummy variables = 4

```

dgeneration1 = 1 if generation = ‘Boomers’, or otherwise dgeneration1 = 0
dgeneration2 = 1 if generation = ‘Generation X’, or otherwise dgeneration2 = 0
dgeneration3 = 1 if generation = ‘Millenials’, or otherwise dgeneration3 = 0
dgeneration4 = 1 if generation = ‘Generation Z’, or otherwise dgeneration4 = 0

```

	Z6 dage1	Z7 dage2	Z8 dage3	Z9 dage4	Z10 dage5
5-14 years	0	0	0	0	0
15-24 years	1	0	0	0	0
25-34 years X	0	1	0	0	0
35-54 years	0	0	1	0	0
55-74 years	0	0	0	1	0
75+ years	0	0	0	0	1

Number of genders = 2

Number of dummy variables = 1

dsex = 1 if SEX = male

dsex = 0 if SEX = female

	Z5 dsex
male	1
female	0

Number of age groups = 6

Number of dummy variables = 5

dage1 = 1 if age = '15-24 years', or otherwise dage1 = 0

dage2 = 1 if age = '25-34 years', or otherwise dage2 = 0

dage3 = 1 if age = '35-54 years', or otherwise dage3 = 0

dage4 = 1 if age = '55-74 years', or otherwise dage4 = 0

dage5 = 1 if age = '75+ years', or otherwise dage5 = 0

	Z1 dgeneration1	Z2 dgeneration2	Z3 dgeneration3	Z4 dgeneration4
Silent	0	0	0	0
Boomers	1	0	0	0
Generation X	0	1	0	0
Millenials	0	0	1	0
Generation Z	0	0	0	1

Number of distinct years = 5 [2006, 2007, 2008, 2009, 2010]

Number of dummy variables = 4

dyear1 = 1 if year = 2007, or otherwise dyear1 = 0

dyear2 = 1 if year = 2008, or otherwise dyear2 = 0

dyear3 = 1 if year = 2009, or otherwise dyear3 = 0

dyear4 = 1 if year = 2010, or otherwise dyear4 = 0

Number of continents = 4 ['Americas', 'Asia', 'Oceania', 'Africa']

Number of dummy variables = 3

dcontinent1 = 1 if continent = 'Africa', or otherwise dcontinent1 = 0

dcontinent2 = 1 if continent = 'Asia', or otherwise dcontinent2 = 0

dcontinent3 = 1 if continent = 'Oceania', or otherwise dcontinent3 = 0

	Z11 dyear1	Z12 dyear2	Z13 dyear3	Z14 dyear4
2006	0	0	0	0
2007	1	0	0	0
2008	0	1	0	0
2009	0	0	1	0
2010	0	0	0	1

Reducing the number of variables to leave the more meaningful ones only:

Since I am predicting suicide rates 'suicideper100kpop' and the predictor suicide\_no = (100k\*population/suicideper100kpop). Therefore, I am not going to use suicide\_no feature in my prediction model because it's unnecessary in this case.

The dependent variable is the suicideper100kpop "Y variable".

	Z15 dcontinent1	Z16 dcontinent2	Z17 dcontinent3
Americas	0	0	0
Africa	1	0	0
Asia	0	1	0
Oceania	0	0	1

#### Methodology:

Note: Ram is a contraction for graphs in the appendix

1. Analyze distribution of suicideper100kpop to see if the distribution is symmetric or

skewed. Ram1

The distribution is positively skewed to the right as we can see the tail in the histogram graph above.

Also, Mean is almost twice the Median which makes it positively skewed. Skewness = 4.1139.

For positive skew, there are the square root transformation, the log transformation, and the inverse/reciprocal transformation (in order of increasing severity).

**2. Create scatterplots to visualize the associations between Suicideper100kpop and the other 3 qualitative variables and check the patterns displayed to see if the associations appear to be linear. Also, compute correlation values of suicide rates per 100k population vs the other variables to interpret the correlation values, and discuss which pairs of variables appear to be strongly associated.**

**Ram2**

According to the correlation factors and the scatterplots matrix above, there is no linear correlation between suicideper100kpop and any of the quantitative predictors (population, gdp\_for\_year\_dollars, and gdp\_per\_capita\_dollars). Since the other predictors (age, generation, continent, sex, and year) were qualitative variables and then have been changed to dummy variables, they will not show any linear relationship and dots will be scattered around 0 and 1.

**3. Regression before Y-transformation:**

**a. Fit a regression model of suicideper100kpop vs the other variables (model M1) then compute the VIF statistics for each x-variable to analyze whether there is a problem of multicollinearity. Ram3**

Using the absolute value of standardized estimate to determine the predictors with significant effect on balance. The strongest predictor is dgeneration3 since the standardized estimate is the highest 0.522. When performing t-test on individual parameters, gdp\_per\_capita\_dollars, dage2, dyear1, dyer2, dyer4, dcontinent1, dcontinent2, and dcontinent3 have p-values that are bigger than 0.05 which make them insignificant X variables. The other variables are significant X variables.

**Suicide rate per 100k population =  $\beta_0 + \beta_1 * \text{population} + \beta_2 * \text{gdp\_for\_year\_dollars} + \beta_3 * \text{gdp\_per\_capita\_dollars} + \beta_4 * \text{dsex} + \beta_5 * \text{dgeneration1} + \beta_6 * \text{dgeneration2} + \beta_7 * \text{dgeneration3} + \beta_8 * \text{dgeneration4} + \beta_9 * \text{dage1} + \beta_{10} * \text{dage2} + \beta_{11} * \text{dage3} + \beta_{12} * \text{dage4} + \beta_{13} * \text{dage5} + \beta_{14} * \text{dyear1} + \beta_{15} * \text{dyear2} + \beta_{16} * \text{dyear3} + \beta_{17} * \text{dyear4} + \beta_{18} * \text{dcontinent1} + \beta_{19} * \text{dcontinent2} + \beta_{20} * \text{dcontinent3} + e$**

Where dummy variables have the values described above in the first page for zeros and ones.

**b. Diagnosing Multicollinearity:**

1. Scatterplot matrix and Pearson correlation matrix for each pair of x variables:

There is no significant collinearity between any of the X variables since all the correlation values are low.

2. Compute Variance Inflation Factor (VIF):

All VIFs < 10. These VIFs suggest no collinearity for any of the independent variables. "The variables with high VIFs are dummy variables that represent a categorical variable with three or more categories. If the proportion of cases in the reference category is small, the indicator variables will necessarily have high VIFs, even if the categorical variable is not associated with other variables in the regression model." Paul Allison said in his paper "When can you safely ignore multicollinearity".

3. Compute Tolerance value (TOL):

All TOLs > 0.1. These TOLs suggest no collinearity for any of the independent variables.

**4. Regression after the log transformation:**

**a. Analyze distribution of Insuicideper100kpop to see if the distribution is symmetric or skewed.**

After we applied log transformation, the distribution became almost normal since the mean is approximately equal to the median. Also, the histogram curve looks symmetrical. The skewness is -0.32 which is very small and can create a normal distribution. Ram4

**b. Create scatterplots to visualize the associations between Insuicideper100kpop and the other 3 qualitative variables and check the patterns displayed to see if the associations appear to be linear.**

Ram5

Collinearity does not change much after the transformation and there is no significant linear correlation between Insuicideper100kpop and any of the quantitative predictors.

**c. Build boxplots to evaluate if suicide rates vary by (age, generation, continent, year, and sex). Ram6**

According to the boxplots above, suicide rates do vary by age, generation, continent, and gender. However, suicide rates do not vary much by year. The mean and the median of suicide rates in different years from 2006 to 2010 do quite similar and do not vary much. Males committed suicide more than females. Furthermore, silent generation and people who are older than 75 years old committed suicide the most. On the other hand, Z-generation and kids who are 5-14 years old committed suicide the least. Africa had the least suicide rates among the 4 continents while Americas and Oceania have the most.

**d. Fit a regression model of Insuicideper100kpop vs the other variables (model M2) then compute the VIF statistics for each x-variable to analyze whether there is a problem of multicollinearity. Ram7**

**e. Comparing both models before and after the Y-variable transformation:**

R<sup>2</sup> and Adj-R<sup>2</sup> increased from (0.2495 – 0.2438) to (0.5225 – 0.5188).

RMSE reduced from 14.482 to 0.9683.

The F-value has increased from 43.80 to 114.18, indicating a stronger support to this model. This indicates that there is at least one independent variable that has a strong influence on suicide rates.

**f. Diagnosing Multicollinearity:** They do not seem to have changed much.

1. Scatterplot matrix and Pearson correlation matrix for each pair of x variables:

There is no collinearity between any of the X variables because all the correlation values are low.

2. Compute Variance Inflation Factor (VIF):

All VIFs < 10. These VIFs suggest no collinearity for any of the independent variables.

3. Compute Tolerance value (TOL):

All TOLs > 0.1. These TOLs suggest no collinearity for any of the independent variables.

**g. Fit a full model (with all independent variables) to predict Suicide rates. Discuss the parameter estimates, significance, goodness-of-fit and AdjR2 values.**

**Insuicideper100kpopulation =  $\beta_0 + \beta_1 * \text{population} + \beta_2 * \text{gdp\_for\_year\_dollars} + \beta_3 * \text{gdp\_per\_capita\_dollars} + \beta_4 * \text{dsex} + \beta_5 * \text{dgeneration1} + \beta_6 * \text{dgeneration2} + \beta_7 * \text{dgeneration3} + \beta_8 * \text{dgeneration4} + \beta_9 * \text{dage1} + \beta_{10} * \text{dage2} + \beta_{11} * \text{dage3} + \beta_{12} * \text{dage4} + \beta_{13} * \text{dage5} + \beta_{14} * \text{dyear1} + \beta_{15} * \text{dyear2} + \beta_{16} * \text{dyear3} + \beta_{17} * \text{dyear4} + \beta_{18} * \text{dcontinent1} + \beta_{19} * \text{dcontinent2} + \beta_{20} * \text{dcontinent3} + e$**

Where dummy variables have the values described above in the first page for zeros and ones

**h. Goodness of Fit:**

**Null hypothesis Ho:**  $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \dots = 0$ . No association between Y and x-variables

**Alternative hypothesis Ha:** At least one coefficient.  $F = 144.18$  and with p-value less than 0.05.

The null hypothesis of no association between y and x is rejected and the F-test gives strong support to the fitted model. Linear regression explains variation in Y because  $SSR >> SSE$ , thus F statistic is large.

Using the absolute value of standardized estimate to determine the predictors with significant effect on `Insuicideper100kpop`. The strongest predictor is `dgeneration3` since the standardized estimate is the highest 0.76. When performing t-test on individual parameters, `gdp_per_capita_dollars`, `dage2`, `dyear1`, `dyear2`, `dyear3`, `dyear4`, and `dcontinent3` have p-values that are higher than 0.05 which make them insignificant X-variables. The other variables' p-values (`population`, `gdp_for_year_dollars`, `dsex`, `dgeneration1`, `dgeneration2`, `dgeneration3`, `dgeneration4`, `dage1`, `dage3`, `dage4`, `dage5`, `dcontinent1`, `dcontinent2`) are less than 0.05 which make them significant X-variables.

The coefficient value of the parameter of X measures the predicted change in Y "log(suicideper100kpop)" for any unit increase in X while the other independent variables stay constant. For example, if the gender variable changed from a female to a male, the logarithmic transformation of the suicide rates per 100k population will increase by 1.12.

Adj-R2 0.5188 does not show a very good model and a higher Adj-R2 will give a better model.

**i. Model assumptions:**

Create residual plots for M2 (standardized residuals vs predicted; standardized residuals vs x-variables; and normal plot of residuals) to analyze the residual plots and check if the regression model assumptions are met by the data.

**i-a. Assumptions of Constant Variance and Independence:**

1. Plot residuals vs predicted values. **Ram8**

Points are randomly scattered, and residual analysis show no concern for the model fit.

2. Plot residuals vs each x-variable. **Ram9**

There is no linearity shown in all scatterplots where Y-variable is plotted against each X-variable as it's shown in the next matrix scatterplot.

**i-b. Assumptions for linearity:**

1. Scatterplot for each x-variable.

There is no linear association shown in the scatterplots. **Ram10**

2. Plot residuals vs each x-variable.

None of them show a straight line as it was shown in the residuals vs each x-variable graphs in the first part above. Therefore, there is no linearity.

**i-c. Assumptions for Normality:**

Plot the normal probability plot of the residuals. It's almost straight which means it's normal. Ram11

**j. Analyze if there are any outliers and/or influential points for M2 model.**

Yes, there are multiple outliers and influential points as they are shown in the figures below.

I would exclude all the outliers and influential points and rerun the model again.

Outliers are the observation numbers (2020, 2153, 2179)

Influential points are the observation numbers (84, 218, 667, 1051, 1132, 1384, 1562, 1631, 1812, 2020, 2153, 2179)

**Ram12**

After removing all these 12 observations and rerun the model, there will be no outliers or influential points in the new dataset. We can also see improvements as R<sup>2</sup> and Adj-R<sup>2</sup> increased from (0.5225 – 0.5188) to (0.5336 – 0.5301) and RMSE decreased from 0.9683 to 0.9484.

**5. Split the new data set to create Test and Training sets. 70% data for training and 30% data set for testing. Ram13**

Create a new variable that's equal to dependent variable for training set.

**6. Apply two variable selection procedures to find an optimal subset of independent variables to predict suicide rates per 100k population.**

The first selection is cp method: Ram14

The best model is the first one with cp = 7.569 and 9 X-variables. Good model contains as few variables as possible.

The second selection is stepwise: Ram15

The best model is the one that has the lowest number of independent variables which was suggested by both stepwise and cp selection.

**7. Fit a regression model M3 for Suicide rates per 100k population based on the selection results.**

**Ram16**

**a. Regression Model M3:**

**Insuicideper100kpopulation** = 1.4862 + -2.18E-8\*population + 6.163E-14\* gdp\_for\_year\_dollars + 1.151\*dsex - 2.2276\* dgeneration3 – 2.2128\* dgeneration4 + 2.0511\* dage1 + .31566\* dage5 – 0.83\* dcontinent1 - 0.3211\* dcontinent2

**b. Diagnosing Multicollinearity:**

1. Scatterplot matrix and Pearson correlation matrix for each pair of x variables:

There is a significant collinearity between dage1 and dgeneration3 and the correlation coefficient is 0.91687. As it's mentioned above, these variables with high collinearity and VIFs are dummy variables

that represent a categorical variable with three or more categories. We can ignore them. Otherwise, there is no any other collinearity between other X variables because all the correlation values are low.

### **Ram17**

#### **2. Compute Variance Inflation Factor (VIF):**

All VIFs < 10. These VIFs suggest no collinearity for any of the independent variables.

#### **3. Compute Tolerance value (TOL):**

All TOLs > 0.1. These TOLs suggest no collinearity for any of the independent variables.

### **c. Model assumptions:**

**Create residual plots for M3 (standardized residuals vs predicted; standardized residuals vs x-variables; and normal plot of residuals) to analyze the residual plots and check if the regression model assumptions are met by the data.**

#### **c-a. Assumptions of Constant Variance and Independence:**

##### **1. Plot residuals vs predicted values. Ram18**

Points are randomly scattered, and residual analysis show no concern for the model fit.

##### **2. Plot residuals vs each x-variable. Ram19**

Points are randomly scattered, and residual analysis show no concern for the model fit.

#### **c-b. Assumptions for linearity:**

##### **1. Scatterplot for each x-variable.**

There is no linear association shown in the scatterplots. Ram20

##### **2. Plot residuals vs each x-variable.**

None of them show a straight line as it was shown in the residuals vs each x-variable graphs in the first part above. Therefore, there is no linearity.

#### **c-c. Assumptions for Normality:**

Plot the normal probability plot of the residuals. It's almost straight which means it's normal. Ram21

### **d. Analyze if there are any outliers and/or influential points for M3 model.**

Yes, there are multiple influential points and an outlier as they are shown in the figures below.

I would exclude all the influential points and rerun the model again.

The only outlier is the observation numbers (2274)

Influential points are the observation numbers (7, 74, 81, 87, 174, 217, 231, 955, 1127, 1207, 1314)

After removing all these 12 observations and rerun the model, there will be no outliers or influential points in the new dataset. Our model improved as R<sup>2</sup> and Adj-R<sup>2</sup> increased from (0.532 – 0.5295) to (0.5438 – 0.5413) and RMSE decreased from 0.9465 to 0.9338. Ram22

**e. Final Fitted regression Model M4:**

**Insuicideper100kpopulation** = 1.47489 + -2.04877E-8\*population + 6.1142E-14\* gdp\_for\_year\_dollars + 1.17262\*dsex - 2.48734\* dgeneration3 - 2.23228\* dgeneration4 + 2.3005\*dage1 + .30628\*dage5 - 0.88739\* dcontinent1 - 0.31239\* dcontinent2

Where dsex = 1 if sex=male, and dsex=0 if sex=female

dgeneration3 = 1 if generation = 'Millenials', or otherwise dgeneration3 = 0

dgeneration4 = 1 if generation = 'Generation Z', or otherwise dgeneration4 = 0

dage1 = 1 if age = '15-24 years', or otherwise dage1 = 0

dage5 = 1 if age = '75+ years', or otherwise dage5 = 0

dcontinent1 = 1 if continent = 'Africa', or otherwise dcontinent1 = 0

dcontinent2 = 1 if continent = 'Asia', or otherwise dcontinent2 = 0

**f. Test performance for the new model M4:**

Compute performance statistics – Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and R<sup>2</sup> for Test set. Ram23

For Test set: RMSE = 0.95536, MAE = 0.7516, R<sup>2</sup> = (yhat<sup>2</sup>) = 0.72762<sup>2</sup> = 0.529

For Training set: RMSE = 0.93381, R<sup>2</sup> = 0.5438

Abs(model R<sup>2</sup> - R<sup>2</sup> cv) = 0.529 - 0.5438 = 0.0148

**g. Use SAS to compute the predicted suicide rates per 100k population and the confidence interval for:**

**g-a. 100000 male population who are Millennials 15-24 years old in Africa, and have GDP = \$150000000000.**

Predicted Insuicidesper100kpop = 1.6654 with confidential intervals (-0.2427 – 3.5735).

That means 95% of the time, the predicted Insuicidesper100kpop would fall within this range.

**Suicide rate per 100k population** = e(1.6654)-1= 4.2877

confidential intervals [(e(-0.2427)-1) – (e(3.5735)-1)] = (0 - 34.64)

**g-b. 100000 female population who are Millennials 15-24 years old in Asia, and have GDP = \$150000000000.**

Predicted Insuicidesper100kpop = 1.036 with confidential intervals (-0.8657 – 2.9377).

That means 95% of the time, the predicted Insuicidesper100kpop would fall within this range.

**Suicide rate per 100k population** = e(1.036)-1= 1.81792

confidential intervals [(e(-0.8657)-1) – (e(2.9377)-1)] = (0 – 18.872)

Ram24

**8. Fitting a Model with an interaction term:**

An interaction occurs when an independent variable has a different effect on the outcome depending on

the values of another independent variable. Thus, I am going to use the product of population and gdp\_for\_year\_dollars.

Check Multicollinearity using Pearson Correlation Coefficients. Then, center the data for of population and gdp\_for\_year\_dollars. Ram25

**a. Run Model Selection**

Ram26

**b. The new model with interaction terms M5:**

Ram27

**c. Model equation with centered data:**

$$\text{Insuicideper100kpop} = 1.61545 + 2.1787E-20 ((2569528-\text{population})*\text{gdp\_for\_year\_dollars}) - 2.4786E-20 ((7.26759E11-\text{gdp\_for\_year\_dollars})*\text{population}) + 4.31323E-8(2569528-\text{population})$$

$$\beta_1 = 2.1787E-20$$

$$\beta_2 = 2.4786E-20$$

$$\beta_3 = 4.31323E-8$$

**d. Compute the predicted suicide rates per 100k population:**

**d-a. 100000 population who have GDP for year = \$150000000000.**

$$\text{Insuicideper100kpop} = 1.61545 + 2.1787E-20 ((2569528-\mathbf{100000})* \mathbf{150000000000}) - 2.4786E-20 ((7.26759E11- \mathbf{150000000000})* \mathbf{100000}) + 4.31323E-8(2569528-\mathbf{100000})$$

$$\text{Insuicideper100kpop} = 1.61545 + 0.00807 - 0.001429555 + 0.106516 = 1.728606445$$

$$\text{Suicide rates per 100k population} = e(1.728606445) - 1 = 4.6327988$$

**d-b. 200000 population who have GDP for year = \$300000000000.**

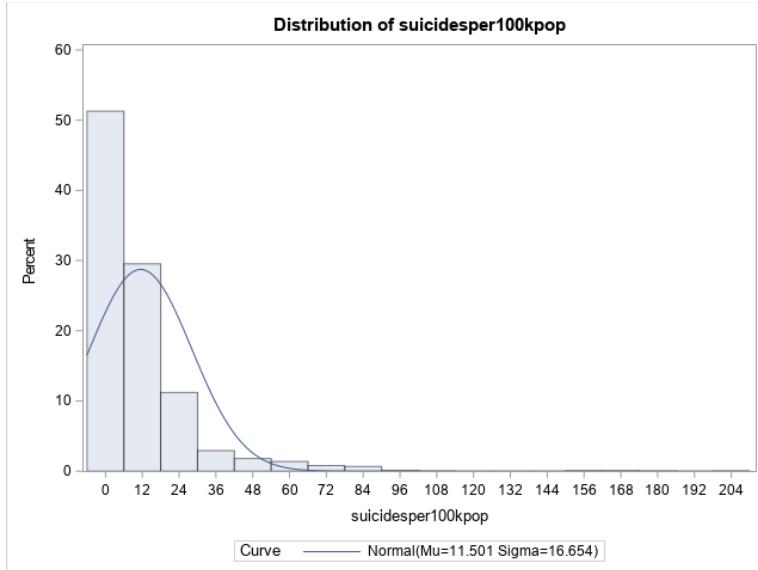
$$\text{Insuicideper100kpop} = 1.61545 + 2.1787E-20 ((2569528-\mathbf{200000})* \mathbf{300000000000}) - 2.4786E-20 ((7.26759E11- \mathbf{300000000000})* \mathbf{200000}) + 4.31323E-8(2569528-\mathbf{200000})$$

$$\text{Insuicideper100kpop} = 1.61545 + 0.015487472 + 0.00148716 + 0.102203 = 1.734627632$$

$$\text{Suicide rates per 100k population} = e(1.734627632) - 1 = 4.66681$$

## Appendix

### Ram1



**The UNIVARIATE Procedure**  
Variable: suicidesper100kpop

Moments			
<b>N</b>	2391	<b>Sum Weights</b>	2391
<b>Mean</b>	11.5010163	<b>Sum Observations</b>	27498.93
<b>Std Deviation</b>	16.6535166	<b>Variance</b>	277.339617
<b>Skewness</b>	4.11398499	<b>Kurtosis</b>	27.9044005
<b>Uncorrected SS</b>	979107.326	<b>Corrected SS</b>	662841.684
<b>Coeff Variation</b>	144.800392	<b>Std Error Mean</b>	0.34057767

Basic Statistical Measures			
<b>Location</b>		<b>Variability</b>	
<b>Mean</b>	11.50102	<b>Std Deviation</b>	16.65352
<b>Median</b>	5.63000	<b>Variance</b>	277.33962
<b>Mode</b>	0.29000	<b>Range</b>	204.88000
		<b>Interquartile Range</b>	12.93000

**Tests for Normality**

Test	Statistic	p Value
Kolmogorov-Smirnov	D	0.245662
Cramer-von Mises	W-Sq	41.38562
Anderson-Darling	A-Sq	227.606

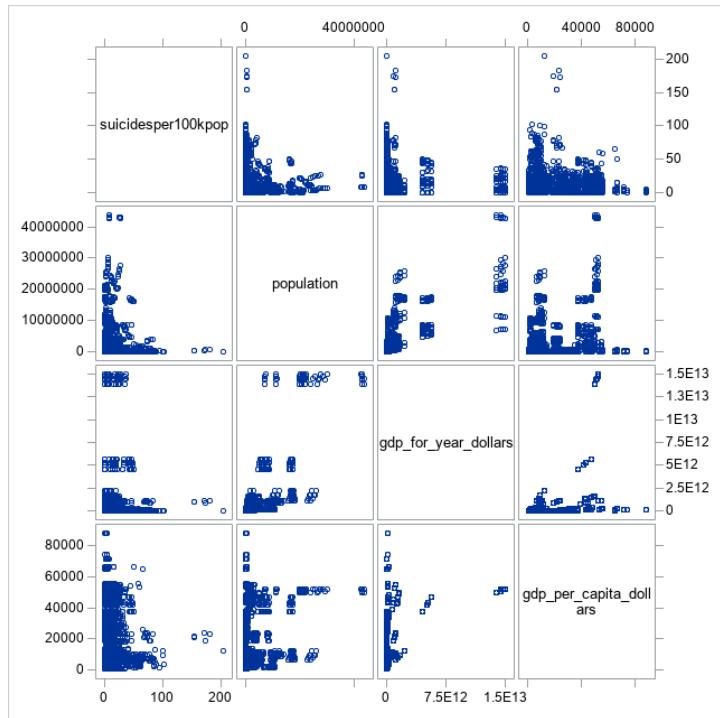
  

**Quantiles (Definition 5)**

Level	Quantile
100% Max	204.92
99%	79.39
95%	42.37
90%	26.19
75% Q3	14.89
50% Median	5.63
25% Q1	1.96
10%	0.75
5%	0.41
1%	0.20
0% Min	0.04

## Ram2

Pearson Correlation Coefficients, N = 2391 Prob >  r  under H0: Rho=0				
	suicidesper100kpop	population	gdp_for_year_dollars	gdp_per_capita_dollars
suicidesper100kpop	1.00000	-0.03458 0.0910	0.04420 0.0307	0.01281 0.5314
population	-0.03458 0.0910	1.00000	0.75403 <.0001	0.20618 <.0001
gdp_for_year_dollars	0.04420 0.0307	0.75403 <.0001	1.00000	0.42895 <.0001
gdp_per_capita_dollars	0.01281 0.5314	0.20618 <.0001	0.42895 <.0001	1.00000



