Diabetes Prevalence In the United States

The Miner League Technical Report

**Prepared By:** Maxwell Carduner,Kalaivani Chandramohan,William Chirciu,Ramkumar Perumal,Charles Saporito

***ABSTRACT***

According to a 2015 American Diabetes Association study (Exhibit 1.13), 30.3 million people had diabetes in United States. This represents a national prevalence rate of 9.4%. The disease problem related to diabetes is high and rising in the United States, fueled by the rise in the prevalence of obesity and unhealthy lifestyles. The objective of this analysis is to determine type 2 diabetes prevalence percentage on a county level using demographics such as employment, education, income, region, obesity and physical inactivity to better target areas with a high diabetic population. The results of our data analysis show that unemployment rate, income, percent of adults with a college degree, obesity prevalence, physical inactivity, and the county region (Northeast, Midwest, South) contribute to a good model which is able to predict with around 80% certainty, the diabetes prevalence in counties throughout the United States.

***Introduction***

Diabetes is a lifelong disease that affects how your body handles glucose, the sugar in your blood. Your pancreas produces insulin to handle the glucose, but the disease causes your cells not to utilize it properly. The insulin tries to get the glucose into the cells to store energy but is unable to keep up, so the glucose keeps building up in the blood which results in symptoms such as hunger, fatigue, dehydration, and blurred vision. Type 2 diabetes is the most popular form of diabetes which plagues more than 27 million people in the United States. Wanting to pinpoint areas with a large presence of individuals with diabetes, we decided to analyze county-level data taken from the Center for Disease Control website, the USDA website, and from the US Census. Understanding situations surrounding a high diabetes prevalence makes it much easier for organizations attempting to combat the disease to do their jobs.

Data provided by the CDC includes physical inactivity percentage, obesity prevalence, and diabetes prevalence. A valid hypothesis would be to say that physical inactivity and obesity are strongly associated with diabetes. However, we wanted to see how things like education, unemployment, income, and median age contributes to the presence of diabetes within a county. Research as cited in Figure 3.15 suggests that counties with a high density of people living under the poverty level had strong associations with diabetes prevalence. They also included things like education, unemployment, population density, percentage non-white, percentage Hispanic, obesity, and physical inactivity in their model. They concluded that poverty level, physical activity, and walking or cycling to work had significant impact on diabetes prevalence in the United States. We will be doing similar analysis with our regression models in determining if physical inactivity, obesity, and other demographic attributes are significant enough to tell us, with as much accuracy as possible, the diabetes prevalence in a county so we can focus more on those areas where diabetes is a common household illness.

**Methodologies**

**Team Member:** Maxwell Carduner

**Data Source:** The dependent variable, county level diagnosed diabetes prevalence, and lifestyle indicators on a county level such as obesity rates and leisure-time physical inactivity rates came from the CDC (<https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html>). The following socio-economic indicators on a county level came from the USDA (<https://www.ers.usda.gov/data-products/county-level-data-sets/>): Unemployment Rate, Household Median Income, Percent of adults with less than a high school diploma, Percent of adults with a high school diploma only, Percent of adults completing some college or associate's degree, and Percent of adults with a bachelor's degree or higher. We obtained Median Age from Census data (<https://datausa.io/map/?level=county&key=age,age_moe,age_rank>).

**Approach**: The CDC diabetes prevalence and lifestyle indicators all downloaded as separate tables so they were merged into one table by F.I.P.S. Code (Zip Code). Because the socio-economic indicators were also available by Zip Code but in separate tables as well, they were merged with the CDC data by Zip Code in order to have all of the predictors and dependent variable, diabetes prevalence, in one dataset for analysis. All of the merging was performed in Excel. After creating our final dataset, we saved the final table as a “.csv” file in order to import into SAS. All variables were numeric so no re-coding was needed.

In order to examine whether transformations were needed, histogram of the dependent variable and scatter plots between the dependent variable and all independent variables were created. The scatter plots between the dependent variable and the independent variables were be evaluated to ensure that the independent variables are linearly associated with the dependent variable.

Interaction variables were created where it made sense: unemployment rate and household median income, obesity prevalence and leisure time physical inactivity rate, median age and median income, unemployment rate and median age, obesity prevalence and median age, leisure time physical inactivity rate and median age, leisure time physical inactivity rate and median income, and Percent of adults with a high school diploma only and unemployment rate.

To satisfy the model assumption that error terms are independent of each other and have constant variance, studentized residual plots were created for the dependent and predicted variables. The normal probability plot of the residuals will be produced to ensure that the error terms are normally distributed. Additionally, observations that are both outliers and influential points will be identified by comparing the studentized residuals and Cook’s D for each point.

**Team Member:** Kalaivani Chandramohan

**Data Source:** We used diabetes prevalence data from the CDC. Data describing the county level is used to create a model to predict the diabetes prevalence within a county by using other county level data.

The variables used are,

1. Unemployment rate

2. Household Median Income

3. Percent of adults with less than a high school diploma

4. Percent of adults with a high school diploma only

5. Percent of adults completing some college or associate degree

6. Percent of adults with a bachelor's degree or higher

7. Obesity Prevalence Percent

8. Leisure Time Physical Inactivity Prevalence Percent

9. Median Age

10. Diagnosed Diabetes Est. Percent

Approach:

1. The distribution of Diagnosed Diabetes Est. Percent is done using the histogram.

2. Examine scatterplots between the dependent and all the independent variables. Linearity is also checked in this phase.

3. Examine the correlation using the Pearson coefficient.

4. To check the significance of the predictors from the regression output. Also, the VIFs/TOL for multicollinearity is examined here.

5. Identify Outliers and Influential Points and remove the observation from the model to improve the performance.

6. Partition data into training and test set.

7. Run model selection on the train dataset and then apply the selection methods (forward/stepwise/backward/adj r^2/r^2).

8. Check for outliers in the selected training models and remove the necessary observations from the model.

9. Rerun the model selection and get the final models for the train data. The residual pattern followed by the predictors have been verified along with the normality tests.

10. The goodness of fit for the test data is determined.

**Team Member:** William Chirciu

**Data Source:** Our dataset is focused on the prevalence of diabetes (in percent) of various counties in the United States. The dataset was integrated from multiple sources. Diabetes prevalence, physical inactivity percentage, and obesity prevalence were all pulled from the CDC website (<https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html>). Household median income,

% of adults with: no highschool diploma, highschool diploma only, associate’s/college degree, bachelor’s degree or higher all came from the USDA website (<https://www.ers.usda.gov/data-products/county-level-data-sets/>. Finally, median age came from US census data<https://datausa.io/map/?level=county&key=age,age_moe,age_rank>.

**Approach:** After integrating the data into a single csv file, I imported the whole thing into SAS. The first thing I sought to do was determine if any transformations needed to be made on any of the variables. First, I checked the distributions of each variable by looking at their histograms. Our dependent variable was perfectly symmetric while the other variables were skewed distributions. However, the variance within those distributions did not seem worth stabilizing. The next thing I checked was the association of our dependent variable with all other variables. Seeing that these associations were basically linear, I concluded that no transformations were required.

Next, I checked for multicollinearity in multiple respects. I first looked at the correlation matrix with our original variables. Because there were no correlation values >0.90, I moved on to create my interaction terms. I then fit a full regression model including interaction terms to determine insignificant terms. After removing the appropriate variables, I once again checked for multicollinearity. Because my remaining interaction terms had high correlation values (>0.90), VIFs >10, and TOLs < 0.10, I decided to center the relevant variables (subtract the mean value of the common variable from each value of the common variable and update each interaction variable appropriately).

Prior to model selection, I wanted to check for heavy outliers and influential points. I found 6 observations that I felt needed to be removed from the dataset to improve our model. I then proceeded to split the data into a training set (80%) and test set (20%).

Wanting to implement a model that maximizes Adjusted R-Square, I used SAS’s ‘adjrsq’ model selection in the regression procedure. I chose the set of variables that maximized Adj. R-Sq while at the same time minimizing the number of predictors in the model. At this point, I wanted to check for significant outliers/influential points. Having found no observations worthy of removing, I decided to evaluate this final model on the test set.

I first calculated the difference between the observed diabetes prevalence in the test set and the predicted value. I used this to calculate the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) of the model. I then created a correlation matrix between the observed and predicted value of diabetes prevalence to determine the overall R-Square of the test set. Finally, I made two predictions with made-up data to see what the model would predict and the appropriate confidence intervals.

**Team Member:** Ramkumar Perumal

**Data Source:**

Our data source was collected from CDC, and USDA websites as mentioned above by my teammates. They are finally merged to include the below variables**.**

**Independent variables:**

*1. Unemployment Rate (Unemp\_Rate)*

*2. Household Median Income (Income)*

*3. Percent of adults with less than a high school diploma (lt\_HighSchool)*

*4. Percent of adults with a high school diploma only (only\_HighSchool)*

*5. Percent of adults completing some college or associate's degree (Col\_Degree)*

*6. Percent of adults with a bachelor's degree or higher (Bach\_Degree)*

*7. Obesity Prev Percent (Obesity)*

*8. Leisure Time Physical Inactivity Prev Perc (Ph\_Inactivity)*

*9. Median Age (Age)*

*10. State*

*11. Region*

**Dependent variable:**

*Diabetes Est. Percent (Diag\_Pct)*

**Approach:**

1. Import the data and add dummy variables for Region (d\_Northeast, d\_Midwest, d\_South)

2. Examine distributions of *Diag\_Pct*

3. Examine scatterplots of all variables in the dataset to:

i. Check for linearity (*Diag\_Pct vs all other independent variables*)

4. Examine the correlation

· Check for possible multicollinearity as well

5. Estimating model parameters:

i. Check the significance of predictors(remove if necessary)

ii. Examine VIFs for multicollinearity

6. Identify Outliers and Influential Points and remove them

7. Re-run linear regression to see if the predictors still valid for alpha (0.05)

8. Partition data into training and test set

9. Run model selection on the train dataset

i. Apply multiple selection methods

ii. Retain the selected models and variables for further validation

10. Check for outliers in the selected training models and remove them

11. Rerun the model selection and get the final models for the train data

i. Check residual and normal plot

12. Get all the predicted values and assess their goodness of fit for the test data

13. Compare RMSE, MAE, R2, and F Value to choose the best model

**Team Member:** Charles Saporito

**Data Source:** Data was retrieved from the CDC website. Data containing those diagnosed with diabetes and their status within those counties were included in the data set provided here (<https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html>). This data set included our dependent variable and independent variables of indicators for diabetes prevalence. Two separate data sets one for median age of those diagnosed with the disease and education levels came from two separate data sets from the USDA and CDC websites and were merged to the original data set. Those sources are provided here.(<https://www.ers.usda.gov/data-products/county-level-data-sets/>) (<https://datausa.io/map/?level=county&key=age,age_moe,age_rank>).

**Approach:** The project consisted of four main steps to properly conduct a regression analysis in order to produce a model that would explain diabetes prevalence among counties in the United States. As a group we decided on focusing our analysis around data provided from government databases. After deciding on focusing our research and analysis on diabetes we queried the CDC and USDA databases for our data. The data provided from the query resulted in our variables for analysis. The data was then extracted and merged into one spreadsheet from the web sources and explored to get a better understanding of what the data meant in terms of our indicators.

After the initial exploration of the raw data the file was then cleaned and transformed into csv file for analysis using SAS. The csv file was then saved and the initial proc import statement was written in SAS to produce a table ready for statistical analysis. There was no need for transformation of the variables into dummy variables and the exploratory stage of analysis was started. In order to understand the distribution of diabetes prevalence a proc means procedure was written. This procedure created the five number summary to analyze the distribution of the data within the percentiles. A proc univariate procedure was written into SAS to visualize the distribution of each variable including our y-variable diabetes diagnosed percentage. The distribution was normal for our y variable, the x-variables were significantly skewed. At this point it was decided by myself that a log transformation was needed to linearly associate them with our y-variable. I then wrote transformations for all x variables. This had solved the issue, and continued my analysis using transformed x-variables.

The next step in the statistical approach was to check for any multicollinearity. Procedures were written into SAS to produce a correlation matrix, pearson correlation tables,and variance inflation rates for each x-variable. After writing those procedures the statistic were analyzed for goodness of fit, error rate, r-squared, and adj r-sq among all the variables in the dataset. A determination that the model could be better explained was concluded and procedures to check for outliers and influential points using the Cooks D statistical graphs were created. Analysis of that output determined removing outliers were recommended and their removal was written into SAS where the proc reg procedure was then re ran to produce a model without outliers including their residual plots in the SAS syntax. After analysis of these outputs it was determined that the model was now ready for test and train to fit a final model

Test and Train was then approached to conclude a final model. First to analyze which variables were the strongest predictors within our data a output was produced to compute standardized estimates of each variable. Then interaction variables were written as well since, through my analysis an association between median income and inactivity variables existed. A proc survey procedure was written and the data was split at .75. After splitting the data a new y was created and the data was ran for model selection using both a stepwise and cp method.The model was then decided on and predictions were ran to see if the model can produce meaningful outputs to determine diabetes prevalence determined by the cross validation of the prediction model created.

***Analysis,Results and Findings***

**Team Member: Maxwell Carduner**

To understand whether or not a linear model would suffice for predicting the dependent variable, diagnosed diabetes prevalence by county, a histogram was created (Exhibit 1.1). Because the histogram had a roughly normal distribution, a linear model did make sense to predict the dependent variable.

To understand whether or not the independent variables needed to be transformed, scatter plots between each independent variable (before interaction terms were created) and the dependent variable were created (Exhibit 1.2). All of the independent variables appeared to have a linear relationship with the dependent variable, so no transformations were needed for the independent variables.

After determining that a linear model made sense, multicollinearity was addressed by examining the correlation table of all of the variables and ran a preliminary linear model on all of the variables to examine VIFs (Exhibit 1.3). After examining the VIFs and correlation table, there were obvious multicollinearity issues between the interaction terms and the independent variables used in the interaction terms. However, several of the interaction variables were insignificant and therefore dropped: median age and median income, obesity prevalence and median age, leisure time physical inactivity rate and median age, leisure time physical inactivity rate and median income, and Percent of adults with a high school diploma only and unemployment rate. Additionally, the Percent of adults with a bachelor's degree or higher variable’s coefficient SAS set to 0 because it was a linear combination of other variables (Exhibit 1.4).

After dropping the insignificant interaction variables and the Percent of adults with a bachelor's degree or higher, another linear model was fitted to examine VIFs (Exhibit 1.5) and a correlation table was produced to understand which variables needed to be centered. After examining the output, it was deemed that the following non-interaction independent variables needed to be centered: unemployment rate, median income, obesity rate, leisure time physical inactivity rate, and median age. Subsequently, the interaction terms containing these variables were created using the centered variables.

After centering the variables, linear models were fit to the centered interaction terms (Exhibit 1.6) and to the original variables in order to compare the models (Exhibit 1.7). Adjusted R squared did improve from 71.73% to 72.99%, so the interaction terms increased the accuracy of the model and will be kept going forward. Additionally, all of the VIFs were lower than 10, so the centering of the variables took care of the multicollinearity concerns.

Of the 3000+ observations used in the analysis, only about 10 were outliers and influential points. Additionally, these observations probably exist because they are counties that are older, have higher obesity rates, are less physically active, have a high unemployment rate, and have lower household income than most counties but that doesn’t mean they are not going to have a high diabetes rate. Because this study is to help identify predictor variables on a county level that lead to higher diabetes prevalence on a county level and less than 0.33% of our points are influential points and outliers, these observations were not removed from the analysis.

To satisfy the model assumptions that the error terms are independent of each other and have constant variance, studentized residuals of the predicted value and the independent variables were created (Exhibit 1.8). After examining the outputs, they all appeared to satisfy this assumption with the exception of a few outliers that may cause you to think there is a funnel shape appearing in the non interaction terms and the interaction terms hovering mostly around 0 because they are products of centered terms. Therefore, all of the variables error terms appear to be independent of each other and have constant variance.

To test that the error terms are normally distributed, a normal probability plot of the studentized residuals was created (Exhibit 1.9). The line from the normal probability plot appears to be roughly at a 45 degree angle so the error terms are normally distributed.

To understand which of the independent variables is the strongest predictor of the dependent variable, standardized coefficients were calculated (Exhibit 1.10). The strongest predictors in order of strength are: Obesity rate, leisure time physical inactivity rate, median age, and unemployment rate.

To fit the final model and validate the results, a 5 fold cross-validation technique was applied. The results are shown in Exhibit 1.11. All eleven variables were left in the model as they were significant. Additionally, adjusted r squared remained at 72.99%.

The final eleven independent variables that predict the dependent variable, diagnosed diabetes prevalence percentage, are: Unemployment rate (centered), household median income (centered), percent of adults with less than a high school degree, percent of adults with a high school degree only, percent of adults with an associate’s degree or some college, obesity rate (centered), leisure time physical inactivity rate (centered), median age (centered), obesity rate (centered) multiplied by leisure time physical inactivity rate (centered), unemployment rate (centered) multiplied by median age (centered), and obesity rate (centered) multiplied by household median income (centered).

Created two cases to predict the diabetes prevalence in two hypothetical counties (Exhibit 1.12):

The first prediction had a county with a centered unemployment rate of 1% (representing 1% lower than average), a centered household median income of -$10,000 (representing $10,000 more than average), 10% of adults with less than a high school degree, 50% of adults with a high school degree only, 40% adults with an associate’s degree or some college, a centered obesity rate of 20% (representing 20% lower than average), a centered leisure time physical inactivity rate of 12% (representing 12% lower than average), a centered median age of 2 (representing 2 years younger than average), and the subsequent interaction terms you get by multiplying the prediction values: 240 for obesity rate (centered) multiplied by leisure time physical inactivity rate (centered), 2 for unemployment rate (centered) multiplied by median age (centered), and -200,000 for obesity rate (centered) multiplied by household median income (centered). This yielded a predicted value of 2.8% for diagnosed diabetes prevalence rate with a 95% prediction confidence interval between 0.18% and 5.44%.

The second prediction had a county with a centered unemployment rate of -5% (representing 5% higher than average), a centered household median income of $10,000 (representing $10,000 less than average), 20% of adults with less than a high school degree, 60% of adults with a high school degree only, 20% adults with an associate’s degree or some college, a centered obesity rate of -10% (representing 10% higher than average), a centered leisure time physical inactivity rate of -12% (representing 12% higher than average), a centered median age of -5 (representing 5 years older than average), and the subsequent interaction terms you get by multiplying the prediction values: 120 for obesity rate (centered) multiplied by leisure time physical inactivity rate (centered), 25 for unemployment rate (centered) multiplied by median age (centered), and -100,000 for obesity rate (centered) multiplied by household median income (centered). This yielded a predicted value of 17.2% for diagnosed diabetes prevalence rate with a 95% prediction confidence interval between 14.65% and 19.82%.

