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, R^2, 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_lanct + 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

		Model 1	**Model 2**
Train	**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-R^2 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-R^2.

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 predictors. To enhance this from 71% to 90% (or above), factors like Gender, Hypertension percentage, prediabetes, smoking habits and the alcohol consumption rate could be added to potentially make a better model.

Appendices

I. Maxwell Carduner

Exhibit 1.1:

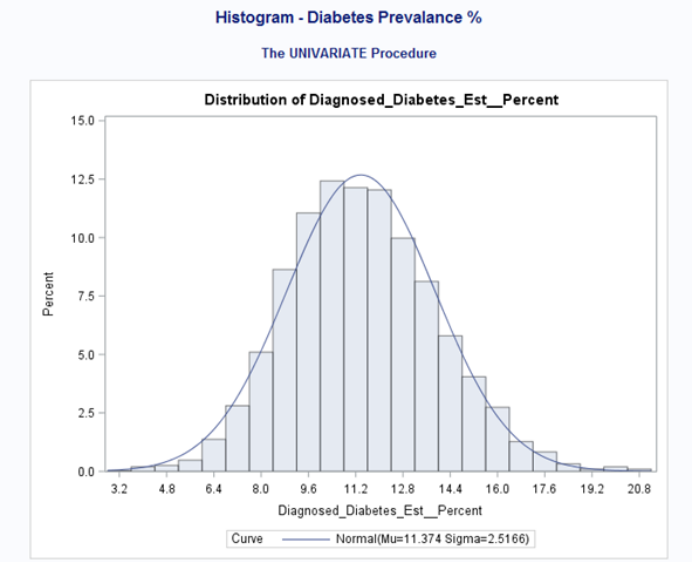
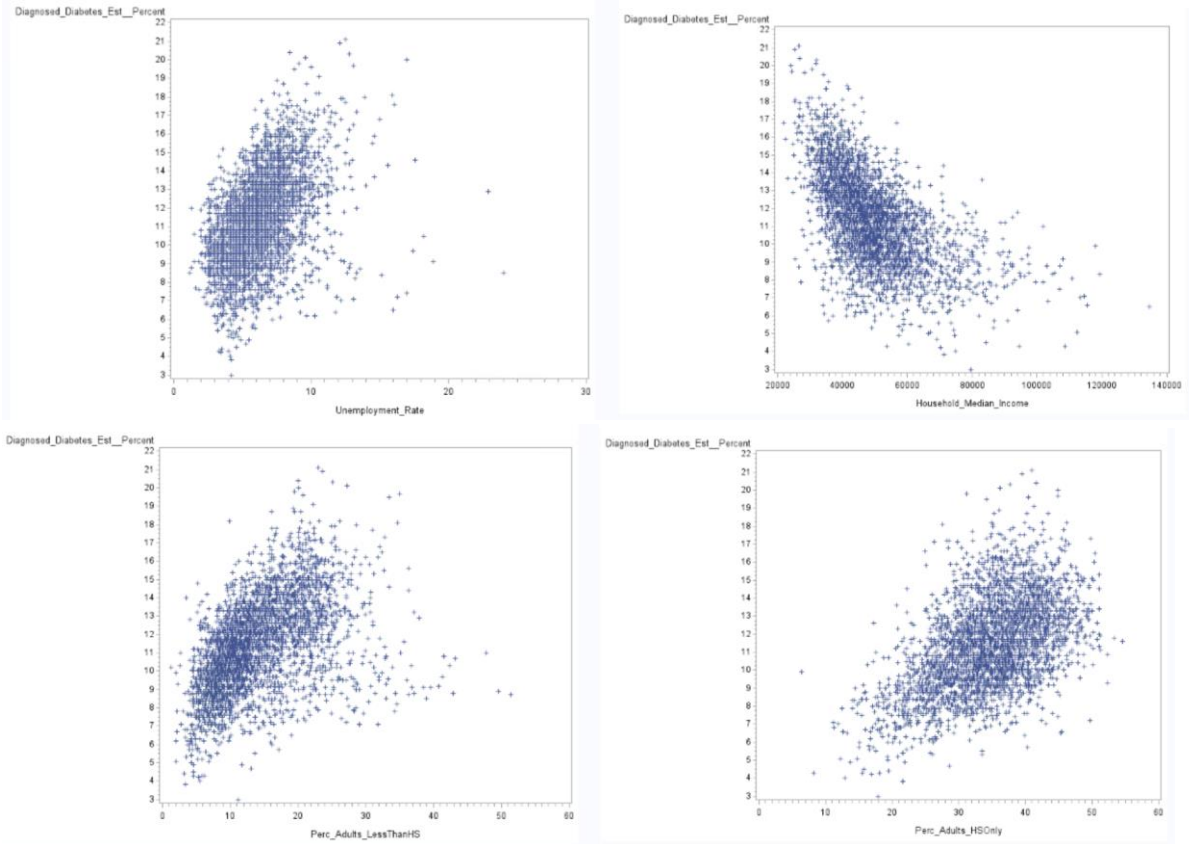


Exhibit 1.2 Scatterplots:



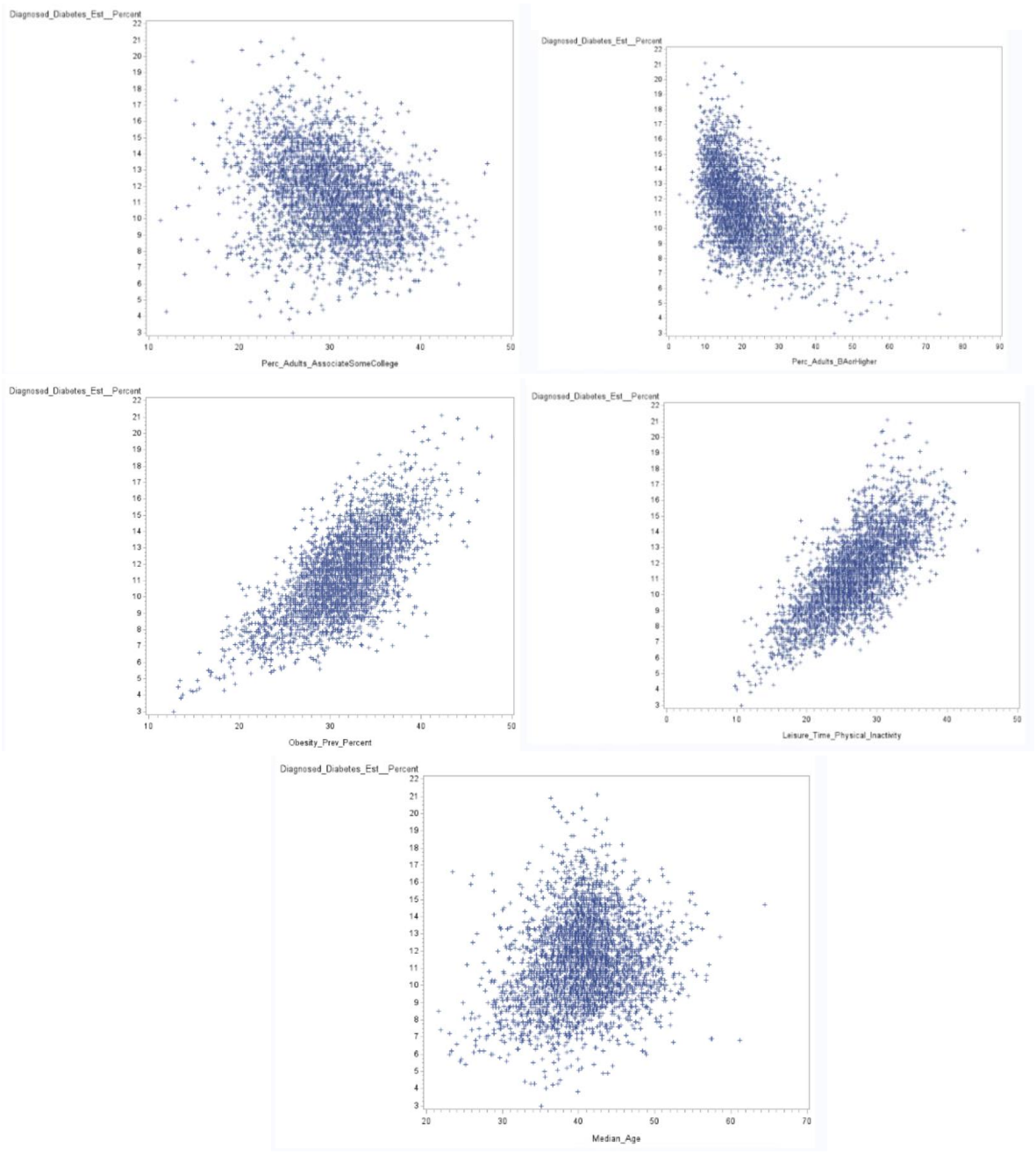


Exhibit 1.3 VIFs:

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	B	-11.01931	2.55456	-4.31	<.0001	0
Unemployment_Rate	B	-0.47485	0.10177	-4.67	<.0001	97.99213
Household_Median_Income	B	0.00007985	0.00002500	3.19	0.0014	190.73665
Perc_Adults_LessThanHS	B	0.01381	0.00582	2.37	0.0177	2.66474
Perc_Adults_HSOnly	B	-0.03281	0.01239	-2.65	0.0082	14.12272
Perc_Adults_Associate SomeCollege	B	-0.05267	0.00646	-8.16	<.0001	2.01177
Perc_Adults_BAorHigher	0	0	.	.	.	.
Obesity_Prev_Percent	B	0.66080	0.07283	9.07	<.0001	198.22599
Leisure_Time_Physical_Inactivity	B	0.38929	0.08045	4.84	<.0001	318.85001
Median_Age	B	0.04997	0.05111	0.98	0.3283	130.61428
unemp_income	B	0.00000189	0.00000105	1.81	0.0711	18.88567
obesity_inactivity	B	-0.00640	0.00110	-5.84	<.0001	144.01349
age_income	B	5.237154E-7	5.148341E-7	1.02	0.3091	142.46810
unemp_age	B	0.01538	0.00191	8.06	<.0001	63.17125
obesity_income	B	-0.00000477	5.888193E-7	-8.10	<.0001	74.33057
obesity_age	B	-0.00127	0.00136	-0.93	0.3514	182.35583
inactivity_age	B	-0.00087617	0.00138	-0.63	0.5257	230.93610
inactivity_income	B	4.235992E-7	5.781943E-7	0.73	0.4638	44.88951
HSOnly_unemp	B	-0.00002681	0.00177	-0.02	0.9879	55.87784

Exhibit 1.4

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

Perc_Adults_BAorHigher =	100 * Intercept + 0.00001 * Unemployment_Rate - 5.61E-9 * Household_Median_Income - 1 * Perc_Adults_LessThanHS - 1 * Perc_Adults_HSOnly - 1 * Perc_Adults_AssociateSomeCollege + 0.00001 * Obesity_Prev_Percent - 5E-6 * Leisure_Time_Physical_Inactivity - 8.84E-7 * Median_Age - 2.27E-8 * obesity_inactivity + 3.1E-7 * unemp_age - 178E-12 * obesity_income

Exhibit 1.5

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-10.18851	1.41341	-7.21	<.0001	0
Unemployment_Rate	1	-0.33682	0.06893	-4.89	<.0001	44.85713
Household_Median_Income	1	0.00010475	0.00001321	7.93	<.0001	53.15418
Perc_Adults_LessThanHS	1	0.01335	0.00571	2.34	0.0195	2.56096
Perc_Adults_HSOnly	1	-0.03212	0.00526	-6.11	<.0001	2.53592
Perc_Adults_Associate SomeCollege	1	-0.05120	0.00637	-8.04	<.0001	1.95548
Obesity_Prev_Percent	1	0.58517	0.04446	13.16	<.0001	73.71046
Leisure_Time_Physical_Inactivity	1	0.37613	0.03420	11.00	<.0001	57.48530
Median_Age	1	0.02288	0.01183	1.93	0.0532	6.98558
obesity_inactivity	1	-0.00652	0.00106	-6.16	<.0001	134.05233
unemp_age	1	0.01410	0.00169	8.35	<.0001	49.30884
obesity_income	1	-0.00000421	4.5334E-7	-9.29	<.0001	43.96521

Exhibit 1.6

The REG Procedure
Model: MODEL1
Dependent Variable: Diagnosed_Diabetes_Est__Percent

Number of Observations Read	3139
Number of Observations Used	3139

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	14524	1320.40447	771.83	<.0001
Error	3127	5349.50959	1.71075		
Corrected Total	3138	19874			

Root MSE	1.30796	R-Square	0.7308
Dependent Mean	11.37400	Adj R-Sq	0.7299
Coeff Var	11.49952		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	13.84762	0.37021	37.41	<.0001	0	0
Unemp_rate_c	1	-0.23715	0.01248	-19.00	<.0001	-0.21378	1.47124
income_c	1	0.00002786	0.00000309	9.02	<.0001	0.14261	2.90518
Perc_Adults_LessThanHS	1	0.01335	0.00571	2.34	0.0195	0.03470	2.56096
Perc_Adults_HSOnly	1	-0.03212	0.00526	-6.11	<.0001	-0.09027	2.53592
Perc_Adults_AssociateSomeCollege	1	-0.05120	0.00637	-8.04	<.0001	-0.10425	1.95548
obesity_c	1	-0.20160	0.00803	-25.10	<.0001	-0.36118	2.40582
inactivity_c	1	-0.17092	0.00740	-23.10	<.0001	-0.35160	2.69078
age_c	1	-0.11092	0.00527	-21.06	<.0001	-0.22986	1.38356
obesity_inactivity_c	1	-0.00652	0.00106	-6.16	<.0001	-0.08220	2.06631
unemp_age_c	1	0.01410	0.00169	8.35	<.0001	0.07876	1.03239
obesity_income_c	1	-0.00000421	4.5334E-7	-9.29	<.0001	-0.13711	2.52879

Exhibit 1.7:

The REG Procedure
Model: MODEL1
Dependent Variable: Diagnosed_Diabetes_Est__Percent

Number of Observations Read	3139
Number of Observations Used	3139

Analysis of Variance

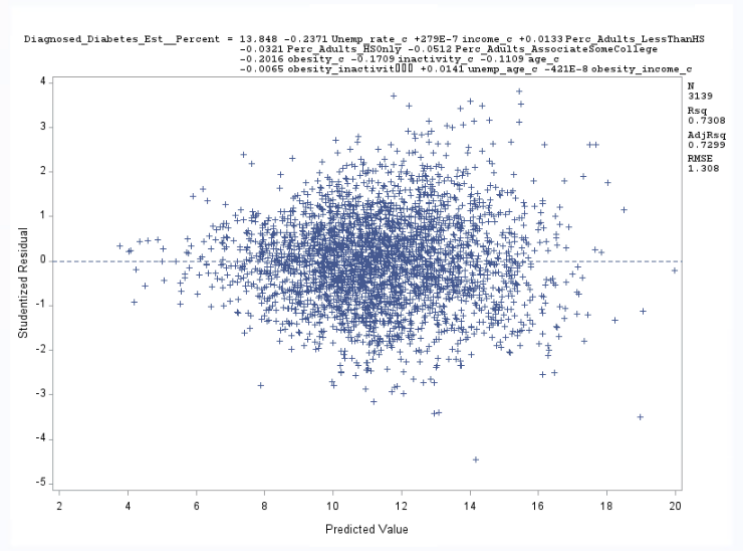
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	14271	1783.81381	996.41	<.0001
Error	3130	5603.44832	1.79024		
Corrected Total	3138	19874			