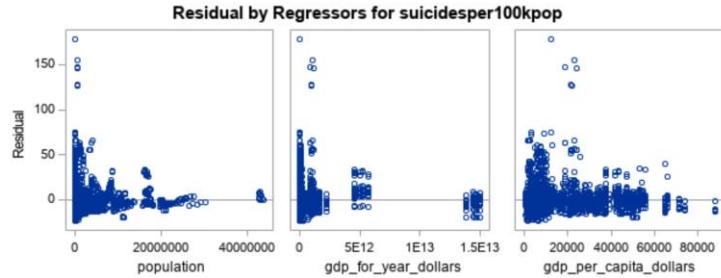
## Ram3

The REG Procedure  
Model: MODEL1  
Dependent Variable: suicidesper100kpop

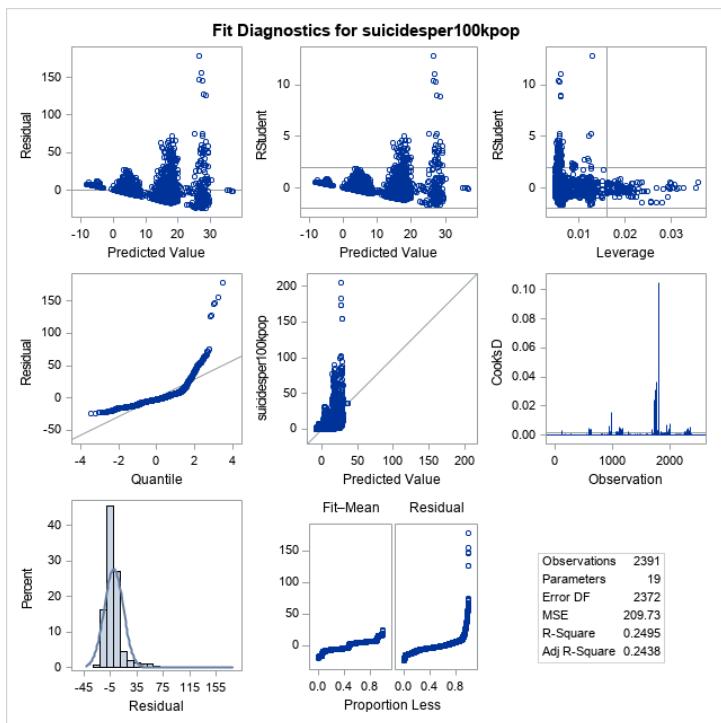
Number of Observations Read	2391
Number of Observations Used	2391

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	165363	9186.83257	43.80	<.0001
Error	2372	497479	209.72964		
Corrected Total	2390	662842			

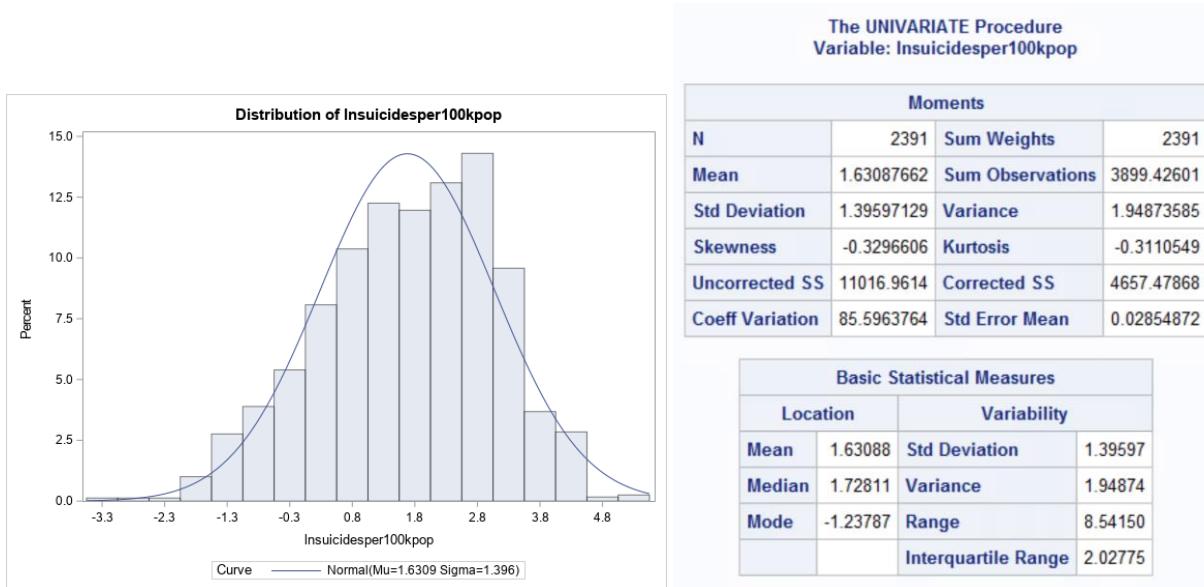
Root MSE	14.48205	R-Square	0.2495
Dependent Mean	11.50102	Adj R-Sq	0.2438
Coeff Var	125.91970		



Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Standardized Estimate	Tolerance	Variance Inflation
Intercept	B	17.46356	1.12029	15.59	<.0001	0	-	0
population	1	-1.86191E-7	9.308322E-8	-2.00	0.0456	-0.05801	0.37618	2.65827
gdp_for_year_dollars	1	7.35842E-13	2.2121E-13	3.33	0.0009	0.10461	0.31996	3.12535
gdp_per_capita_dollars	1	-0.00002697	0.00002152	-1.25	0.2101	-0.02767	0.64947	1.53971
dsex	1	11.76773	0.59455	19.79	<.0001	0.35264	0.99676	1.00325
dgeneration1	B	-9.61350	1.15412	-8.33	<.0001	-0.20578	0.51846	1.92878
dgeneration2	B	-9.52091	1.84429	-5.16	<.0001	-0.23752	0.14947	6.69026
dgeneration3	B	-21.51656	2.06251	-10.43	<.0001	-0.52282	0.12598	7.93773
dgeneration4	B	-20.24836	1.22524	-16.53	<.0001	-0.37667	0.60908	1.64183
dage1	1	9.17208	2.02353	4.53	<.0001	0.21099	0.14603	6.84778
dage2	B	-0.41815	1.78659	-0.23	0.8150	-0.00971	0.18398	5.43524
dage3	0	0	-	-	-	-	-	-
dage4	B	-9.24097	1.07246	-8.62	<.0001	-0.20940	0.53576	1.86649
dage5	0	0	-	-	-	-	-	-
dyear1	1	-1.31230	0.98250	-1.34	0.1818	-0.03147	0.56990	1.75468
dyear2	1	-0.98567	0.99989	-0.99	0.3243	-0.02312	0.57522	1.73846
dyear3	1	-2.10420	0.97373	-2.16	0.0308	-0.05136	0.56019	1.78510
dyear4	1	-1.66732	1.04803	-1.59	0.1118	-0.04039	0.49088	2.03718
dcontinent1	1	-2.52446	1.36005	-1.86	0.0636	-0.03388	0.94993	1.05271
dcontinent2	1	-0.67810	0.66794	-1.02	0.3101	-0.01969	0.84138	1.18853
dcontinent3	1	-0.64698	1.33514	-0.48	0.6280	-0.00942	0.83686	1.19495

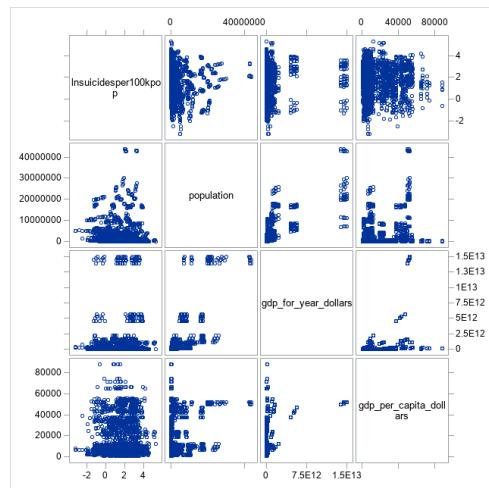


## Ram4

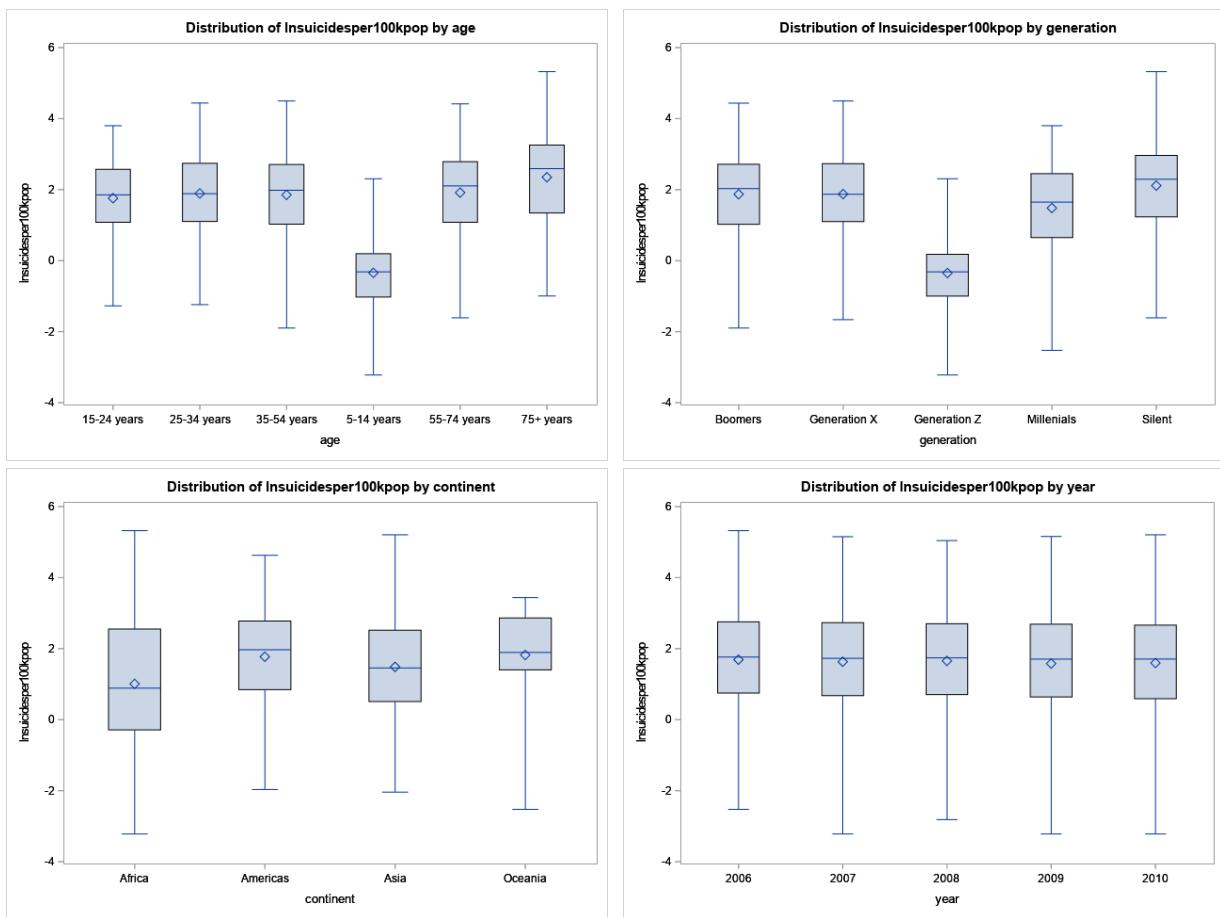


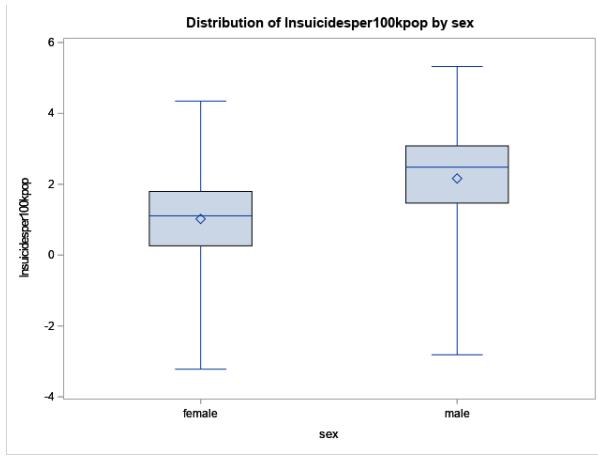
## Ram5

Pearson Correlation Coefficients, N = 2391 Prob >  r  under H0: Rho=0				
	Insuicidesper100kpop	population	gdp_for_year_dollars	gdp_per_capita_dollars
Insuicidesper100kpop	1.00000	-0.03964 0.0526	0.05480 0.0074	0.05283 0.0098
population	-0.03964 0.0526	1.00000	0.75403 <.0001	0.20618 <.0001
gdp_for_year_dollars	0.05480 0.0074	0.75403	1.00000	0.42895 <.0001
gdp_per_capita_dollars	0.05283 0.0098	0.20618 <.0001	0.42895 <.0001	1.00000



## Ram6





**Ram7**

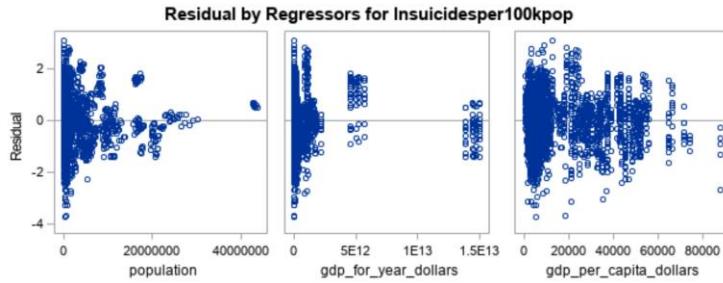
The REG Procedure  
Model: MODEL1  
Dependent Variable: Insuicidesper100kpop

Number of Observations Read	2391
Number of Observations Used	2391

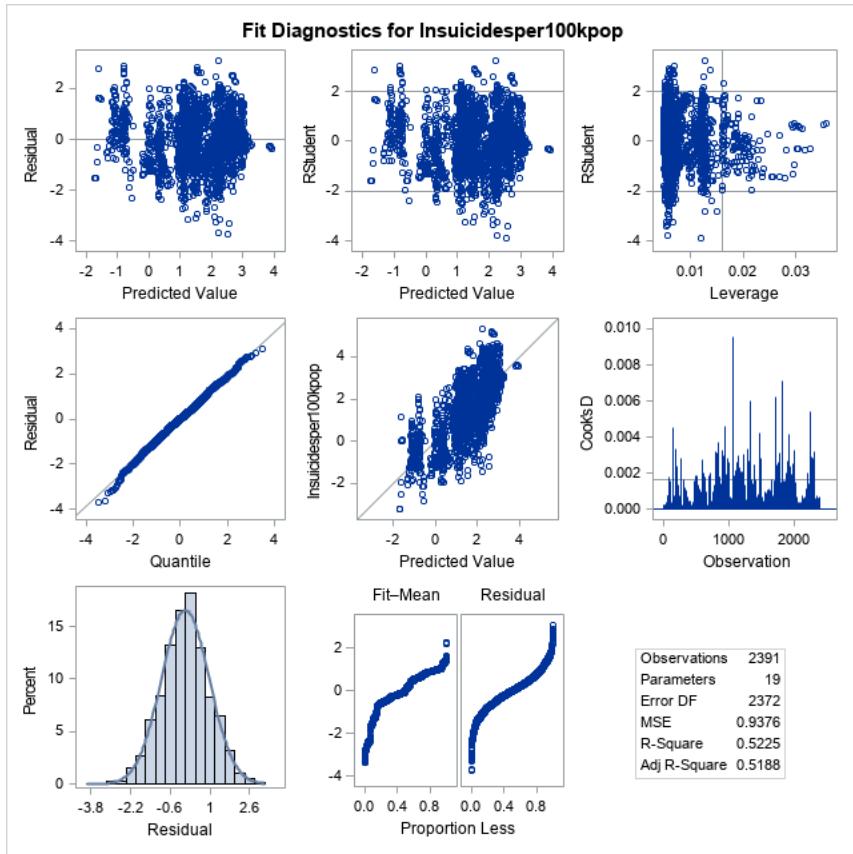
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	2433.39107	135.18839	144.18	<.0001
Error	2372	2224.08761	0.93764		
Corrected Total	2390	4657.47868			

Root MSE	0.96832	R-Square	0.5225
Dependent Mean	1.63088	Adj R-Sq	0.5188
Coeff Var	59.37416		

Quantiles (Definition 5)	
Level	Quantile
100% Max	5.322620
99%	4.374372
95%	3.746441
90%	3.265378
75% Q3	2.700690
50% Median	1.728109
25% Q1	0.672944
10%	-0.287682
5%	-0.891598
1%	-1.609438
0% Min	-3.218876

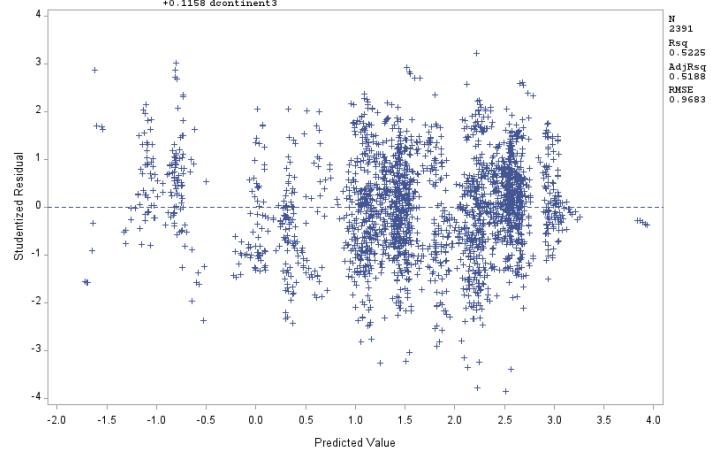


Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Standardized Estimate	Tolerance	Variance Inflation
Intercept	B	1.92352	0.07491	25.68	<.0001	0	.	0
population	1	-2.40457E-8	6.223864E-9	-3.86	0.0001	-0.08938	0.37618	2.65827
gdp_for_year_dollars	1	8.0893E-14	1.47909E-14	5.47	<.0001	0.13719	0.31996	3.12535
gdp_per_capita_dollars	1	-0.00000167	0.00000144	-1.16	0.2465	-0.02041	0.64947	1.53971
dsex	1	1.12233	0.03975	28.23	<.0001	0.40123	0.99676	1.00325
dgeneration1	B	-0.36360	0.07717	-4.71	<.0001	-0.09285	0.51846	1.92878
dgeneration2	B	-0.41357	0.12332	-3.35	0.0008	-0.12308	0.14947	6.69026
dgeneration3	B	-2.64732	0.13791	-19.20	<.0001	-0.76738	0.12598	7.93773
dgeneration4	B	-2.60432	0.08192	-31.79	<.0001	-0.57795	0.60908	1.64183
dage1	1	2.15443	0.13530	15.92	<.0001	0.59122	0.14603	6.84778
dage2	B	0.05168	0.11946	0.43	0.6653	0.01431	0.18398	5.43524
dage3	0	0	-	-	-	-	-	-
dage4	B	-0.35426	0.07171	-4.94	<.0001	-0.09577	0.53576	1.86649
dage5	0	0	-	-	-	-	-	-
dyear1	1	-0.09950	0.06569	-1.51	0.1300	-0.02847	0.56990	1.75468
dyear2	1	-0.04890	0.06686	-0.73	0.4646	-0.01368	0.57522	1.73846
dyear3	1	-0.12130	0.06511	-1.86	0.0626	-0.03532	0.56019	1.78510
dyear4	1	-0.10493	0.07007	-1.50	0.1344	-0.03032	0.49088	2.03718
dcontinent1	1	-0.80749	0.09094	-8.88	<.0001	-0.12927	0.94993	1.05271
dcontinent2	1	-0.29588	0.04466	-6.63	<.0001	-0.10248	0.84138	1.18853
dcontinent3	1	0.11580	0.08927	1.30	0.1947	0.02012	0.83686	1.19495

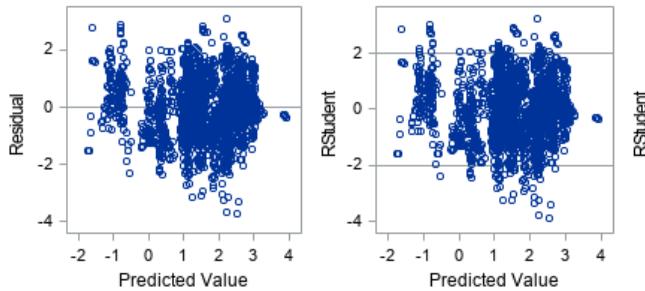


## Ram8

```
lnsuicidesper100kpop = 1.9235 -24E-9 population +81E-15 gdp_for_year_dollars -167E-8 gdp_per_capita_dollars
+1.1223 dsex -0.3636 dgeneration1 -0.4139 dgeneration2 -2.6473 dgeneration3
-2.6026 dgeneration4 +2.1544 dayel +0.0517 dayel2 -0.3543 dayel4 -0.0995 dyear1
+0.0489 dyear2 +1.0113 dyear3 -0.1049 dyear4 -0.8075 decontinent1 -0.3959 decontinent2
+0.11568 decontinent3
```

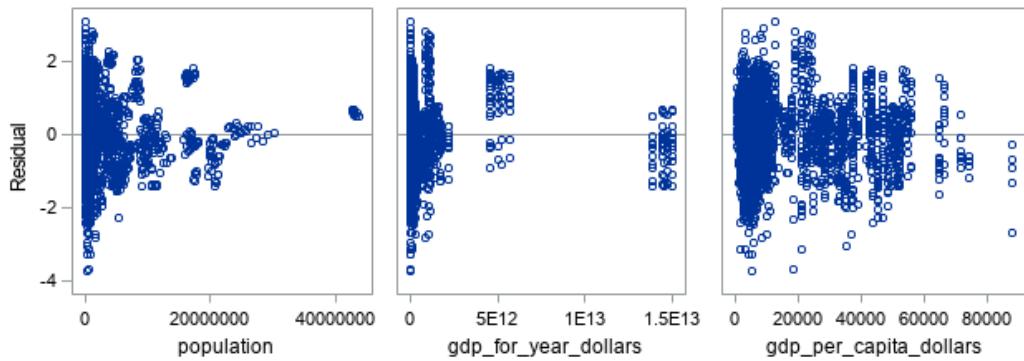


### Fit Diagnostics for Insuicidesper100kpop

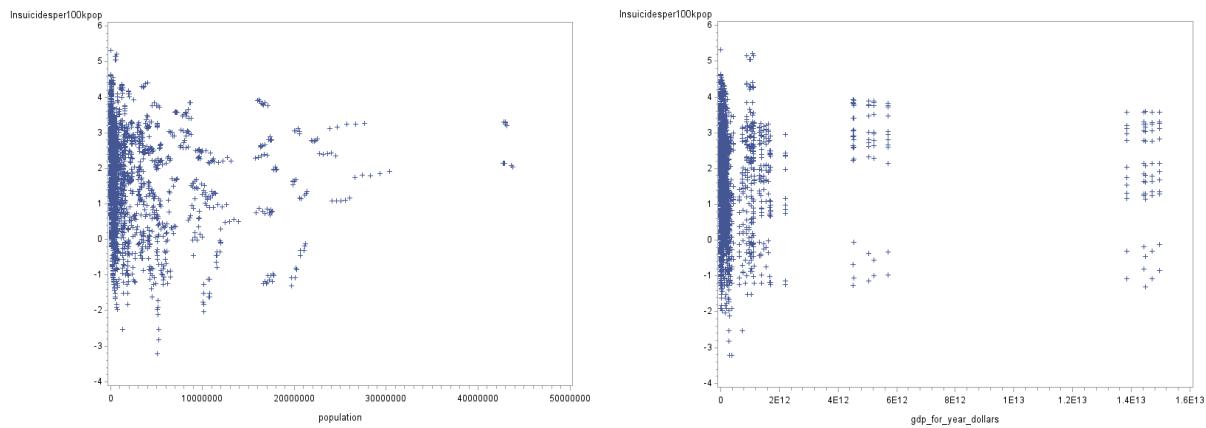


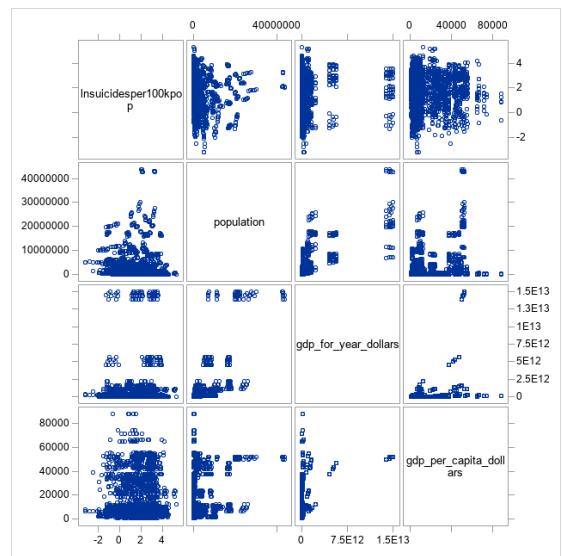
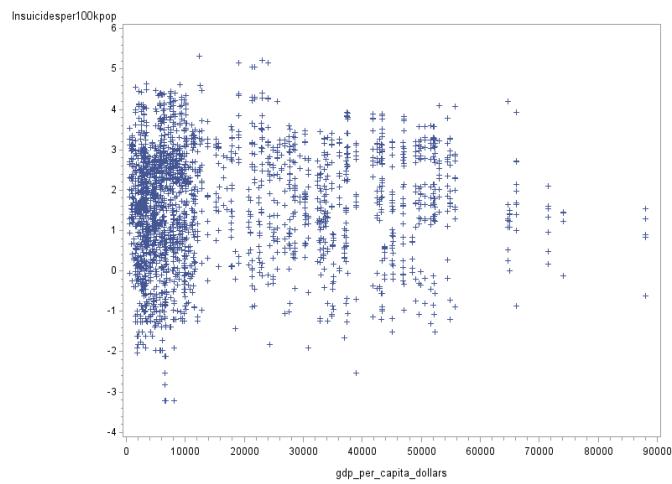
Ram9