**Team Member**: Kalaivani Chandramohan

**Exploratory data analysis**

The data has been imported from the csv file (appendix 2.1)

**Distribution**

To analyze the distribution of Diagnosed Diabetes Est. Percent using histogram

Here from the histogram, the distribution is normal and symmetric (Appendix 2.2). Therefore, transformation is not required for this dataset.

**Correlation**

The strength of a linear association between 2 variables is defined as correlation. In this dataset the relationship between Diagnosed\_percent vs the other variables have been calculated.

For the Pearson correlation, the larger the absolute value, the stronger the linear association: a correlation of –1.0 indicates a perfectly negative linear association, 0.0 indicates no linear association, and +1.0 indicates a perfectly positive linear association.7 Illustrating the overall concept of correlation, Appendix (4.3) is a graphical representation of hypothetical scatter plots of data for all the variables.

**Positive Correlation**

Weak Relationship - Age

Moderate Relationship - Unemp\_Rate, HighSch\_only, less\_HighSchool\_Pct, Bach\_Deg\_Pct

Strong Relationship - Obes\_Per, Leisure\_Inact

**Negative Correlation**

Weak Relationship - College\_Deg

Moderate Relationship - Income

Strong Relationship – None

The relationship between the diagnosed percentage and other variable are available at appendix (2.3). Obesity and the Physical inactivity has strong correlation relationship with the diagnose percent.

**Multicollinearity**

Multicollinearity exists when two or more of the predictors in a regression model are moderately or highly correlated.

From the Pearson correlation Coefficients table, we can say that there is no multicollinearity exists in this model. The predictors are not highly correlated with each other. The variance inflation factor (VIF) has been computed to verify the collinearity once again. From the output we can say that, none of the values are greater than 10. This can be referred to (Appendix 2.5)

The predictor bachelor degree has been removed from the model as the data is biased.

**Outliers and influential points**

In linear regression, an outlier is an observation with large residuals.The affected observation are removed from the model.

The outliers with the standardized residuals ≥ 3 are removed.

An observation is said to influential if removing the observation substantially changes the estimate of coefficients. The Influential points are detected based on the Diffts, Dfbetas and the cook’s distance D of larger values.

| Diffts |>2 => 0.107

After removing the outliers and the influential points, the final model is created. The adj r2 value has been improved from 0.72 to 0. 764.The parameter estimates has been significantly changed (Appendix 2.6).

The data issues have been fixed in this stage for the further analysis.

**Partition of data**

All the data issue has been fixed in the exploratory step and then the data split is done for both the train and test (75% and 25% respectively). A new\_y variable which represents the diagnosed percentage (Diagnose\_percent) has been added. (Appendix 2.7)

**Training dataset with the Model selection**

The model selection methods applied to the train data set are backwards (Appendix 2.8) and the CP (Appendix 2.9)

The outcome of both the methods are the same which ended up with the below 8 predictors.

Unemp\_Rate Income less\_HighSch\_Pct HighSch\_only College\_Deg Obes\_Per Leisure\_Inact Med\_Ag

**Fitness of the model**

All the variables left in the model are significant (Appendix 2.10)

From the overall goodness of the fit, we can say that, F-value is high, p-value is smaller than 0.05. Hence, alternative hypothesis is being rejected and null hypothesis is accepted. The root mse value is high, the r-square and the adj r2 value is also high. The predictors are significant.

The Residual plot follows some patterns (Appendix 2.11)

The Unemp\_Rate,Less\_Highsch\_Pct,College\_deg,Leisure\_Inact and the Median\_age doesn’t follow any pattern. The residual shows the data are randomly scattered around the zero line and shows Constance.

The Highsch\_only and the Obes\_per shows an increasing trend and follows a funnel shaped pattern.

The Household Median income shows a decreasing trend and the points are not randomly scattered.

The residual plot shows a linear trend and the data is normally distributed. The assumption of normality is satisfied.

**Test data validation**

During the selection process, models are fit on the training data, and the prediction error for the models so obtained is found by using the validation data.

The linear regression model gives the predicted values for the new\_y variable (Appendix 2.12). The performance of the prediction is seen in the appendix (2.13)

**Here comes the final model,**

|  |  |  |
| --- | --- | --- |
|  | Model | |
| Train | RMSE | 1.15341 |
| R square | 0.7672 |
| Adj-R-Square | 0.7663 |
| GOF | ok |
| Residuals | ok |
|  |  |  |
| Test | RMSE | 1.35015 |
| MAE | 0.98005 |
| R square | 0.7146 |
| Adj-R-Square | 0.7136 |
| CV-R-square | 0.038020634 |

There is no M2 model exists here. So, the train and the test performance are compared.

The CV-R-Square value is less than 0.3. Hence, the above one is considered as a good predictive model. The RMSE values of the test is not high then the train. Therefore, we can say that, the above obtained model is good one

The model equation is the following,

**Diagnose\_Percent = -2.94 + 0.28 Unemp\_Rate - 0.00001 Income +0.036 less\_Highsch\_Pct -0.04 highsch\_only – 0.05Col\_Degree + 0.20 Obes\_per + 0.1735 Leisure\_Ianct + 0.11 Med\_Age**

In summary, we can say that the predictors obesity and the physical inactivity influences the diagnosed diabetes Estimate percent in high level when compared to the other variables.

**Team Member:** William Chirciu

**Checking if Transformations are Required**

Created histograms for each one of our original variables. Our dependent variable was symmetric with mean = 11.374% and Std.Dev = 2.51661. The rest of our independent variables had slightly skewed distributions, but we can see in **Figure 3.2** that the variation within their distributions was not worth stabilizing.

Next, I checked to see if linearity was violated. I created a scatterplot in SAS using all our original variables. We can see in **Figure 3.1** that all associations involving our dependent variable Diabetes\_Prevalence, are linear. Therefore, I concluded that no transformations were needed.

**Interaction Terms**

To help me understand my data further, I created several interaction terms. I decided to combine Age+Obesity, Age+Inactivity, Unemployment+Age, Unemployment+Obesity, and Unemployment+Inactivity. After checking for insignificant terms and looking at the results in **Figure 3.3**, Unemployment+Age and Unemployment+Obesity were the only interaction variables that remained.

**Checking for multicollinearity**

In **Figure 3.4**, we see that Unemployment\_Rate, Unemployment\_Age, and Unemployment\_Obesity all have VIFs > 10 and TOLs < 0.10. To counter this, I centered Unemployment\_Rate by subtracting each of its values from the mean and recalculated the interaction terms. Our other variables show no problems with multicollinearity.

**Initial Full Model**

Initial analysis of a full regression model in **Figure 3.5** shows an Adj. R-Sq of 0.7255 with an F-score of 830.45. We see almost all of our variables are now significant with p-values of <0.05. The only exception is Adults\_Bachelors\_Or\_Higher which was removed from our model since it is a linear combination of our other variables according to SAS.

**Checking for Outliers and Influential Points and Residual Plots**

I wanted to improve my full model as best as I could prior to performing model selection. To do this, I checked for outliers and Influential points by looking at the Studentized Residuals and Cook’s distance for each observation. The only ones I considered removing were those that had arrow heads on both ends (heavy outliers+influencers). In **Figure 3.6**, we can see the 6 observations I removed from the dataset.

Finally, I looked at the residual plots to determine if any patterns were present. The residual plot for our dependent variable looks to have a funnel pattern as seen in **Figure 3.7**. The same is true for our other predictors. So, it is safe to conclude that constant variance and independence is violated in this model

Fortunately, normality looks to be satisfied as seen in the quantile residual plot.

**Splitting my dataset**

Before I perform model selection, it was important that I split the data into a training set and a test set. **Figure 3.8** shows the parameters I used to do this. 80% (2507 observations) of my data made it into the training set and the other 20% (626 observations) into the test set. The new dataset was titled ‘diabetes\_split’.

**Model Selection**

For model selection, I wanted to do my best to maximize the Adjusted R-Square of the model. Performing adjrsq during the regression procedure in SAS returns many models from highest adjusted r-square to lowest adjusted r-square. In **Figure 3.9** I’ve shown the top six results from this analysis. Now, several of these models include the variable Adults\_Bachelors\_Or\_Higher. However, as we’ve seen before, this predictor is a linear combination of all the other predictors, so I am going to ignore all the results that contain it. Because of this I am left with 2 models: one has 10 predictors with an adjusted r-square of ~0.7316 and the other has 9 predictors with an adjusted r-square of ~0.7308. Because the model with 9 predictors has an Adjusted R-square value only slightly less than that of the model with 10 predictors, I am going to favor the former. Not having to record the percent of adults in a county without a high school diploma will save a lot more money and time.

**Final Model**

To ensure the model I had was finalized, I once again took a look to see if there were any significant influential points and outliers. I did not find any and decided to fit my final model. In **Figure 3.10**, we can see that I ended up with adjusted R-square of 0.7308 with an F-statistic of 756.75. This signifies a good model. All the variables are significant with p-values <0.0001. Looking at the standardized estimates, we can see that Unemployment\_Rate\_c along with its associated interaction terms (unemployment+age, unemployment+obesity) have a significant effect on our model with values of 0.52, -0.45,-0.32 respectively. It looks like obesity\_prevalence and physical\_inactivity\_prevalence also have a significant impact on the model with estimates of 0.38 and 0.37 respectively. Median\_Income appears to have the least impact with a standardized estimate of -0.09. Lastly, I wanted to see if anything changed with the residuals of my final model. Looking at **Figure 3.11,** our model still violates constant variance and independence as seen in the patterns in our residual plots. Our final model equation is as follows:

Obesity\_Prevalence (%) = -0.45459 + (6.24562 - Unemployment\_Rate) \* 0.58497 – Median\_Income\*0.00001793 – Adults\_HighSchool\_Diploma\_Only\*0.03957 – Adults\_Associates\_Degree\*0.06875 + Obesity\_Prevalence\*0.21275 + Physical\_Inactivity\_Prevalence\*0.17862 + Median\_Age\*0.11424 – ((6.24562 - Unemployment\_Rate)\*Median\_Age)\*-0.01250 – ((6.24562 - Unemployment\_Rate)\*Obesity\_Prevalence)\*0.01091

**Model Evaluation on Test Set**

I had my model predict Diabetes Prevalence on the test set. To compute the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE), I had to find the difference between the predicted value and the observed value in SAS. In **Figure 3.12**, we can see the results of our error calculation. In our test set with 626 observations, my model was able to predict the diabetes prevalence with RMSE = 1.30504 and MAE = 1.03168. We get a bit more detail in **Figure 3.13** from the descriptives of our observed and predicted target variables. Our predictions had a mean of 11.36 which is only slightly higher than our observed values with a mean of 11.30. I also calculated the R-Square of my model on the test set, which came out to be 0.85272^2 = 0.72713 which means we have an Adjusted R-Square of 0.72314. Looking at both our Adjusted R-Square and error values, I conclude that I have a decent model.

**Predictions**

The last thing I did was make 2 predictions. I made up values for 2 counties and ran my model on them. **Figure 3.14** shows the results of these predictions.

County 1:

Diabetes Prevalence = ? //Value to be predicted

Unemployment\_Rate = 6.0%

Household Median Income = $55192

Percent of Adults with no High School Diploma = 15.786% //Data not needed for our model

Percent of Adults with a High School Diploma only = 39.91%

Percent of Adults with a college/Associate’s Degree = 30.014%

Percent of Adults with Bachelor’s Degree or higher = 25.336% //Data not needed for our model

Obesity Prevalence = 32.5%

Physical Inactivity Prevalence = 27.3%

Median Age = 39.1

Predicted value: 11.1072%

95% C.I. Interval: 11.0188 – 11.1955

95% P.I. Interval: 8.5537 – 13.6607

County 2

Diabetes Prevalence = ? //Value to be predicted

Unemployment\_Rate = 5.7%

Household Median Income = $53467

Percent of Adults with no High School Diploma = 17.989% //Data not needed for our model

Percent of Adults with a High School Diploma only = 35.76%

Percent of Adults with a college/Associate’s Degree = 28.650%

Percent of Adults with Bachelor’s Degree or higher = 23.548% //Data not needed for our model

Obesity Prevalence = 30.6%

Physical Inactivity Prevalence = 29.6%

Median Age = 37.8

Predicted value: 11.1968

95% C.I. Interval: 11.1073 – 11.2863

95% P.I. Interval: 8.6432 – 13.7503

**Team Member: Ramkumar Perumal**

**Analysis, Results, & Findings:**

Importing the following 10 column data to my analysis, with 3139 observations

1. *Diag\_Pct (Dependent Variable)*

2. Unemp\_Rate

3. Income

4. lt\_HighSchool

5. only\_HighSchool

6. Col\_Degree

7. Bach\_Degree

8. Obesity

9. Ph\_Inactivity

10. Age

11. State

12. Region

Created Region variable based on state, to group them all into South, West, Midwest, and Northeast. Based on regions the dummy variables d\_South, d\_West, d\_Midwest, d\_Northeast are created(Appendix 4.1).

**Distribution of Estimated Diagnosis Percentage**

Here from the histogram (*Appendix 4.2*) we can see that the distribution of Diag\_Pct looks normal and symmetric. So this doesn’t need a transformation.

**Correlation Diag\_Pct vs Other variables**

|  |  |  |
| --- | --- | --- |
| **Strength\Polarity** | **Positive** | **Negative** |
| **Weak** | Age | Col\_Degree, d\_Northeast, d\_Midwest |
| **Moderate** | Unemp\_Rate,lt\_HighSchool, only\_HighSchool, Bach\_Degree, d\_South | Income |
| **Strong** | Obesity, Ph\_Inactivity | None |

From the scatterplot (Appendix 4.3) and correlation outputs (Appendix 4.4), we can observe that the relationship of Diag\_Pct with all other variables are as shown in the above table. It is important to point that Obseity and Ph\_Inactivity has strong positive association with Diag\_Pct.

**Multicollinearity Elimination**

Correlation coefficients do not indicate any multicollinearity among the independent variables, although to make sure ran the linear regression with VIF.

(Appendix 4.5) is the first output, which indicates Bach\_Degree is a linear combination of other variables leaving the model biased. We can go ahead and remove the Bach\_Degree.

The new model gives 11 variables without Bach\_Degree, running linear regression on this model confirms there aren’t any other interrelation among the X variables(Appendix 4.6).

**Outliers and Influential Points**

Running Linear regression with influence and r shows the observations’ Cook’s distance and DFFITS. For this sample the DFFITS limit is=> |DFFITS| > 2 \* root(p/n)

= 0.12

Removed all the observations with DFFITS above 0.12 and studentized residual above +/- 3

Fitting this data again with a linear model gives Adj R-Sq 0.80 which is better than earlier (Appendix 4.7). It’s noted that lt\_HighSchool and only\_HighSchool aren’t significant, however they are left to observe in the Training phase.

**Train and Test Data**

Split the data into 60% and 40% sampling sizes for train and test respectively. This is done by adding a new\_y variable representing the Diag\_Pct and it’s made empty for the test dataset (Appendix 4.8).

**Train data Model Estimation**

Using backward and stepwise selection method the linear model for the train dataset is estimated.

**M1:** The result of backward selection gives new\_y = Unemp\_Rate Income Col\_Degree Obesity Ph\_Inactivity Age d\_Northeast d\_Midwest d\_South. (Appendix 4.9)

**M2**: The result of stepwise selection gives new\_y = Unemp\_Rate Income Col\_Degree Obesity Ph\_Inactivity Age d\_South. (Appendix 4.10)

**Model fitness:**

As expected both model selection methods removed the lt\_HighSchool and only\_HighSchool as insignificant. Also the stepwise method removed d\_Northeast and d\_West as well.

As for the residuals of both models Income shows a pattern but it’s not so extreme. Others have a few extreme values but not too dangerous. Refer to Appendix (4.13) and Appendix (4.14)

**Test Data Validation**

The selected test data rows have no values on new\_y column. Running linear regression with the two models M1 and M2 gives the predicted values for the new\_y variable (Appendix 4.15 & Appendix 4.16). With the predicted and observed variables in hand, calculating the difference between them gives the accuracy/performance of the prediction here (Appendix 4.17 & Appendix 4.18)

Here are the performance statistics of M1 & M2

|  |  |  |  |
| --- | --- | --- | --- |
| **Train** |  | **Model 1** | **Model 2** |
| **RMSE** | 1.05 | 1.05 |
| **R-Square** | 0.808 | 0.807 |
| **Adj-R-Square** | 0.807 | 0.806 |
| **GOF** | (OK)F=832/<0.0001 | (OK)F=1064/<0.0001 |
| **Residuals** | OK | OK |
| **Test** | **RMSE** | 1.07739 | 1.0765 |
| **MAE** | 0.8645 | 0.8624 |
| **R-Square** | 0.805344708 | 0.80575757 |
| **Adj-R-Square** | 0.80385782 | 0.80460156 |
| **CV-R-Square** | 0.002655292 | 0.00124243 |

For the train dataset M1 and M2 performed almost similar, however the F test’s F-Value is slightly higher for M2.

For the test data CV-R2 is less than 0.3 for both models which shows both are good predictive models. When compared M2 is slightly doing better than M1 in terms of RMSE, MAE, and Adj-R2.

**Conclusion**

Based on all the above results Model 2 with seven independent variables is a better model for prediction. The linear equation of that model goes like

***Diag\_Pct = -3.66 + 0.28 Unemp\_Rate - 0.00001 Income – 0.027 Col\_Degree + 0.20 Obseity + 0.12 Ph\_Inactivity + 0.11 Age +1.23 d\_South***

The significance of the predictors goes in this order

***Obesity > Ph\_Inactivity > d\_South > Unemp\_Rate > Age > Col\_Degree > Income***

We can safely say Obesity, Leisure Time Physical Inactivity, living in the southern states significantly influences the Diagnosed Diabetes Est. Percent. This analysis somewhat arrives at similar discussion referred here at Appendix 4.19.

**Team Member:** Charles Saporito

After importing the data into SAS the variables were chosen for analysis. Our y-variable, since we wanted to know if diabetes was prevalent among counties in the United States was the percentage of those diagnosed with diabetes. Our x-variables for analysis were as follows unemployment rate,median income, education levels such as those with less than a high school diploma, those with some college, and those with Bachelor degrees. Our remaining variables were obesity prevalence percentage, inactivity percentage, and the median age.