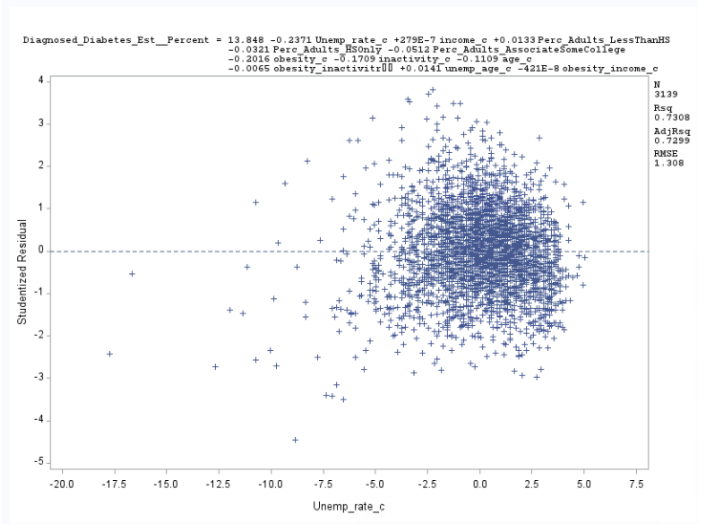
Root MSE	1.33800	R-Square	0.7181
Dependent Mean	11.37400	Adj R-Sq	0.7173
Coeff Var	11.76365		

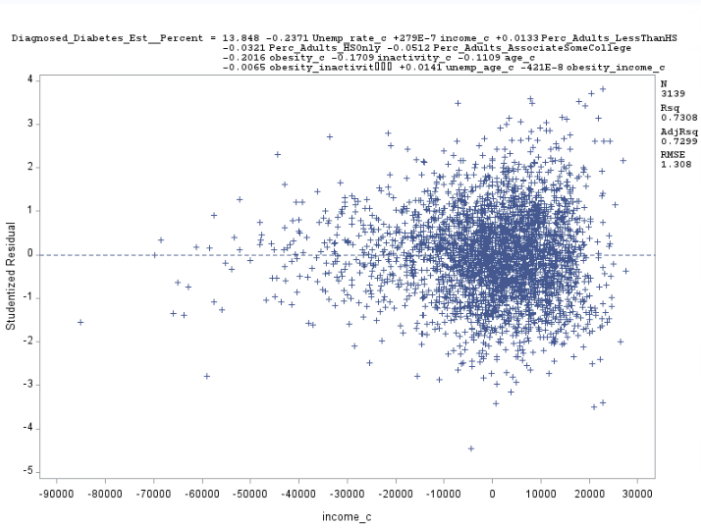
Parameter Estimates

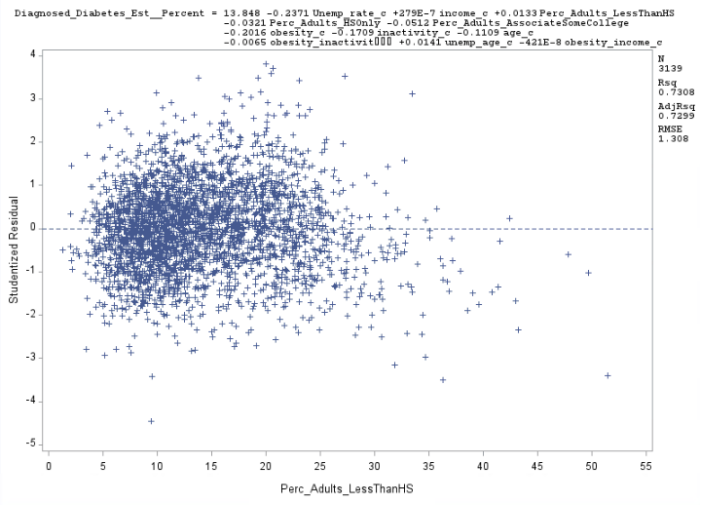
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-2.36608	0.48585	-4.87	<.0001	0	0
Unemployment_Rate	1	0.24640	0.01234	19.96	<.0001	0.22212	1.37414
Household_Median_Income	1	-0.00001553	0.00000284	-5.47	<.0001	-0.07948	2.34394
Perc_Adults_LessThanHS	1	0.01505	0.00578	2.61	0.0092	0.03912	2.50186
Perc_Adults_HSOnly	1	-0.03854	0.00504	-7.64	<.0001	-0.10832	2.22920
Perc_Adults_AssociateSomeCollege	1	-0.06687	0.00605	-11.06	<.0001	-0.13614	1.88367
Obesity_Prev_Percent	1	0.21720	0.00799	27.17	<.0001	0.38914	2.27702
Leisure_Time_Physical_Inactivity	1	0.16997	0.00749	22.69	<.0001	0.34965	2.83665
Median_Age	1	0.11613	0.00535	21.69	<.0001	0.24066	1.36702