### Residual by Regressors for Insuicidesper100kpop



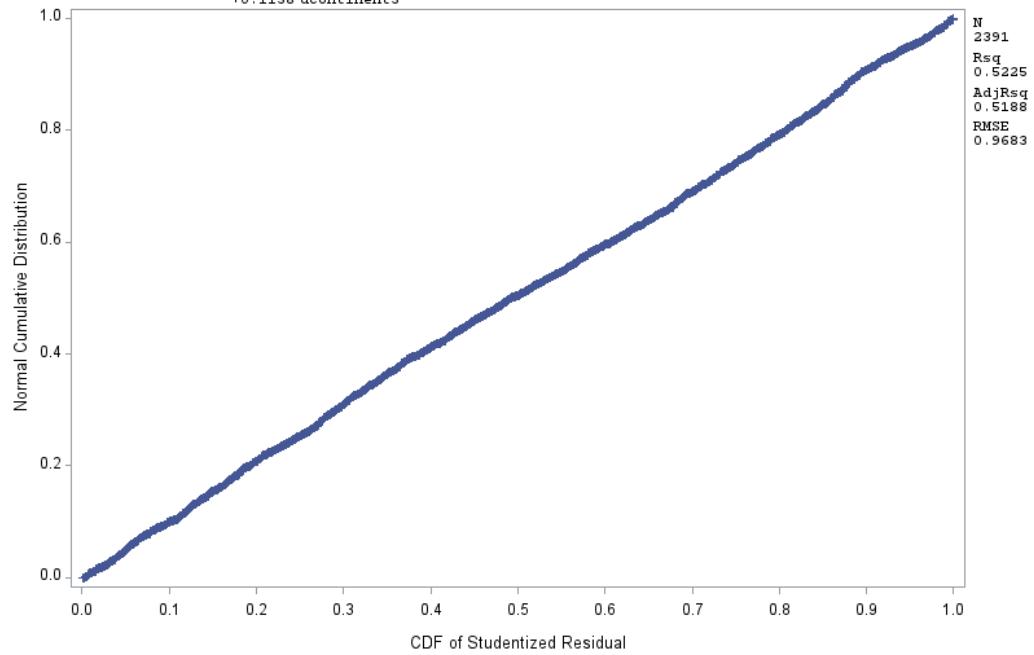
Ram10



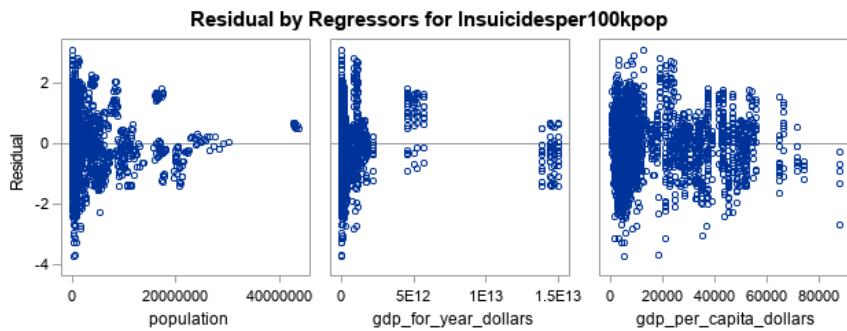
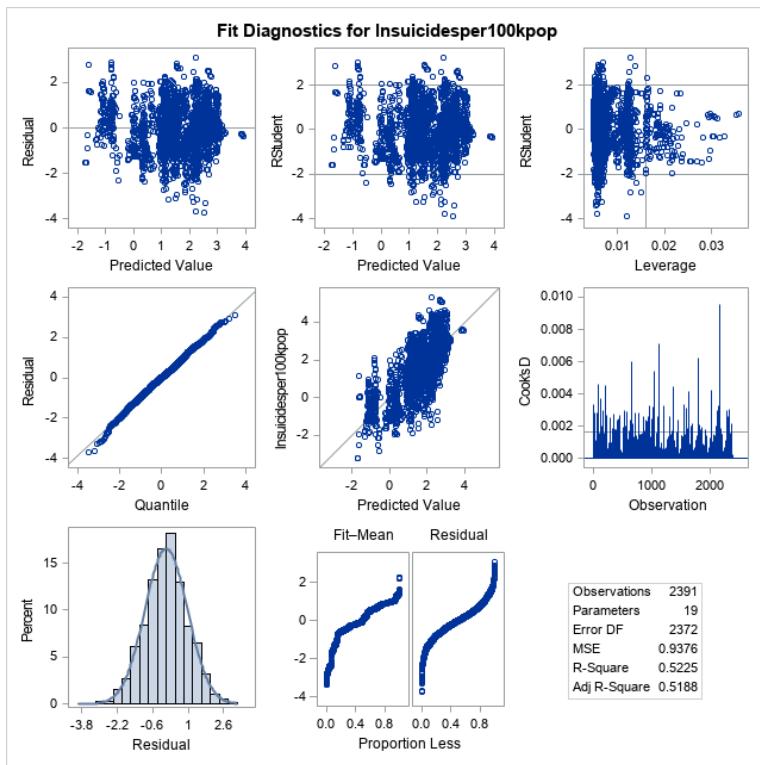


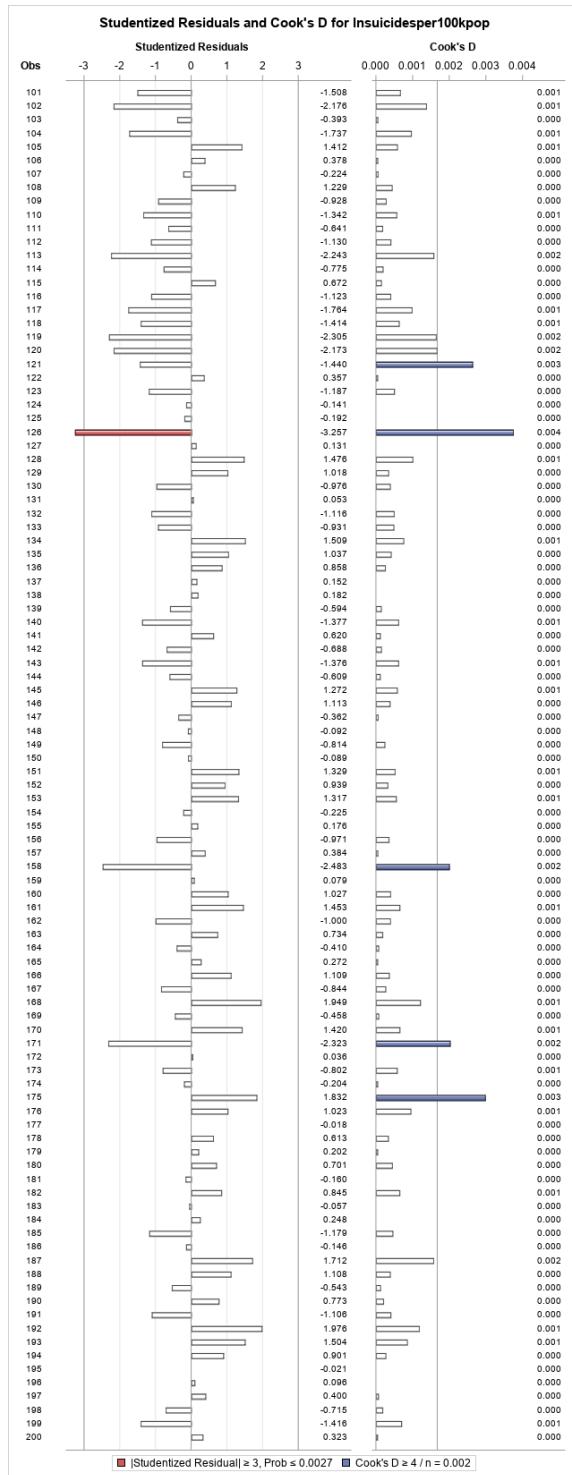
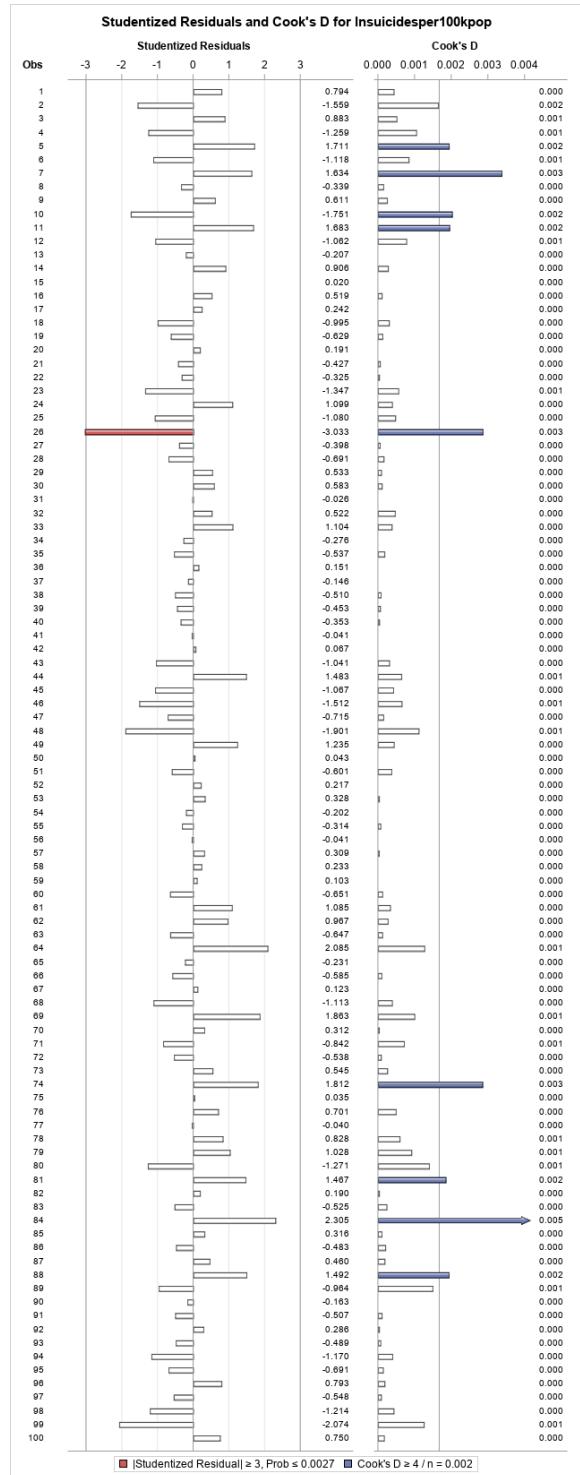
## Ram11

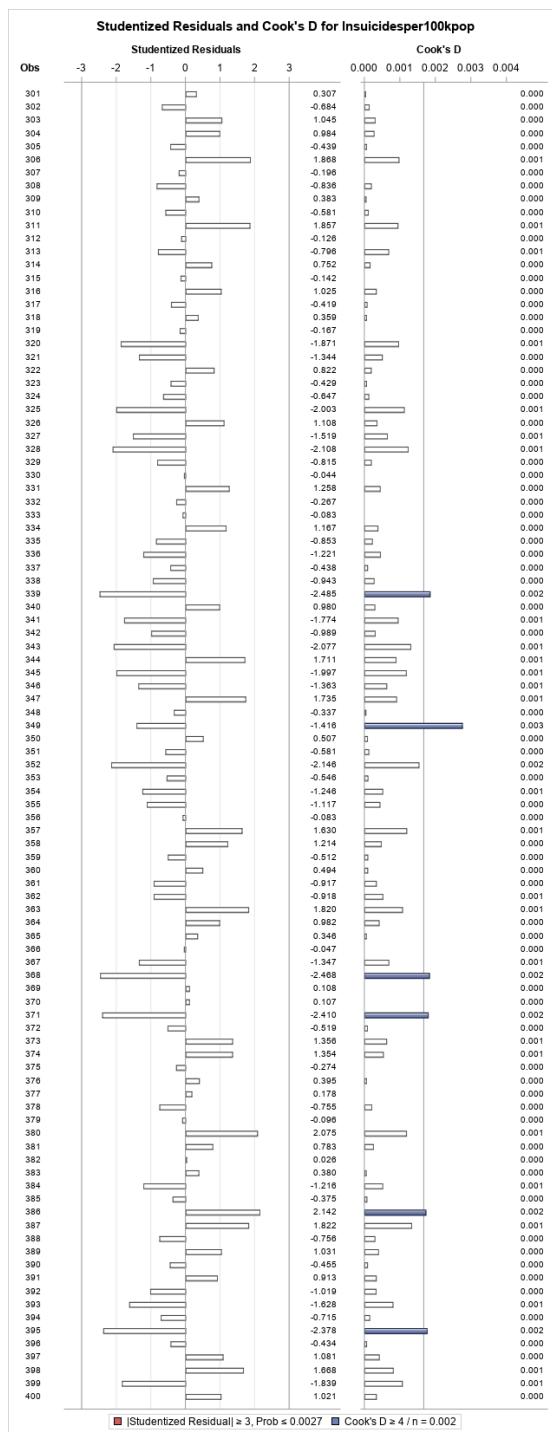
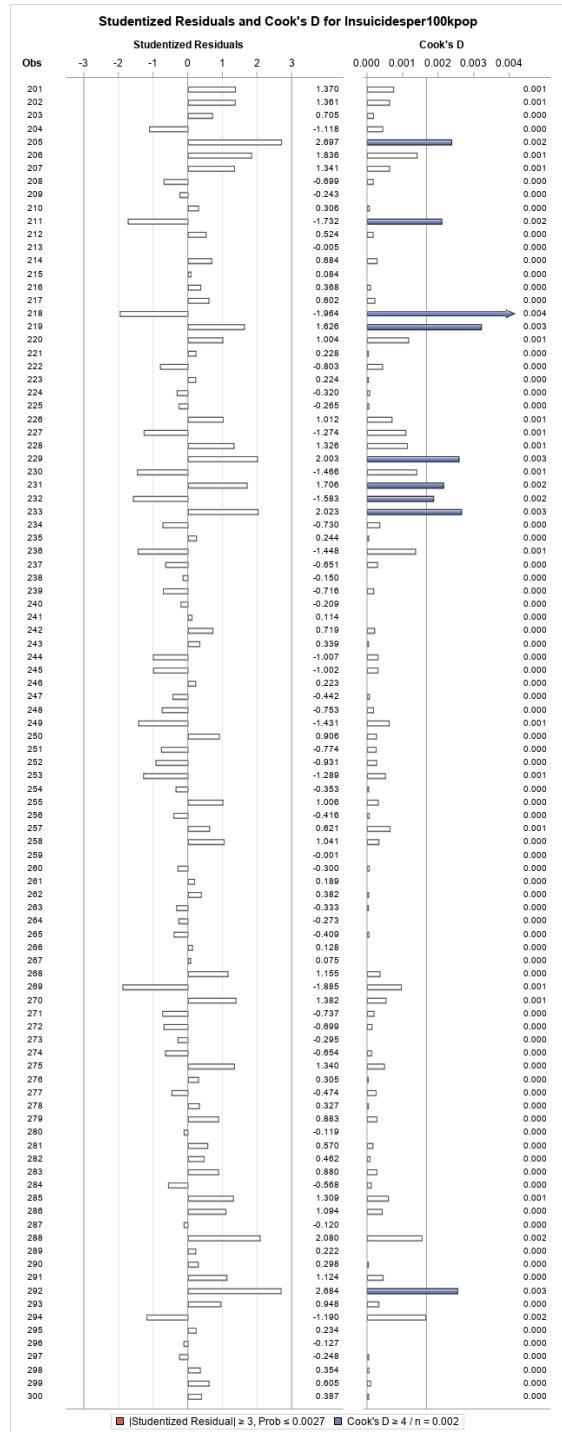
```
lnsuicidesper100kpop = 1.9235 -24E-9 population +81E-15 gdp_for_year_dollars -167E-8 gdp_per_capita_dollars
+1.1223 dsex -0.3636 dgeneration1 -0.4136 dgeneration2 -2.6473 dgeneration3
-2.6043 dgeneration4 +2.1544 dage1 +0.0517 dage2 -0.3543 dage4 -0.0995 dyear1
-0.0489 dyear2 -0.1213 dyear3 -0.1049 dyear4 -0.8075 dcontinent1 -0.2959 dcontinent2
+0.1158 dcontinent3
```

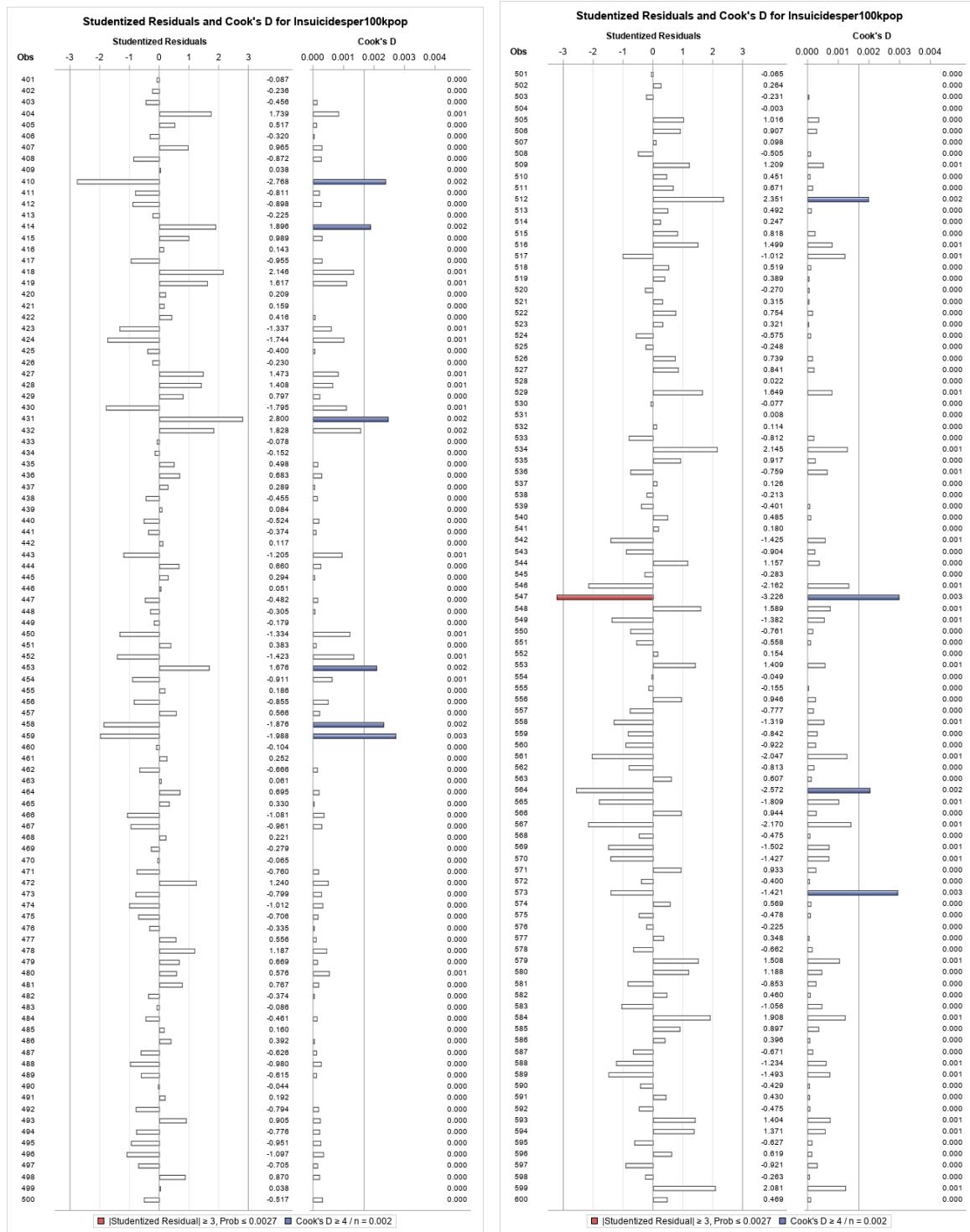


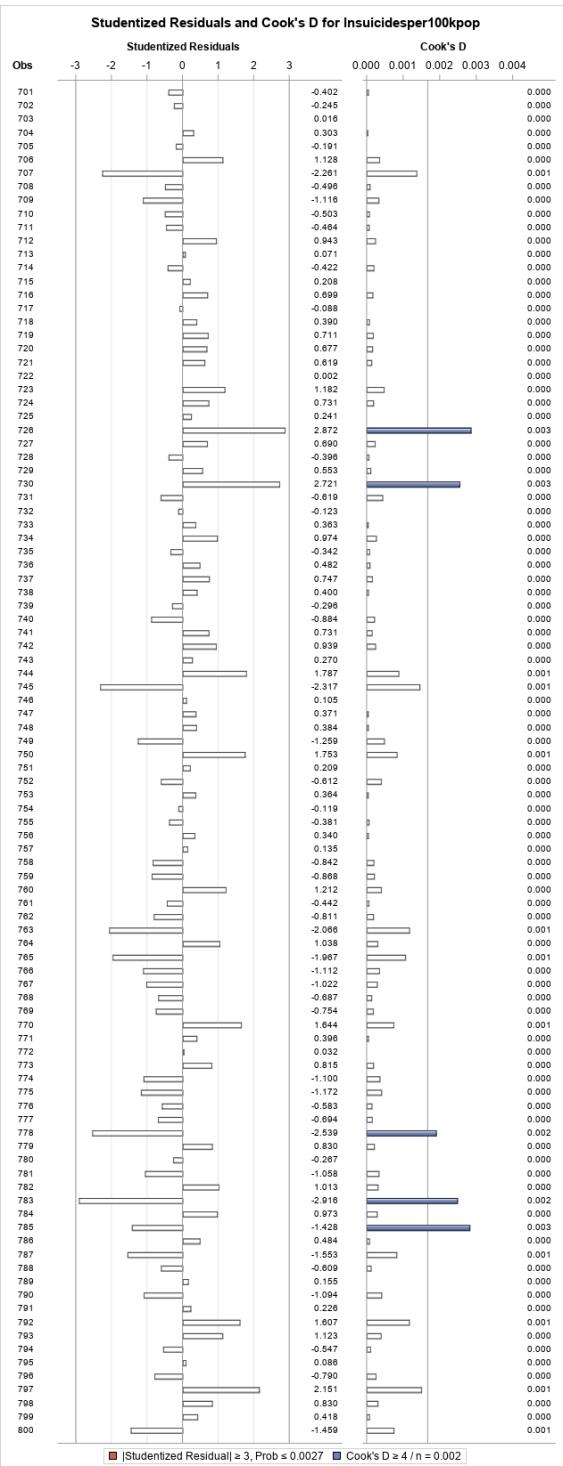
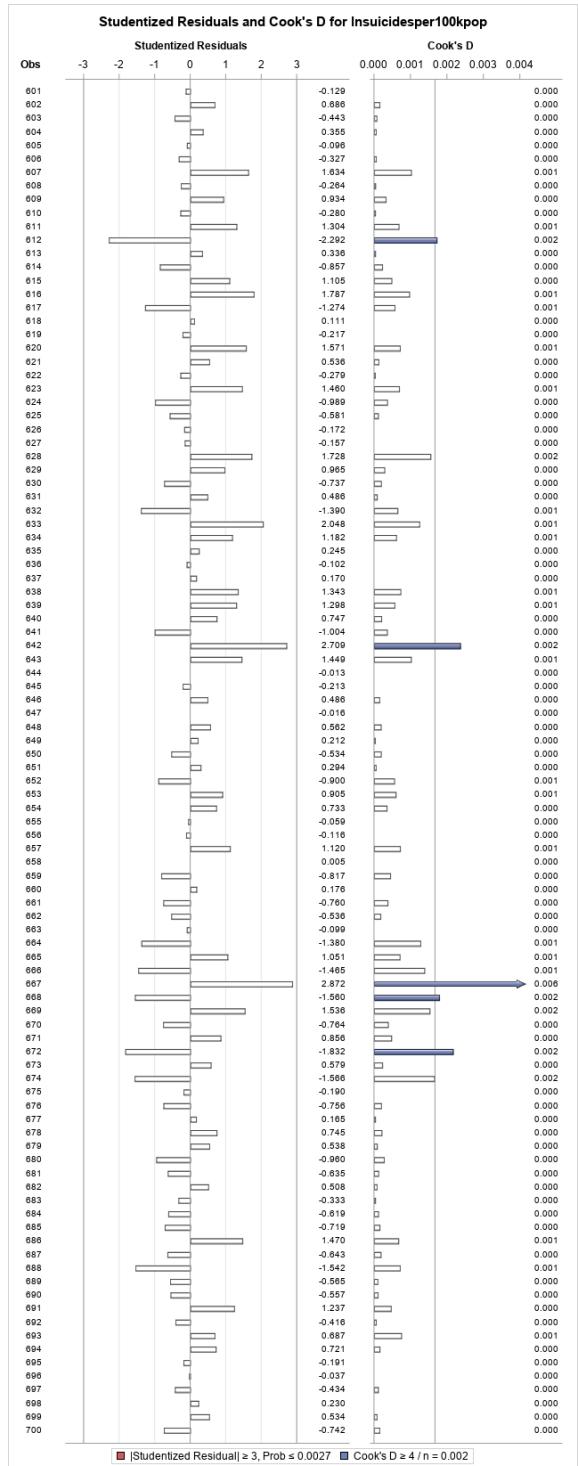
## Ram12