After analysis of the data to choose variables a five number summary was ran to show an output of the mean within the percentiles among the data of those diagnosed with diabetes(figure 5.1). In the 25th percentile 9.6% carry a diagnosis of diabetes the median of those diagnosed is 11.3% and the 75th percentile have a estimated 13% diagnosis percentage. The upper quartile range carries the highest percent of those diagnosed with diabetes. While the lower percentile does not carry as hgh of prevalence of the disease.

When continuing the exploratory analysis the y-variable appeared normally distributed when creating histograms for the variables(figure 5.2). The x-variables were originally skewed. SAS output showed biases in the x-variables. I felt it best to do a log transformation on the x-variables for the rest of the analysis. This solved the issue of biases and normally distributed the x-variables. To check for multicollinearity the VIF was analyzed. None existed(Figure 5.3)as all VIFS were <10. Residual plots were created to check for distribution of errors and normality. The residual plot among the predictors showed there was no violation of assumptions and constant variance existed. This meant that these variables were good for predicting our y-variable. The normality plot showed a normal distribution of errors and not violating normality(Figure 5.4)

WIth all variables included in the data for exploratory analysis the model produced a r-square of 73.04% variation of the variables on diabetes prevalence and a adj-r-sq of .7296 suggest this model explains the variation among the variables decently. There was a goodness of fit of 941.7 suggesting a good model for analyzing diabetes prevalence. The correlation matrix showed moderately strong associations between the Y-variable and the x-variables. This can all be seen in figure 5.5 of the appendix.

Outliers and influential points were determined of the 3,139 observations in the sample. The Cooks D graph showed outliers and influential points at 5 points within the data. These observations were removed and the analysis was ran again. A improvement in goodness of fit to 948.63, r-sq to 73.20% and adj-r-sq to .7312 resulted from the removal of outliers(Figure 5.6).

After analyzing and exploring the original data set, time came for analyzing test and train data and running predictions. It was determined at the time of setting up the data that the strongest predictors were obesity prevalence percentage and leisure time inactivity. Seeing this was quite interesting, it was determined to analyze their interaction on diabetes prevalence as these two predictors accounted for a high rate of diagnosis of the disease.The interaction terms resulted in p<.05 with no multicollinearity. It proved to be true that these two predictors had an interaction on determining a diagnosis of diabetes leading to higher rates of prevalence(figure 5.7). Model selection was set up in test and train data sets. The data sampling size was set at .75 with a selection probability of .75. Test and train used 2352 observations. Stepwise and CP methods ran on the test data fitted a model with r-sq of 74.74% and C(p) of 8.1129. It was determined that the best model was resulted from the stepwise procedure and did not include the education variable of those with a bachelor degree. The model was then validated and produced a f-value of 948.63 with a r- square of 73.20% and adj-r-sq of .7312 with a root mse of 1.29925. This model can explain the variation among the variables well in the test data to predict which counties have high diabetes prevalence rates and which do not. Furthermore adj-r-sq explains this variation well with a low error rate in the model, and a high goodness of fit making the model a strong predictor of the y variable among the test data and the final model selected. All outputs from test and train data and selection methods can be seen in figure 5.8 of the appendix.

The findings of this analysis concluded that diabetes prevalence,type 2 can be determined by socioeconomic factors such as employment and education.It is uncertain what exactly causes the mechanisms that start the process of diabetes, but a strong correlation is attributed to being overweight, the environment, and genetics (Mayo Clinic 2018). This model can prove the methods that such as institutions like Mayo clinic can use in their method for determining its prevalence. See appendix 5.9 for reference to the Mayo Clinic site.

***Future Work***

The prevalence of diabetes in united states has increased considerably and this increase could be explained by many factors including some individual-level and environmental risk factors as the evolution of the disease and global landscape changes. The dataset which is being used here currently shows 71% of the variation in diabetes prevalence between counties is explained by the rest of the predictors. To Enhance the diabetes prevalence from 71% to 90% (or above) can be achieved by including a couple of more factors like Gender, Hypertension percentage, prediabetes, smoking habits and the alcohol consumption rate.This addition of predictors could potentially make a better model.

**Appendices**

1. Maxwell Carduner

Exhibit 1.1:

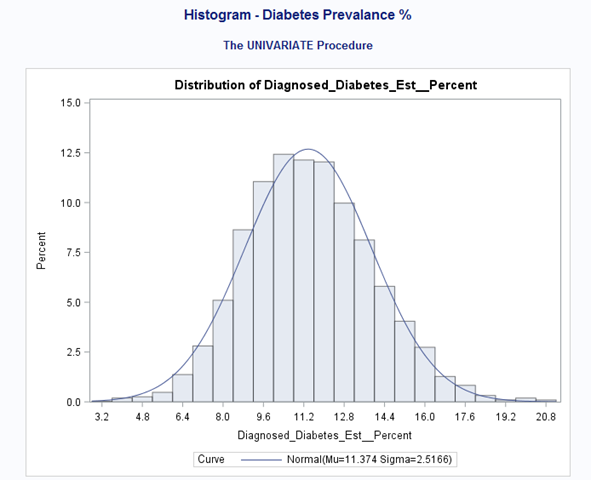
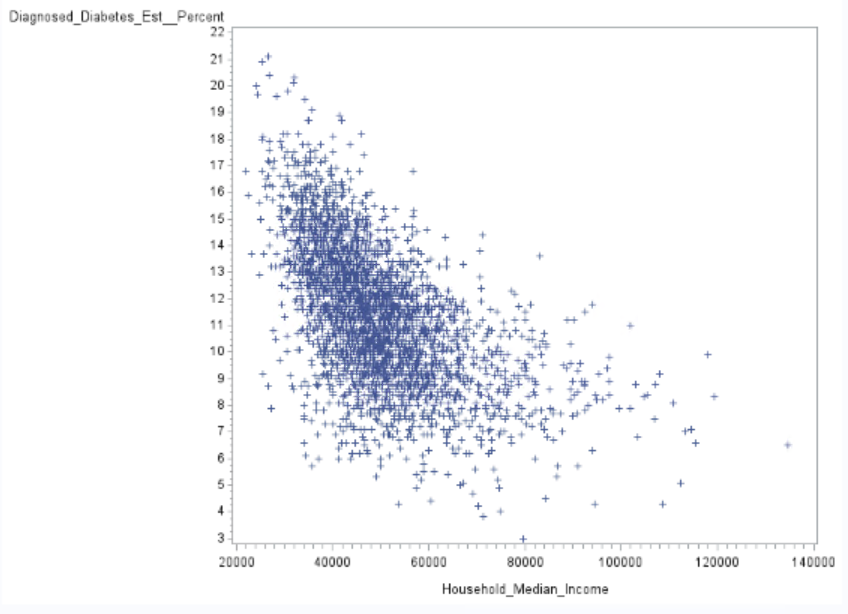
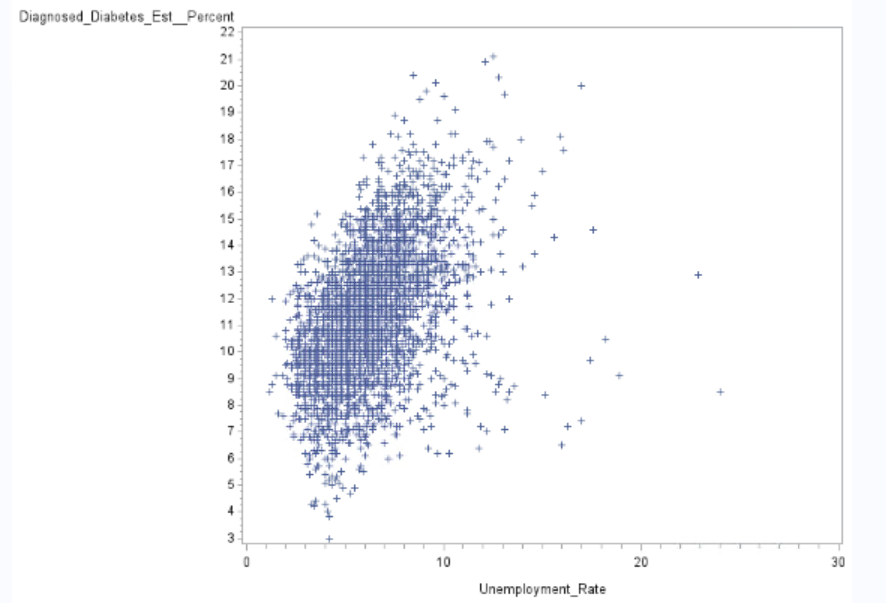
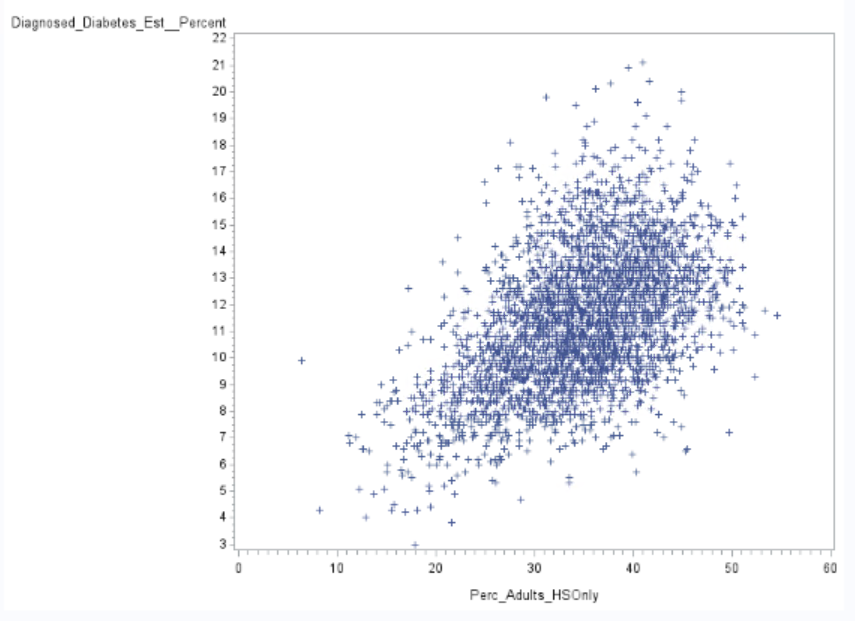
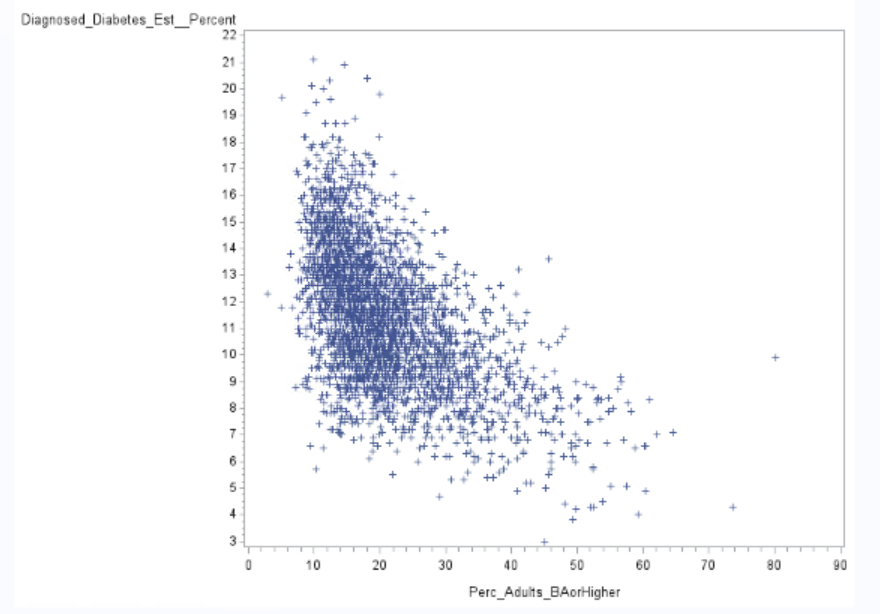
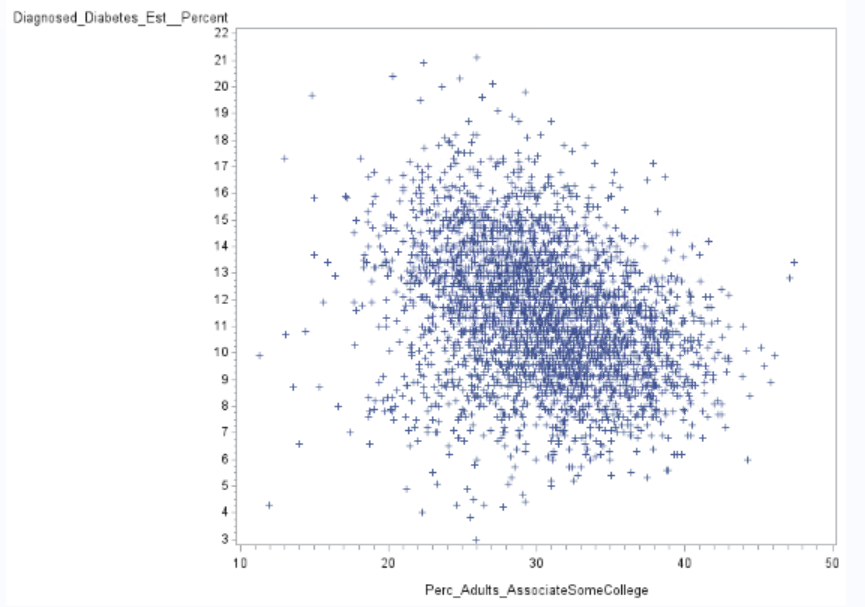
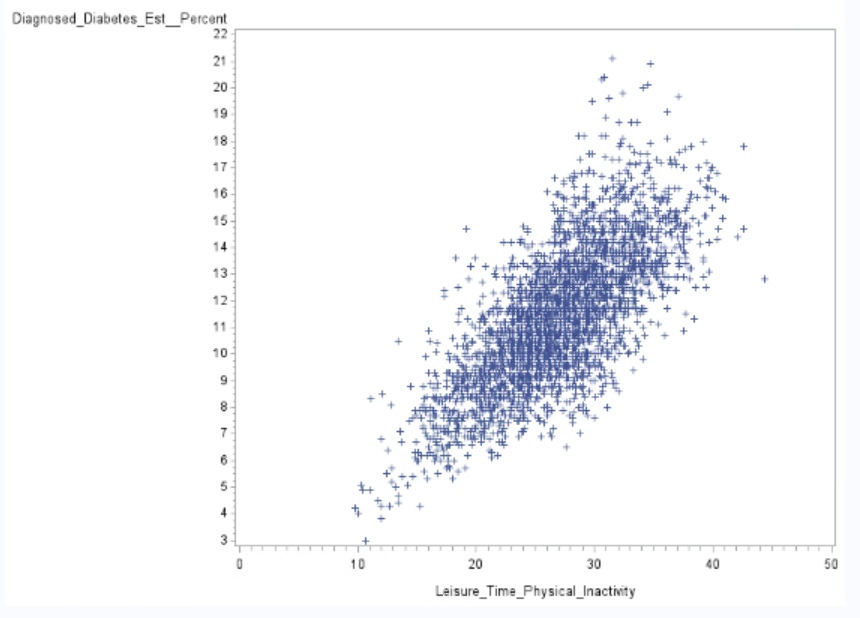
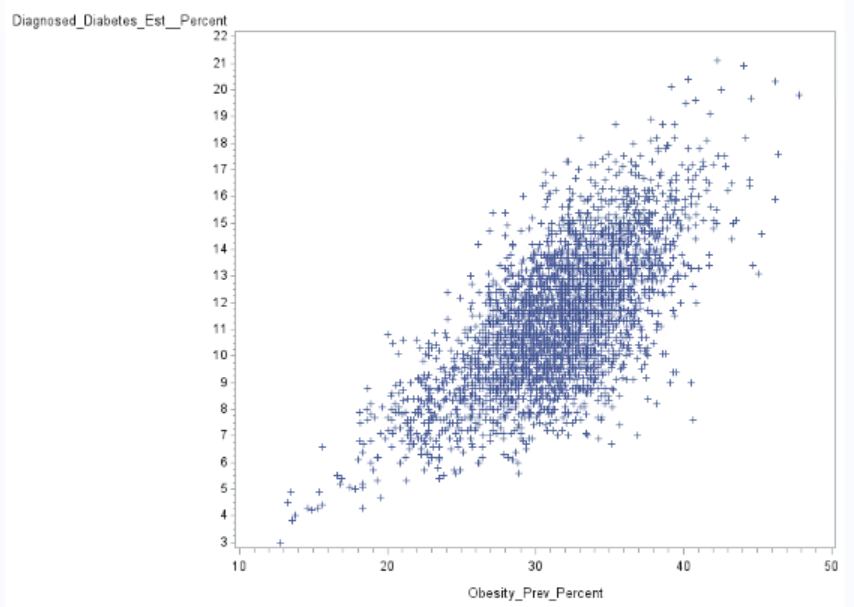


Exhibit 1.2 Scatterplots:









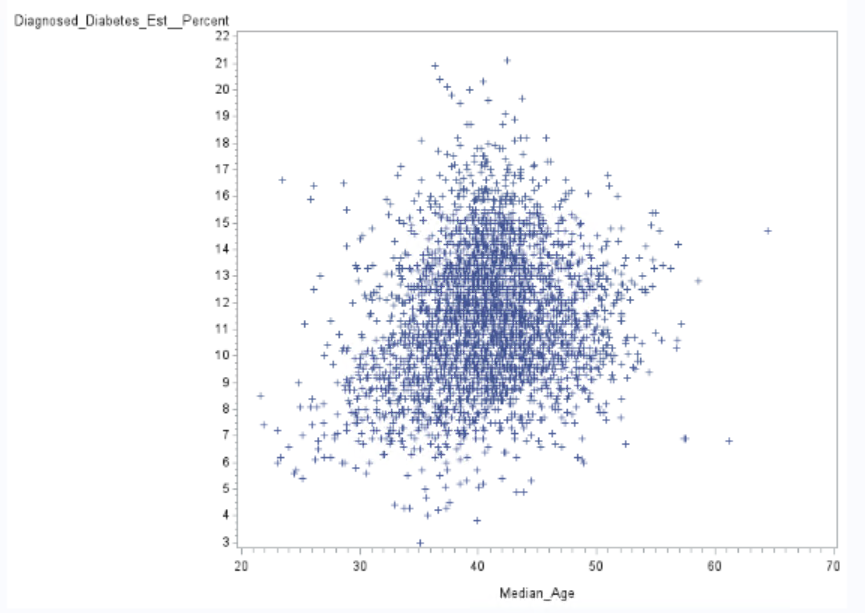


Exhibit 1.3 VIFs:



Exhibit 1.4

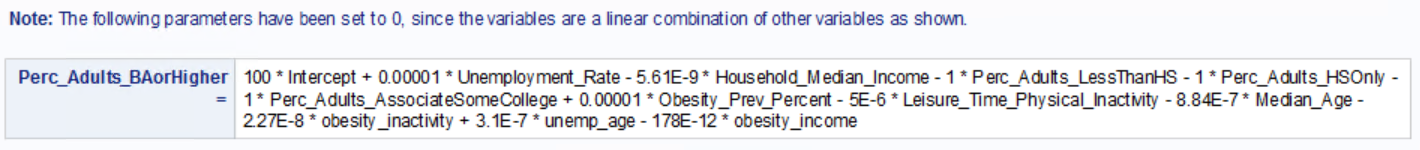


Exhibit 1.5

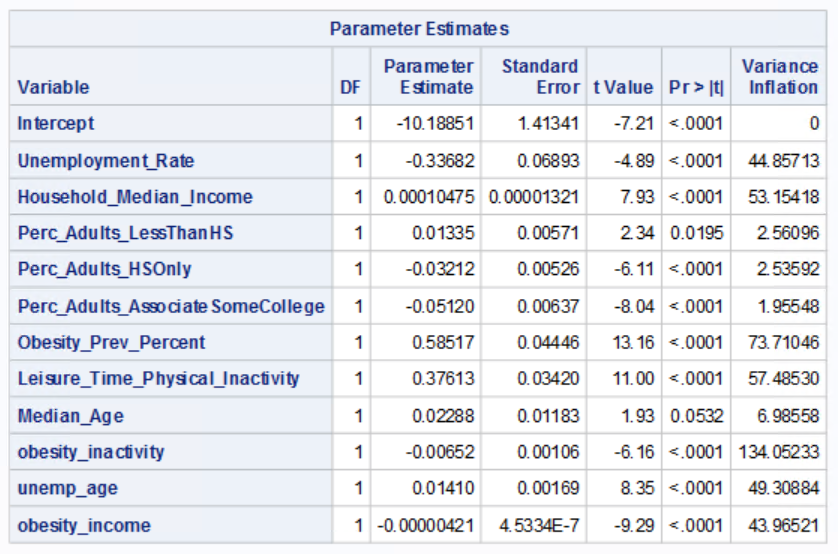


Exhibit 1.6

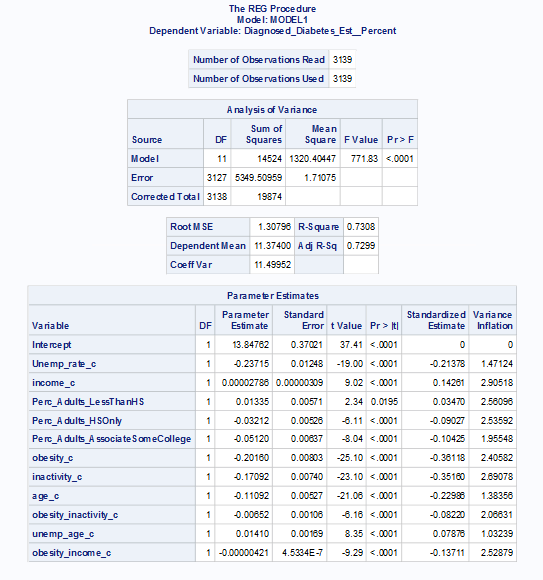


Exhibit 1.7:

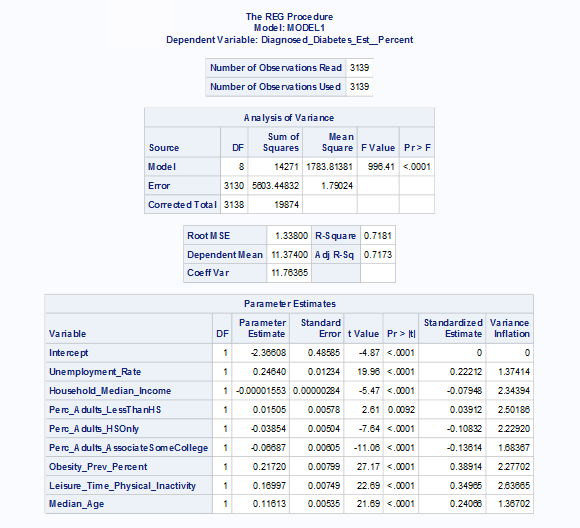
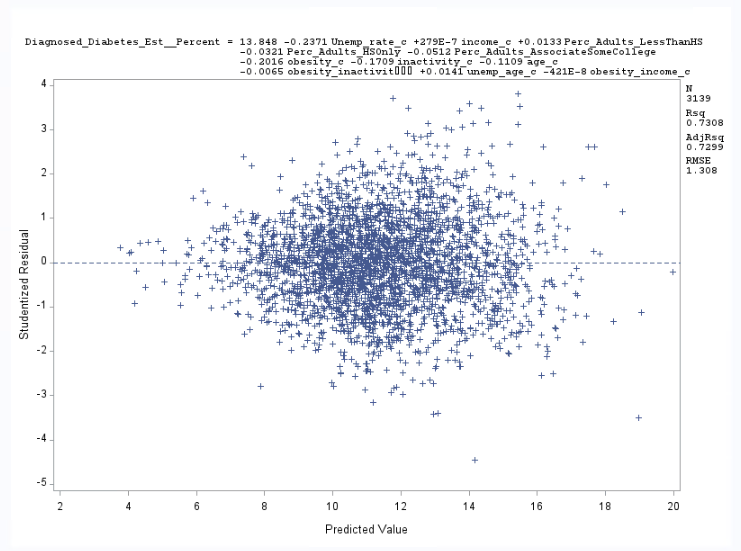
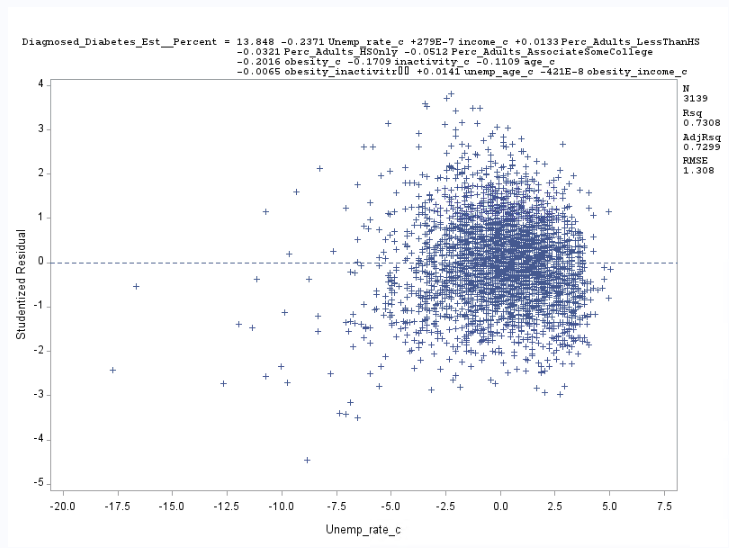
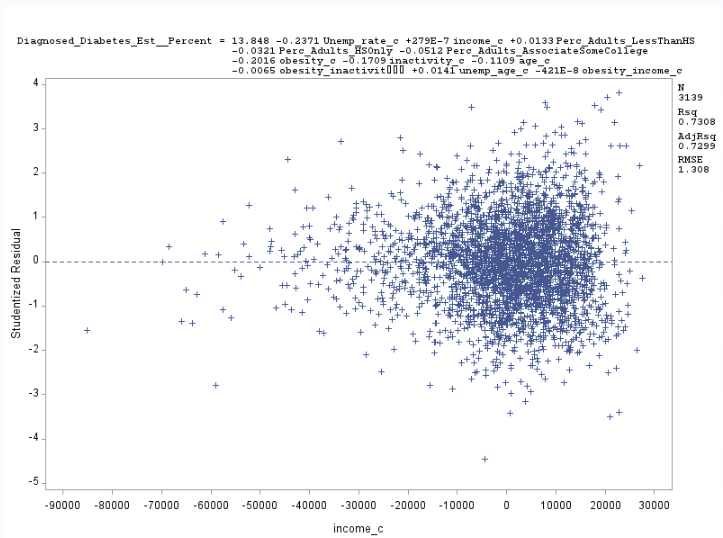
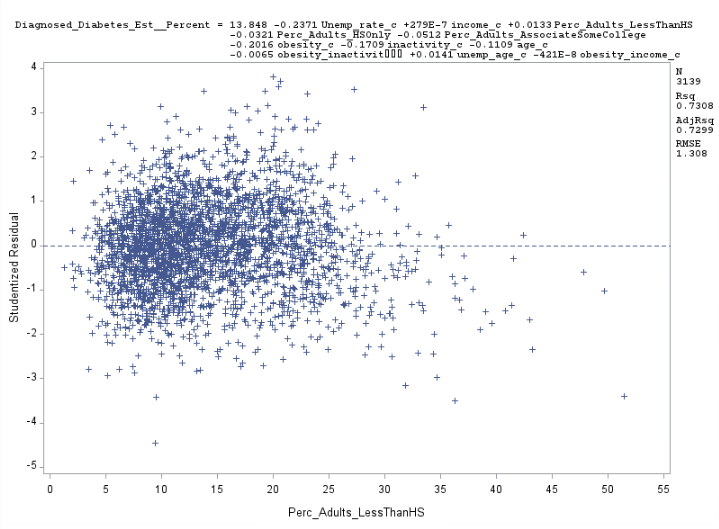


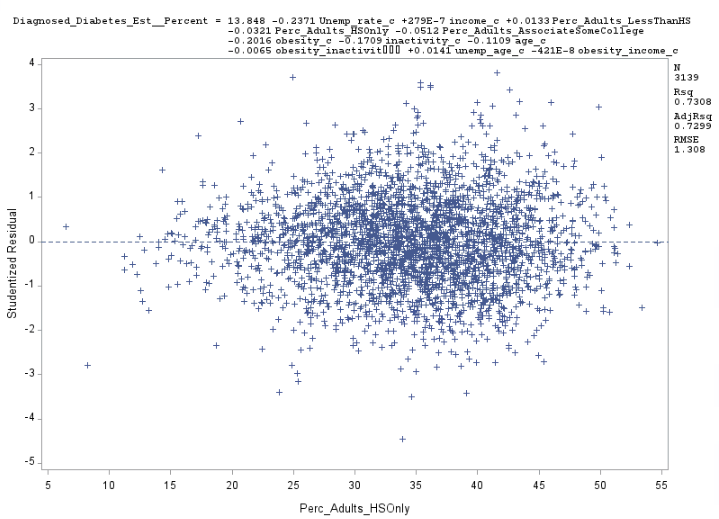
Exhibit 1.8



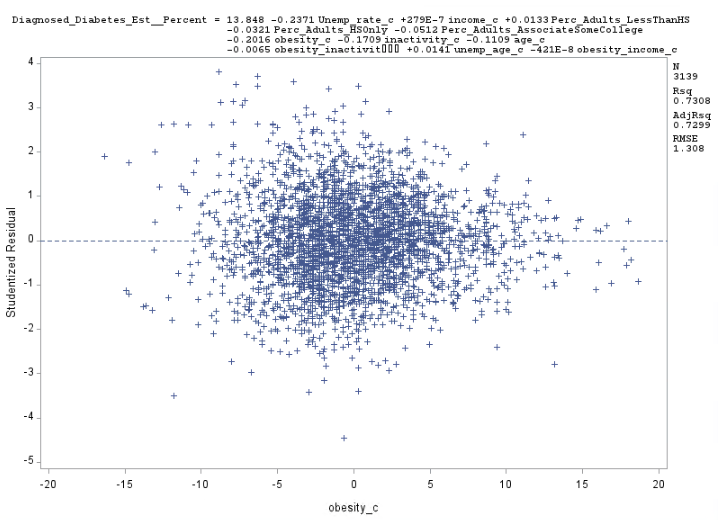


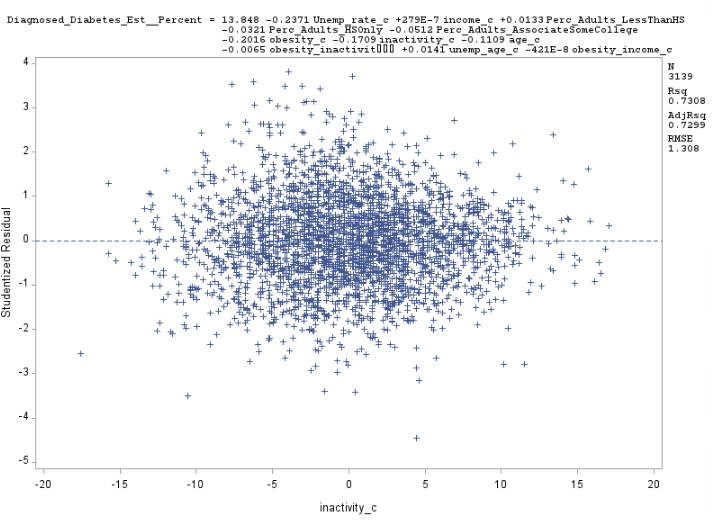


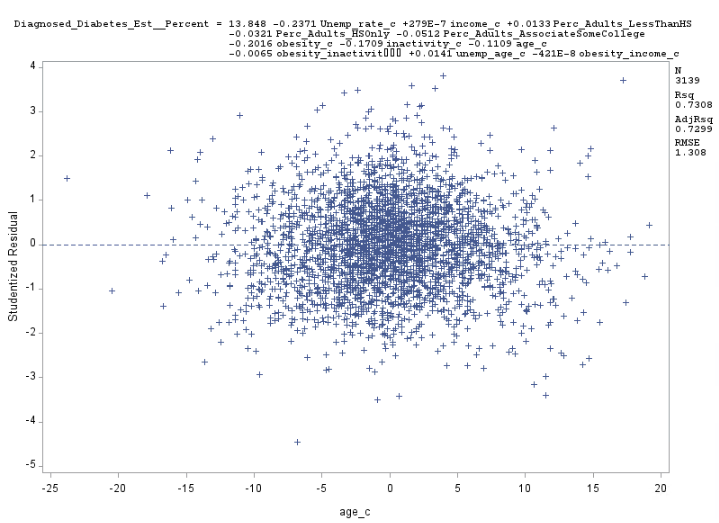


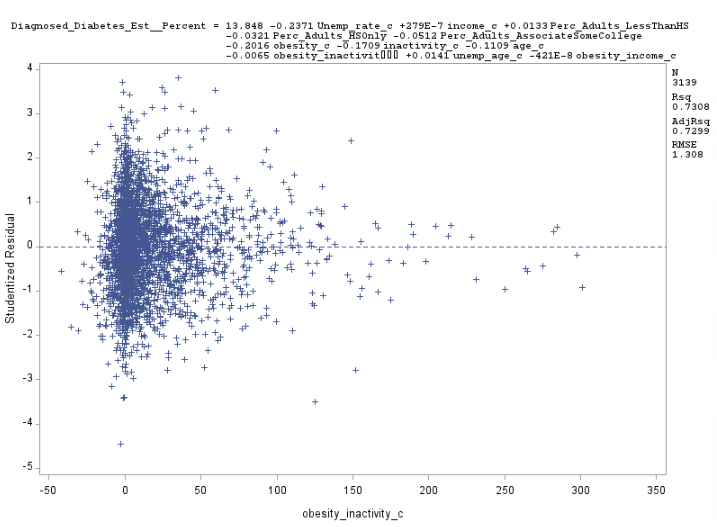


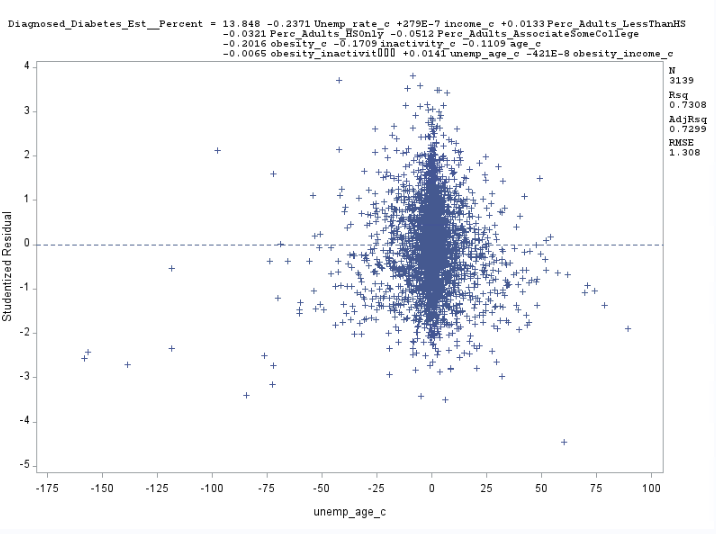












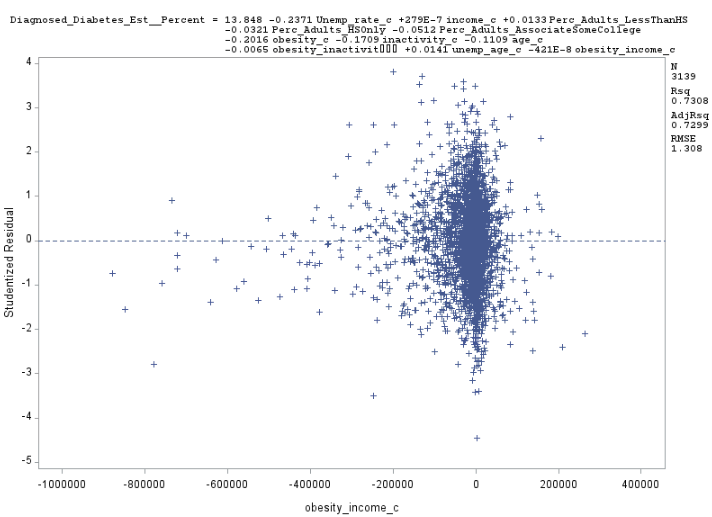


Exhibit 1.9:

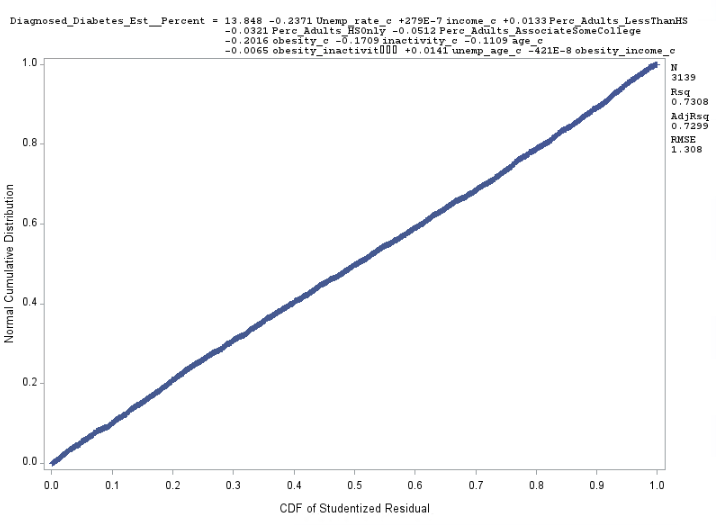


Exhibit 1.10:

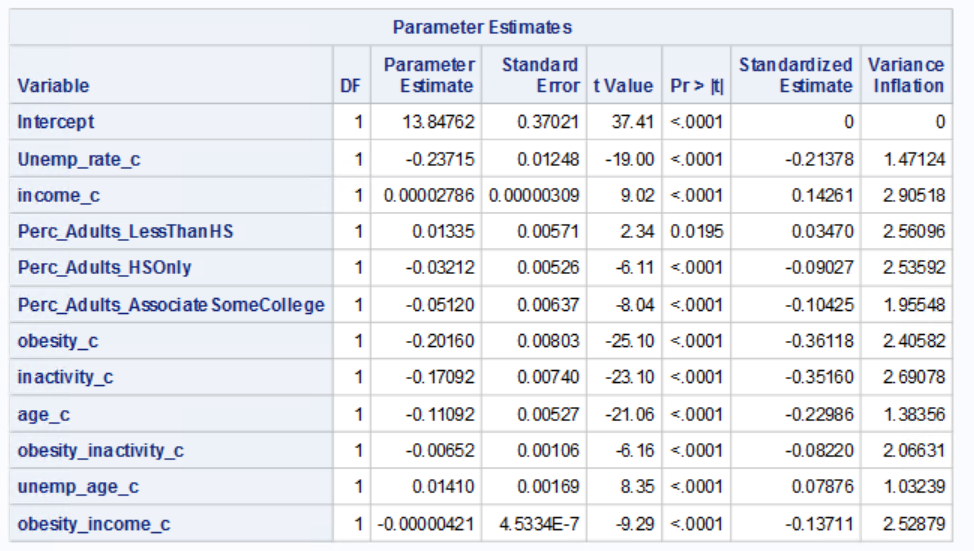
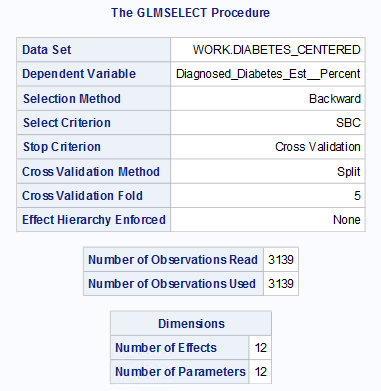
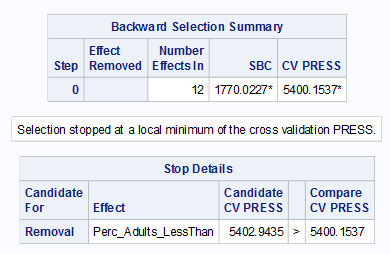
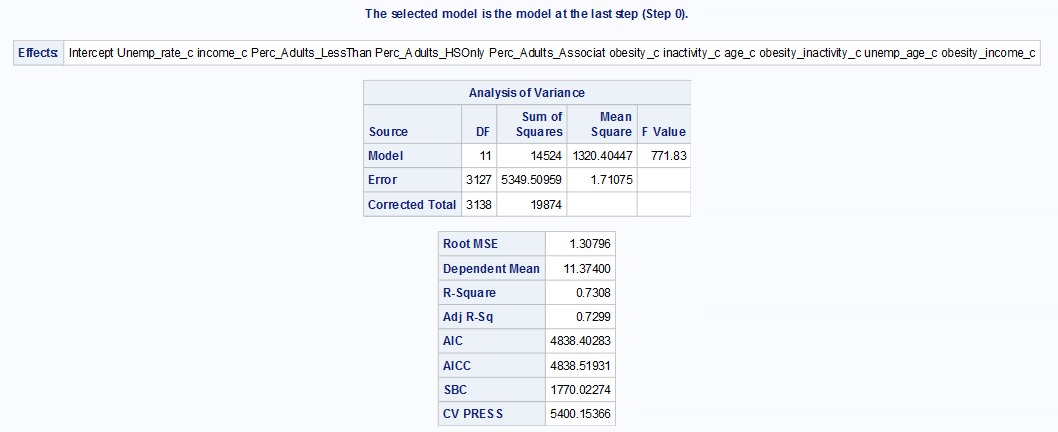


Exhibit 1.11:







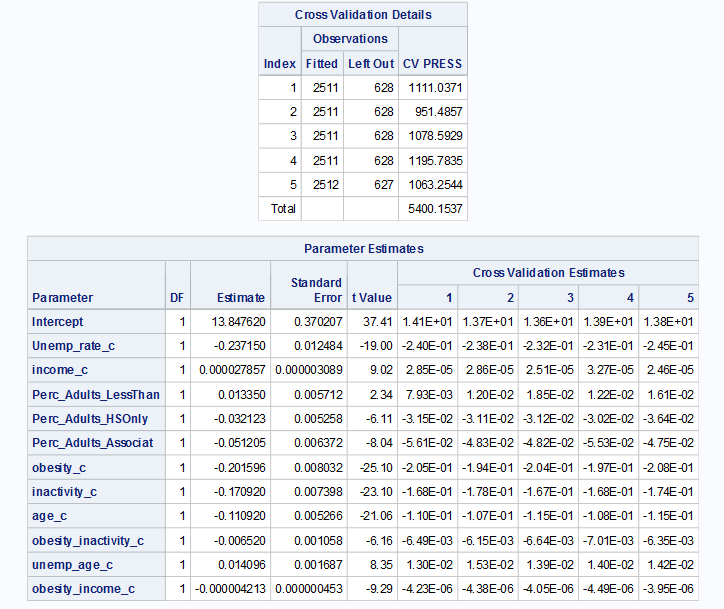


Exhibit 1.12

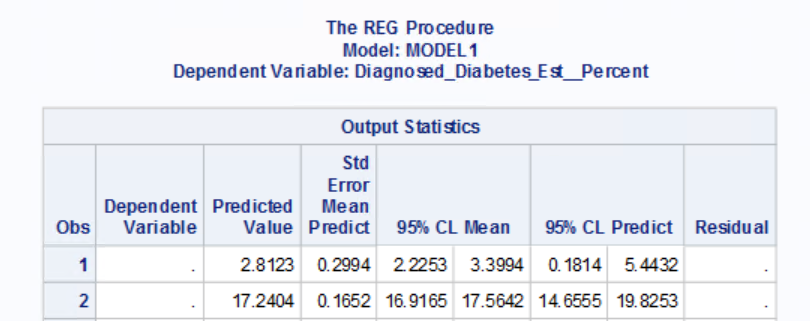
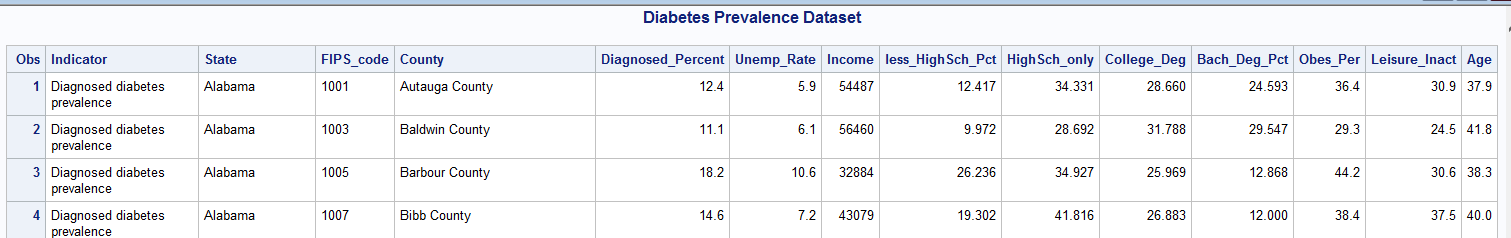


Exhibit 1.13 References Used:

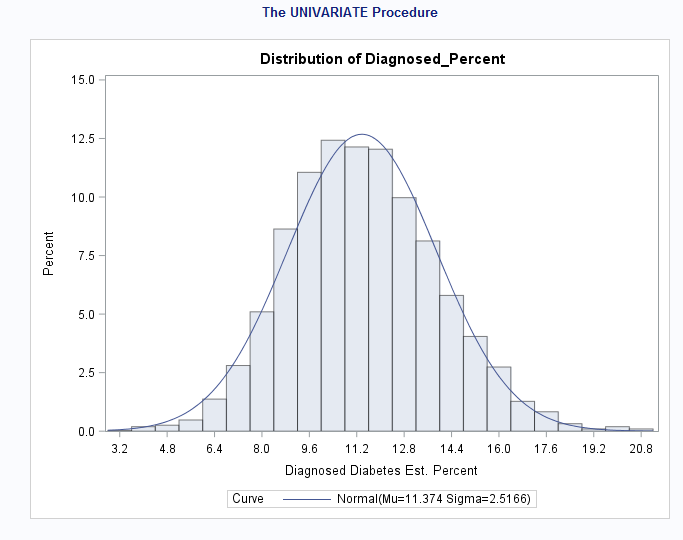
Statistics About Diabetes. 2018. <http://www.diabetes.org/diabetes-basics/statistics/> Accessed May 13, 2018.