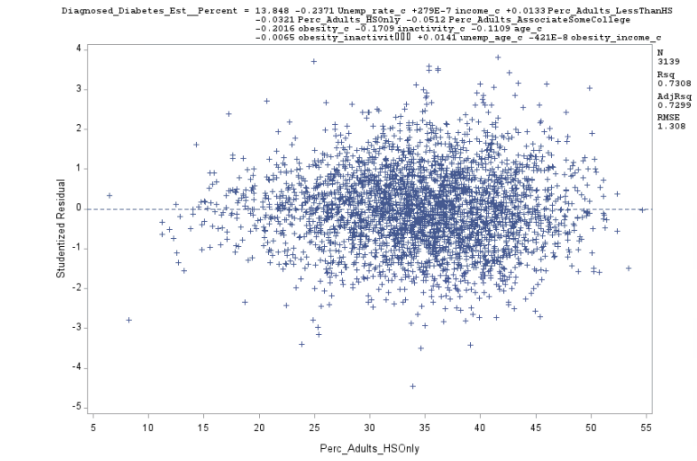
Exhibit 1.8

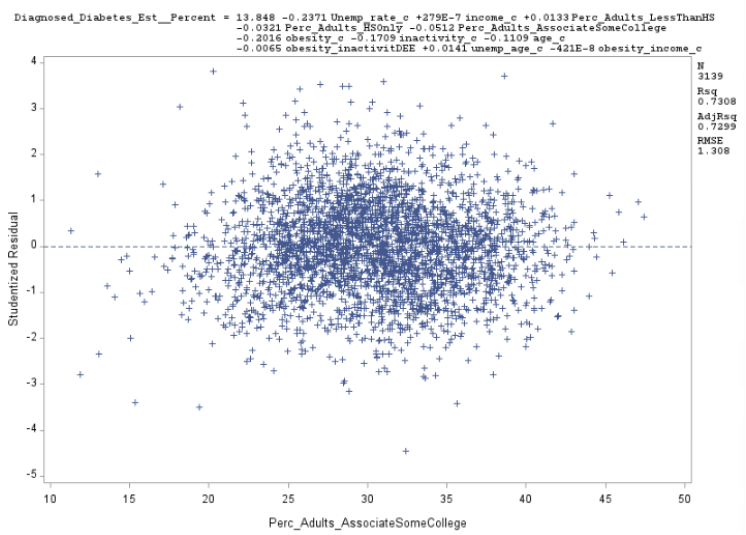


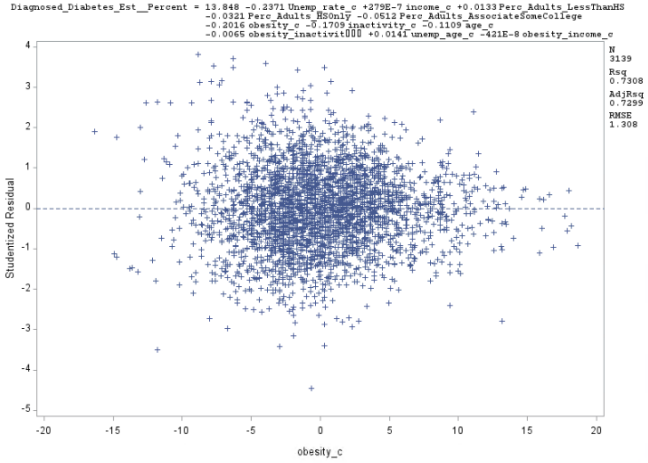

Diagnosed_Diabetes_Est__Percent = 13.848 -0.2371 Unemp_rate_c +279E-7 income_c +0.0133 Perc_Adults_LessThanHS
-0.0321 Perc_Adults_HSOnly -0.0512 Perc_Adults_AssociateSomeCollege
-0.2016 obesity_c -0.1709 inactivity_c -0.1109 age_c
-0.0065 obesity_inactivitr +0.0141 unemp_age_c -421E-8 obesity_income_c
N
3139
Rsq
0.7308
AdjRsq
0.7299
RMSE
1.308
Studentized Residual
Unemp_rate_c

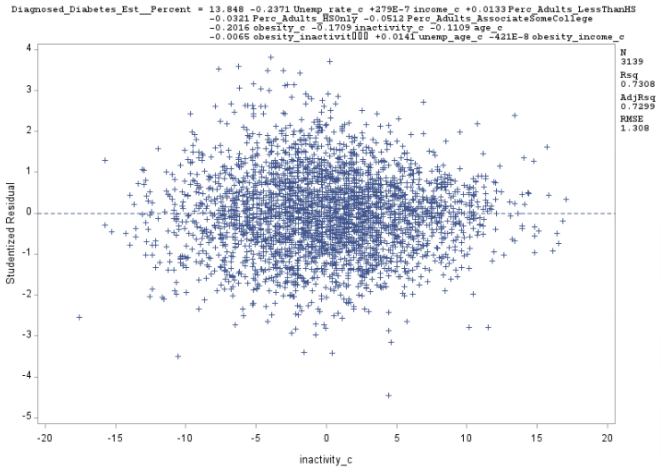

Diagnosed_Diabetes_Est__Percent = 13.848 -0.2371 Unemp_rate_c +279E-7 income_c +0.0133 Perc_Adults_LessThanHS
-0.0321 Perc_Adults_HSOnly -0.0512 Perc_Adults_AssociateSomeCollege
-0.2016 obesity_c -0.1709 inactivity_c -0.1109 age_c
-0.0065 obesity_inactivit +0.0141 unemp_age_c -421E-8 obesity_income_c
N
3139
Rsq
0.7308
AdjRsq
0.7299
RMSE
1.308
Studentized Residual
income_c

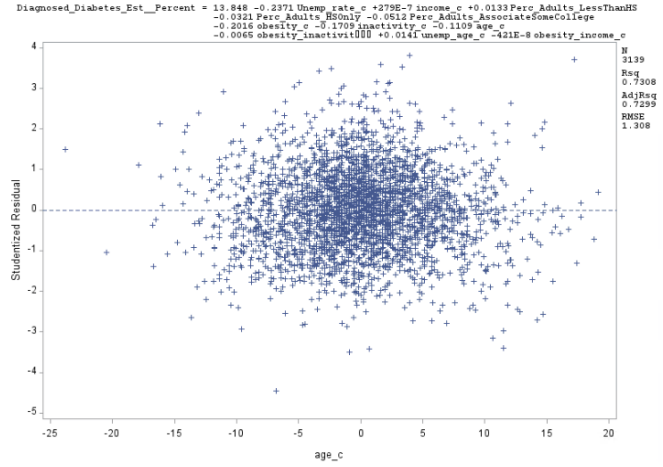

Diagnosed_Diabetes_Est__Percent = 13.848 -0.2371 Unemp_rate_c +279E-7 income_c +0.0133 Perc_Adults_LessThanHS
-0.0321 Perc_Adults_HSOnly -0.0512 Perc_Adults_AssociateSomeCollege
-0.2016 obesity_c -0.1709 inactivity_c -0.1109 age_c
-0.0065 obesity_inactivit +0.0141 unemp_age_c -421E-8 obesity_income_c
N
3139
Rsq
0.7308
AdjRsq
0.7299
RMSE
1.308
Studentized Residual
Perc_Adults_LessThanHS

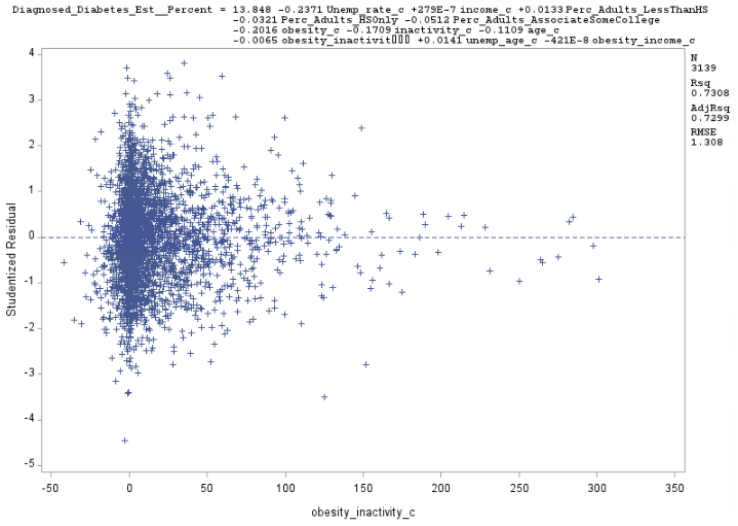

Diagnosed_Diabetes_Est__Percent = 13.848 -0.2371 Unemp_rate_c +279E-7 income_c +0.0133 Perc_Adults_LessThanHS
-0.0321 Perc_Adults_HSOnly -0.0512 Perc_Adults_AssociateSomeCollege
-0.2016 obesity_c -0.1709 inactivity_c -0.1109 age_c
-0.0065 obesity_inactivitDDD +0.0141 unemp_age_c -421E-8 obesity_income_c
N
3139
Rsq
0.7308
AdjRsq
0.7299
RMSE
1.308
Studentized Residual
Perc_Adults_HSOnly