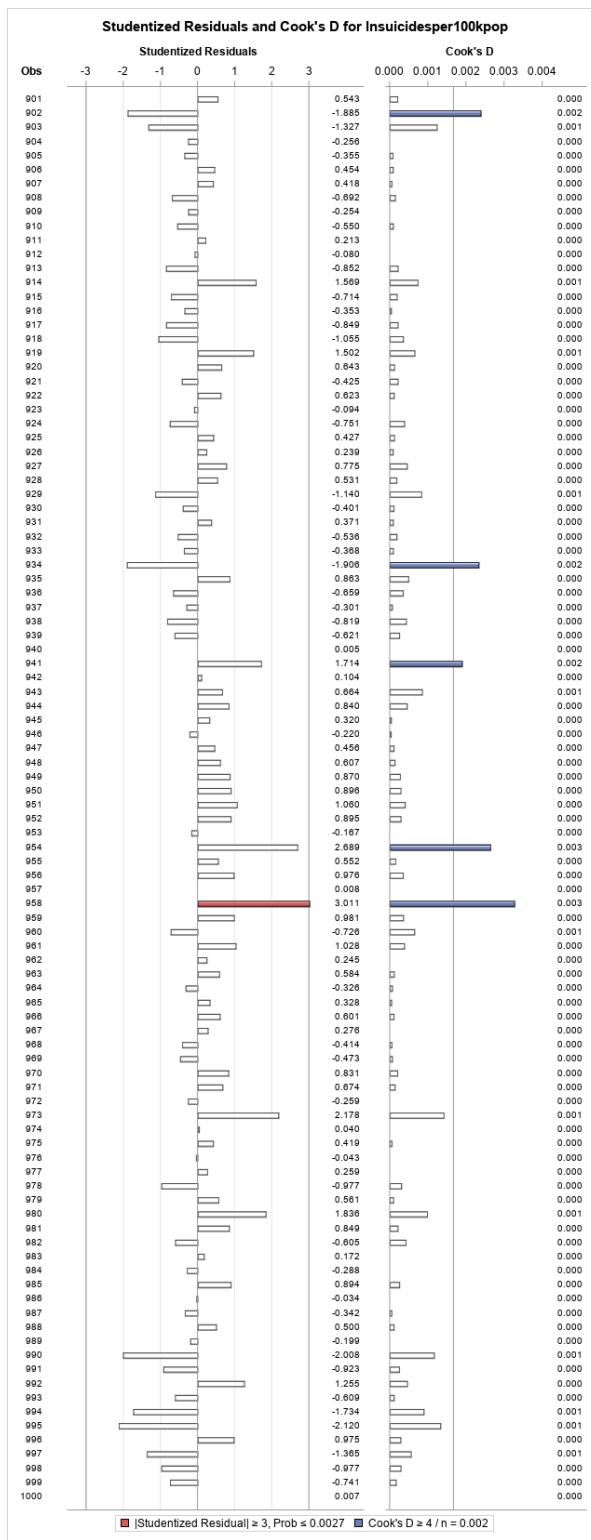
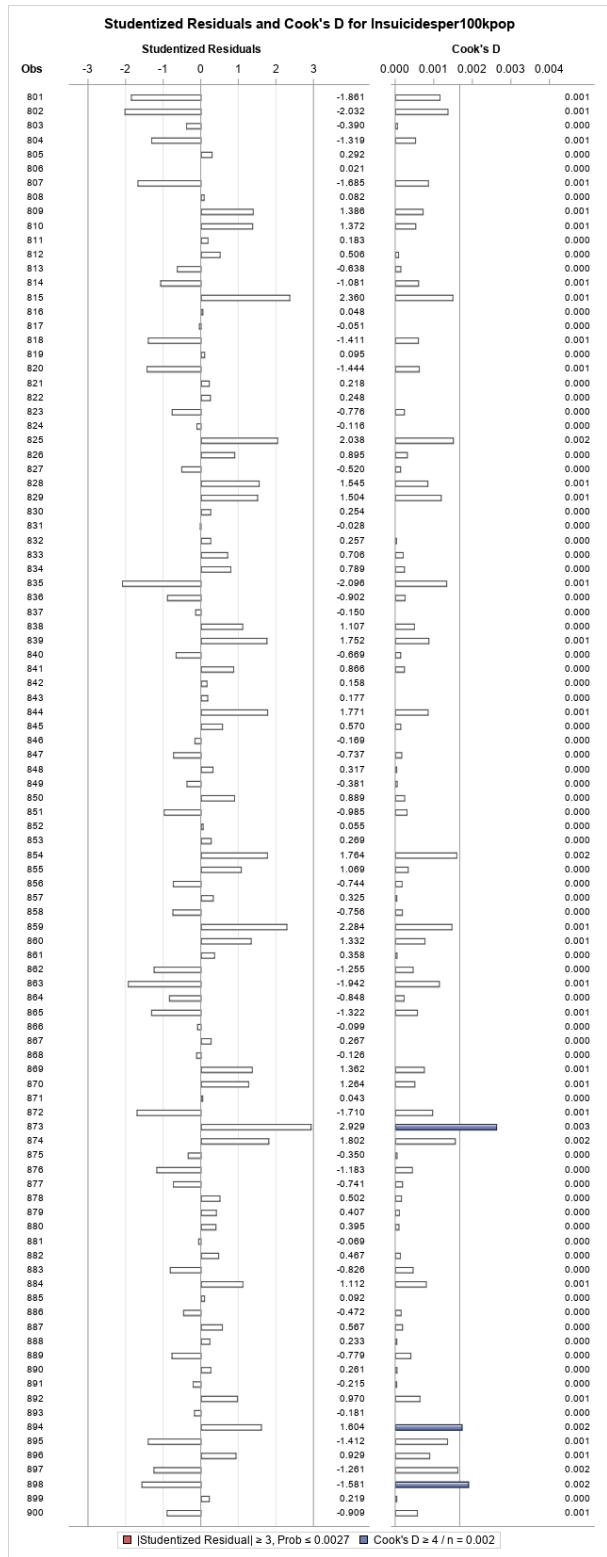


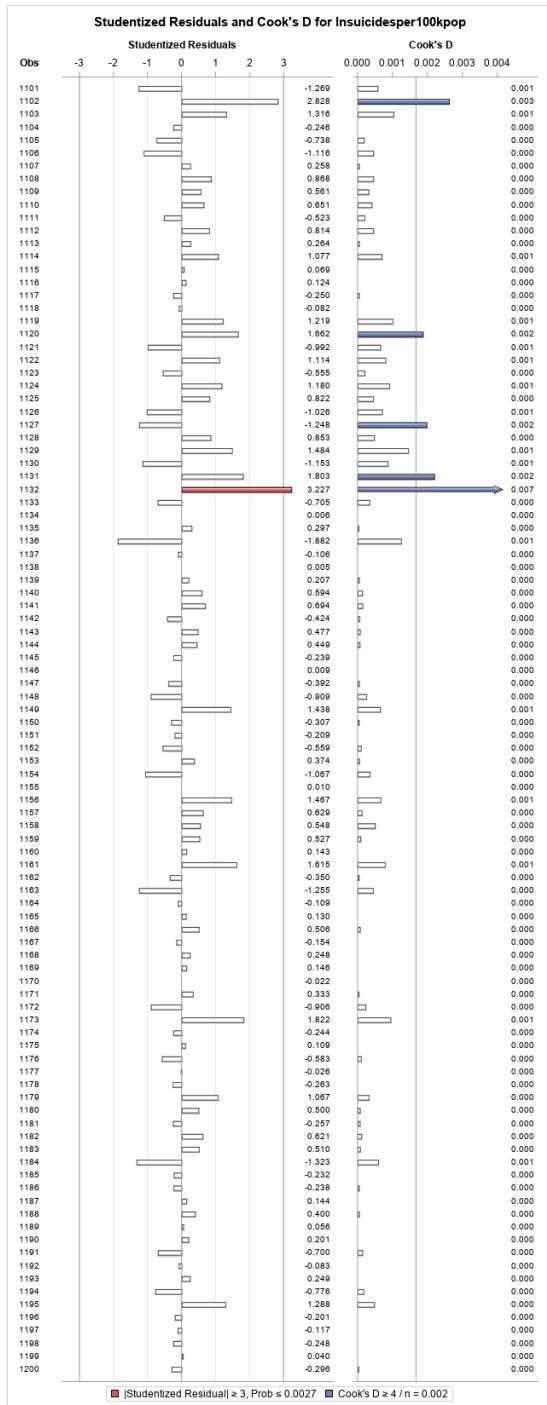
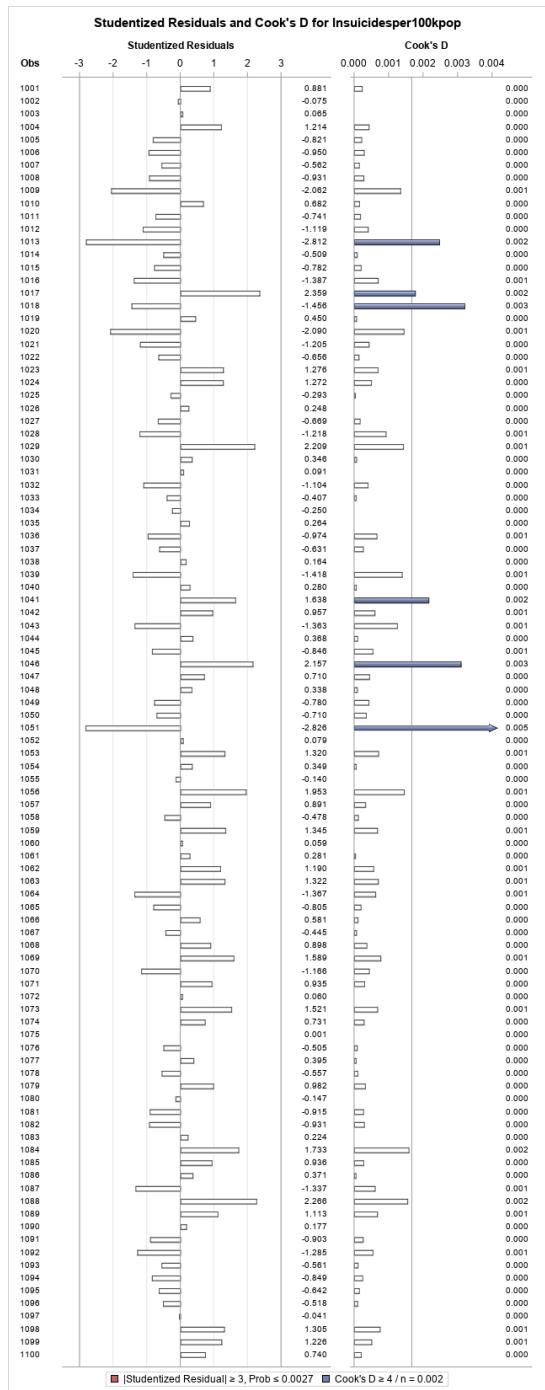


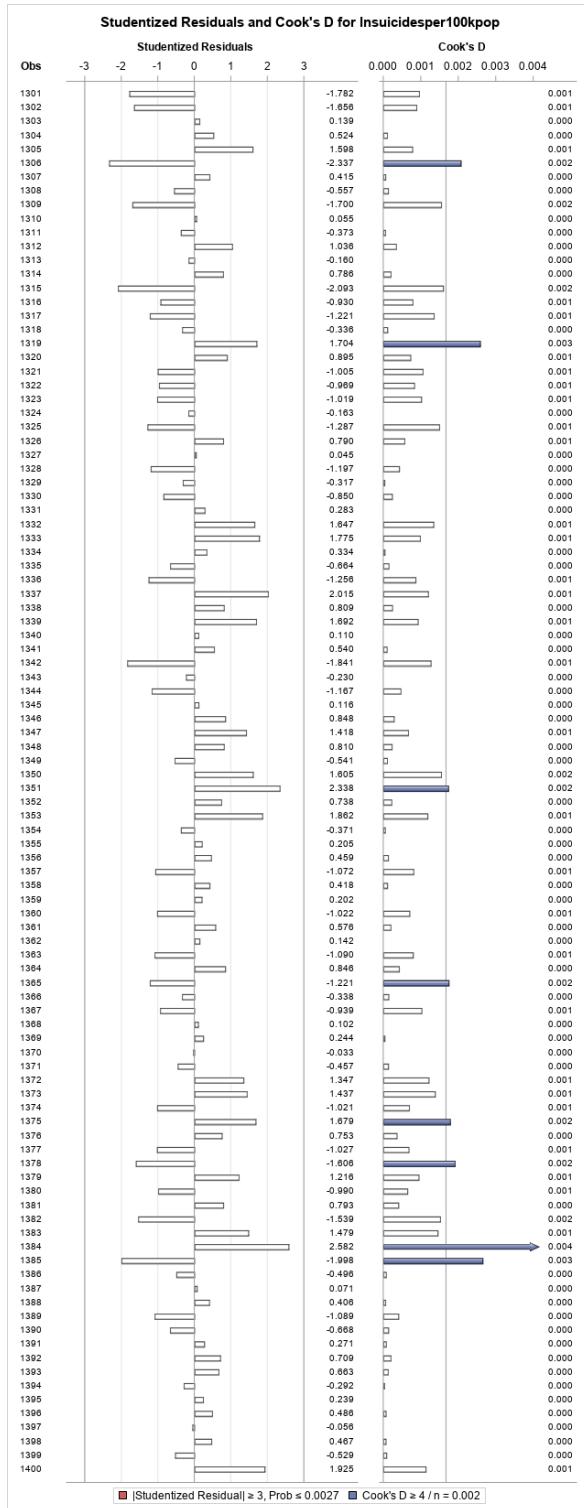
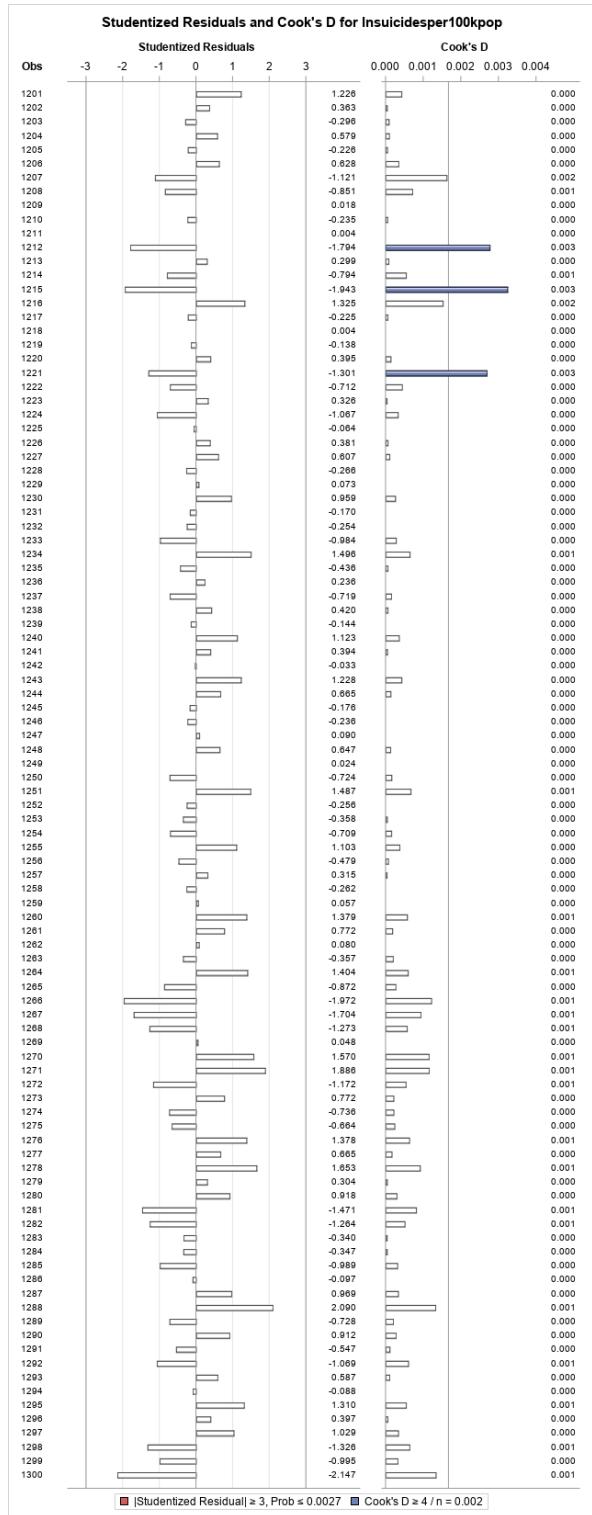


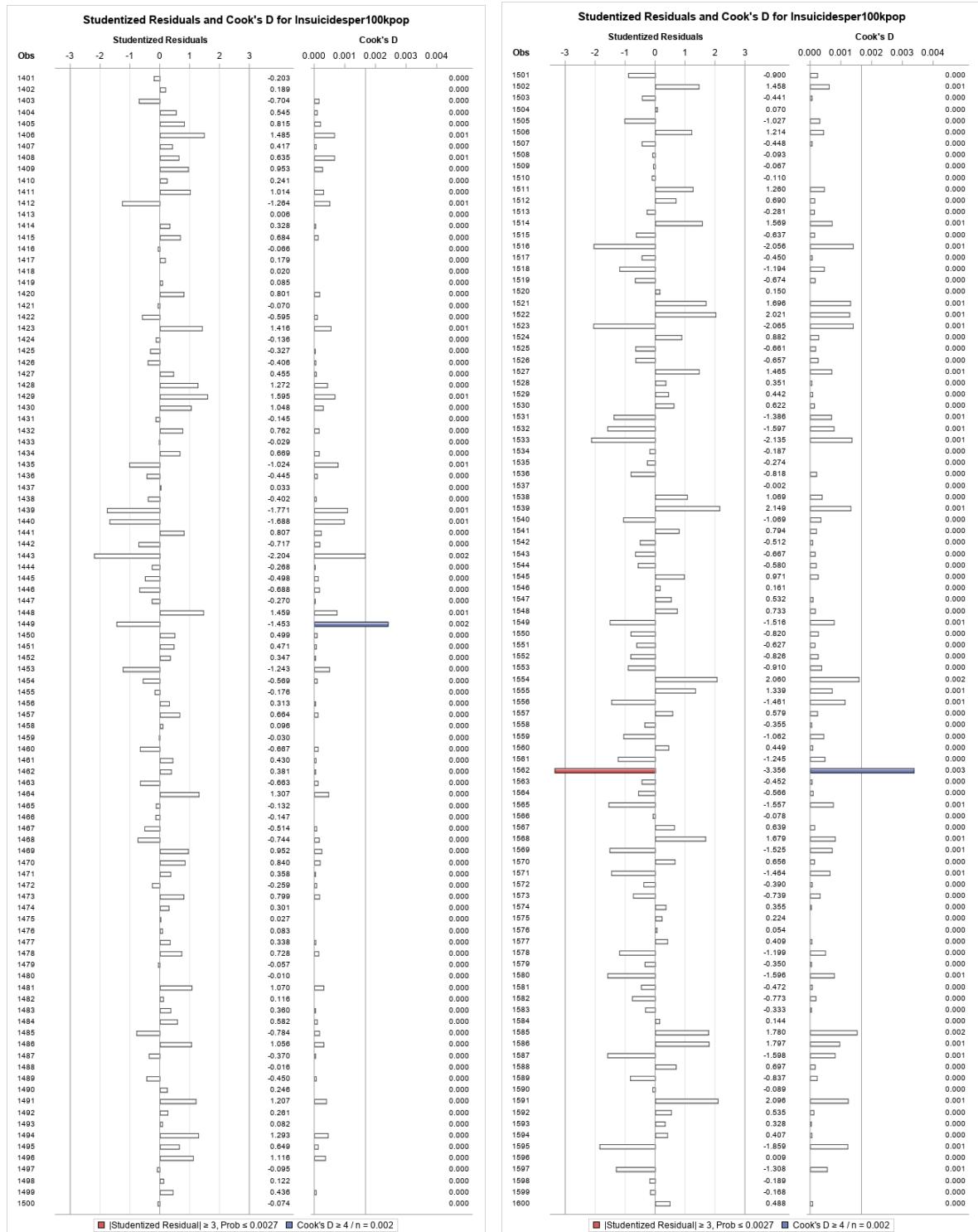


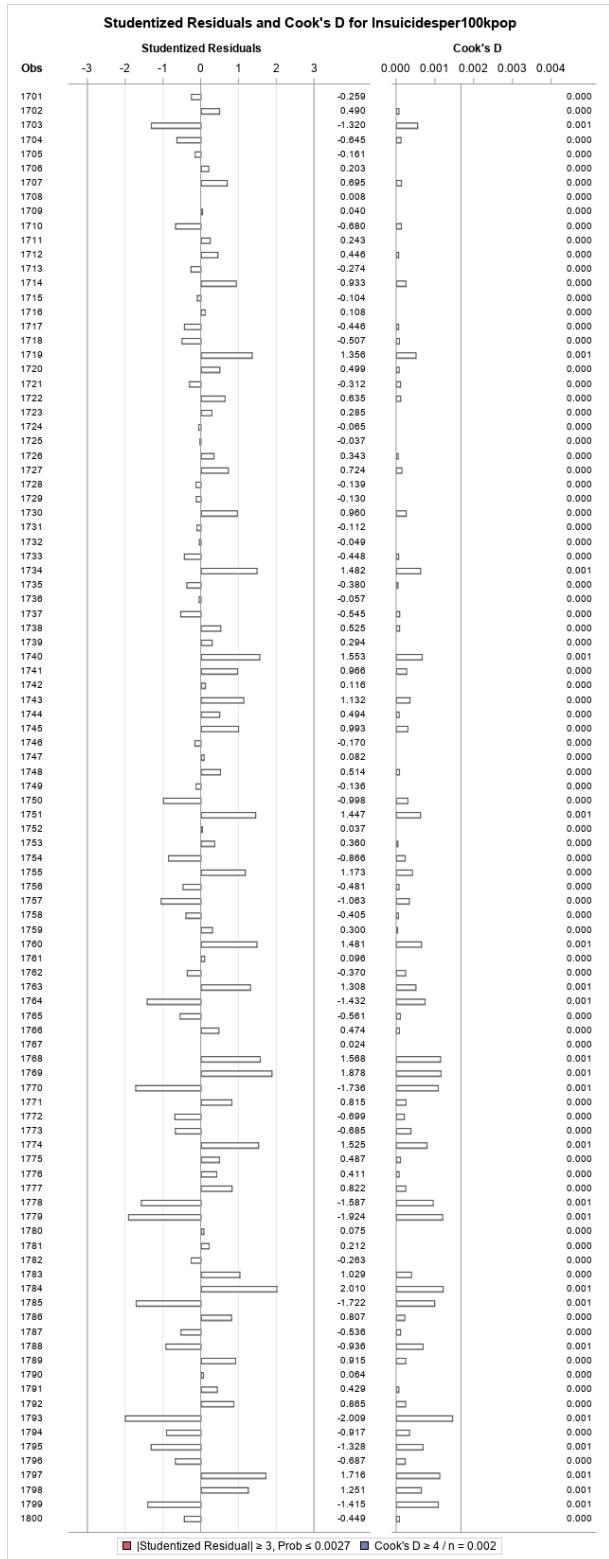
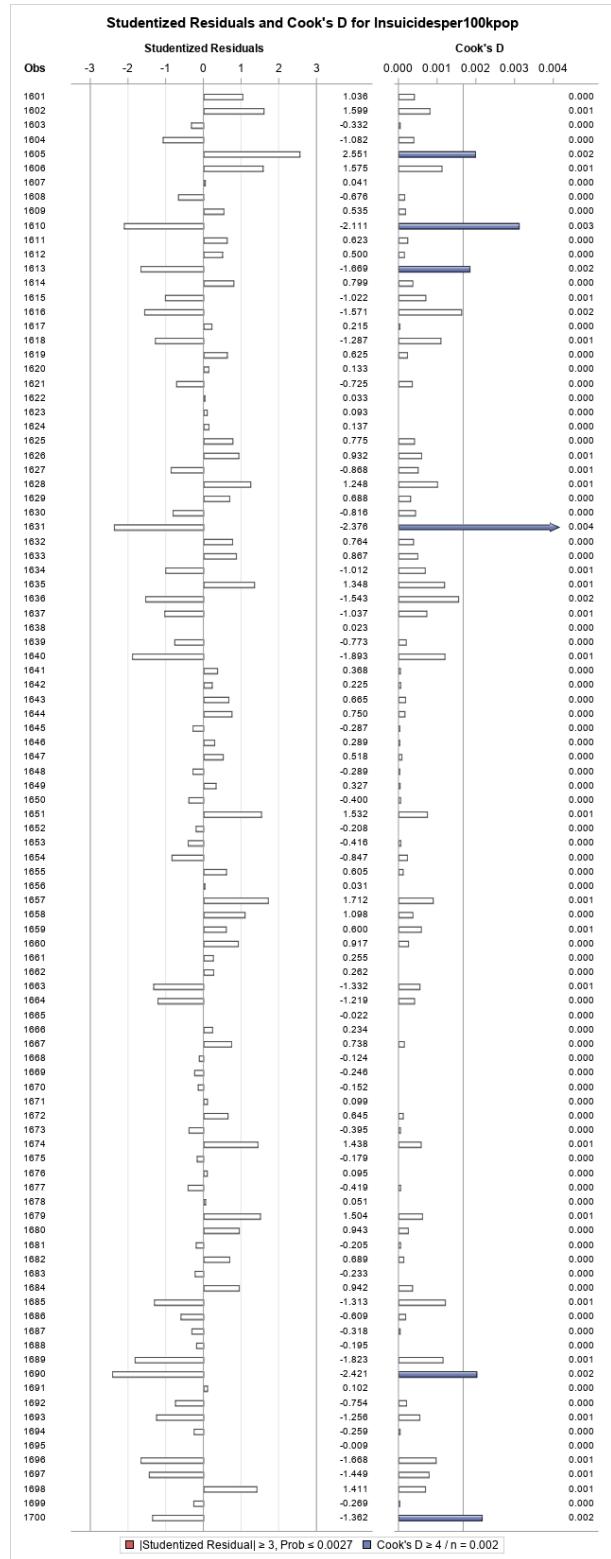