1. Kalaivani Chandramohan

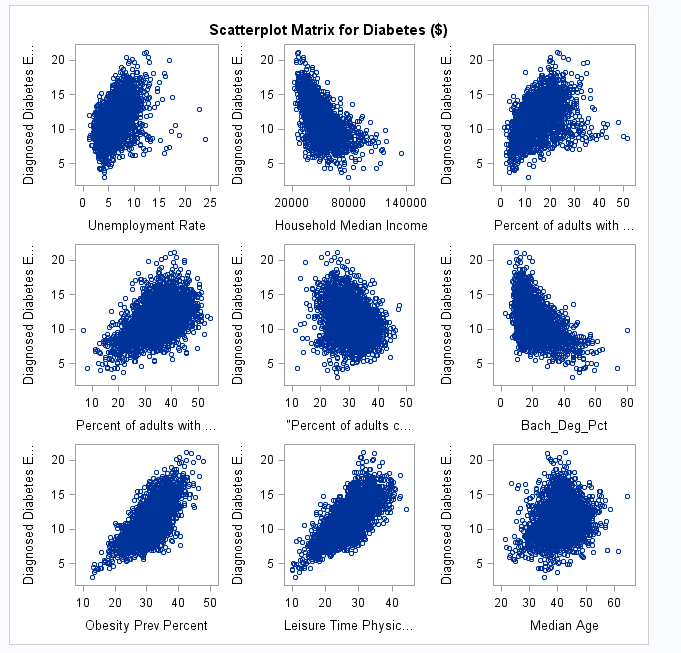
2.1: Imported Data



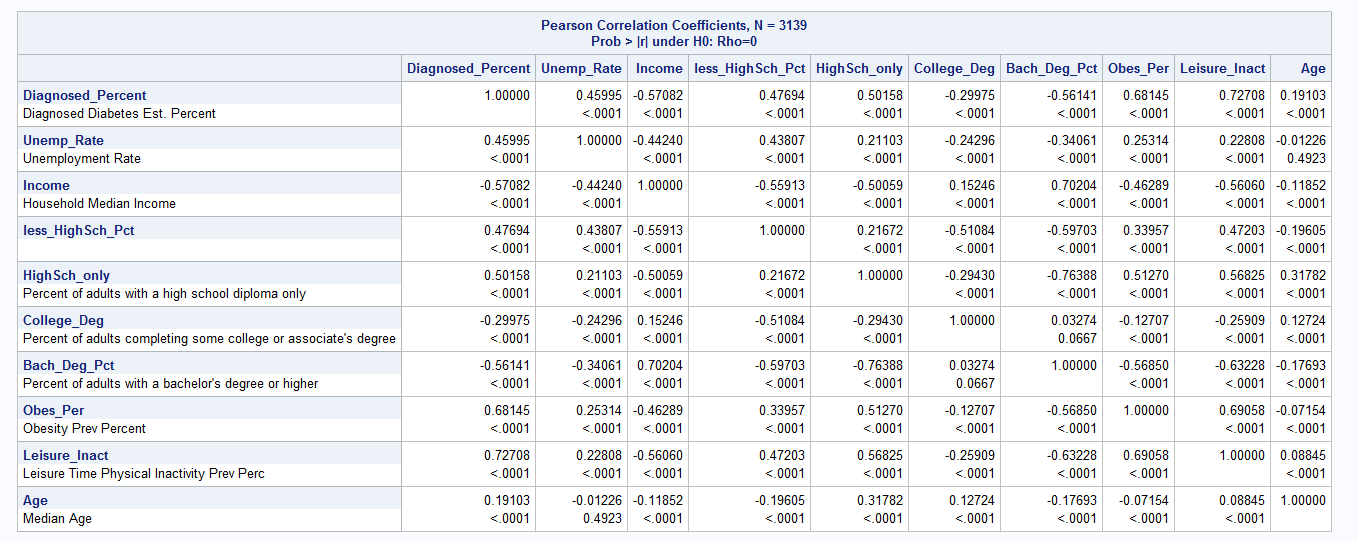
2.2



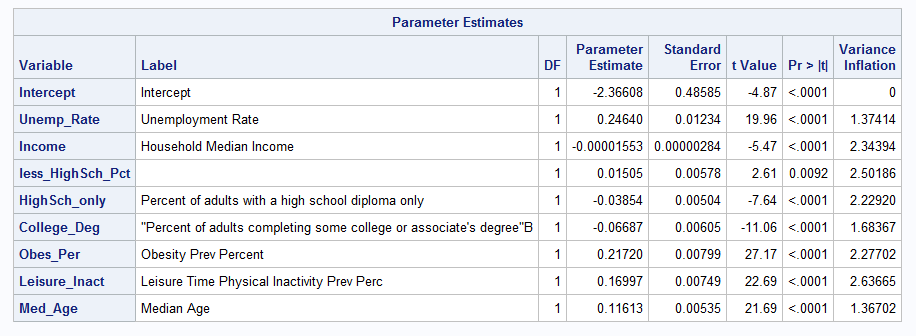
2.3 scatter plot



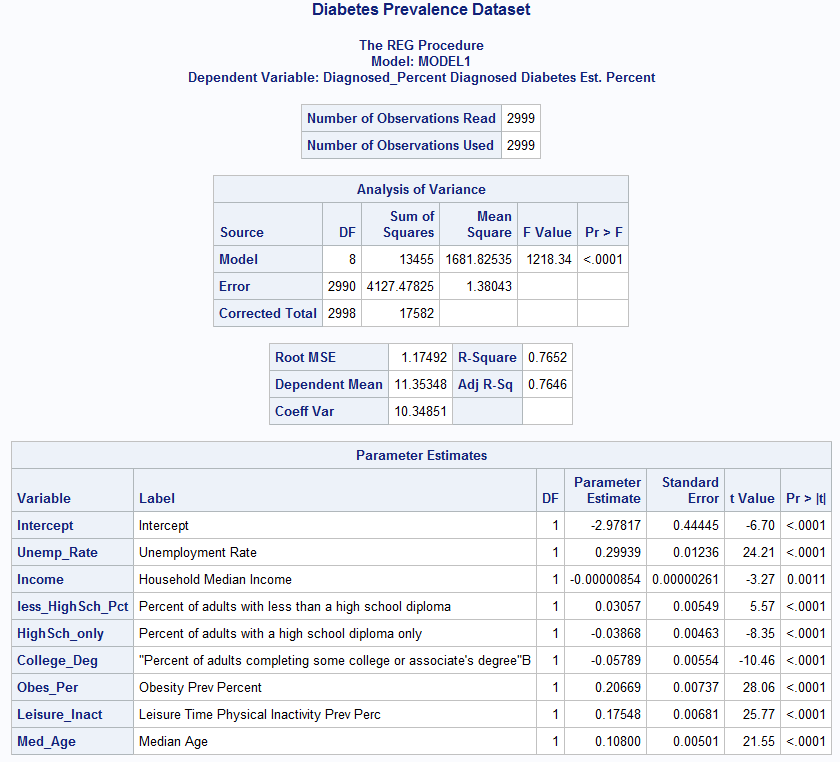
2.4 Correlation



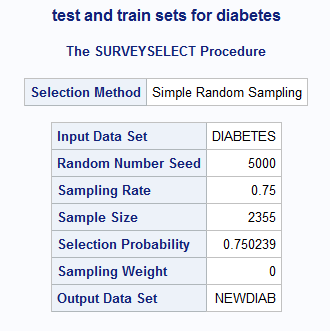
2.5 VIF

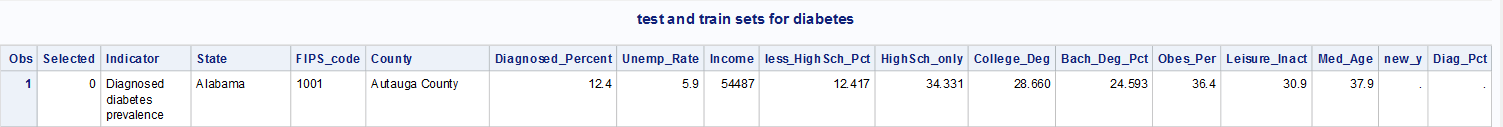


2.6 Final model

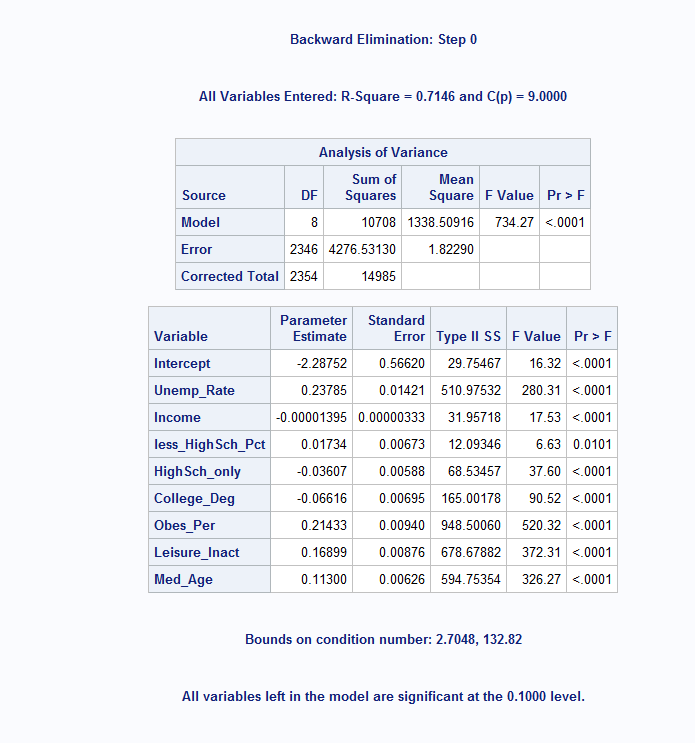


2.7

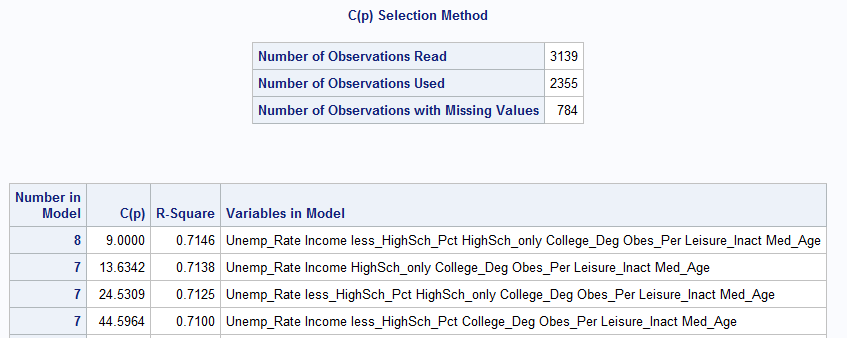




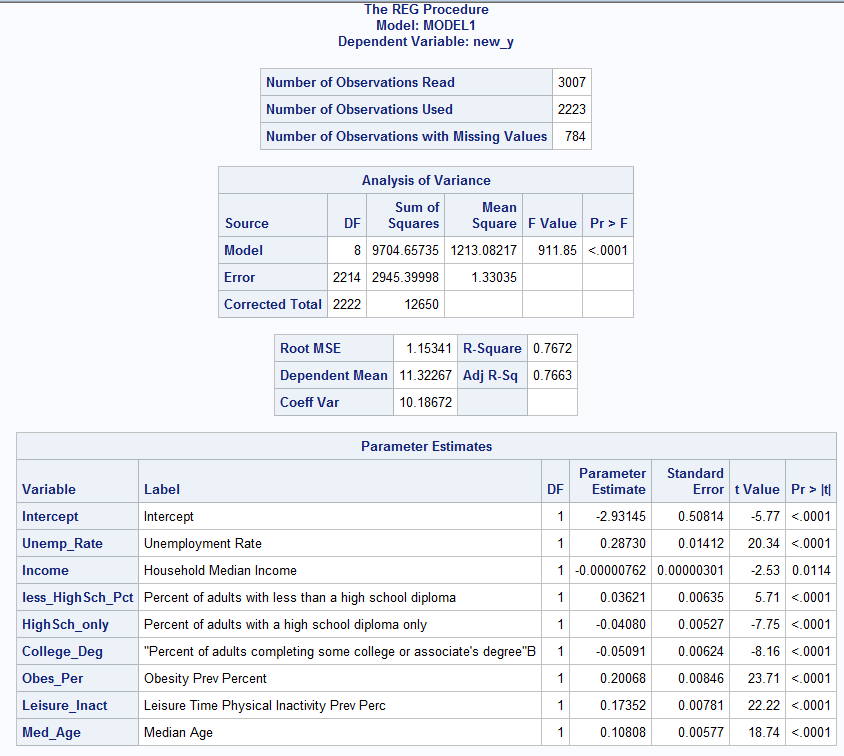
2.8 selection Method



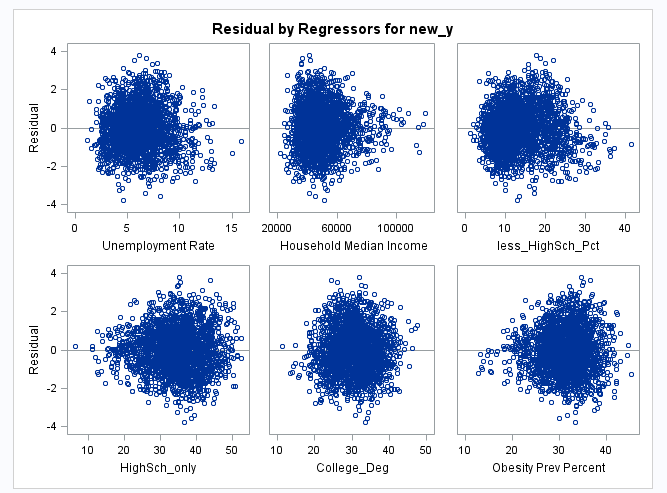
2.9



2.10 Regression analysis

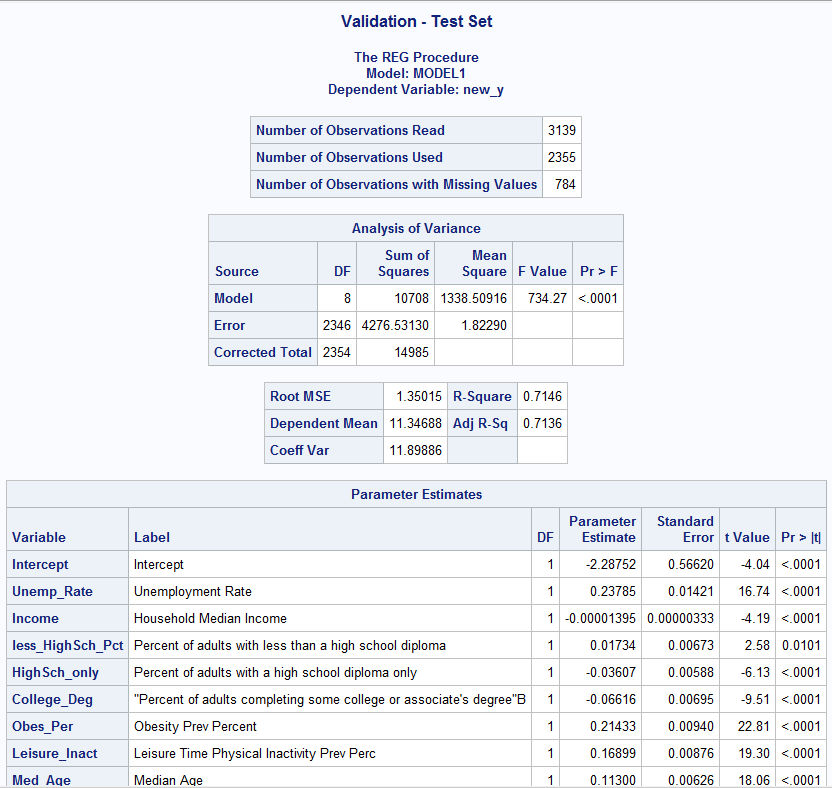


2.11

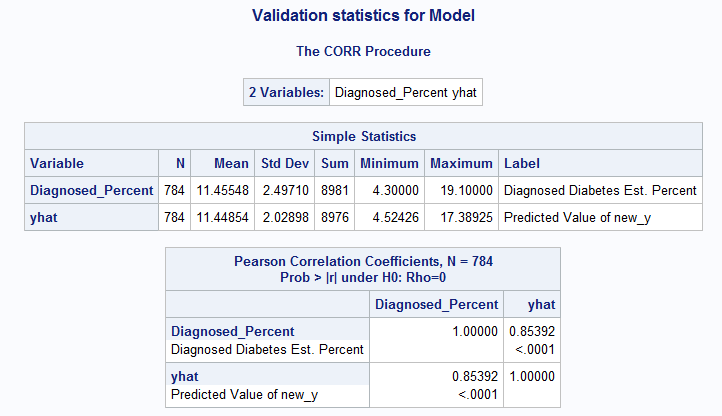




2.12



2.13

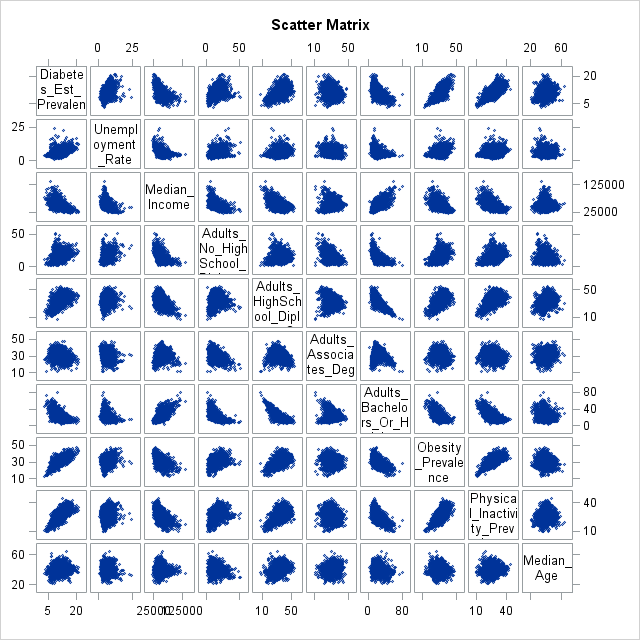


2.14 Reference Used

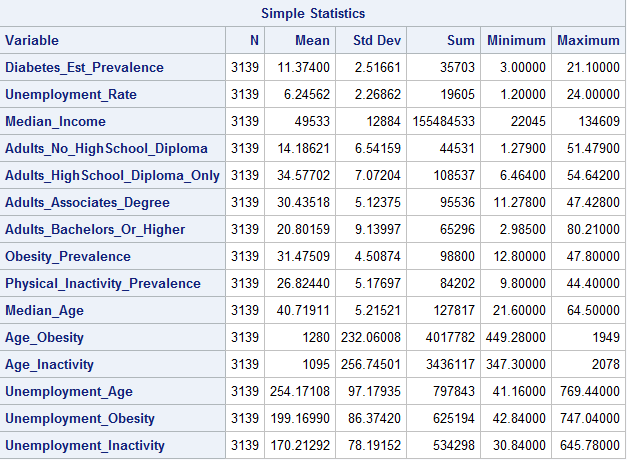
Yelena Bird, Mark Lemstra, Marla Rogers and John Moraros. October 2015.The relationship between socioeconomic status/income and prevalence of diabetes and associatedconditions:<https://equityhealthj.biomedcentral.com/articles/10.1186/s12939-015-02370>. Accesses October 12, 2015

1. William Chirciu

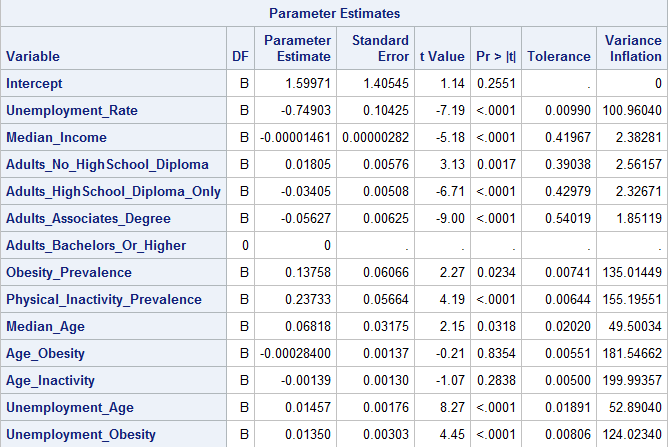
3.1: Scatterplots



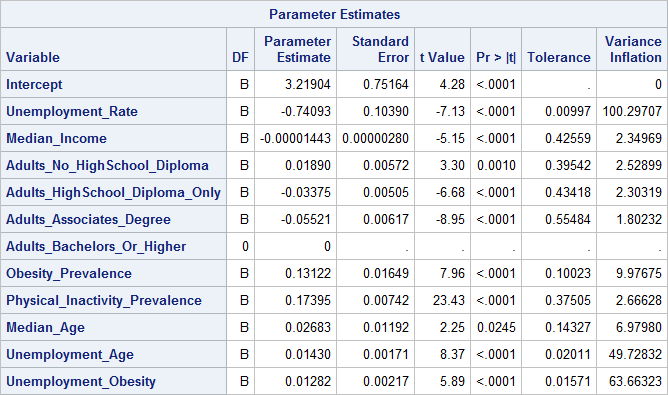
3.2: Descriptives



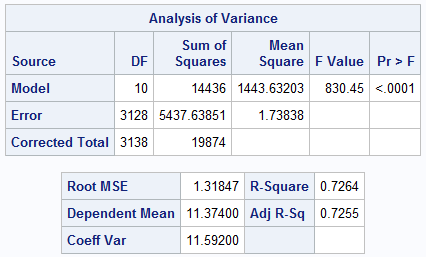
3.3: Regression Results All Variables + Interaction Terms (Including Insignificant Variables)

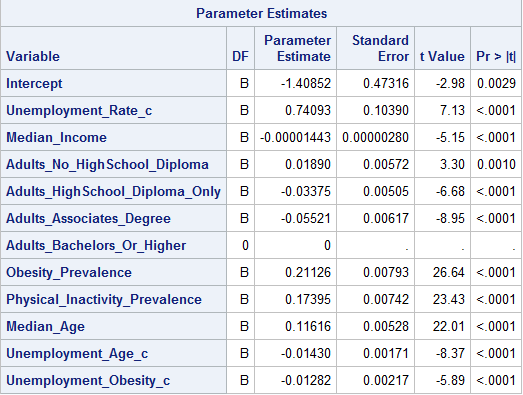


3.4:Regression Results All Variables + Interaction Terms



3.5: Regression Results All Variables + Interaction Terms (Post-Centering)





3.6: Heavy Outliers/Influential Points Prior to Train/Test Split







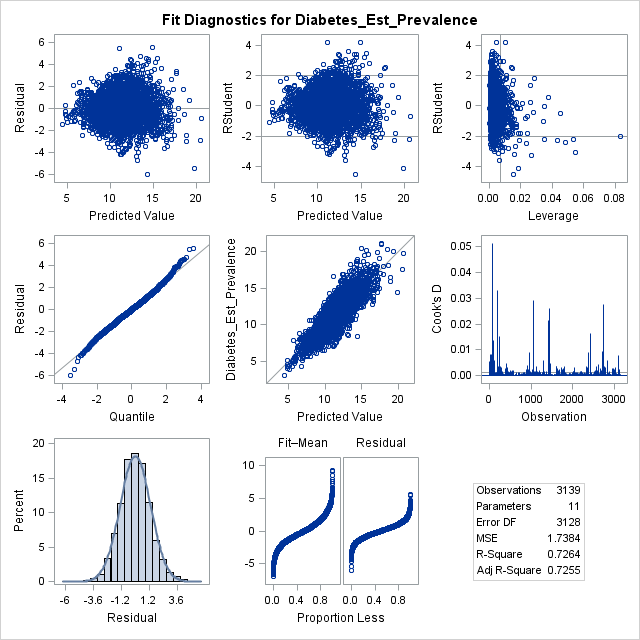


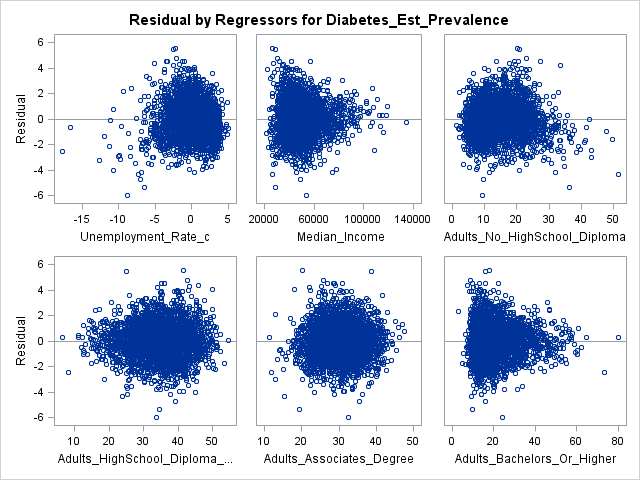


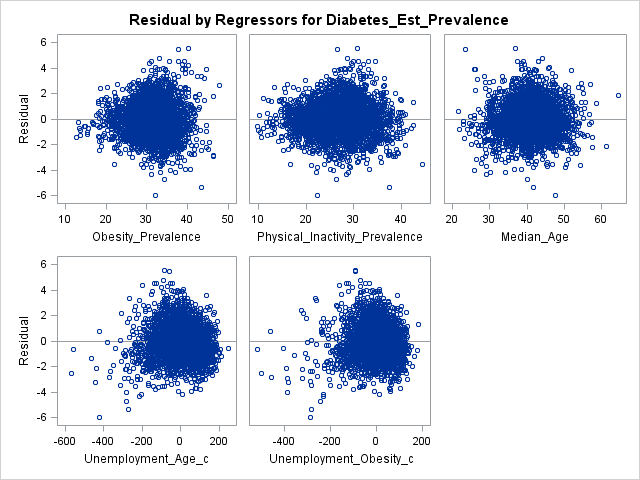




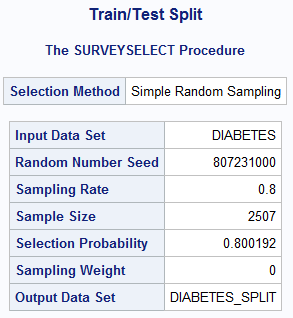
3.7: Residual Plots and Diagnostics Prior to Model Selection