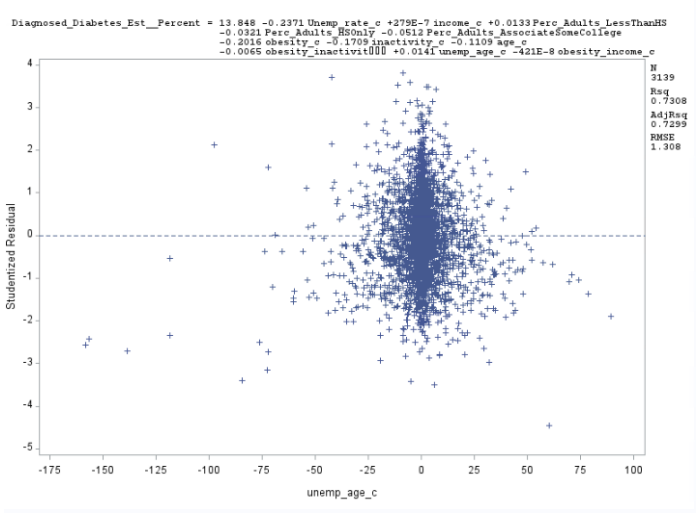

Diagnosed_Diabetes_Est__Percent = 13.848 -0.2371 Unemp_rate_c +279E-7 income_c +0.0133 Perc_Adults_LessThanHS
-0.0321 Perc_Adults_HSOnly -0.0512 Perc_Adults_AssociateSomeCollege
-0.2016 obesity_c -0.1709 inactivity_c -0.1109 age_c
-0.0065 obesity_inactivitDEE +0.0141 unemp_age_c -421E-8 obesity_income_c
N
3139
Rsq
0.7308
AdjRsq
0.7299
RMSE
1.308
Studentized Residual
Perc_Adults_AssociateSomeCollege


Diagnosed_Diabetes_Est__Percent = 13.848 -0.2371 Unemp_rate_c +279E-7 income_c +0.0133 Perc_Adults_LessThanHS
-0.0321 Perc_Adults_HSOnly -0.0512 Perc_Adults_AssociateSomeCollege
-0.2016 obesity_c -0.1709 inactivity_c -0.1109 age_c
-0.0065 obesity_inactivitDDD +0.0141 unemp_age_c -421E-8 obesity_income_c
N
3139
Rsq
0.7308
AdjRsq
0.7299
RMSE
1.308
Studentized Residual
obesity_c


Diagnosed_Diabetes_Est__Percent = 13.848 -0.2371 Unemp_rate_c +279E-7 income_c +0.0133 Perc_Adults_LessThanHS
-0.0321 Perc_Adults_HSOnly -0.0512 Perc_Adults_AssociateSomeCollege
-0.2016 obesity_c -0.1709 inactivity_c -0.1109 age_c
-0.0065 obesity_inactivit -0.0141 unemp_age_c -421E-8 obesity_income_c
N
3139
Rsq
0.7308
AdjRsq
0.7299
RMSE
1.308
Studentized Residual
inactivity_c


Diagnosed_Diabetes_Est__Percent = 13.848 -0.2371 Unemp_rate_c +279E-7 income_c +0.0133 Perc_Adults_LessThanHS
-0.0321 Perc_Adults_HSOnly -0.0512 Perc_Adults_AssociateSomeCollege
-0.2016 obesity_c -0.1709 inactivity_c -0.1109 age_c
-0.0065 obesity_inactivit -0.0141 unemp_age_c -421E-8 obesity_income_c
N
3139
Rsq
0.7308
AdjRsq
0.7299
RMSE
1.308
Studentized Residual
age_c


Diagnosed_Diabetes_Est__Percent = 13.848 -0.2371 Unemp_rate_c +279E-7 income_c +0.0133 Perc_Adults_LessThanHS
-0.0321 Perc_Adults_HSOnly -0.0512 Perc_Adults_AssociateSomeCollege
-0.2016 obesity_c -0.1709 inactivity_c -0.1109 age_c
-0.0065 obesity_inactivit -0.0141 unemp_age_c -421E-8 obesity_income_c
N
3139
Rsq
0.7308
AdjRsq
0.7299
RMSE
1.308
Studentized Residual
obesity_inactivity_c



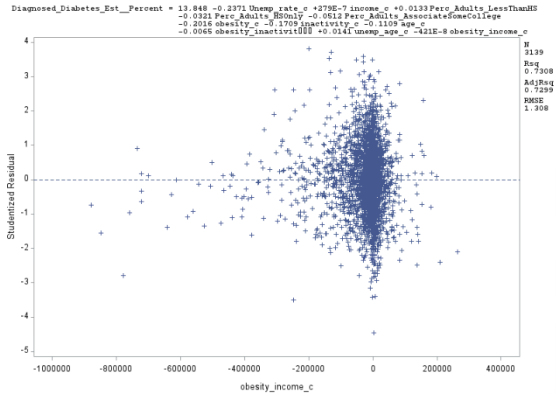


Exhibit 1.9:

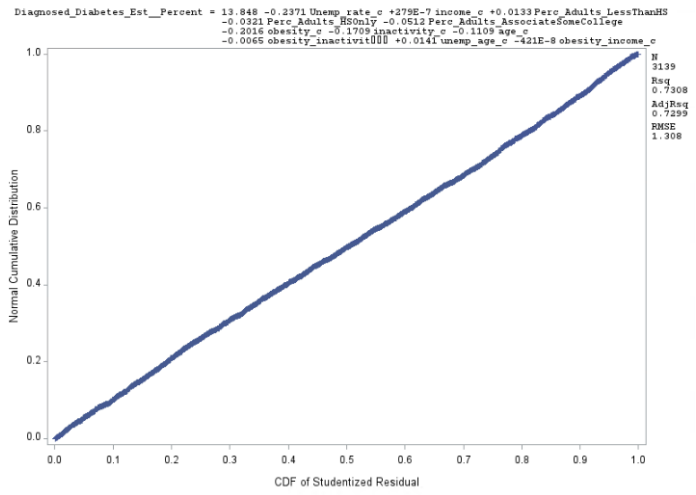


Exhibit 1.10:

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	13.84762	0.37021	37.41	<.0001	0	0
Unemp_rate_c	1	-0.23715	0.01248	-19.00	<.0001	-0.21378	1.47124
income_c	1	0.00002786	0.00000309	9.02	<.0001	0.14261	2.90518
Perc_Adults_LessThanHS	1	0.01335	0.00571	2.34	0.0195	0.03470	2.56096
Perc_Adults_HSOnly	1	-0.03212	0.00526	-6.11	<.0001	-0.09027	2.53592
Perc_Adults_Associate SomeCollege	1	-0.05120	0.00637	-8.04	<.0001	-0.10425	1.95548
obesity_c	1	-0.20160	0.00803	-25.10	<.0001	-0.36118	2.40582
inactivity_c	1	-0.17092	0.00740	-23.10	<.0001	-0.35160	2.69078
age_c	1	-0.11092	0.00527	-21.06	<.0001	-0.22986	1.38356
obesity_inactivity_c	1	-0.00652	0.00106	-6.16	<.0001	-0.08220	2.06631
unemp_age_c	1	0.01410	0.00169	8.35	<.0001	0.07876	1.03239
obesity_income_c	1	-0.00000421	4.5334E-7	-9.29	<.0001	-0.13711	2.52879

Exhibit 1.11:

The GLMSELECT Procedure

Data Set	WORK.DIABETES_CENTERED
Dependent Variable	Diagnosed_Diabetes_Est__Percent
Selection Method	Backward
Select Criterion	SBC
Stop Criterion	Cross Validation
Cross Validation Method	Split
Cross Validation Fold	5
Effect Hierarchy Enforced	None

Number of Observations Read	3139
Number of Observations Used	3139

Dimensions	
Number of Effects	12
Number of Parameters	12

Backward Selection Summary				
Step	Effect Removed	Number Effects In	SBC	CV PRESS
0		12	1770.0227*	5400.1537*

Selection stopped at a local minimum of the cross validation PRESS.

Stop Details				
Candidate For	Effect	Candidate CV PRESS		Compare CV PRESS
Removal	Perc_Adults_LessThan	5402.9435	>	5400.1537

The selected model is the model at the last step (Step 0).