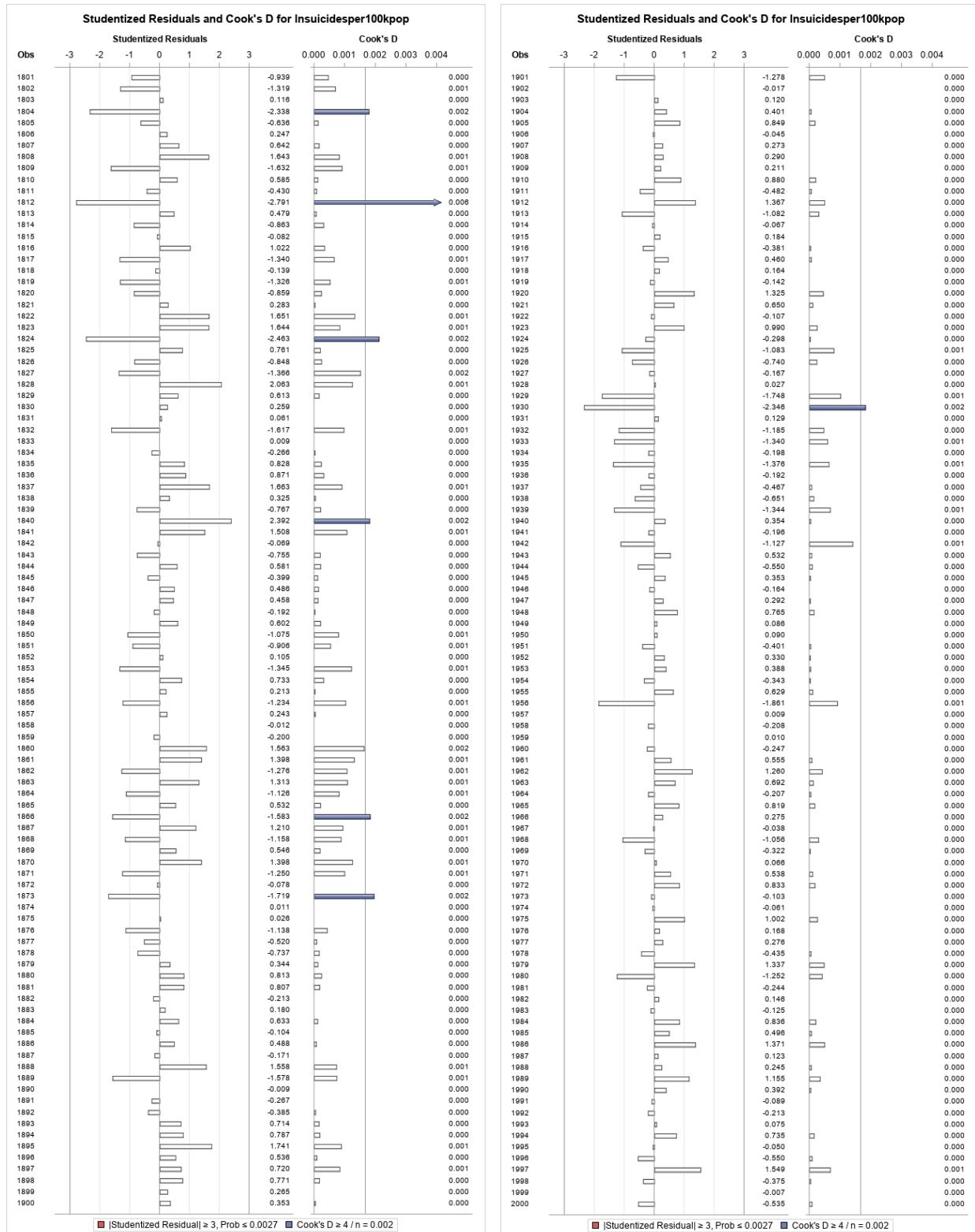


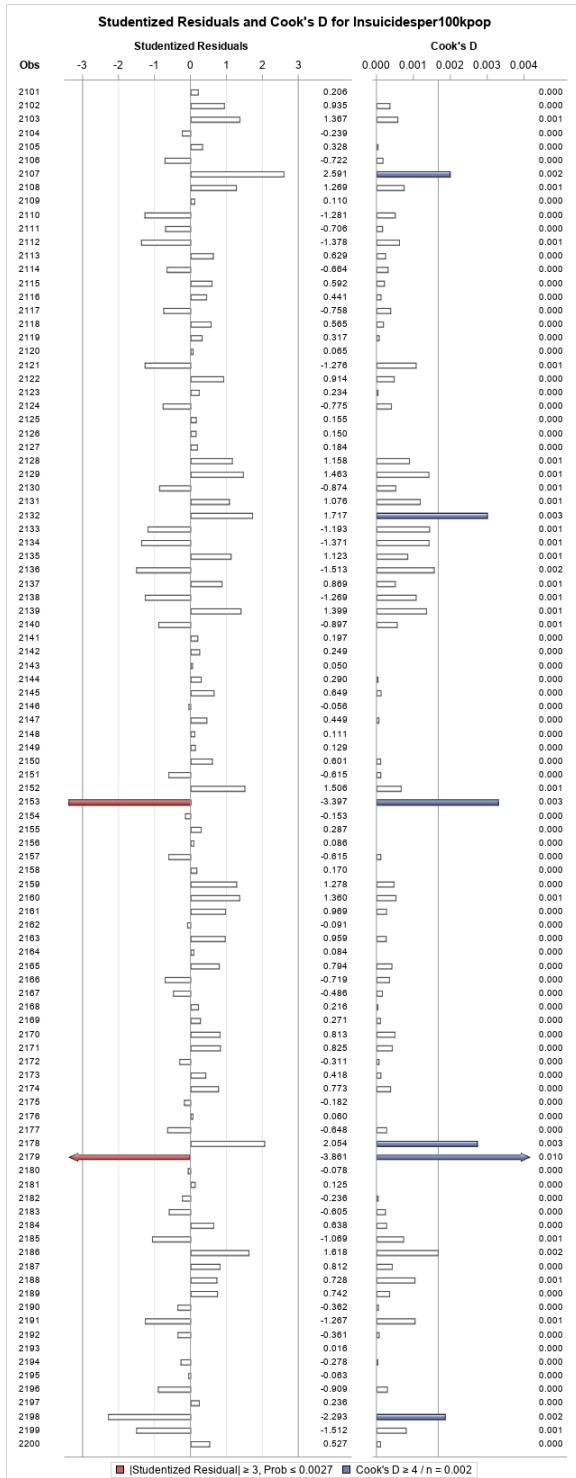
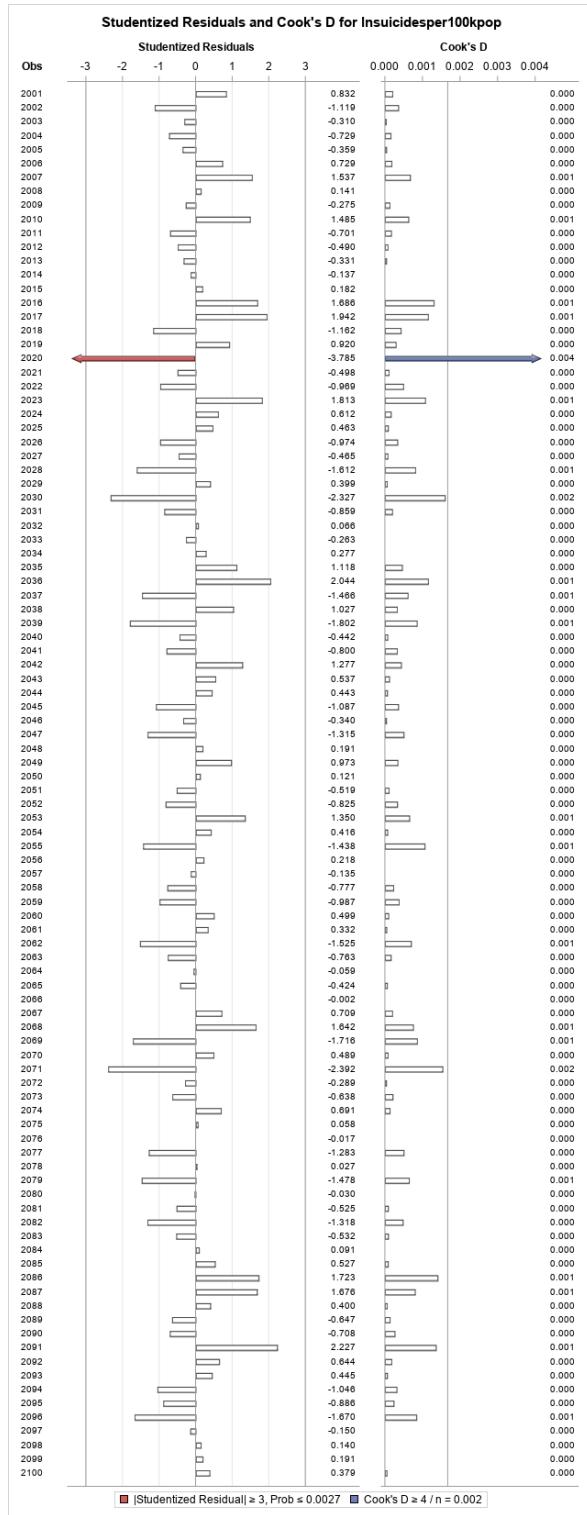


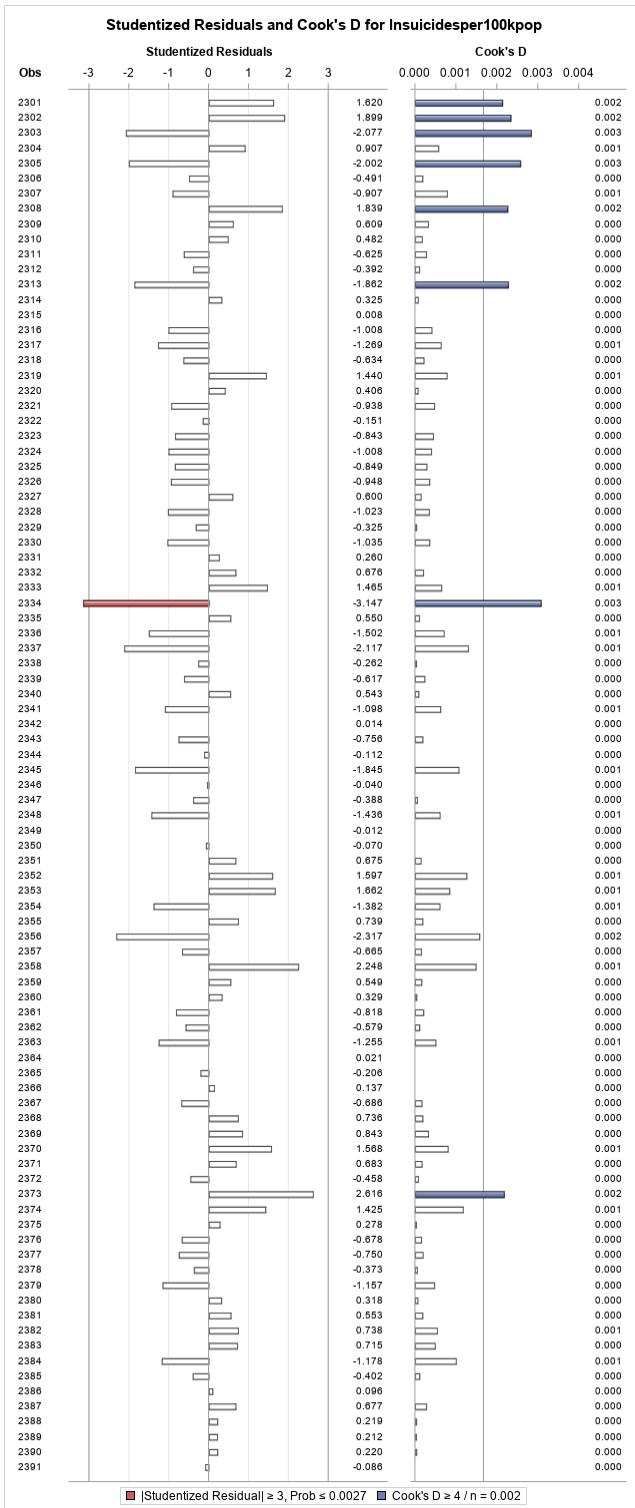
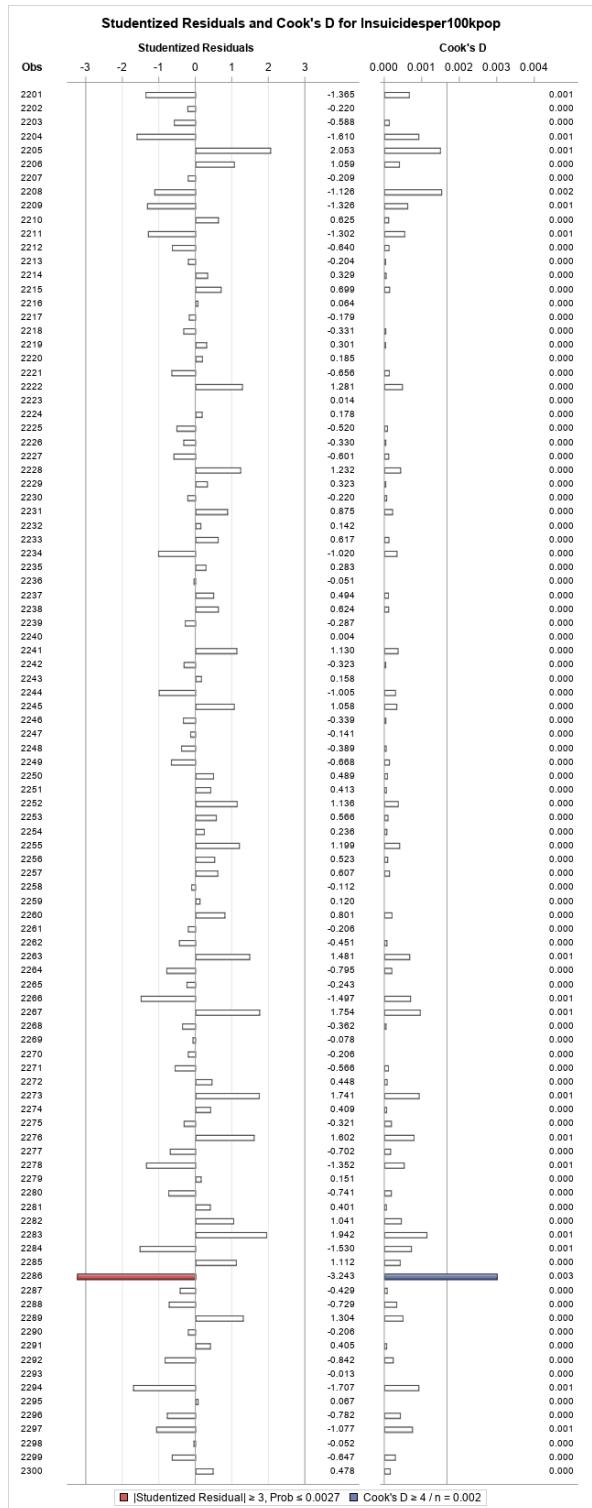












## Ram13

The SURVEYSELECT Procedure	
Selection Method Simple Random Sampling	
Input Data Set	NEW_SUICIDE
Random Number Seed	12345
Sampling Rate	0.7
Sample Size	1666
Selection Probability	0.700294
Sampling Weight	0
Output Data Set	ALL_SUICIDE

## Ram14

Number in Model	C(p)	R-Square	Variables in Model
9	7.5690	0.5320	population gdp_for_year_dollars dsex dgénération3 dgénération4 dage1 dage5 dcontinent1 dcontinent2
10	8.1279	0.5324	population gdp_for_year_dollars dsex dgénération3 dgénération4 dage1 dage5 dyear4 dcontinent1 dcontinent2
10	9.0365	0.5322	population gdp_for_year_dollars dsex dgénération3 dgénération4 dage1 dage5 dcontinent1 dcontinent2 dcontinent3
11	9.0489	0.5328	population gdp_for_year_dollars dsex dgénération3 dgénération4 dage1 dage5 dyear1 dyear4 dcontinent1 dcontinent2
10	9.0920	0.5322	population gdp_for_year_dollars dsex dgénération3 dgénération4 dage1 dage5 dyear1 dcontinent1 dcontinent2
10	9.1302	0.5322	population gdp_for_year_dollars gdp_per_capita_dollars dsex dgénération3 dgénération4 dage1 dage5 dcontinent1 dcontinent2
10	9.2867	0.5321	population gdp_for_year_dollars dsex dage1 dage2 dage3 dage4 dage5 dcontinent1 dcontinent2
10	9.3160	0.5321	population gdp_for_year_dollars dsex dgénération3 dgénération4 dage1 dage2 dage5 dcontinent1 dcontinent2
10	9.3855	0.5321	population gdp_for_year_dollars dsex dgénération3 dgénération4 dage1 dage3 dage5 dcontinent1 dcontinent2
10	9.5383	0.5321	population gdp_for_year_dollars dsex dgénération2 dgénération3 dgénération4 dage1 dage5 dcontinent1 dcontinent2
10	9.5454	0.5320	population gdp_for_year_dollars dsex dgénération3 dgénération4 dage1 dage5 dyear3 dcontinent1 dcontinent2
10	9.5464	0.5320	population gdp_for_year_dollars dsex dgénération1 dgénération2 dage1 dage4 dage5 dcontinent1 dcontinent2
10	9.5464	0.5320	population gdp_for_year_dollars dsex dgénération3 dgénération4 dage1 dage5 dyear2 dcontinent1 dcontinent2
10	9.5558	0.5320	population gdp_for_year_dollars dsex dgénération1 dgénération3 dgénération4 dage1 dage5 dcontinent1 dcontinent2
10	9.5633	0.5320	population gdp_for_year_dollars dsex dgénération3 dgénération4 dage1 dage4 dage5 dcontinent1 dcontinent2
11	9.6565	0.5326	population gdp_for_year_dollars dsex dgénération3 dgénération4 dage1 dage5 dyear4 dcontinent1 dcontinent2 dcontinent3

## Ram15

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	dgeneration4		1	0.2352	0.2352	1040.40	511.80	<.0001
2	dsex		2	0.1800	0.4152	406.353	511.89	<.0001
3	dgeneration3		3	0.0226	0.4378	328.597	66.72	<.0001
4	dage1		4	0.0528	0.4906	143.926	172.26	<.0001
5	dcontinent1		5	0.0141	0.5047	96.0257	47.33	<.0001
6	dcontinent2		6	0.0136	0.5184	49.8132	47.00	<.0001
7	dage5		7	0.0086	0.5270	21.2490	30.32	<.0001
8	gdp_for_year_dollars		8	0.0024	0.5294	14.9215	8.30	0.0040
9	population		9	0.0026	0.5320	7.5690	9.37	0.0022

Variable population Entered: R-Square = 0.5320 and C(p) = 7.5690

Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	9	1686.82013	187.42446	209.20	<.0001	
Error	1656	1483.64772	0.89592			
Corrected Total	1665	3170.46785				

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.48623	0.04766	871.39685	972.63	<.0001
population	-2.18082E-8	7.125849E-9	8.39148	9.37	0.0022
gdp_for_year_dollars	6.16337E-14	1.46516E-14	15.85382	17.70	<.0001
dsex	1.15099	0.04655	547.70381	611.33	<.0001
dgeneration3	-2.22765	0.14313	217.02900	242.24	<.0001
dgeneration4	-2.21285	0.07864	709.44153	791.86	<.0001
dage1	2.05111	0.15025	166.95630	186.35	<.0001
dage5	0.31566	0.07011	18.16334	20.27	<.0001
dcontinent1	-0.83026	0.10470	56.33860	62.88	<.0001
dcontinent2	-0.32115	0.04890	38.63597	43.12	<.0001

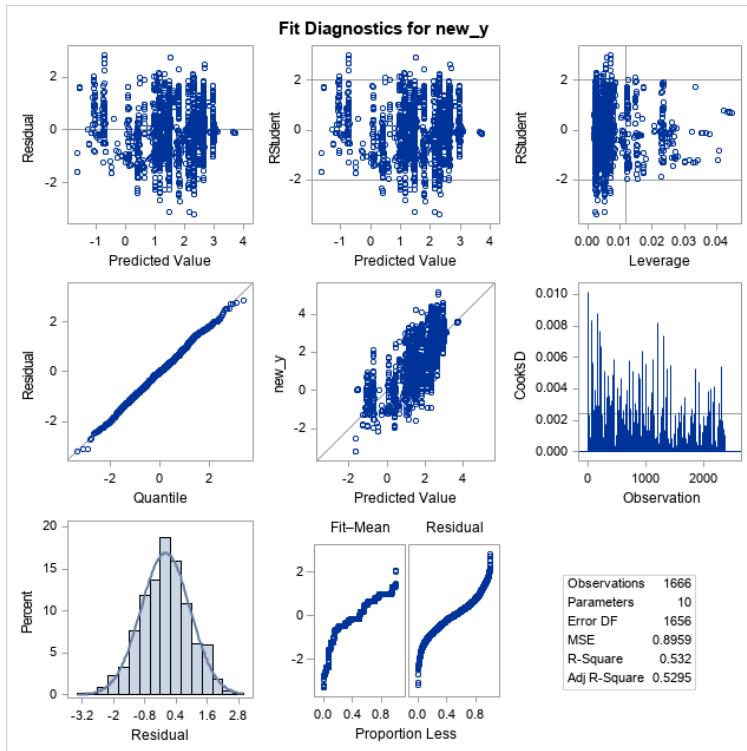
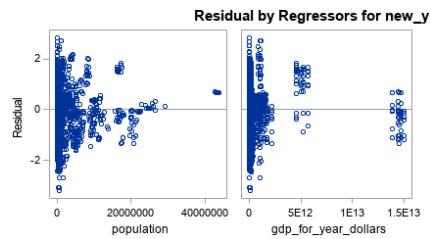
## Ram16

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Standardized Estimate	Tolerance	Variance Inflation
Intercept	1	1.48623	0.04766	31.19	<.0001	0	.	0
population	1	-2.18082E-8	7.125849E-9	-3.06	0.0022	-0.07793	0.43583	2.29449
gdp_for_year_dollars	1	6.16337E-14	1.46516E-14	4.21	<.0001	0.10658	0.44019	2.27173
dsex	1	1.15099	0.04655	24.73	<.0001	0.41651	0.99581	1.00421
dgeneration3	1	-2.22765	0.14313	-15.56	<.0001	-0.64939	0.16232	6.16052
dgeneration4	1	-2.21285	0.07864	-28.14	<.0001	-0.48933	0.93451	1.07008
dage1	1	2.05111	0.15025	13.65	<.0001	0.56529	0.16479	6.06832
dage5	1	0.31566	0.07011	4.50	<.0001	0.08118	0.86939	1.15023
dcontinent1	1	-0.83026	0.10470	-7.93	<.0001	-0.13605	0.95997	1.04170
dcontinent2	1	-0.32115	0.04890	-6.57	<.0001	-0.11286	0.95678	1.04517

Number of Observations Read	2379
Number of Observations Used	1666
Number of Observations with Missing Values	713

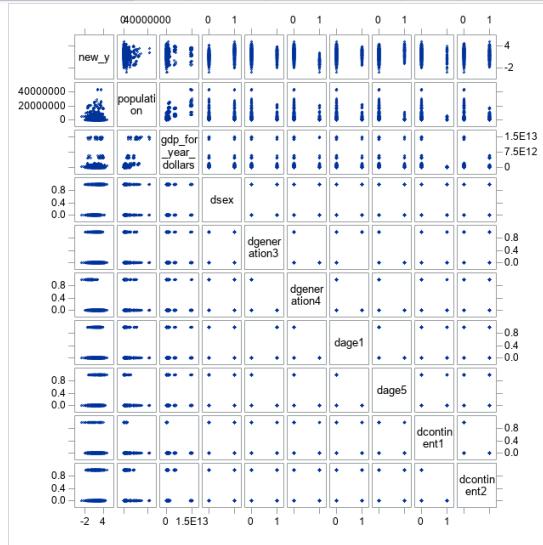
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	1686.82013	187.42446	209.20	<.0001
Error	1656	1483.64772	0.89592		
Corrected Total	1665	3170.46785			

Root MSE	0.94653	R-Square	0.5320
Dependent Mean	1.64352	Adj R-Sq	0.5295
Coeff Var	57.59169		

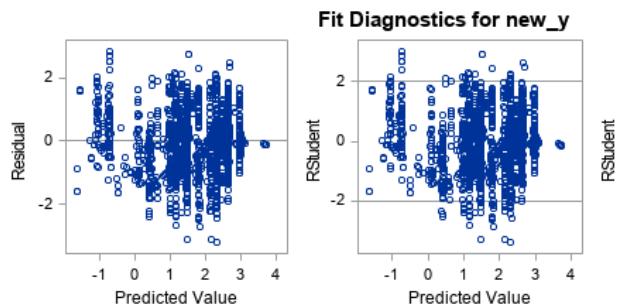
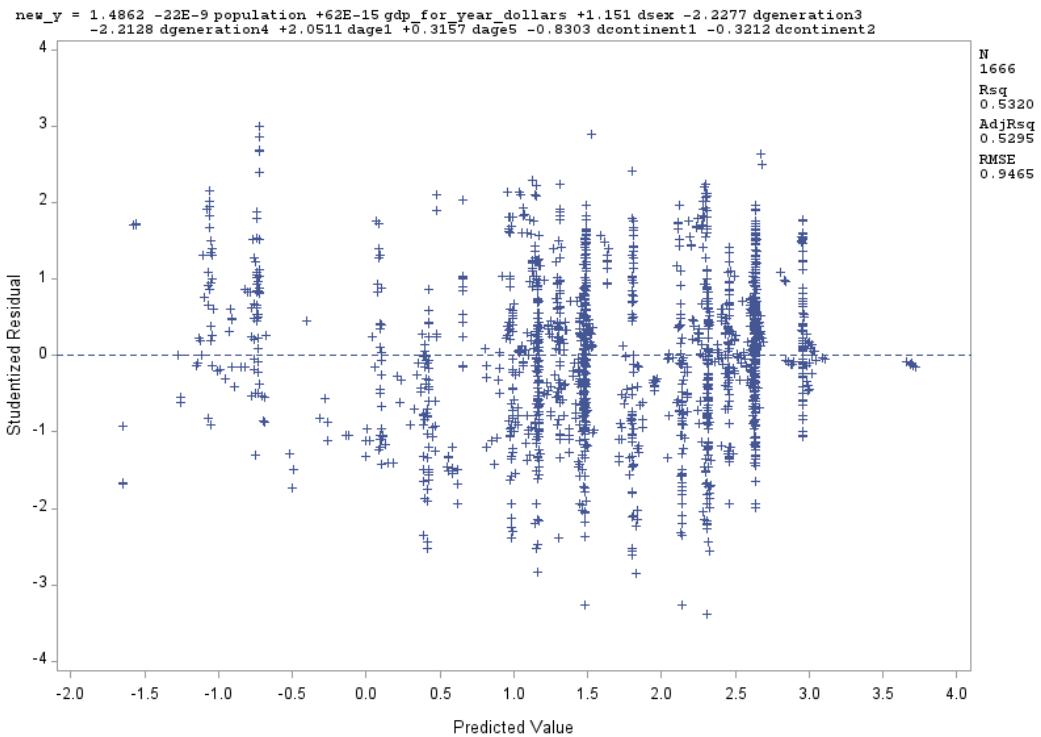


## Ram17

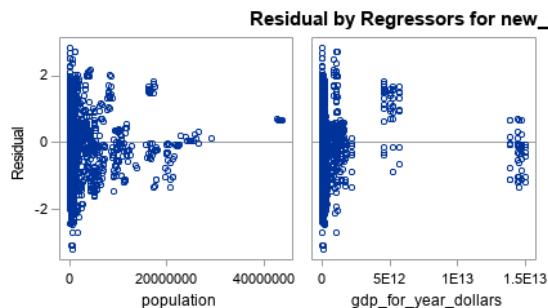
Pearson Correlation Coefficients Prob >  r  under H0: Rho=0 Number of Observations											
	new_y	population	gdp_for_year_dollars	dsex	dgeneration3	dgeneration4	dage1	dage5	dcontinent1	dcontinent2	
new_y	1.00000	-0.05436		0.04252	0.43633	-0.07228	-0.48500	0.02823	0.20374	-0.11154	-0.08510
		0.0265		0.0827	<.0001	0.0032	<.0001	0.2495	<.0001	<.0001	0.0005
	1666	1666		1666	1666	1666	1666	1666	1666	1666	1666
population	-0.05436	1.00000		0.75405	-0.03615	0.00589	0.04796	-0.00236	-0.13742	-0.03231	-0.04598
	0.0265			<.0001	0.0780	0.7741	0.0193	0.9084	<.0001	0.1152	0.0249
	1666	2379		2379	2379	2379	2379	2379	2379	2379	2379
gdp_for_year_dollars	0.04252	0.75405		1.00000	-0.01869	-0.00632	0.02631	-0.00841	0.01902	-0.05599	-0.06715
	0.0827	<.0001		0.3622	0.7582	0.1995	0.6819	0.3539	0.0063	0.0010	
	1666	2379		2379	2379	2379	2379	2379	2379	2379	2379
dsex	0.43633	-0.03615		-0.01869	1.00000	-0.01046	-0.00434	-0.00877	0.01846	0.00431	-0.00718
	<.0001	0.0780		0.3622	0.6100	0.8326	0.8326	0.6690	0.3682	0.8336	0.7265
	1666	2379		2379	2379	2379	2379	2379	2379	2379	2379
dgeneration3	-0.07228	0.00589		-0.00632	-0.01046	1.00000	-0.17624	0.91687	-0.20806	-0.00036	0.00405
	0.0032	0.7741		0.7582	0.6100	2379	2379	2379	2379	0.9859	0.8433
	1666	2379		2379	2379	2379	2379	2379	2379	2379	2379
dgeneration4	-0.48500	0.04796		0.02631	-0.00434	-0.17624	1.00000	-0.16159	-0.14173	-0.01895	-0.02521
	<.0001	0.1995		0.1995	0.8326	<.0001	2379	2379	2379	<.0001	0.3554
	1666	2379		2379	2379	2379	2379	2379	2379	2379	2379
dage1	0.02823	-0.00236		-0.00841	-0.00877	0.91687	-0.16159	1.00000	-0.19077	0.00102	0.00844
	0.2495	0.9084		0.6819	0.6690	<.0001	2379	2379	2379	<.0001	0.9603
	1666	2379		2379	2379	2379	2379	2379	2379	2379	2379
dage5	0.20374	-0.13742		0.01902	0.01846	-0.20806	-0.14173	-0.19077	1.00000	-0.00809	-0.00496
	<.0001	<.0001		0.3539	0.3682	<.0001	2379	2379	2379	0.6934	0.8091
	1666	2379		2379	2379	2379	2379	2379	2379	2379	2379
dcontinent1	-0.11154	-0.03231		-0.05599	0.00431	-0.00036	-0.01895	0.00102	-0.00809	1.00000	-0.17910
	<.0001	0.1152		0.0063	0.8336	0.9859	0.3554	0.9603	0.6934	<.0001	
	1666	2379		2379	2379	2379	2379	2379	2379	2379	2379
dcontinent2	-0.08510	-0.04598		-0.06715	-0.00718	0.00405	-0.02521	0.00844	-0.00496	-0.17910	1.00000
	0.0005	0.0249		0.0010	0.7265	0.8433	0.2191	0.6807	0.8091	<.0001	
	1666	2379		2379	2379	2379	2379	2379	2379	2379	2379