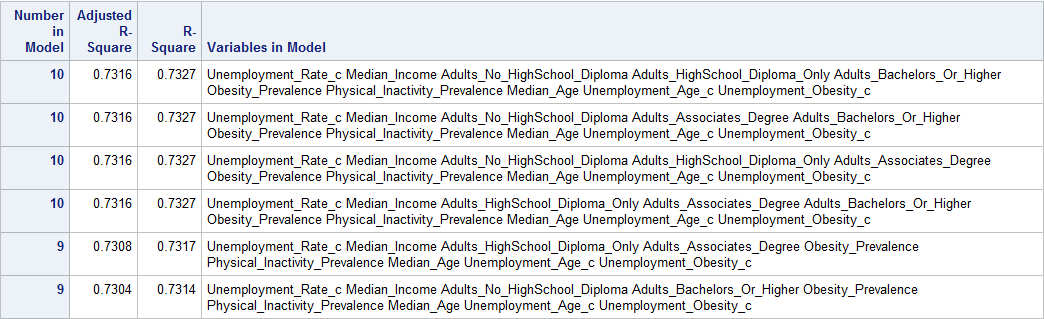




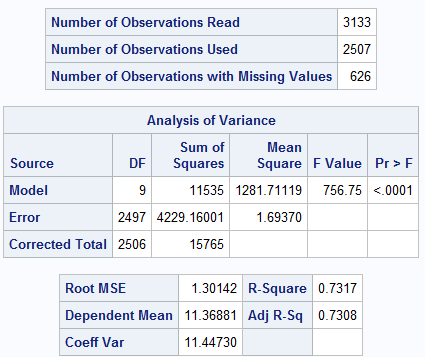
3.8 Training and Test Split

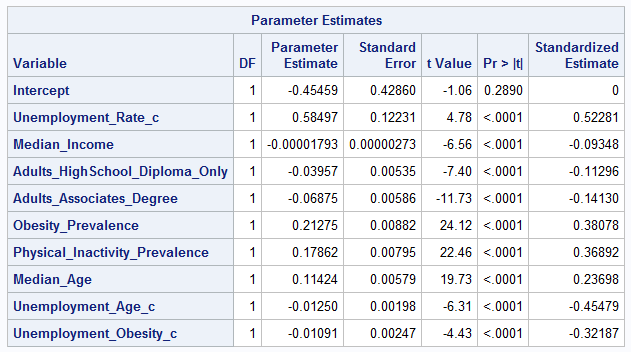


3.9 Adjusted R-Square Model Selection Top Results

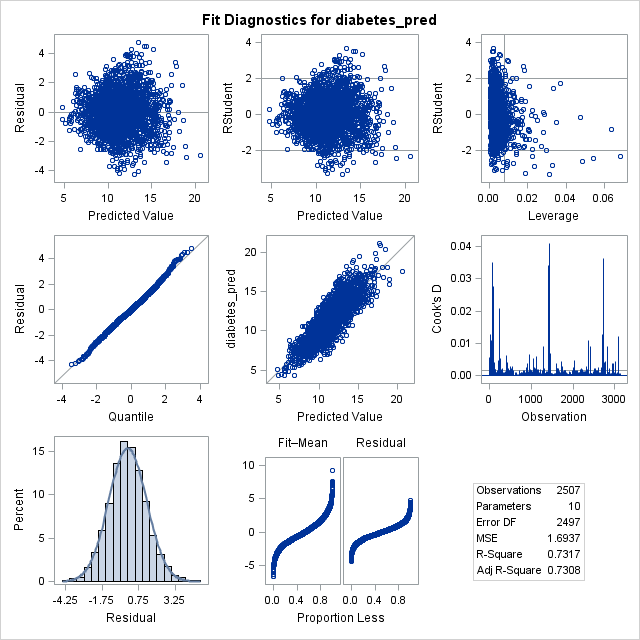


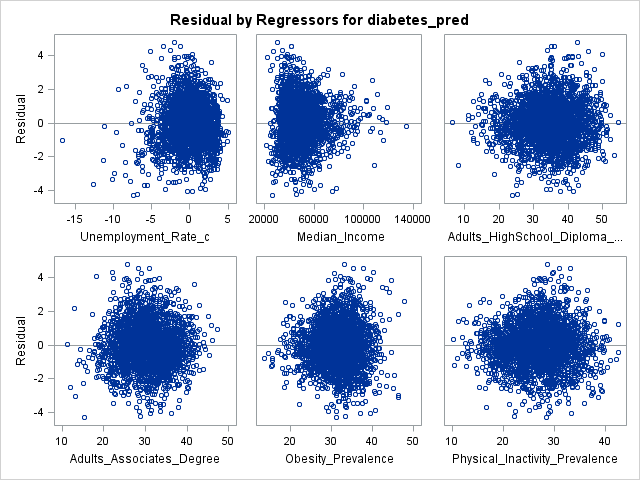
3.10: Final Model Regression Results

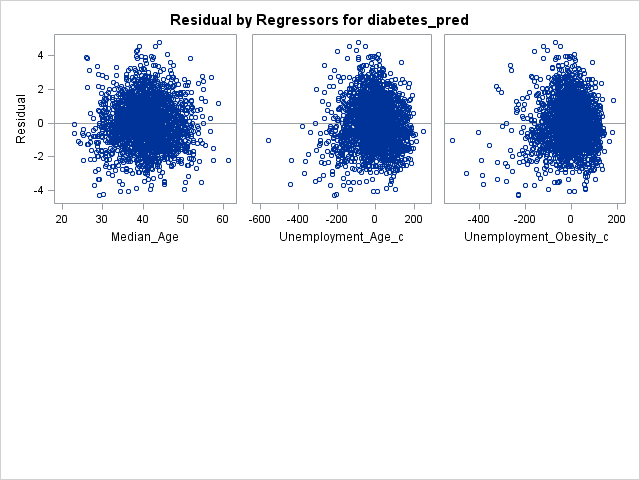




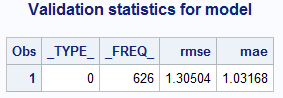
3.11: Final Model Diagnostics/Residuals



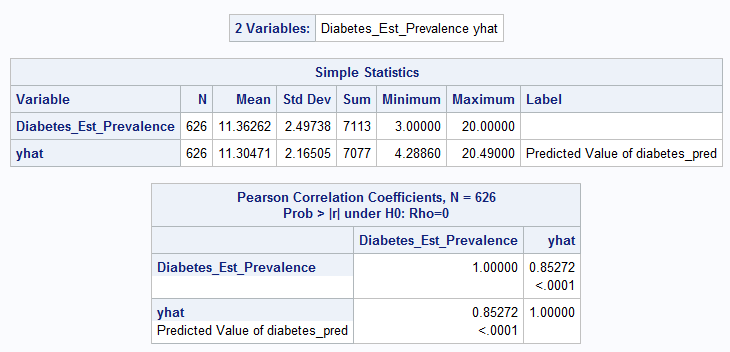




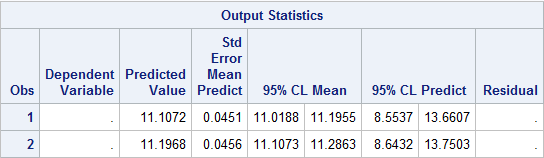
3.12: Final Model Error



3.13: Test Set R-Square



3.14: Two Predictions Results

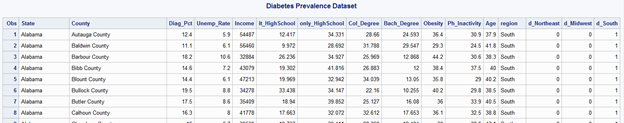


3.15 References

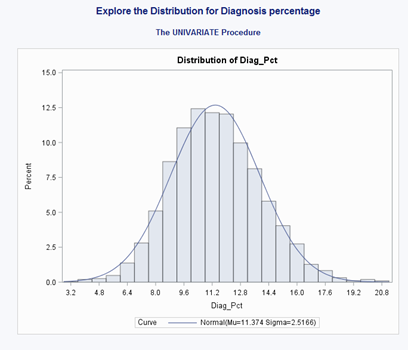
Hipp JA, Chalise N. Spatial Analysis and Correlates of County-Level Diabetes Prevalence, 2009–2010. Prev Chronic Dis 2015;12:140404. DOI: <http://dx.doi.org/10.5888/pcd12.140404>.