Effects	Intercept Unemp_rate_c income_c Perc_Adults_LessThan Perc_Adults_HSOnly Perc_Adults_Associat obesity_c inactivity_c age_c obesity_inactivity_c unemp_age_c obesity_income_c

Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	11	14524	1320.40447	771.83
Error	3127	5349.50959	1.71075	
Corrected Total	3138	19874		

Root MSE	1.30796
Dependent Mean	11.37400
R-Square	0.7308
Adj R-Sq	0.7299
AIC	4838.40283
AICC	4838.51931
SBC	1770.02274
CV PRESS	5400.15366

Cross Validation Details			
	Observations		
Index	Fitted	Left Out	CV PRESS
1	2511	628	1111.0371
2	2511	628	951.4857
3	2511	628	1078.5929
4	2511	628	1195.7835
5	2512	627	1063.2544
Total			5400.1537

Parameter Estimates									
					Cross Validation Estimates				
Parameter	DF	Estimate	Standard Error	t Value	1	2	3	4	5
Intercept	1	13.847620	0.370207	37.41	1.41E+01	1.37E+01	1.36E+01	1.39E+01	1.38E+01
Unemp_rate_c	1	-0.237150	0.012484	-19.00	-2.40E-01	-2.38E-01	-2.32E-01	-2.31E-01	-2.45E-01
income_c	1	0.000027857	0.000003089	9.02	2.85E-05	2.86E-05	2.51E-05	3.27E-05	2.46E-05
Perc_Adults_LessThan	1	0.013350	0.005712	2.34	7.93E-03	1.20E-02	1.85E-02	1.22E-02	1.61E-02
Perc_Adults_HSOnly	1	-0.032123	0.005258	-6.11	-3.15E-02	-3.11E-02	-3.12E-02	-3.02E-02	-3.64E-02
Perc_Adults_Associat	1	-0.051205	0.006372	-8.04	-5.61E-02	-4.83E-02	-4.82E-02	-5.53E-02	-4.75E-02
obesity_c	1	-0.201596	0.008032	-25.10	-2.05E-01	-1.94E-01	-2.04E-01	-1.97E-01	-2.08E-01
inactivity_c	1	-0.170920	0.007398	-23.10	-1.68E-01	-1.78E-01	-1.67E-01	-1.68E-01	-1.74E-01
age_c	1	-0.110920	0.005266	-21.06	-1.10E-01	-1.07E-01	-1.15E-01	-1.08E-01	-1.15E-01
obesity_inactivity_c	1	-0.006520	0.001058	-6.16	-6.49E-03	-6.15E-03	-6.64E-03	-7.01E-03	-6.35E-03
unemp_age_c	1	0.014096	0.001687	8.35	1.30E-02	1.53E-02	1.39E-02	1.40E-02	1.42E-02
obesity_income_c	1	-0.000004213	0.000000453	-9.29	-4.23E-06	-4.38E-06	-4.05E-06	-4.49E-06	-3.95E-06

Exhibit 1.12

The REG Procedure
Model: MODEL1
Dependent Variable: Diagnosed_Diabetes_Est__Percent

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	2.8123	0.2994	2.2253	3.3994	0.1814	5.4432	.
2	.	17.2404	0.1652	16.9165	17.5642	14.6555	19.8253	.

Exhibit 1.13 References Used:

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

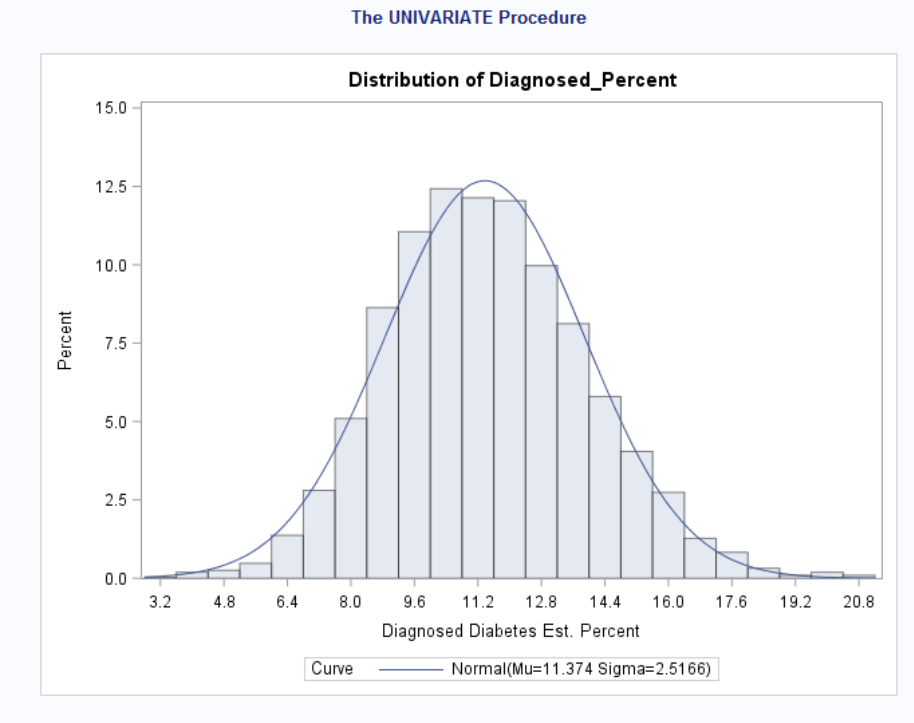
II. Kalaivani Chandramohan

2.1: Imported Data

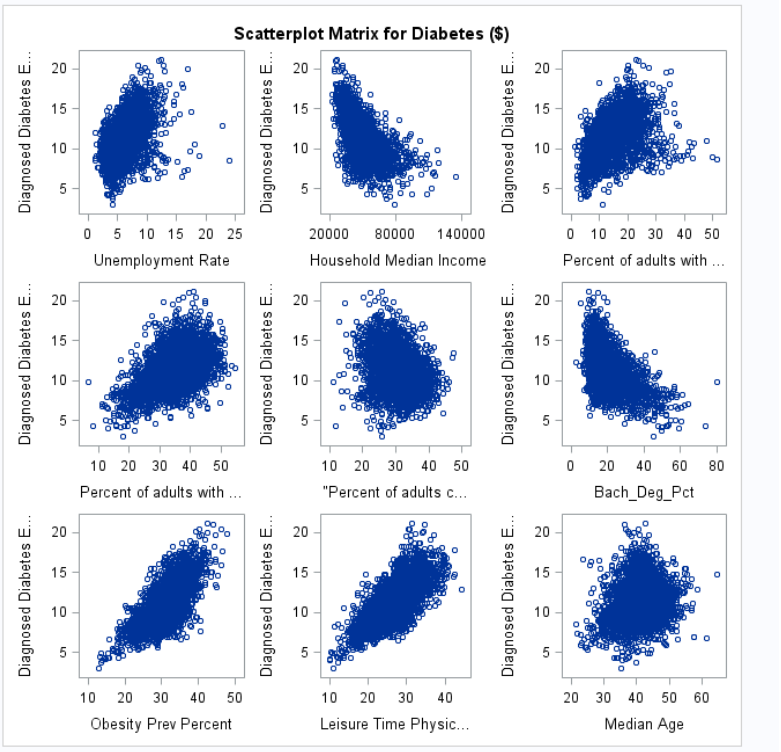
Diabetes Prevalence Dataset

Obs	Indicator	State	FIPS_code	County	Diagnosed_Percent	Unemp_Rate	Income	less_HighSch_Pct	HighSch_only	College_Deg	Bach_Deg_Pct	Obes_Per	Leisure_Inact	Age
1	Diagnosed diabetes prevalence	Alabama	1001	Autauga County	12.4	5.9	54487	12.417	34.331	28.660	24.593	36.4	30.9	37.9
2	Diagnosed diabetes prevalence	Alabama	1003	Baldwin County	11.1	6.1	56460	9.972	28.692	31.788	29.547	29.3	24.5	41.8
3	Diagnosed diabetes prevalence	Alabama	1005	Barbour County	18.2	10.6	32884	26.236	34.927	25.969	12.868	44.2	30.6	38.3
4	Diagnosed diabetes prevalence	Alabama	1007	Bibb County	14.6	7.2	43079	19.302	41.816	26.883	12.000	38.4	37.5	40.0