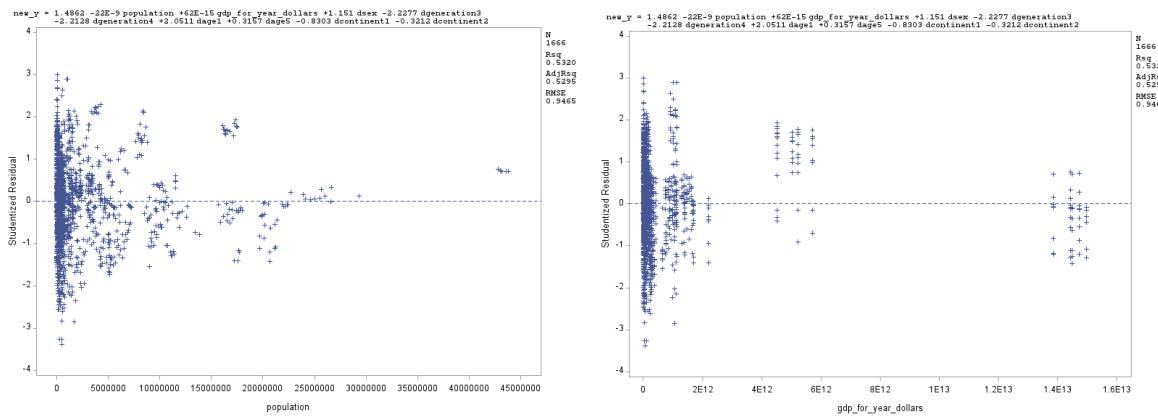


## Ram18

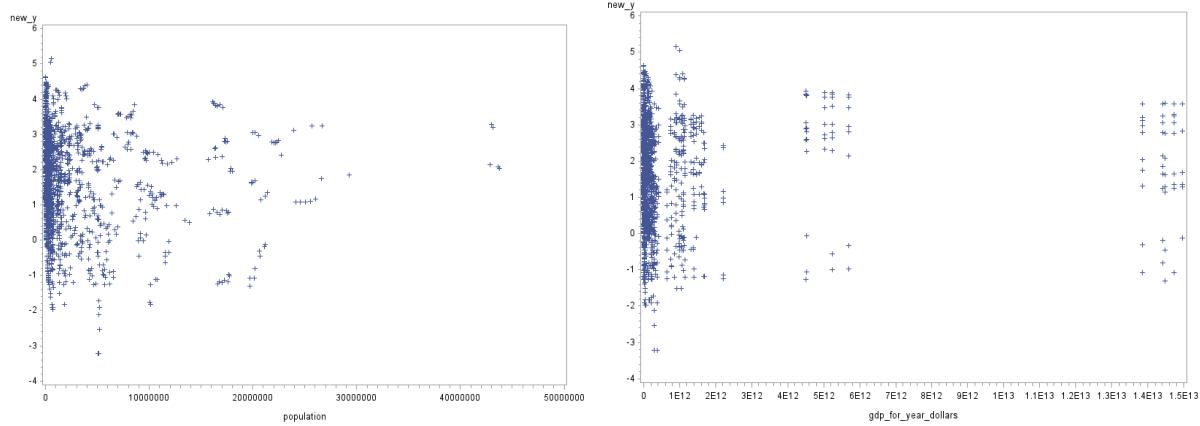


## Ram19

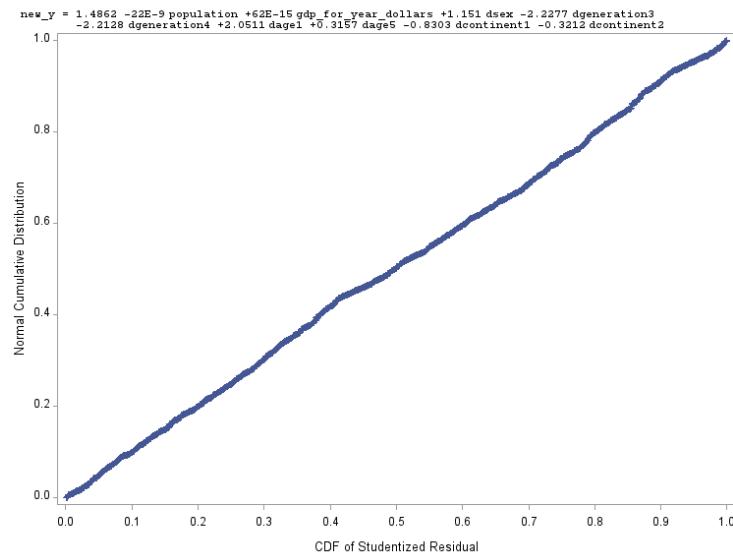




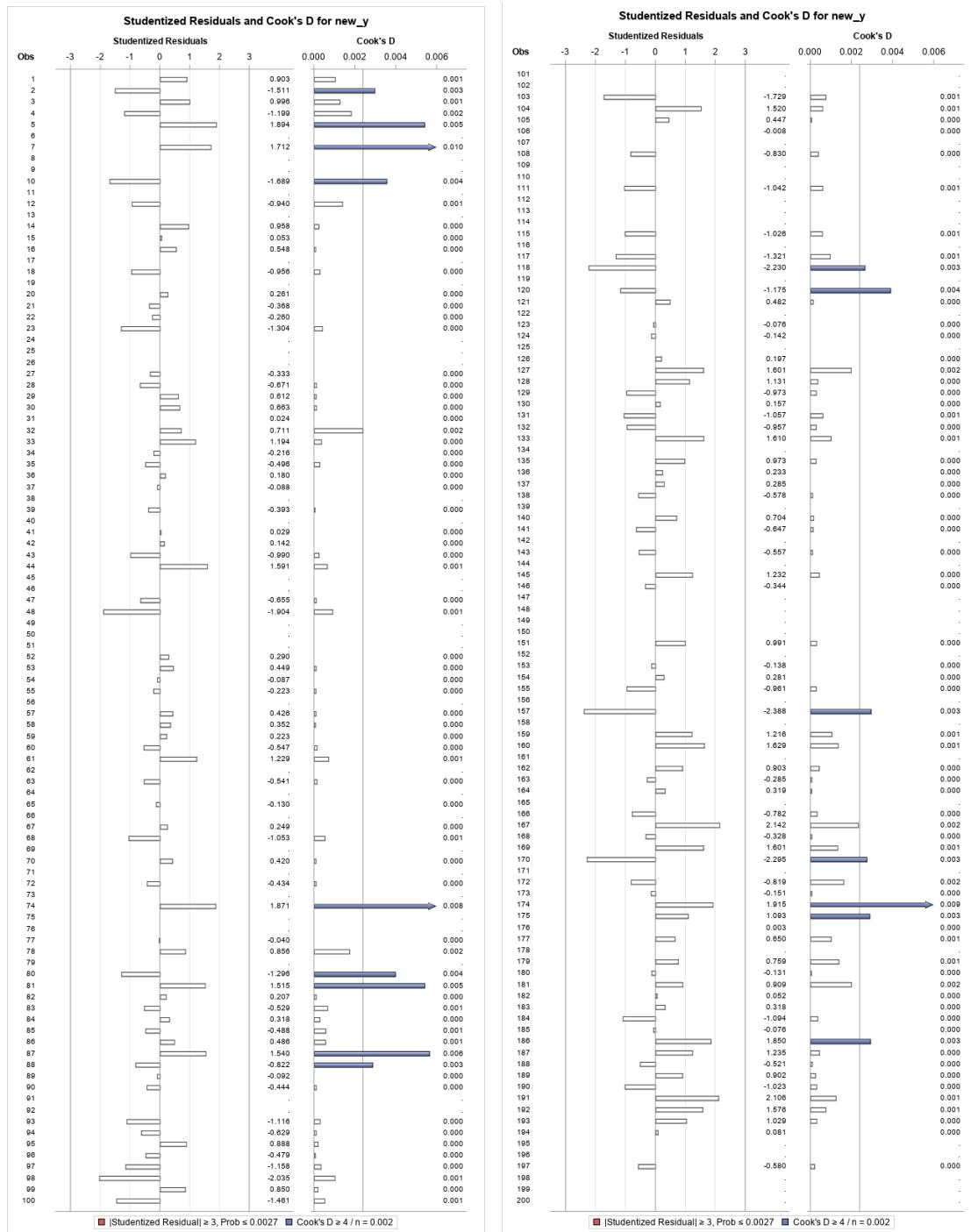
## Ram20

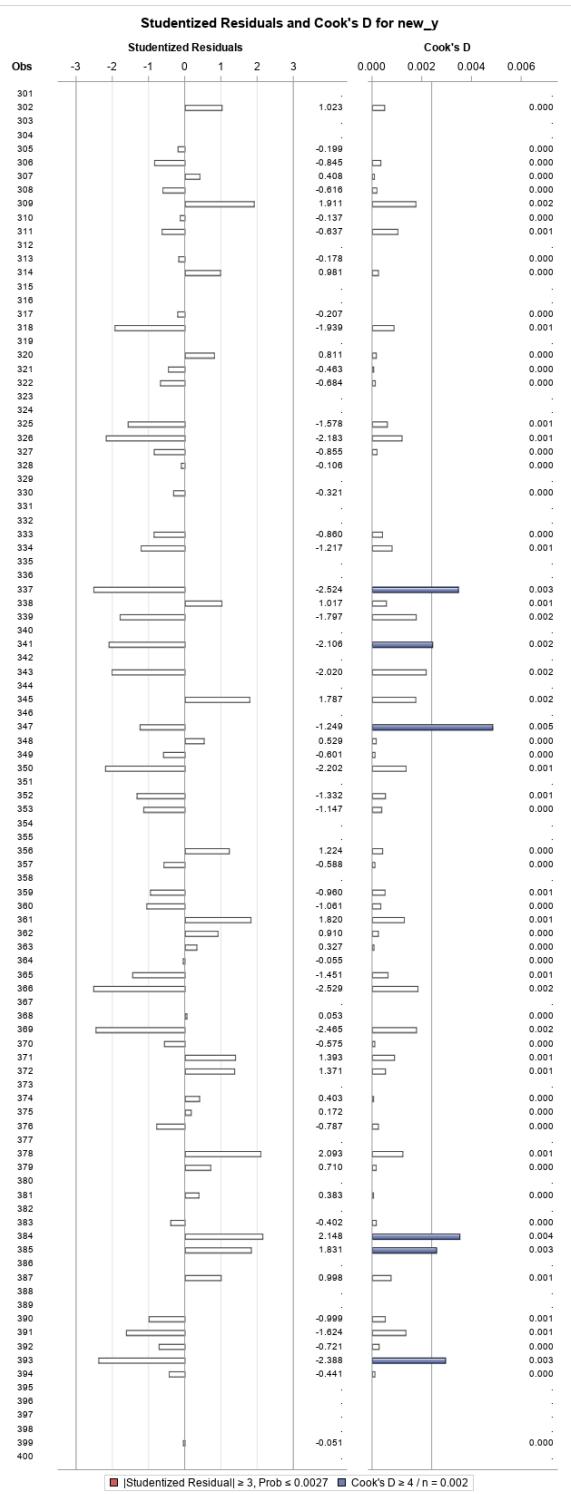
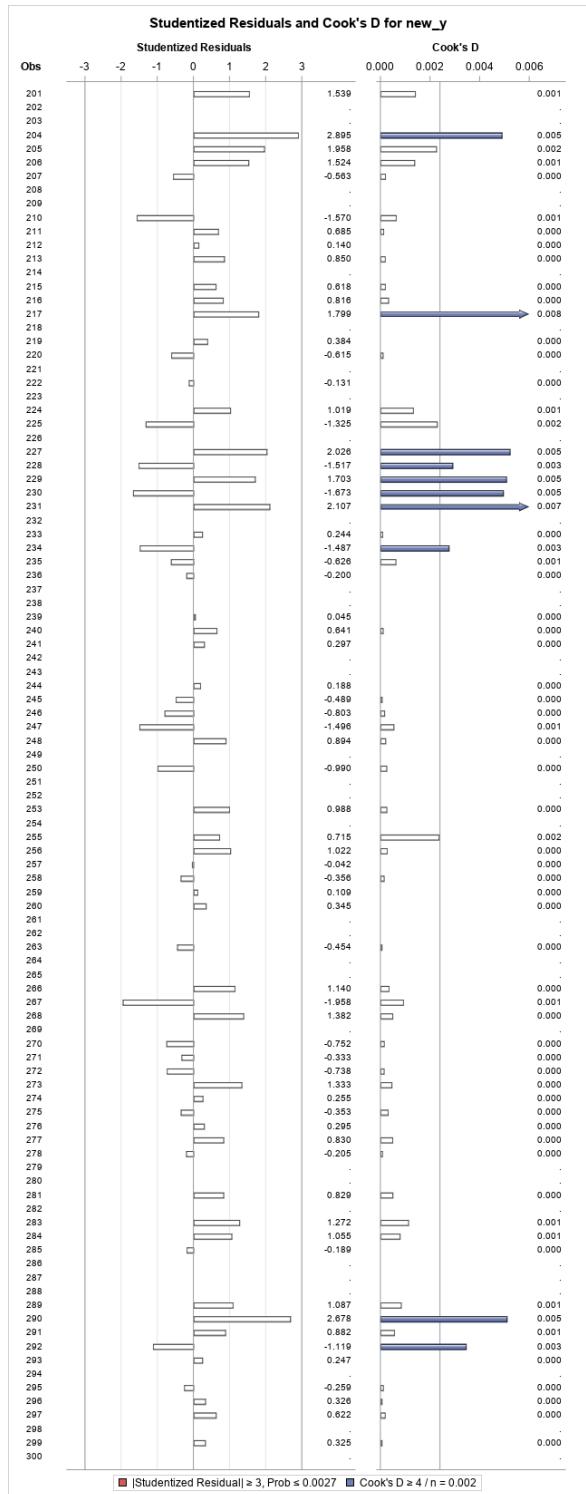


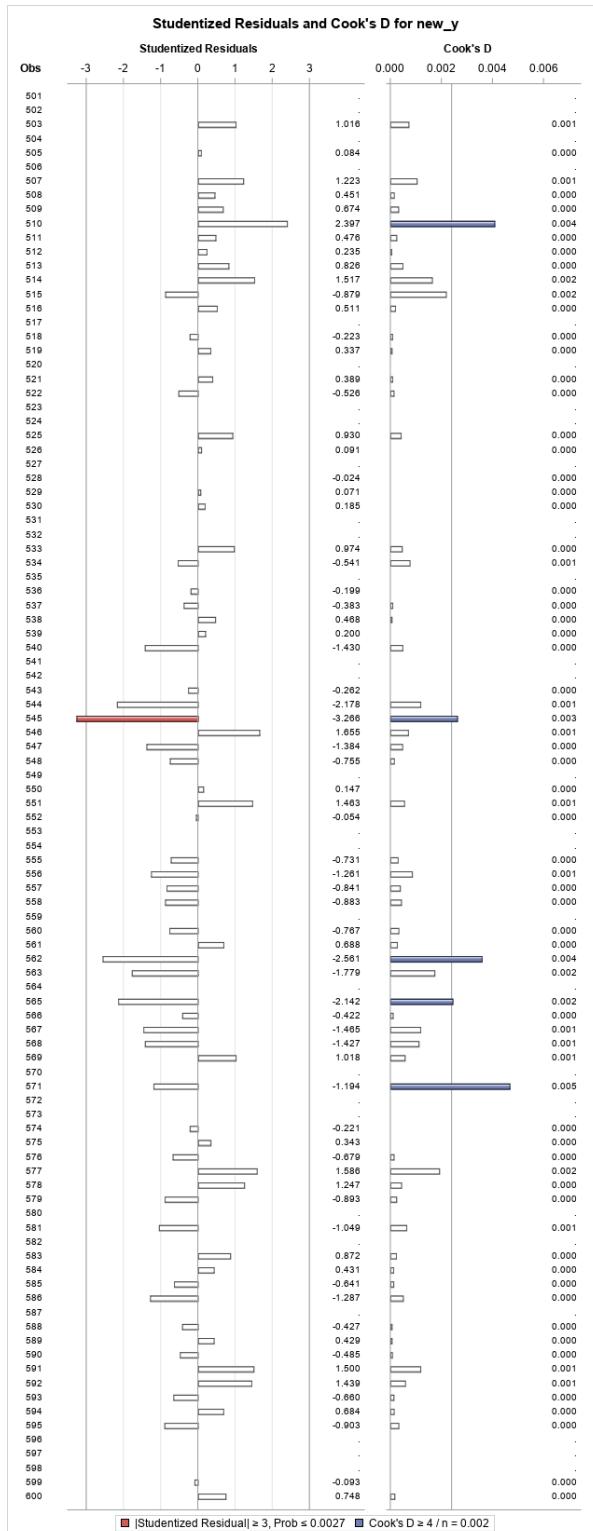
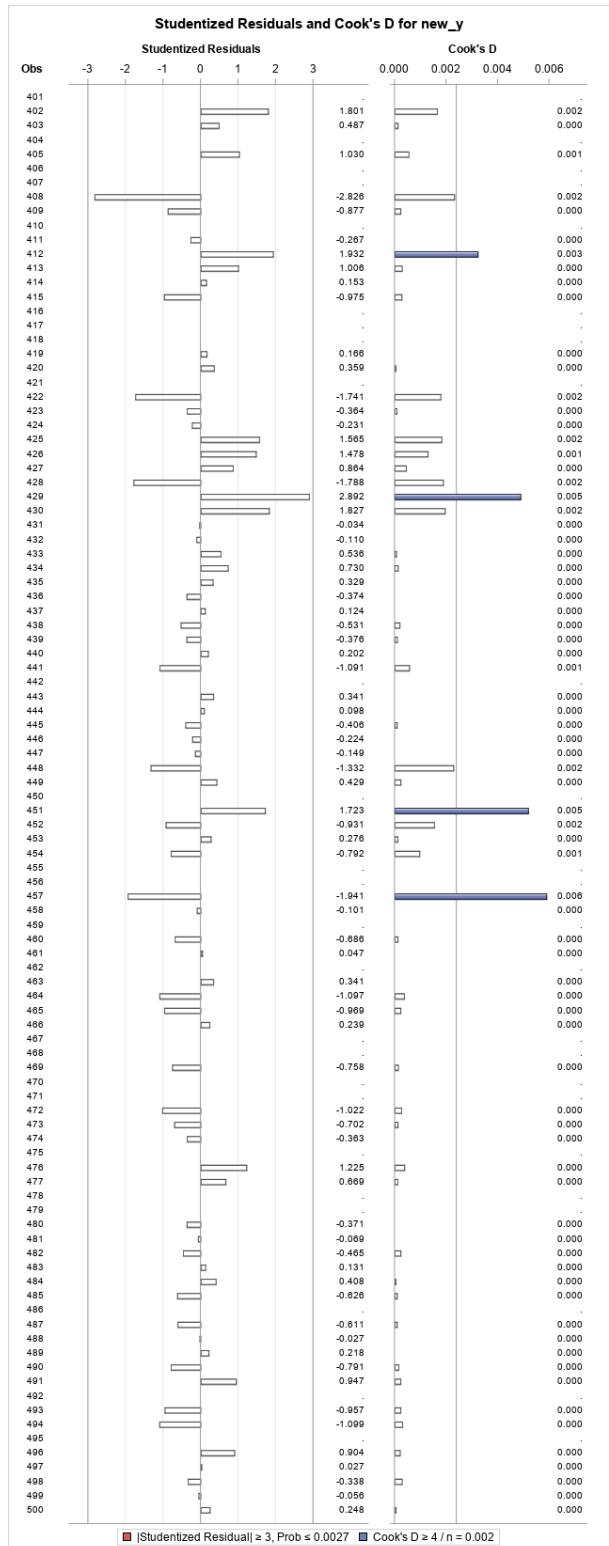
## Ram21

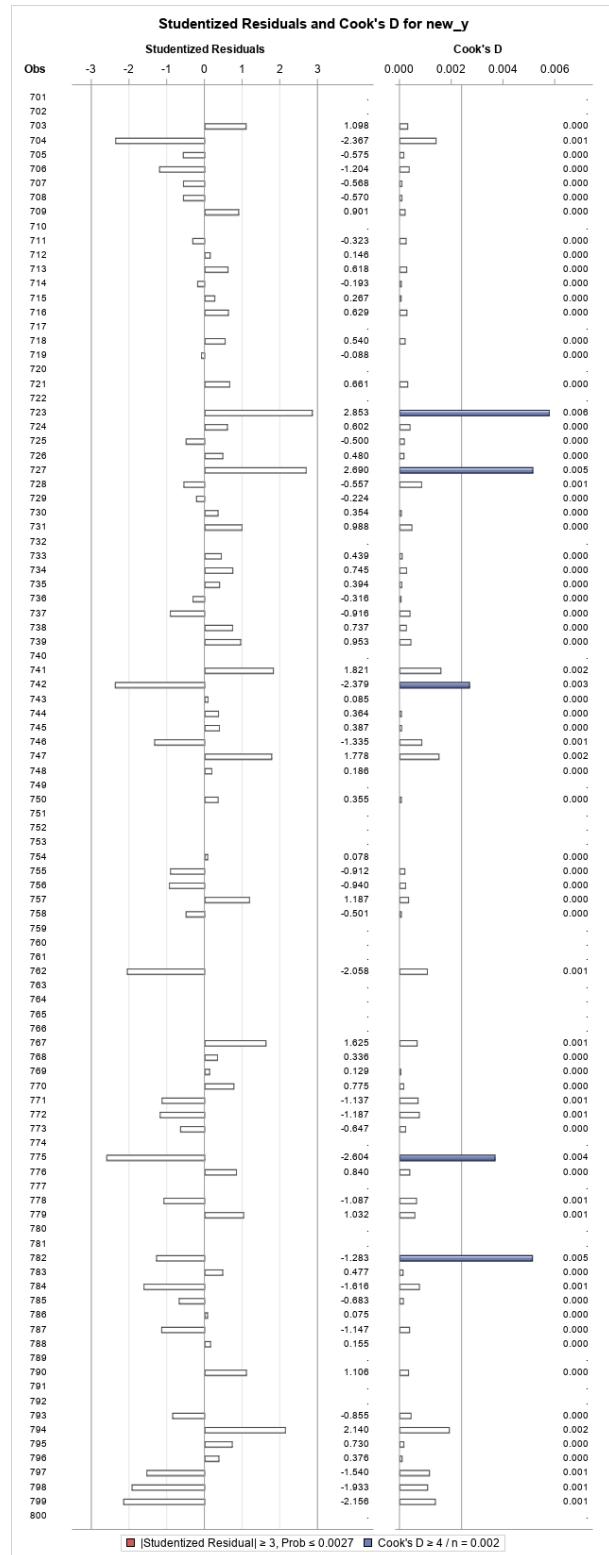
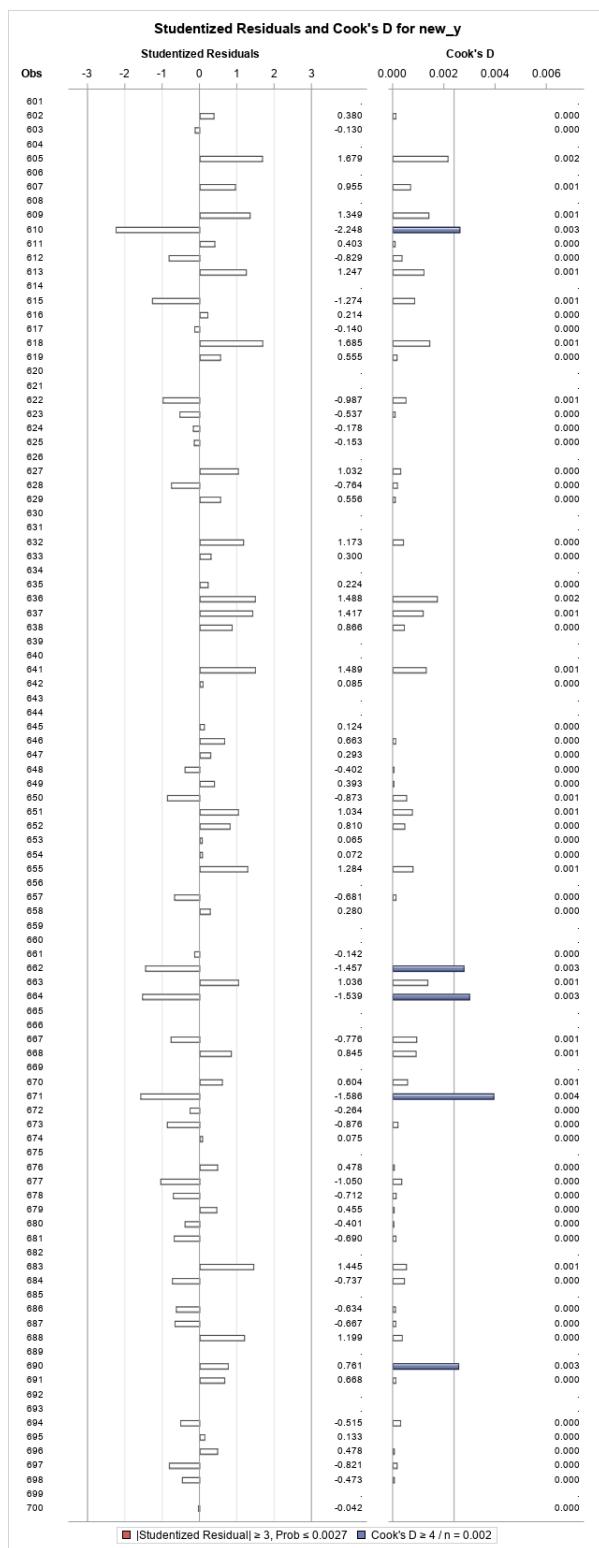