1. Ramkumar Perumal
2. Outputs

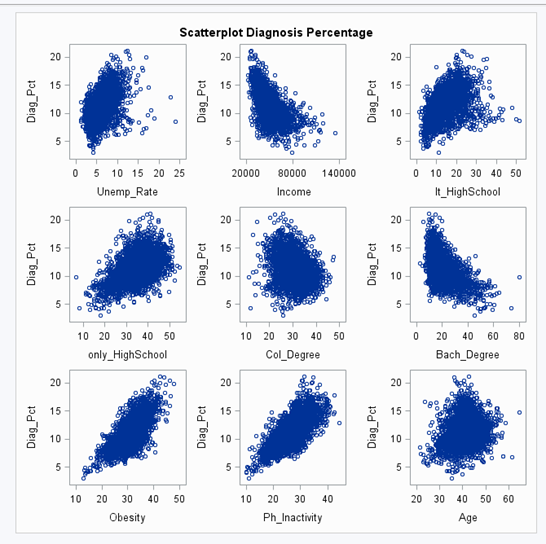
4.1



4.2



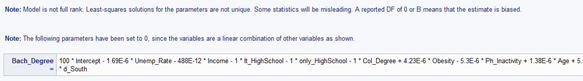
4.3



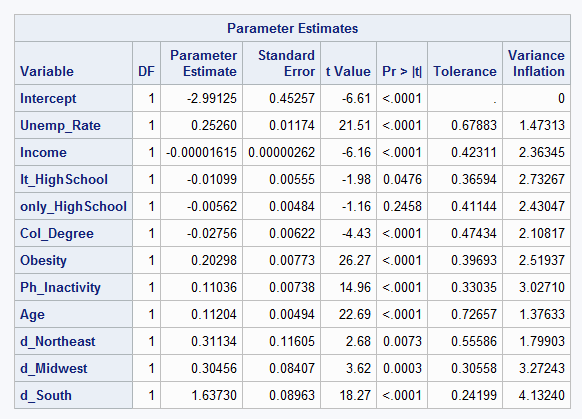
4.4



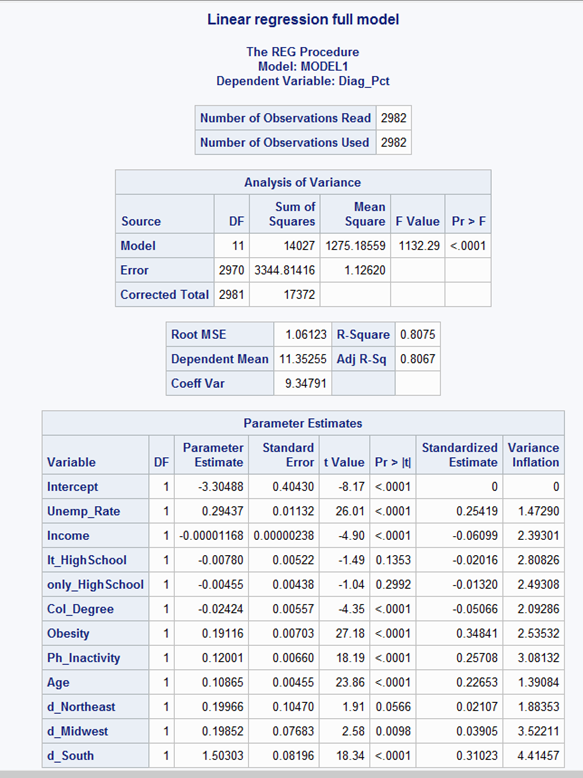
4.5



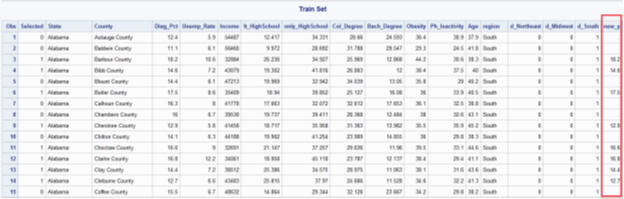
4.6



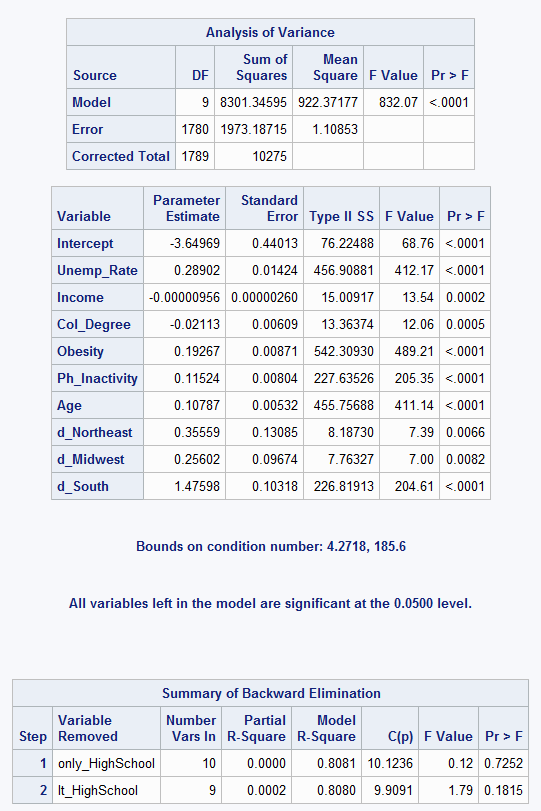
4.7



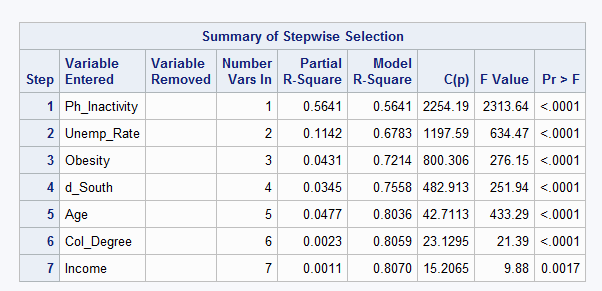
4.8



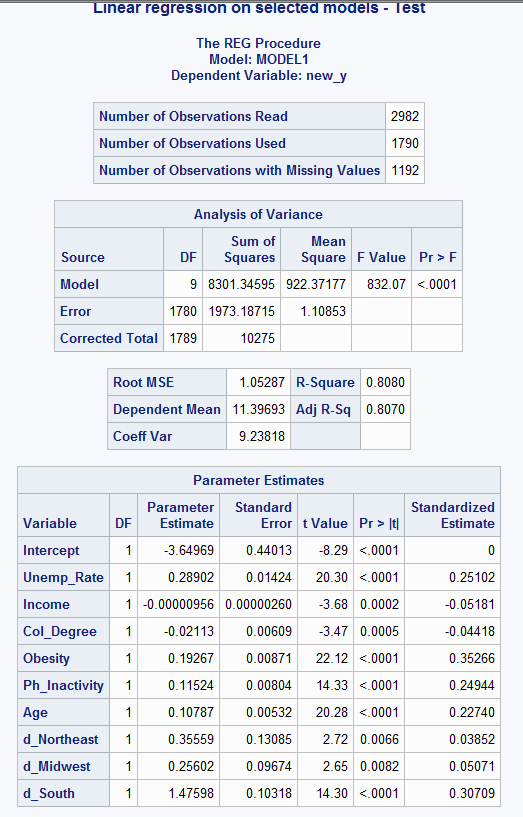
4.9



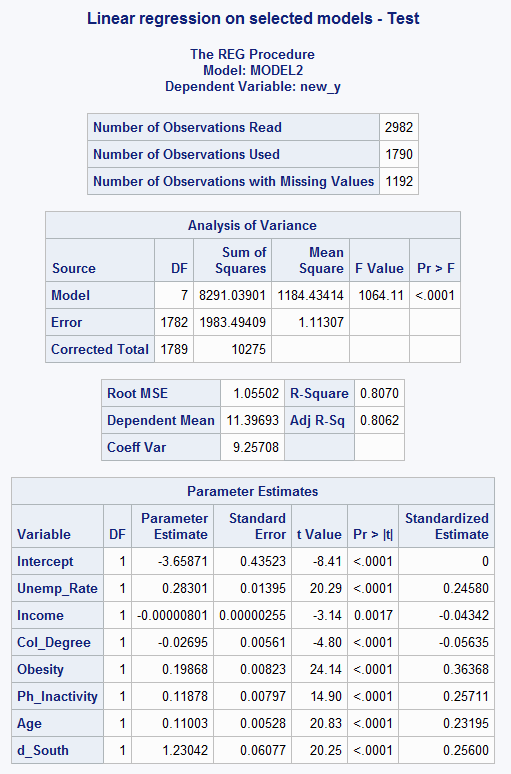
4.10



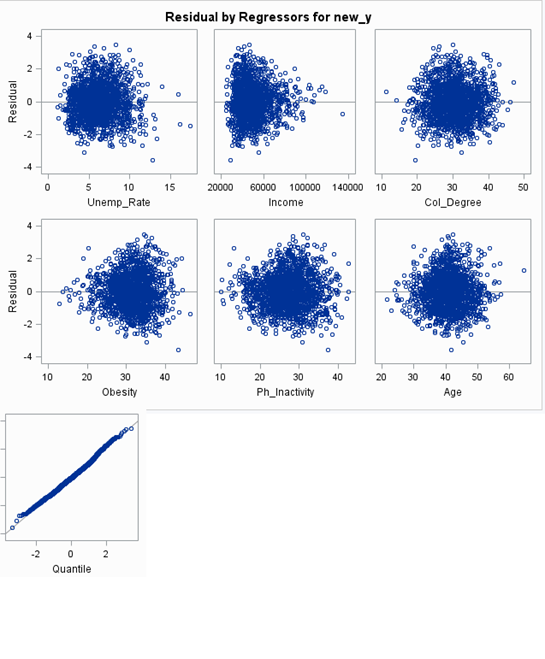
4.11



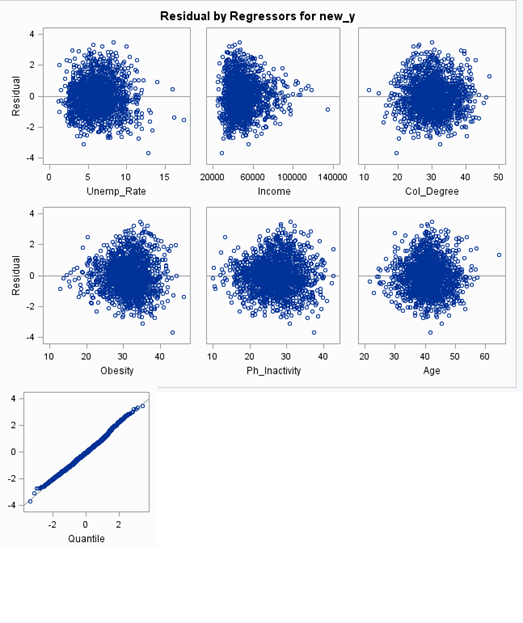
4.12



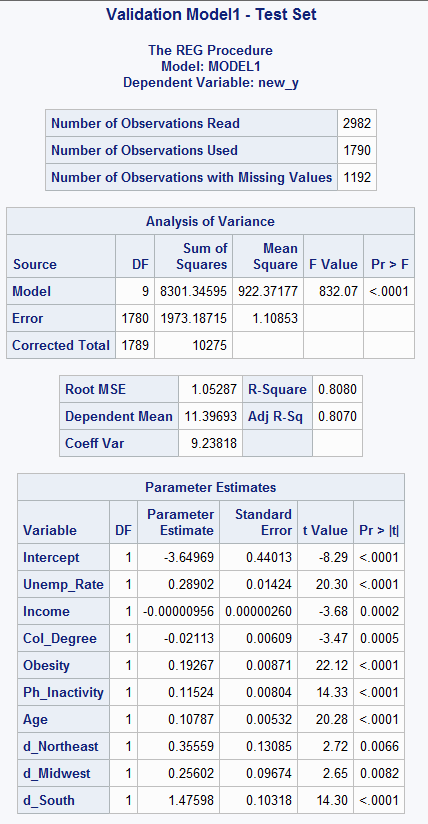
4.13



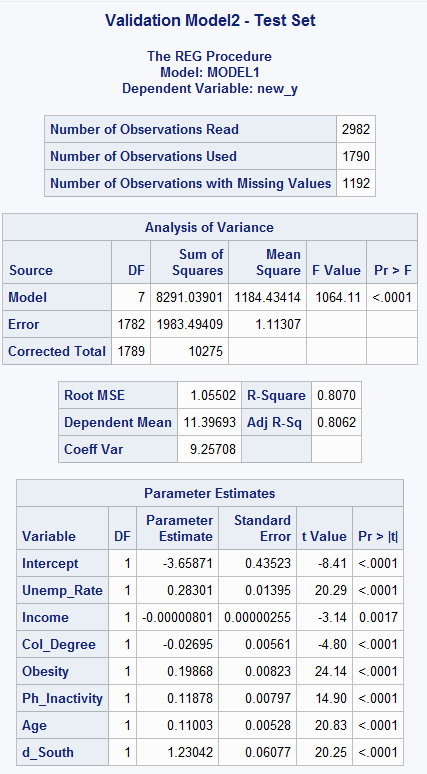
4.14



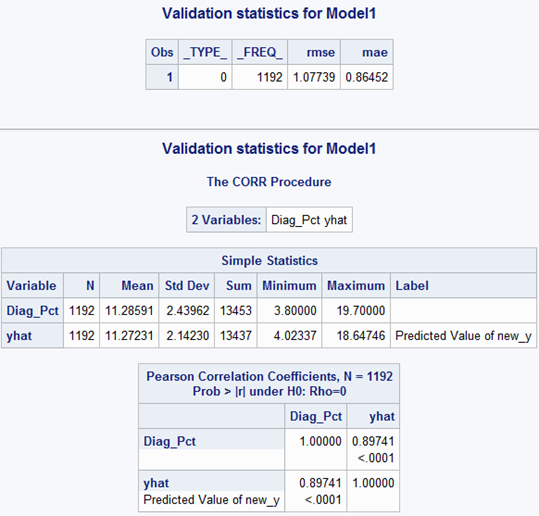
4.15



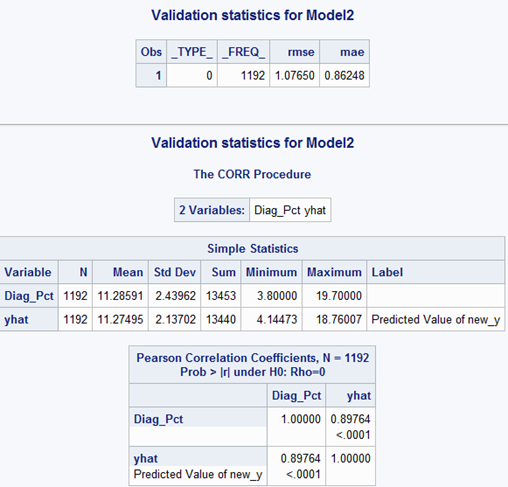
4.16



4.17



4.18

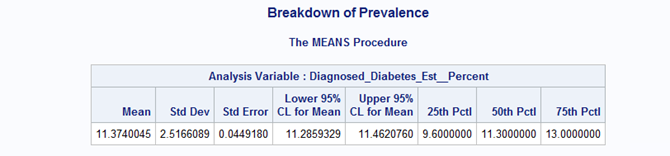


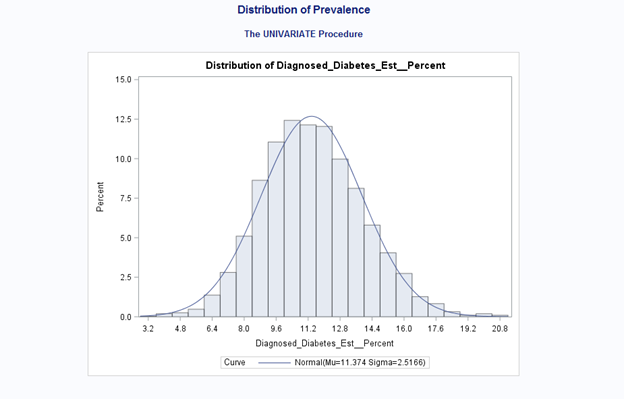
4.19

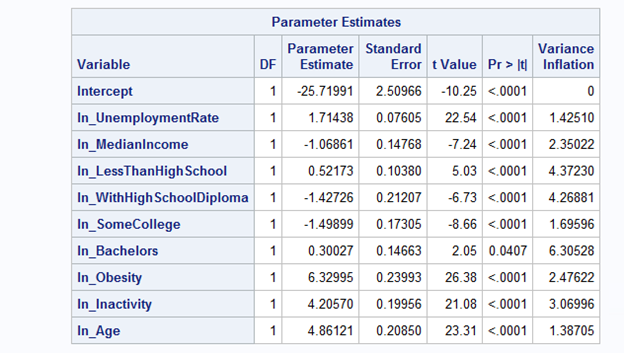
*Worku Animaw , Yeshaneh Seyoum :* ***Increasing prevalence of diabetes mellitus in a developing country and its related factors****,* [*http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0187670*](http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0187670)

*Published: November 7, 2017*

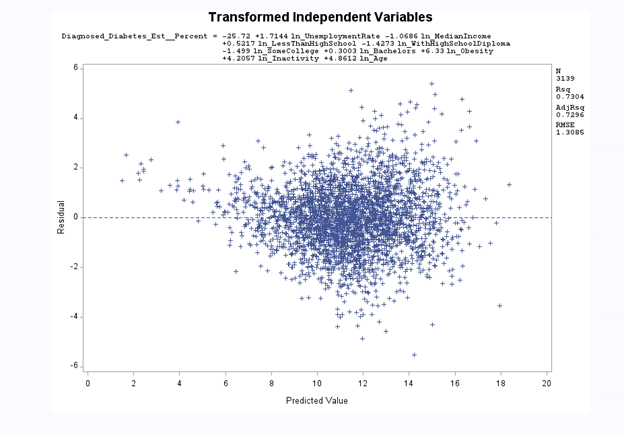
1. Charles Saporito

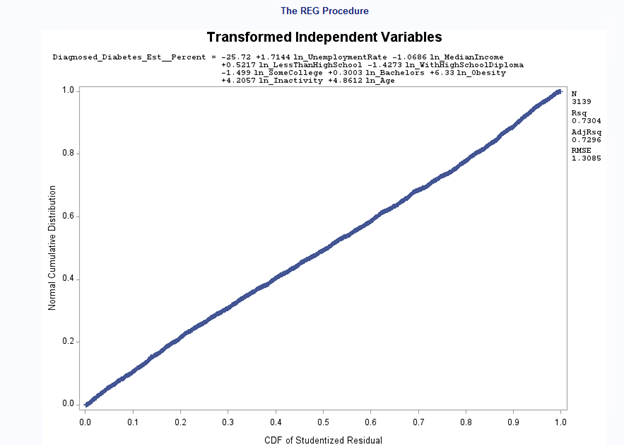
5.1

5.2

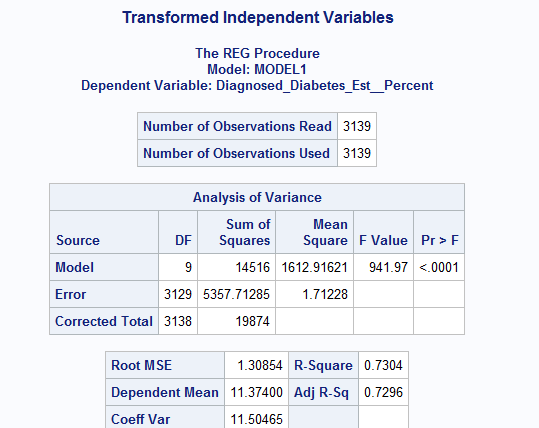
5.3 VIF shown in the above output

5.4



****

5.5



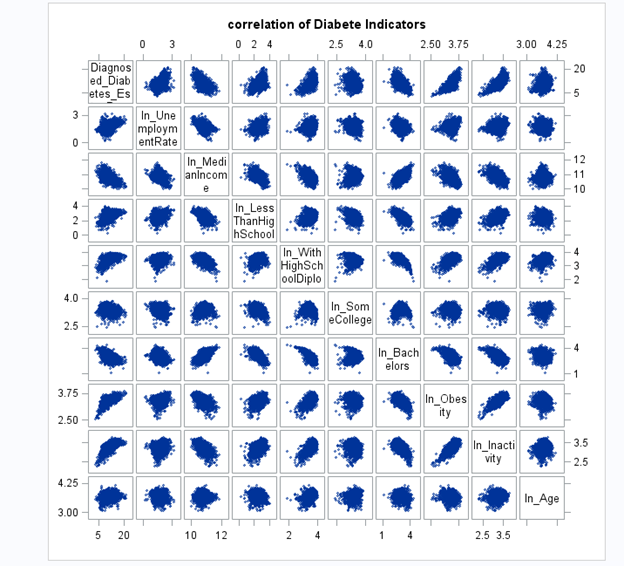
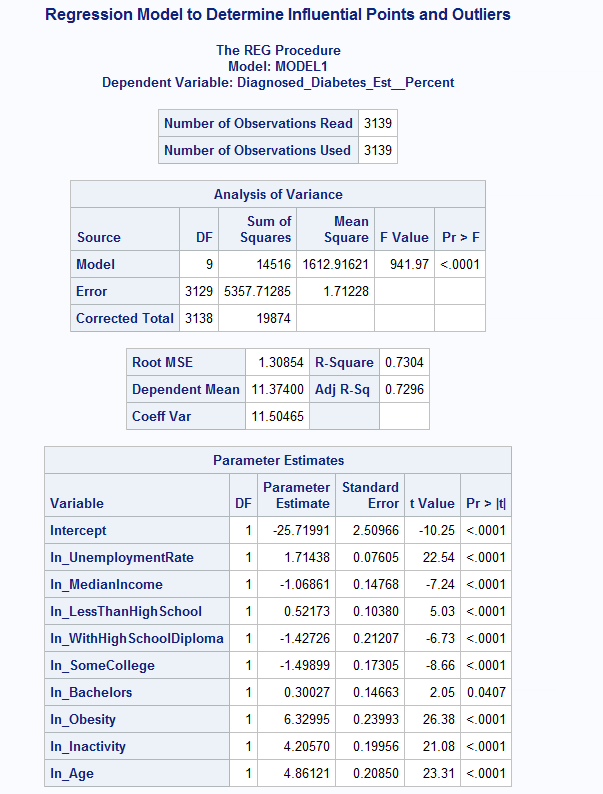


Figure 5.6













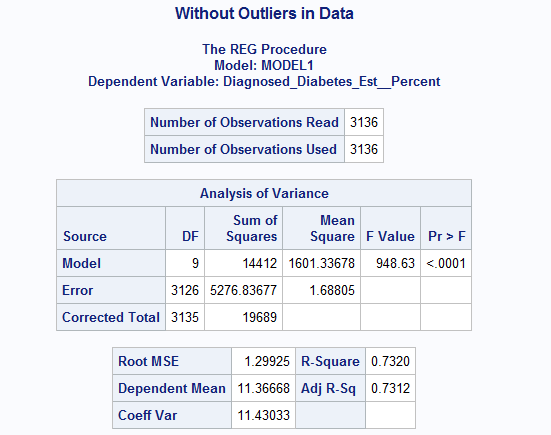
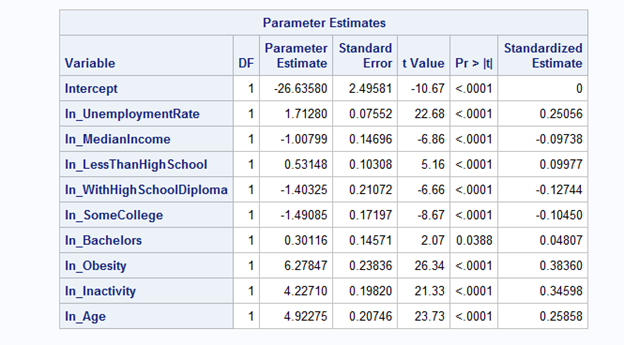
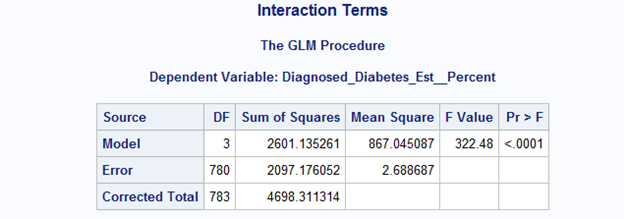
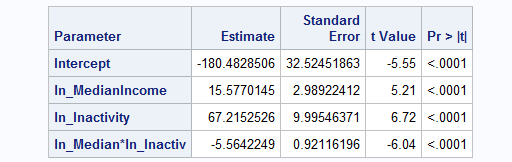
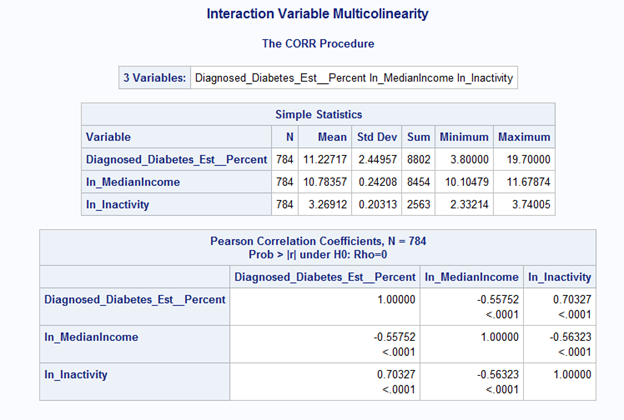


Figure 5.7

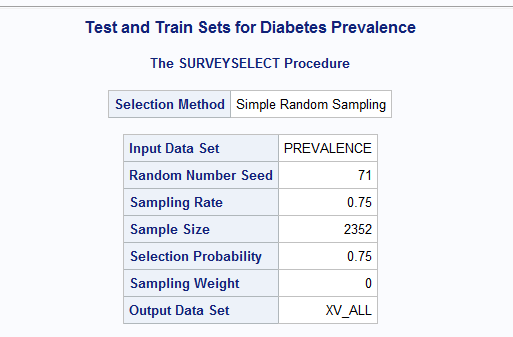


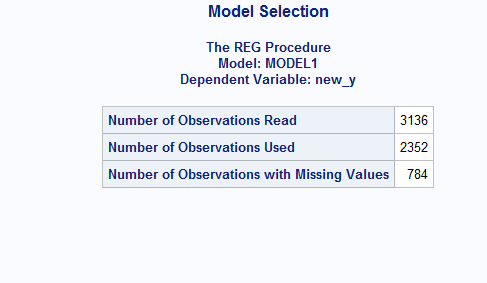


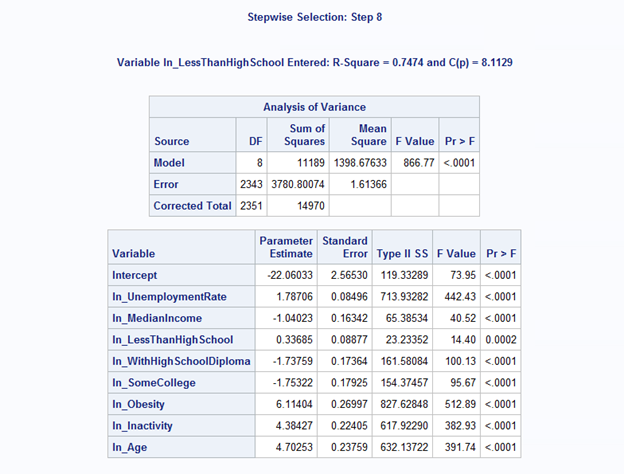


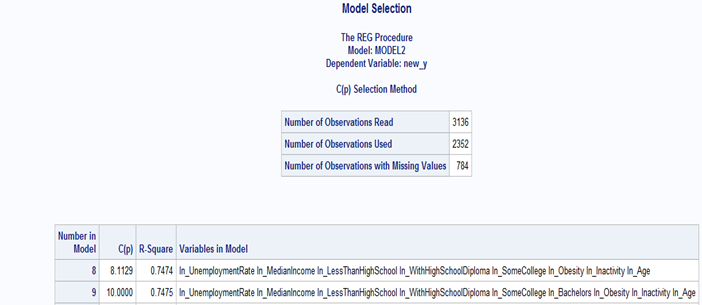


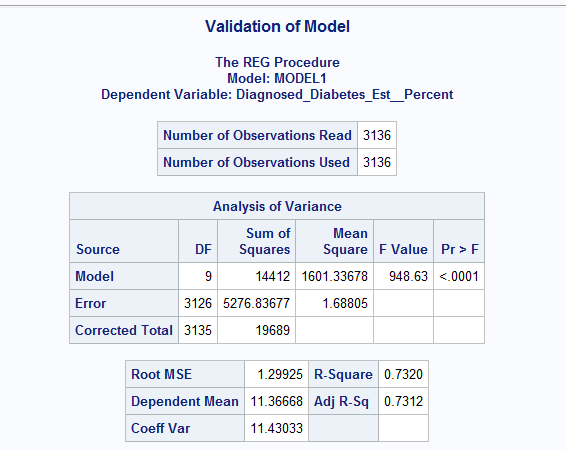
5.8











5.9 references

Mayo Clinic <https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444><Accessed 5/29/2018>