2.2



2.3 scatter plot



2.4 Correlation

Pearson Correlation Coefficients, N = 3139
Prob > |r| under H0: Rho=0

	Diagnosed_Percent	Unemp_Rate	Income	less_HighSch_Pct	HighSch_only	College_Deg	Bach_Deg_Pct	Obes_Per	Leisure_Inact	Age
Diagnosed_Percent Diagnosed Diabetes Est. Percent	1.00000	0.45995 <.0001	-0.57082 <.0001	0.47694 <.0001	0.50158 <.0001	-0.29975 <.0001	-0.56141 <.0001	0.68145 <.0001	0.72708 <.0001	0.19103 <.0001
Unemp_Rate Unemployment Rate	0.45995 <.0001	1.00000	-0.44240 <.0001	0.43807 <.0001	0.21103 <.0001	-0.24296 <.0001	-0.34061 <.0001	0.25314 <.0001	0.22808 <.0001	-0.01226 0.4923
Income Household Median Income	-0.57082 <.0001	-0.44240 <.0001	1.00000	-0.55913 <.0001	-0.50059 <.0001	0.15246 <.0001	0.70204 <.0001	-0.46289 <.0001	-0.56060 <.0001	-0.11852 <.0001
less_HighSch_Pct	0.47694 <.0001	0.43807 <.0001	-0.55913 <.0001	1.00000	0.21672 <.0001	-0.51084 <.0001	-0.59703 <.0001	0.33957 <.0001	0.47203 <.0001	-0.19605 <.0001
HighSch_only Percent of adults with a high school diploma only	0.50158 <.0001	0.21103 <.0001	-0.50059 <.0001	0.21672 <.0001	1.00000	-0.29430 <.0001	-0.76388 <.0001	0.51270 <.0001	0.56825 <.0001	0.31782 <.0001
College_Deg Percent of adults completing some college or associate's degree	-0.29975 <.0001	-0.24296 <.0001	0.15246 <.0001	-0.51084 <.0001	-0.29430 <.0001	1.00000	0.03274 0.0667	-0.12707 <.0001	-0.25909 <.0001	0.12724 <.0001
Bach_Deg_Pct Percent of adults with a bachelor's degree or higher	-0.56141 <.0001	-0.34061 <.0001	0.70204 <.0001	-0.59703 <.0001	-0.76388 <.0001	0.03274 0.0667	1.00000	-0.56850 <.0001	-0.63228 <.0001	-0.17693 <.0001
Obes_Per Obesity Prev Percent	0.68145 <.0001	0.25314 <.0001	-0.46289 <.0001	0.33957 <.0001	0.51270 <.0001	-0.12707 <.0001	-0.56850 <.0001	1.00000	0.69058 <.0001	-0.07154 <.0001
Leisure_Inact Leisure Time Physical Inactivity Prev Perc	0.72708 <.0001	0.22808 <.0001	-0.56060 <.0001	0.47203 <.0001	0.56825 <.0001	-0.25909 <.0001	-0.63228 <.0001	0.69058 <.0001	1.00000	0.08845 <.0001
Age Median Age	0.19103 <.0001	-0.01226 0.4923	-0.11852 <.0001	-0.19605 <.0001	0.31782 <.0001	0.12724 <.0001	-0.17693 <.0001	-0.07154 <.0001	0.08845 <.0001	1.00000

2.5 VIF

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-2.36608	0.48585	-4.87	<.0001	0
Unemp_Rate	Unemployment Rate	1	0.24640	0.01234	19.96	<.0001	1.37414
Income	Household Median Income	1	-0.00001553	0.00000284	-5.47	<.0001	2.34394
less_HighSch_Pct		1	0.01505	0.00578	2.61	0.0092	2.50186
HighSch_only	Percent of adults with a high school diploma only	1	-0.03854	0.00504	-7.64	<.0001	2.22920
College_Deg	"Percent of adults completing some college or associate's degree"B	1	-0.06687	0.00605	-11.06	<.0001	1.68367
Obes_Per	Obesity Prev Percent	1	0.21720	0.00799	27.17	<.0001	2.27702
Leisure_Inact	Leisure Time Physical Inactivity Prev Perc	1	0.16997	0.00749	22.69	<.0001	2.63665
Med_Age	Median Age	1	0.11613	0.00535	21.69	<.0001	1.36702

2.6 Final model

Diabetes Prevalence Dataset

The REG Procedure
Model: MODEL1
Dependent Variable: Diagnosed_Percent Diagnosed Diabetes Est. Percent

Number of Observations Read	2999
Number of Observations Used	2999

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	13455	1681.82535	1218.34	<.0001
Error	2990	4127.47825	1.38043		
Corrected Total	2998	17582			

Root MSE	1.17492	R-Square	0.7652
Dependent Mean	11.35348	Adj R-Sq	0.7646
Coeff Var	10.34851		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2.97817	0.44445	-6.70	<.0001
Unemp_Rate	Unemployment Rate	1	0.29939	0.01236	24.21	<.0001
Income	Household Median Income	1	-0.00000854	0.00000261	-3.27	0.0011
less_HighSch_Pct	Percent of adults with less than a high school diploma	1	0.03057	0.00549	5.57	<.0001
HighSch_only	Percent of adults with a high school diploma only	1	-0.03868	0.00463	-8.35	<.0001
College_Deg	"Percent of adults completing some college or associate's degree"B	1	-0.05789	0.00554	-10.46	<.0001
Obes_Per	Obesity Prev Percent	1	0.20669	0.00737	28.06	<.0001
Leisure_Inact	Leisure Time Physical Inactivity Prev Perc	1	0.17548	0.00681	25.77	<.0001
Med_Age	Median Age	1	0.10800	0.00501	21.55	<.0001

2.7

test and train sets for diabetes

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	DIABETES
Random Number Seed	5000
Sampling Rate	0.75
Sample Size	2355
Selection Probability	0.750239
Sampling Weight	0
Output Data Set	NEWDIAB

test and train sets for diabetes

Obs	Selected	Indicator	State	FIPS_code	County	Diagnosed_Percent	Unemp_Rate	Income	less_HighSch_Pct	HighSch_only	College_Deg	Bach_Deg_Pct	Obes_Per	Leisure_Inact	Med_Age	new_y	Diag_Pct
1	0	Diagnosed diabetes prevalence	Alabama	1001	Autauga County	12.4	5.9	54487	12.417	34.331	28.660	24.593	36.4	30.9	37.9	.	.

2.8 selection Method

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7146 and C(p) = 9.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	10708	1338.50916	734.27	<.0001
Error	2346	4276.53130	1.82290		
Corrected Total	2354	14985			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-2.28752	0.56620	29.75467	16.32	<.0001
Unemp_Rate	0.23785	0.01421	510.97532	280.31	<.0001
Income	-0.00001395	0.00000333	31.95718	17.53	<.0001
less_HighSch_Pct	0.01734	0.00673	12.09346	6.63	0.0101
HighSch_only	-0.03607	0.00588	68.53457	37.60	<.0001
College_Deg	-0.06616	0.00695	165.00178	90.52	<.0001
Obes_Per	0.21433	0.00940	948.50060	520.32	<.0001
Leisure_Inact	0.16899	0.00876	678.67882	372.31	<.0001
Med_Age	0.11300	0.00626	594.75354	326.27	<.0001

Bounds on condition number: 2.7048, 132.82

All variables left in the model are significant at the 0.1000 level.