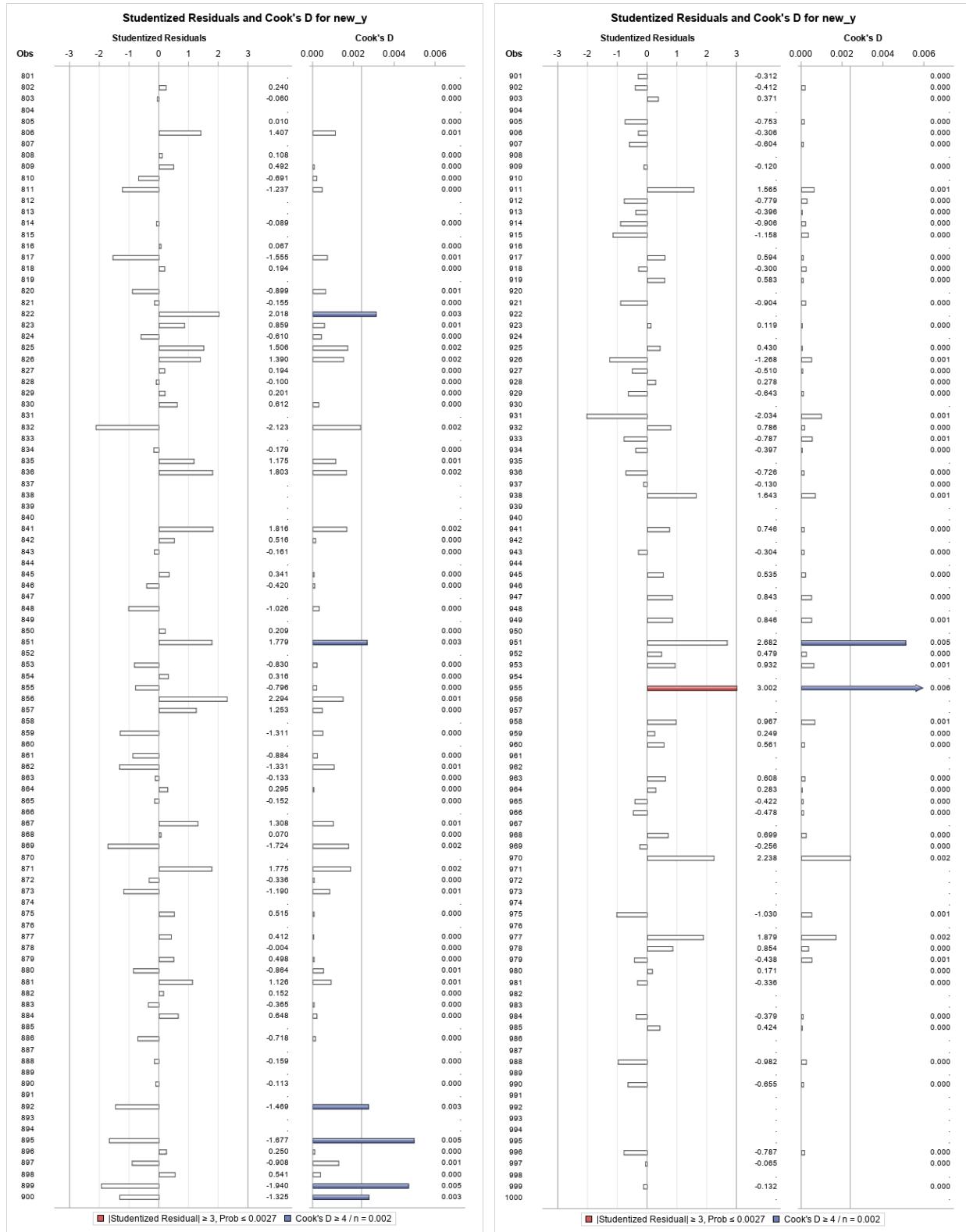
## Ram22

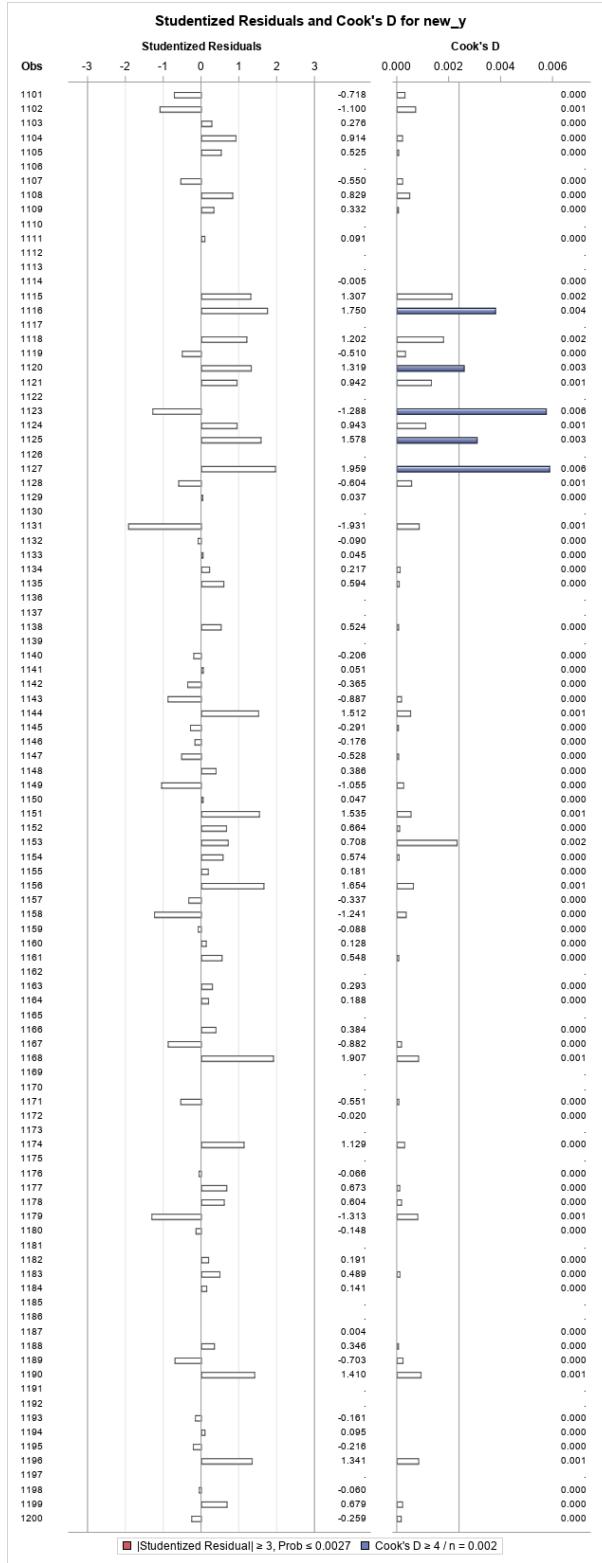
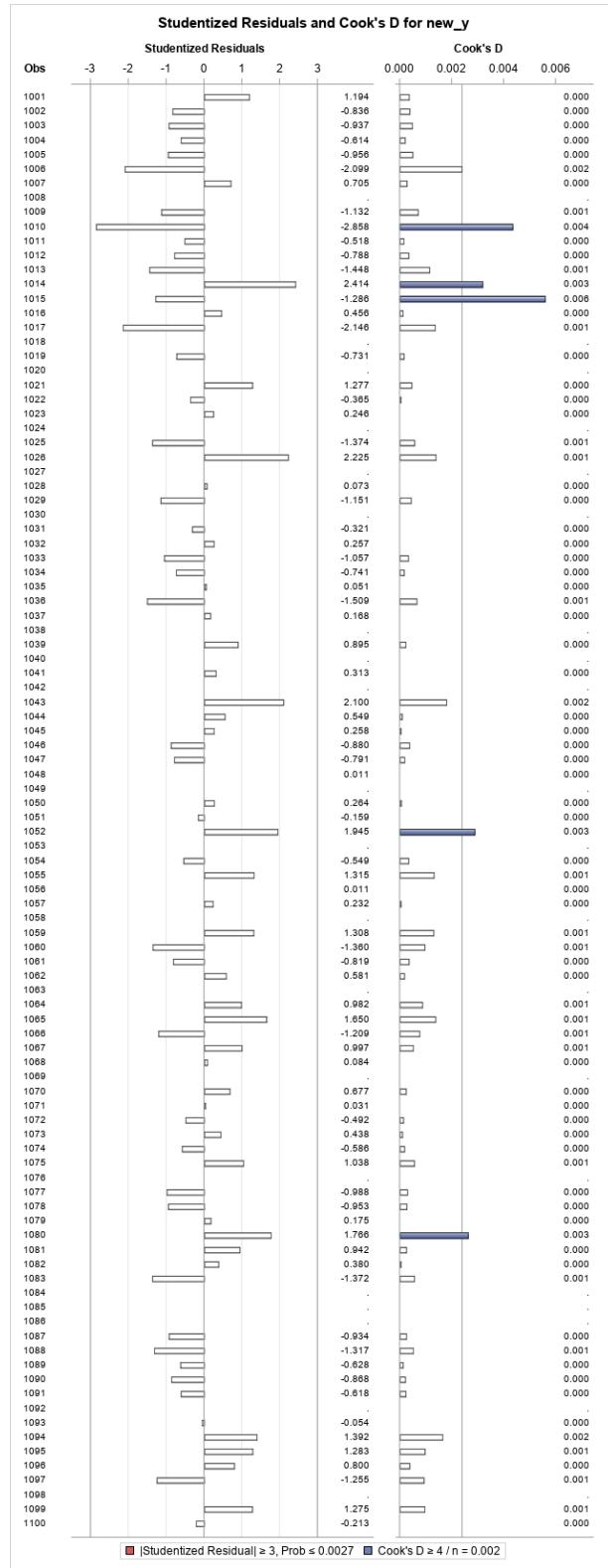


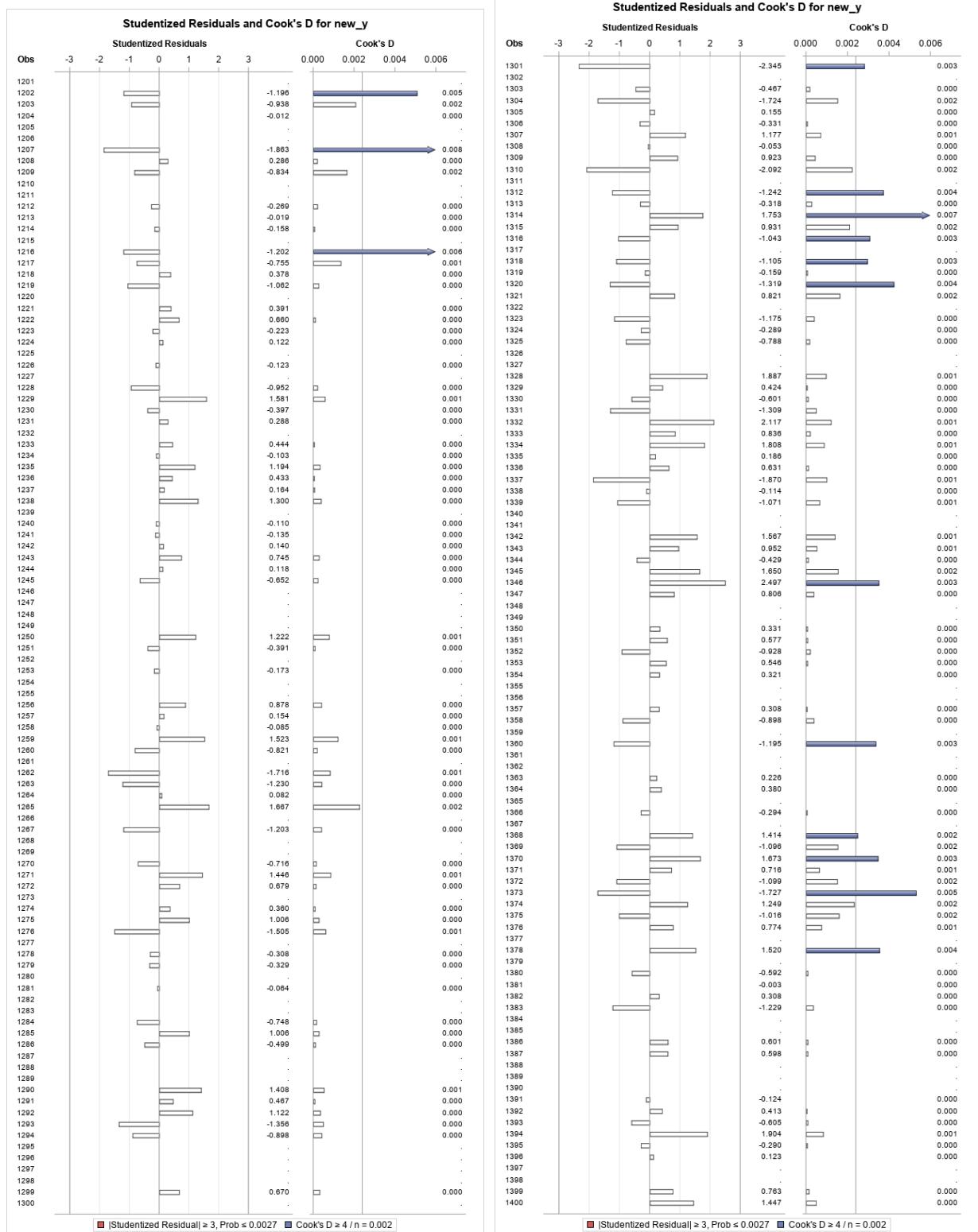


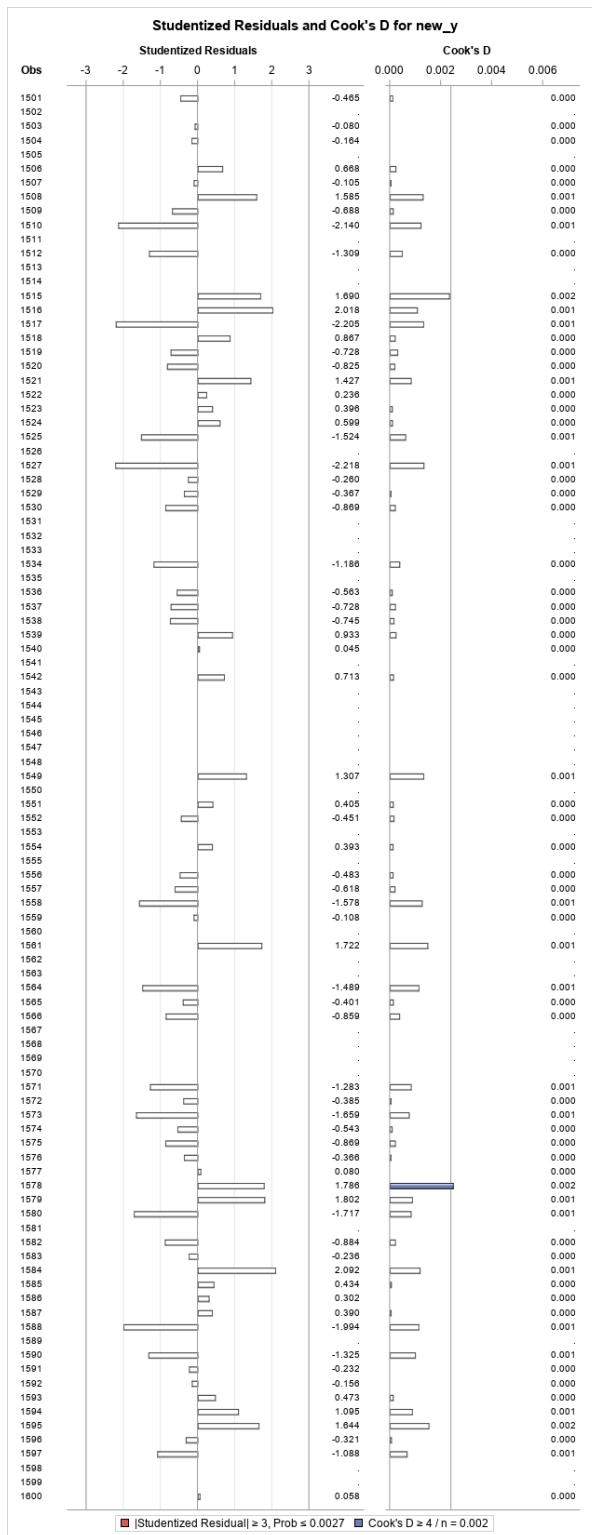
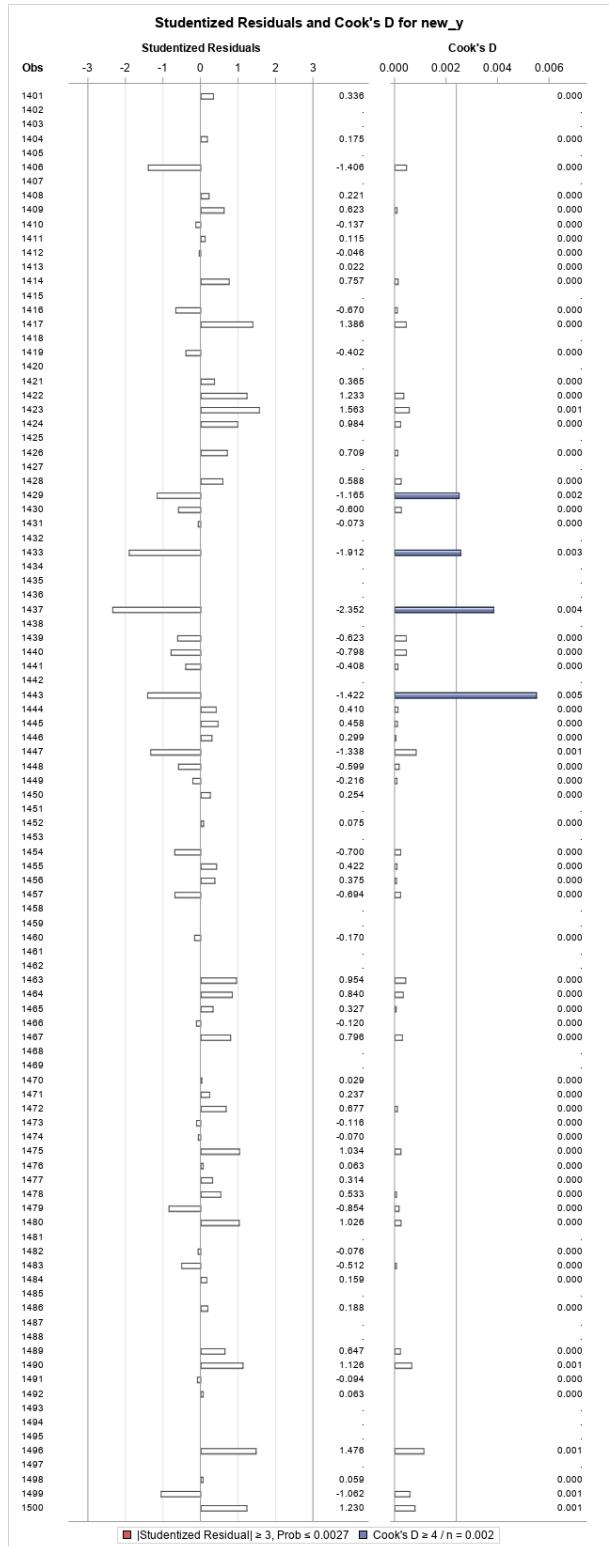


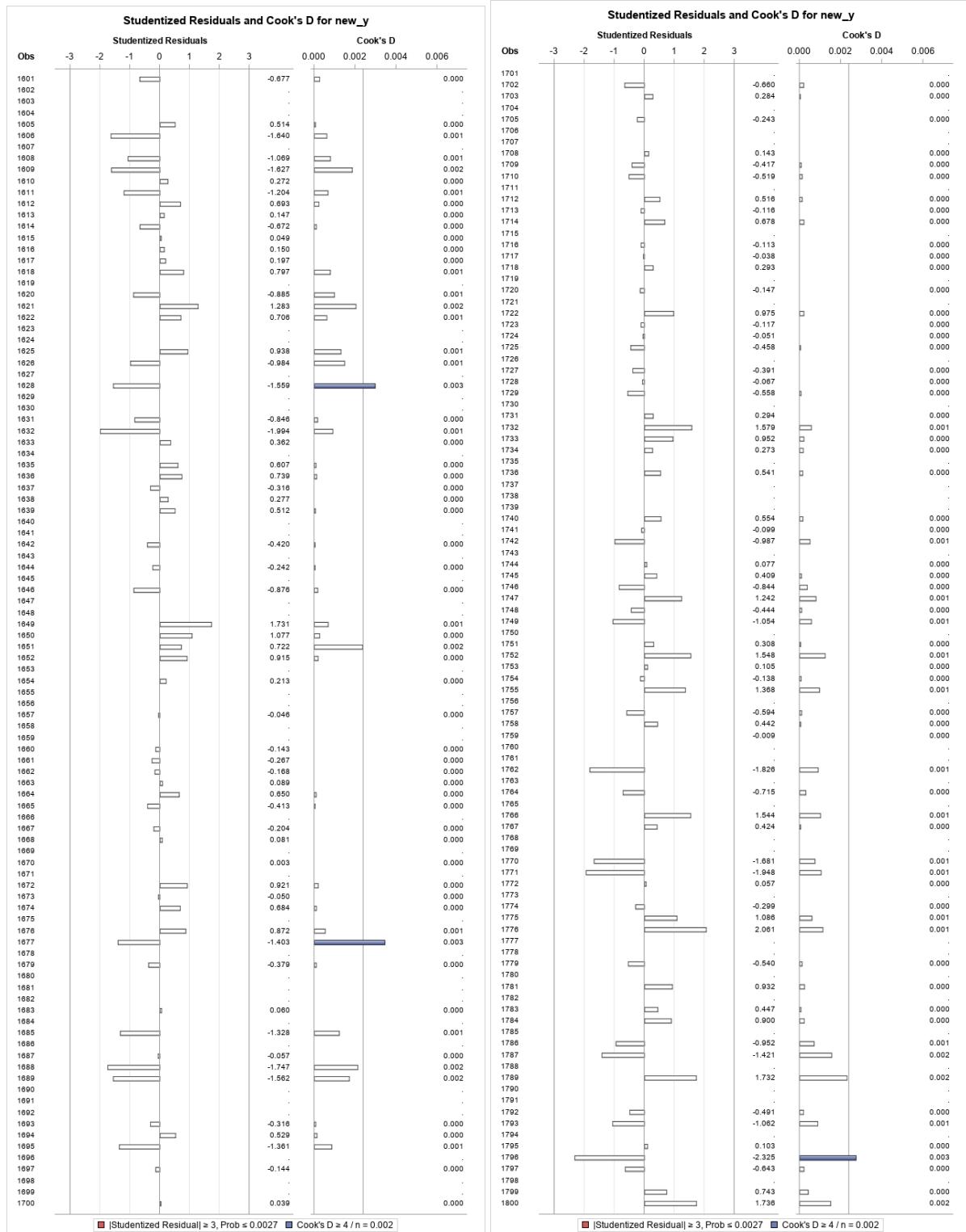


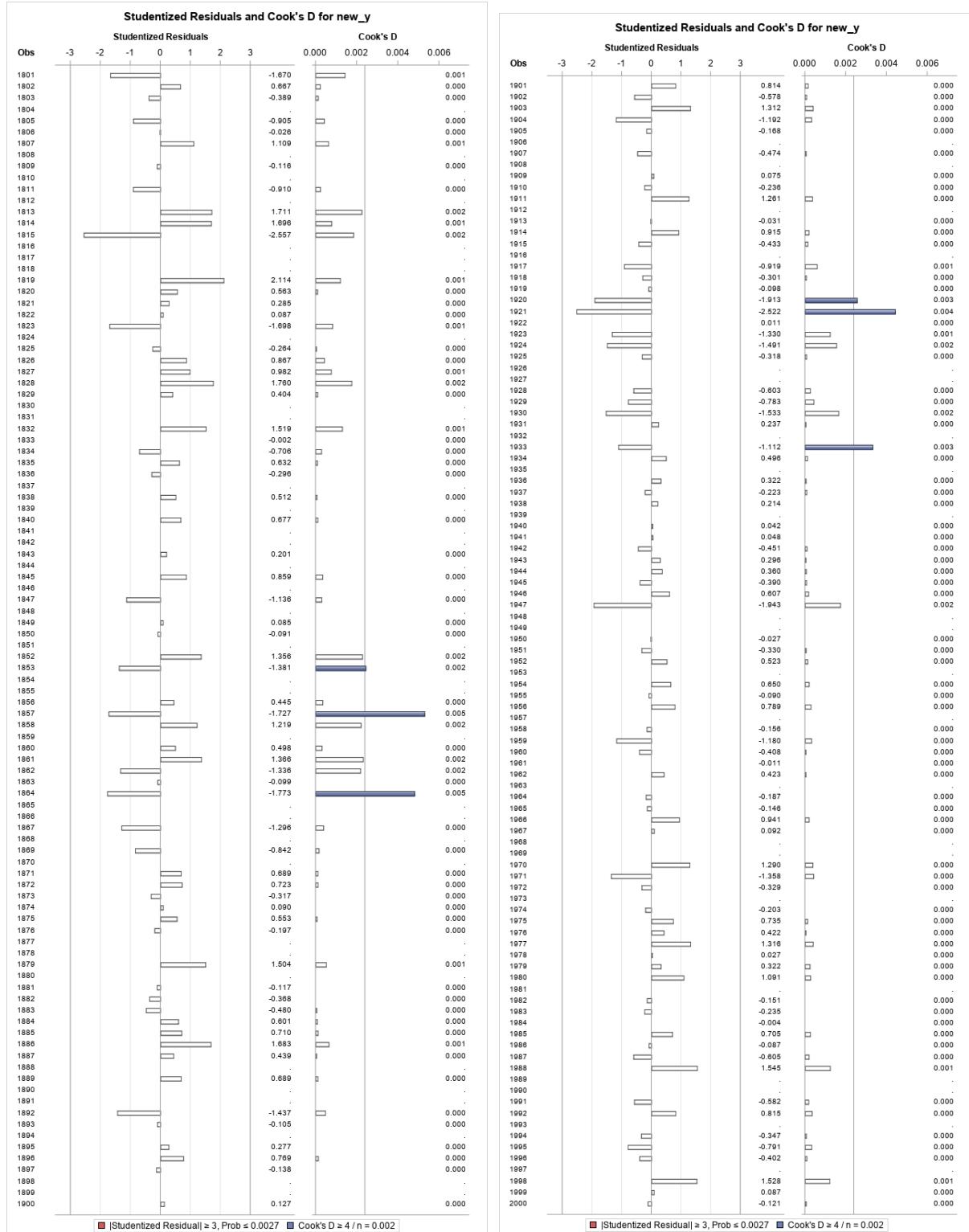


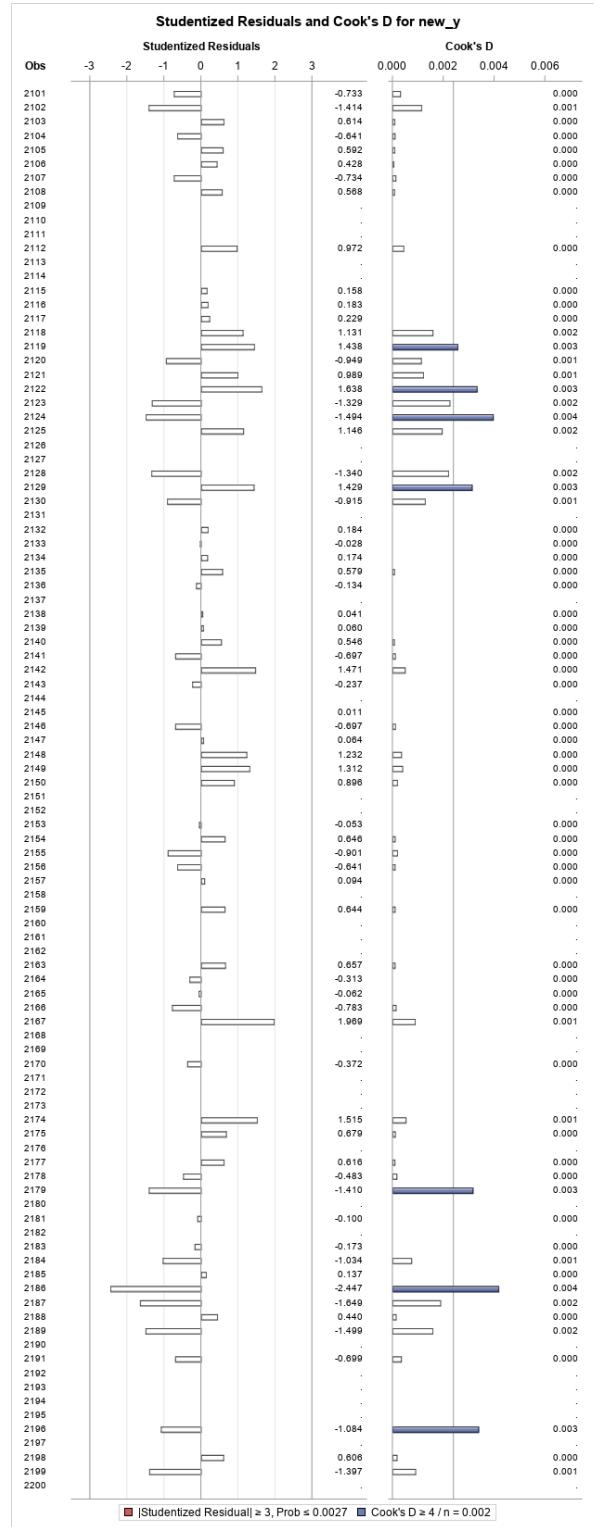
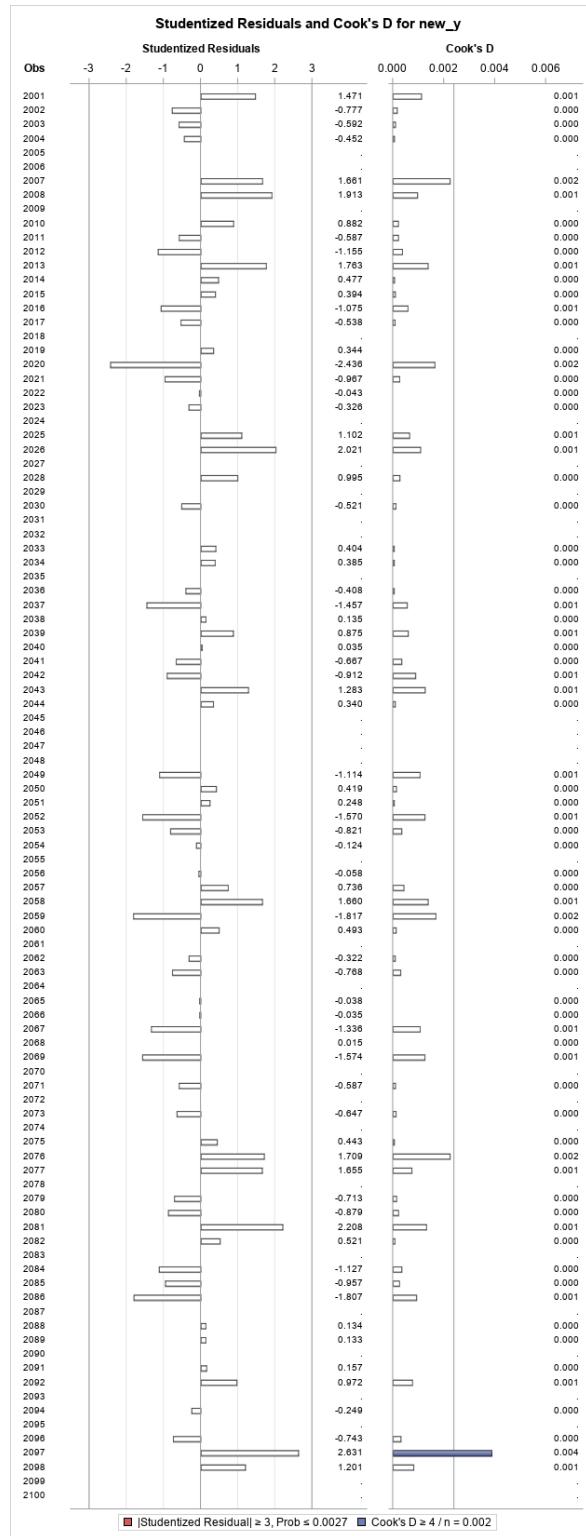


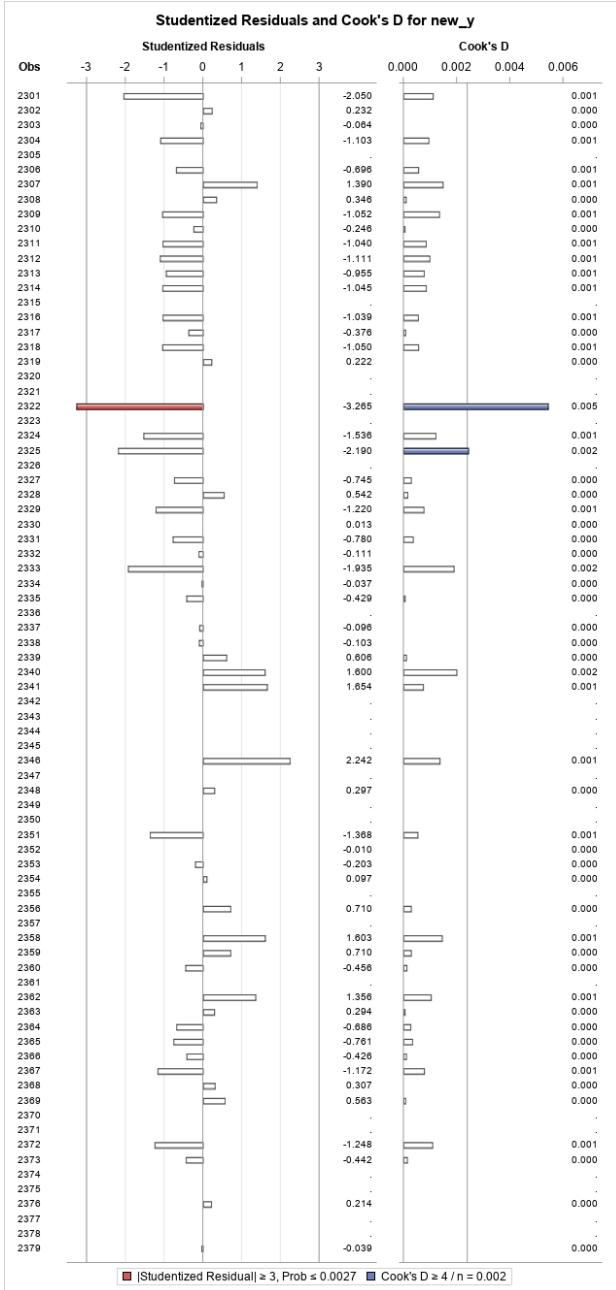
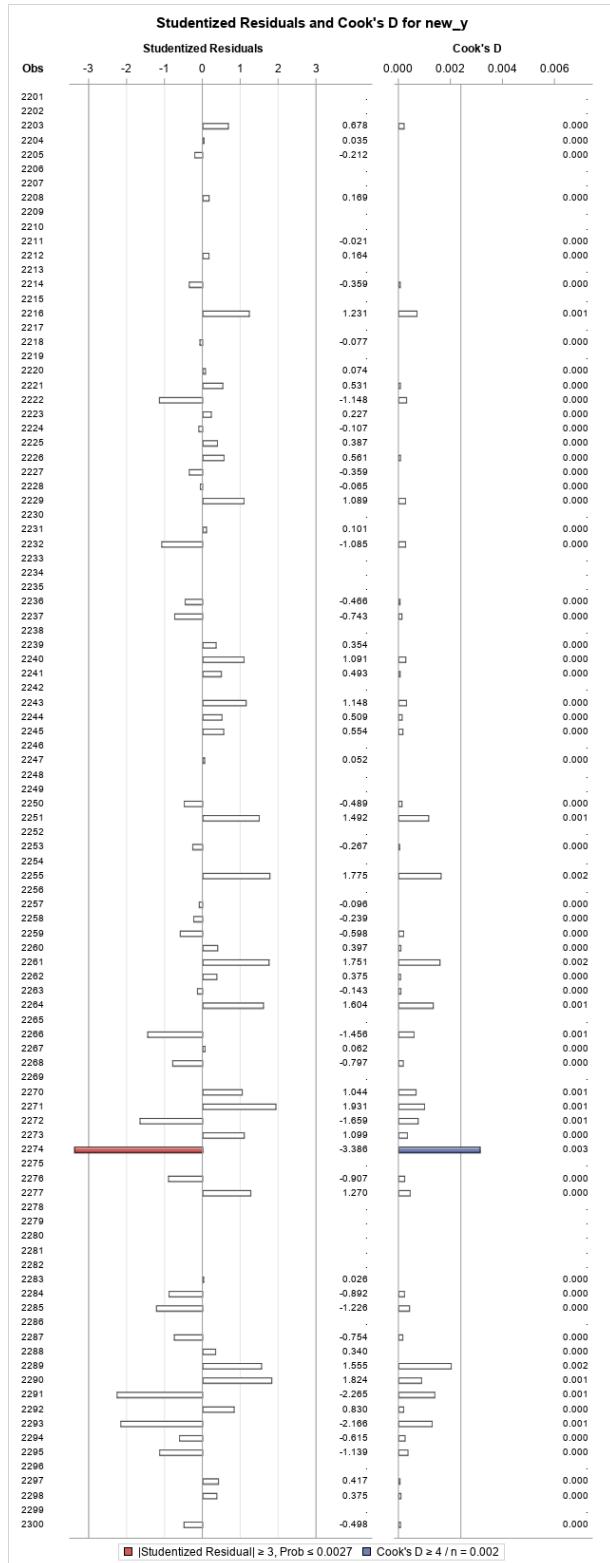


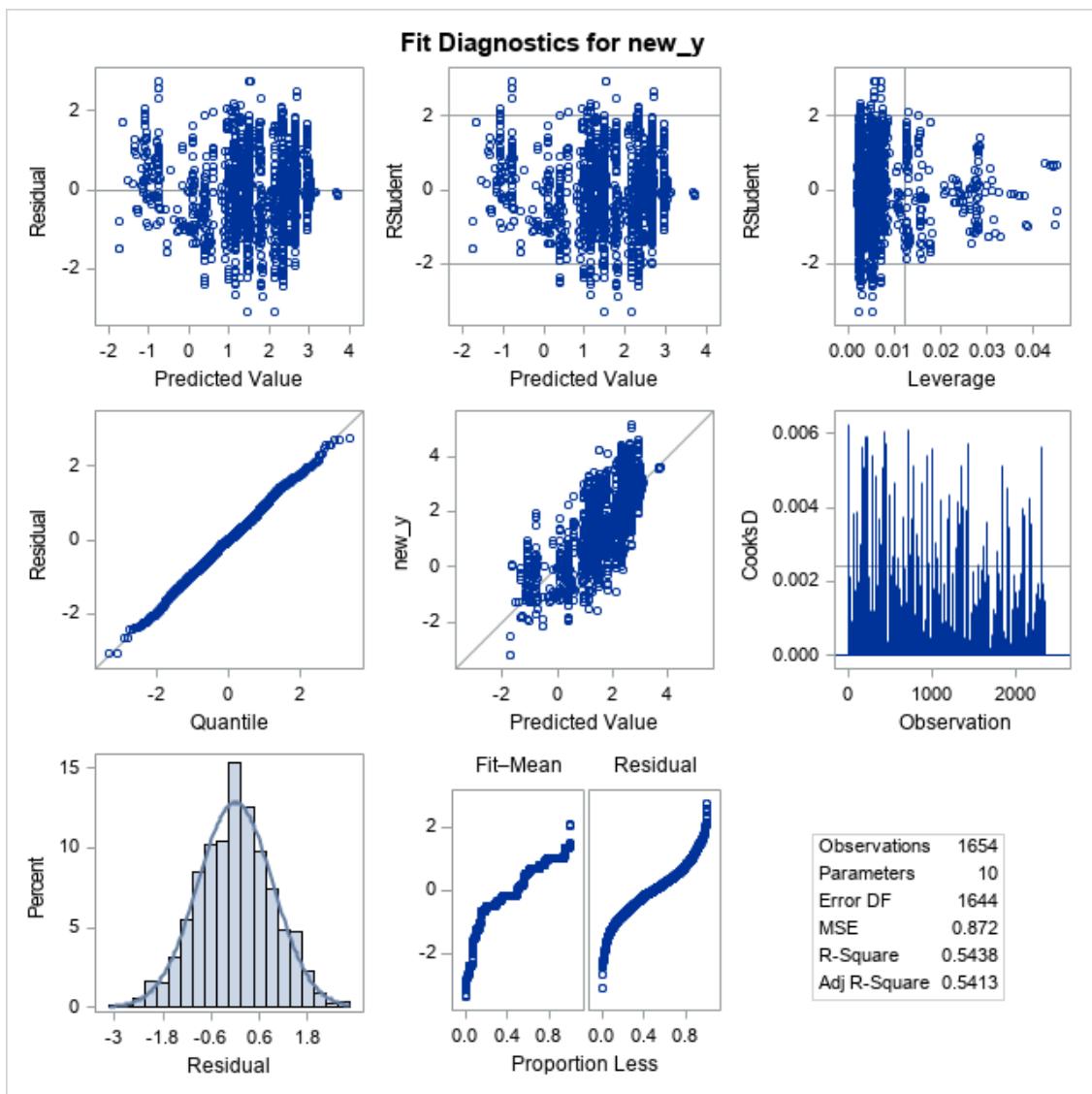












Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	1.47489	0.04714	31.29	<.0001
population	1	-2.04877E-8	7.037048E-9	-2.91	0.0036
gdp_for_year_dollars	1	6.11426E-14	1.44567E-14	4.23	<.0001
dsex	1	1.17262	0.04610	25.43	<.0001
dgeneration3	1	-2.48734	0.15483	-16.06	<.0001
dgeneration4	1	-2.23228	0.07777	-28.70	<.0001
dage1	1	2.30050	0.16132	14.26	<.0001
dage5	1	0.30628	0.06929	4.42	<.0001
dcontinent1	1	-0.88739	0.10494	-8.46	<.0001
dcontinent2	1	-0.31239	0.04838	-6.46	<.0001

The REG Procedure  
Model: MODEL1  
Dependent Variable: new\_y

Number of Observations Read	2367
Number of Observations Used	1654
Number of Observations with Missing Values	713

Analysis of Variance					
	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	1708.93825	189.88203	217.75	<.0001
Error	1644	1433.56599	0.87200		
Corrected Total	1653	3142.50424			

Root MSE	0.93381	R-Square	0.5438
Dependent Mean	1.64818	Adj R-Sq	0.5413
Coeff Var	56.65689		

Ram23

2 Variables:	Insuicidesper100kpop yhat						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Insuicidesper100kpop	713	1.62811	1.39255	1161	-3.21888	5.20576	
yhat	713	1.62890	1.00018	1161	-1.98710	3.14432	Predicted Value of new_y
Pearson Correlation Coefficients, N = 713							
Prob >  r  under H0: Rho=0				Insuicidesper100kpop		yhat	
Insuicidesper100kpop				1.00000		0.72762 <.0001	
yhat				0.72762		1.00000 <.0001	
Predicted Value of new_y							

The SAS System

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	713	0.95536	0.75161

Ram24

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Predict	95% CL Mean		95% CL Predict		Residual
1	.	1.6654	0.0985	1.4723	1.8584	-0.2427	3.5735	.
2	.	1.0360	0.0580	0.9222	1.1498	-0.8657	2.9377	.
3	1.50185	0.6619	0.0913	0.4829	0.8410	-1.2448	2.5687	0.8399
4	-0.86750	0.5568	0.0923	0.3757	0.7378	-1.3501	2.4637	-1.4243
5	1.59127	0.6637	0.0914	0.4846	0.8429	-1.2430	2.5705	0.9275
6	-0.54473	0.5885	0.0911	0.4098	0.7672	-1.3182	2.4951	-1.1332
7	2.25759	0.5353	0.0991	0.3410	0.7296	-1.3729	2.4435	1.7223
8	-0.56212	0.4382	0.1001	0.2420	0.6344	-1.4702	2.3466	-1.0003
9	0.01980	-1.5326	0.1481	-1.8230	-1.2422	-3.4530	0.3878	1.5524
10	-1.96611	-1.6302	0.1483	-1.9209	-1.3394	-3.5506	0.2903	-0.3360
11	1.33763	0.6643	0.0914	0.4851	0.8435	-1.2424	2.5710	0.6733
12	-0.96758	0.6306	0.0906	0.4529	0.8083	-1.2760	2.5372	-1.5982

## Ram25

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
new_y	1654	1.64818	1.37880	2726	-3.21888	5.15825
population	2367	2569528	5210182	6082072935	1269	43805214
gdp_for_year_dollars	2367	7.26759E11	2.37846E12	1.72024E15	610930037	1.49644E13
dsex	2367	0.53274	0.49903	1261	0	1.00000
dgeneration3	2367	0.20279	0.40216	480.00000	0	1.00000
dgeneration4	2367	0.10731	0.30957	254.00000	0	1.00000
dage1	2367	0.17913	0.38354	424.00000	0	1.00000
dage5	2367	0.14364	0.35080	340.00000	0	1.00000
dcontinent1	2367	0.05027	0.21856	119.00000	0	1.00000
dcontinent2	2367	0.37305	0.48372	883.00000	0	1.00000
pop_gdp	2367	1.12064E19	5.91921E19	2.65256E22	7.7527E11	6.40453E20

Pearson Correlation Coefficients												
	Prob >  r  under H0: Rho=0											
	Number of Observations											
	new_y	population	gdp_for_year_dollars	dsex	dgeneration3	dgeneration4	dage1	dage5	dcontinent1	dcontinent2	pop_gdp	
new_y	1.00000	-0.05576 0.0233 1654	0.04178 0.0894 1654	0.43910 <.0001 1654	-0.06635 <.0001 1654	-0.49042 <.0001 1654	0.02591 0.2922 1654	0.20085 <.0001 1654	-0.11763 <.0001 1654	-0.08531 0.0005 1654	0.03308 0.1787 1654	
		1.00000	0.75394 <.0001 2367	-0.03718 0.0705 2367	0.00854 0.6779 2367	0.04803 0.0194 2367	-0.00280 0.8919 2367	-0.13796 <.0001 2367	-0.03041 0.1392 2367	-0.04687 0.0226 2367	0.79583 <.0001 2367	
			1.00000 <.0001 2367	-0.01936 0.3463 2367	-0.00436 0.8321 2367	0.02627 0.2014 2367	-0.00878 0.6694 2367	0.01881 0.3604 2367	-0.05527 0.0072 2367	-0.06774 0.0010 2367	0.89186 <.0001 2367	
dsex	0.43910 <.0001 1654	-0.03718 0.0705 2367	-0.01936 0.3463 2367	1.00000 0.7809 2367	-0.00572 0.8610 2367	-0.00360 0.6767 2367	-0.00857 0.4201 2367	0.01658 0.7625 2367	0.00621 0.4201 2367	-0.00947 0.6450 2367	-0.01557 0.4489 2367	
dgeneration3	-0.06635 0.0070 1654	0.00854 0.6779 2367	0.00436 0.8321 2367	-0.00572 0.7809 2367	1.00000 0.2367 2367	-0.17486 0.2367 2367	0.92622 <.0001 2367	-0.20656 0.7913 2367	-0.00544 0.7913 2367	0.00638 0.7563 2367	-0.01168 0.5701 2367	
dgeneration4	-0.49042 <.0001 1654	0.04803 0.0194 2367	0.02627 0.2014 2367	0.00360 0.8610 2367	-0.17486 0.2367 2367	1.00000 0.2367 2367	-0.16196 0.2367 2367	-0.14200 0.4001 2367	-0.01730 0.2295 2367	-0.02471 0.7610 2367	0.00626 0.7610 2367	
dage1	0.02591 0.2922 1654	-0.00280 0.8919 2367	-0.00878 0.6694 2367	-0.00857 0.6767 2367	0.92622 <.0001 2367	-0.16196 0.2367 2367	0.00000 0.2367 2367	-0.19132 0.2367 2367	-0.00160 0.9382 2367	0.00872 0.6715 2367	-0.01357 0.5095 2367	
dage5	0.20085 <.0001 1654	-0.13796 <.0001 2367	0.01881 0.3604 2367	0.01658 0.4201 2367	-0.20656 0.2367 2367	-0.14200 0.2367 2367	-0.19132 0.2367 2367	1.00000 0.2367 2367	-0.01154 0.5747 2367	-0.00457 0.8241 2367	-0.04190 0.0415 2367	
dcontinent1	-0.11763 <.0001 1654	-0.03041 0.1392 2367	-0.05527 0.0072 2367	0.00621 0.7625 2367	-0.00544 0.7913 2367	-0.01730 0.4001 2367	-0.00160 0.9382 2367	-0.01154 0.5747 2367	1.00000 0.2367 2367	-0.17748 0.2367 2367	-0.04137 0.0442 2367	
dcontinent2	-0.08531 0.0005 1654	-0.04687 0.0226 2367	-0.06774 0.0010 2367	-0.00947 0.6450 2367	0.00638 0.7563 2367	-0.02471 0.2295 2367	0.00872 0.6715 2367	-0.00457 0.8241 2367	-0.17748 0.2367 2367	1.00000 0.0415 2367	-0.09388 0.0415 2367	
pop_gdp	0.03308 0.1787 1654	0.79583 <.0001 2367	0.89186 <.0001 2367	-0.01557 0.4489 2367	-0.01168 0.5701 2367	0.00626 0.7610 2367	-0.01357 0.5095 2367	-0.04190 0.0415 2367	-0.04137 0.0442 2367	-0.09388 0.0415 2367	1.00000 0.0415 2367	

## Ram26

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	63.48780	21.16260	11.34	<.0001
Error	1650	3079.01644	1.86607		
Corrected Total	1653	3142.50424			

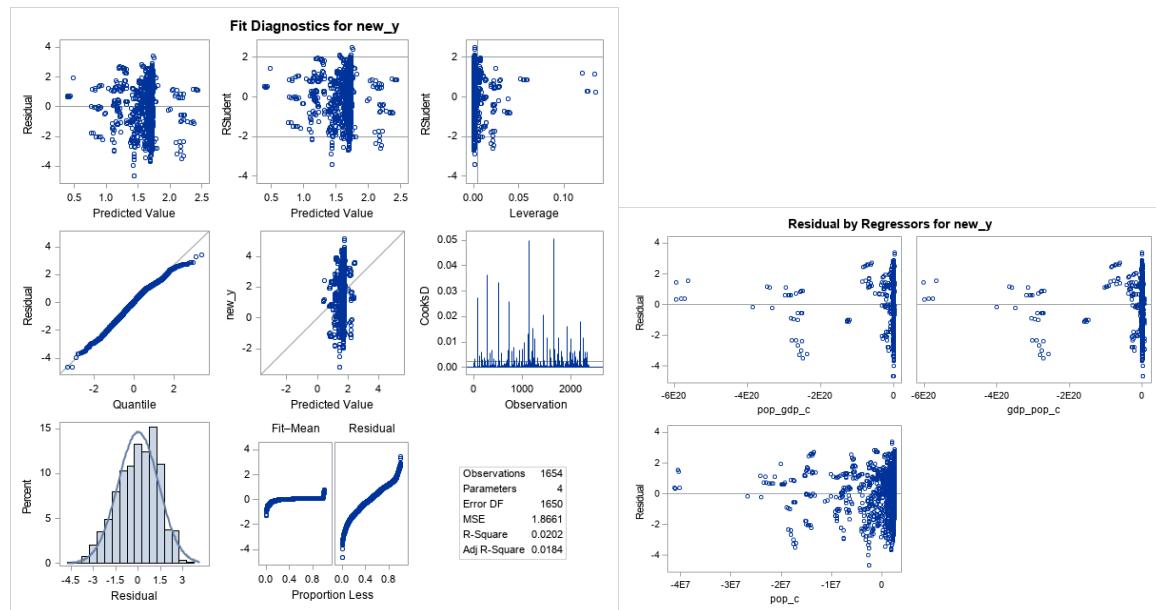
  

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.61545	0.03618	3720.56284	1993.80	<.0001
gdp_pop_c	-2.4786E-20	1.09432E-20	9.57316	5.13	0.0236
pop_gdp_c	2.17877E-20	1.20045E-20	6.14696	3.29	0.0697
pop_c	4.313237E-8	1.222008E-8	23.24804	12.46	0.0004

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	pop_c		1	0.0031	0.0031	28.7857	5.15	0.0233
2	gdp_pop_c		2	0.0151	0.0182	5.2941	25.46	<.0001
3	pop_gdp_c		3	0.0020	0.0202	4.0000	3.29	0.0697

## Ram27



Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	63.48780	21.16260	11.34	<.0001
Error	1650	3079.01644	1.86607		
Corrected Total	1653	3142.50424			

Root MSE	R-Square	0.0202
Dependent Mean	Adj R-Sq	0.0184
Coeff Var		82.88173

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	1	1.61545	0.03618	44.65	<.0001
pop_gdp_c	1	2.17877E-20	1.20045E-20	1.81	0.0697
gdp_pop_c	1	-2.4786E-20	1.09432E-20	-2.26	0.0236
pop_c	1	4.313237E-8	1.222008E-8	3.53	0.0004