2.9

C(p) Selection Method

Number of Observations Read	3139
Number of Observations Used	2355
Number of Observations with Missing Values	784

Number in Model	C(p)	R-Square	Variables in Model
8	9.0000	0.7146	Unemp_Rate Income less_HighSch_Pct HighSch_only College_Deg Obes_Per Leisure_Inact Med_Age
7	13.6342	0.7138	Unemp_Rate Income HighSch_only College_Deg Obes_Per Leisure_Inact Med_Age
7	24.5309	0.7125	Unemp_Rate less_HighSch_Pct HighSch_only College_Deg Obes_Per Leisure_Inact Med_Age
7	44.5964	0.7100	Unemp_Rate Income less_HighSch_Pct College_Deg Obes_Per Leisure_Inact Med_Age

2.10 Regression analysis

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	3007
Number of Observations Used	2223
Number of Observations with Missing Values	784

Analysis of Variance

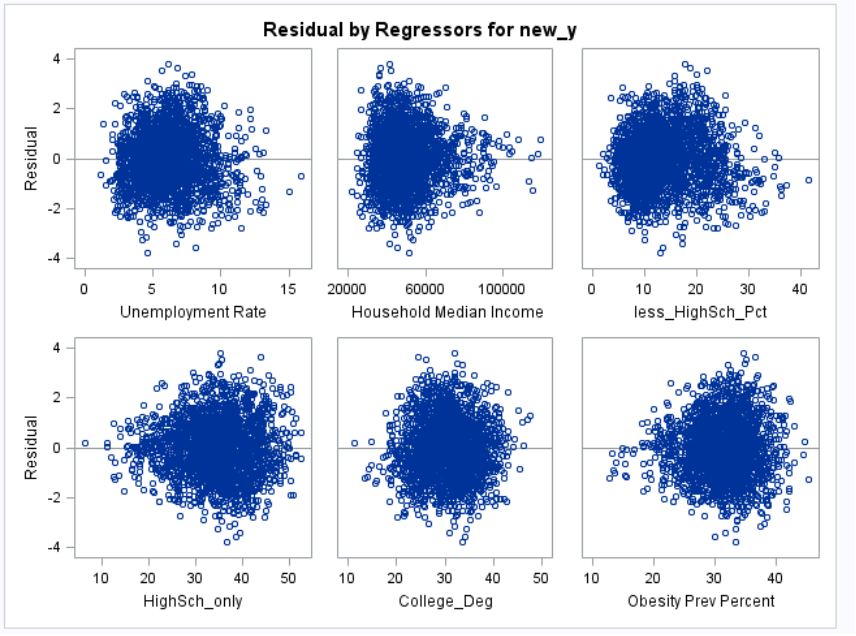
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	9704.65735	1213.08217	911.85	<.0001
Error	2214	2945.39998	1.33035		
Corrected Total	2222	12650			

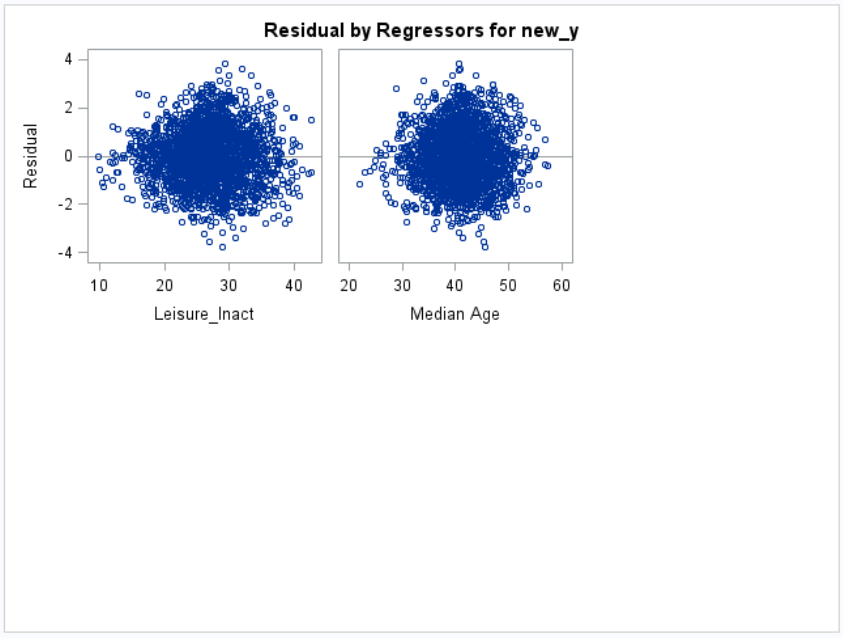
Root MSE	1.15341	R-Square	0.7672
Dependent Mean	11.32267	Adj R-Sq	0.7663
Coeff Var	10.18672		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2.93145	0.50814	-5.77	<.0001
Unemp_Rate	Unemployment Rate	1	0.28730	0.01412	20.34	<.0001
Income	Household Median Income	1	-0.00000762	0.00000301	-2.53	0.0114
less_HighSch_Pct	Percent of adults with less than a high school diploma	1	0.03621	0.00635	5.71	<.0001
HighSch_only	Percent of adults with a high school diploma only	1	-0.04080	0.00527	-7.75	<.0001
College_Deg	"Percent of adults completing some college or associate's degree"B	1	-0.05091	0.00624	-8.16	<.0001
Obes_Per	Obesity Prev Percent	1	0.20068	0.00846	23.71	<.0001
Leisure_Inact	Leisure Time Physical Inactivity Prev Perc	1	0.17352	0.00781	22.22	<.0001
Med_Age	Median Age	1	0.10808	0.00577	18.74	<.0001

2.11





2.12

Validation - Test Set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	3139
Number of Observations Used	2355
Number of Observations with Missing Values	784

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	10708	1338.50916	734.27	<.0001
Error	2346	4276.53130	1.82290		
Corrected Total	2354	14985			

Root MSE	1.35015	R-Square	0.7146
Dependent Mean	11.34688	Adj R-Sq	0.7136
Coeff Var	11.89886		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-2.28752	0.56620	-4.04	<.0001
Unemp_Rate	Unemployment Rate	1	0.23785	0.01421	16.74	<.0001
Income	Household Median Income	1	-0.00001395	0.00000333	-4.19	<.0001
less_HighSch_Pct	Percent of adults with less than a high school diploma	1	0.01734	0.00673	2.58	0.0101
HighSch_only	Percent of adults with a high school diploma only	1	-0.03607	0.00588	-6.13	<.0001
College_Deg	"Percent of adults completing some college or associate's degree"B	1	-0.06616	0.00695	-9.51	<.0001
Obes_Per	Obesity Prev Percent	1	0.21433	0.00940	22.81	<.0001
Leisure_Inact	Leisure Time Physical Inactivity Prev Perc	1	0.16899	0.00876	19.30	<.0001
Med_Age	Median Age	1	0.11300	0.00626	18.06	<.0001

2.13

Validation statistics for Model

The CORR Procedure

2 Variables:	Diagnosed_Percent yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Diagnosed_Percent	784	11.45548	2.49710	8981	4.30000	19.10000	Diagnosed Diabetes Est. Percent
yhat	784	11.44854	2.02898	8976	4.52426	17.38925	Predicted Value of new_y

Pearson Correlation Coefficients, N = 784
Prob > |r| under H0: Rho=0

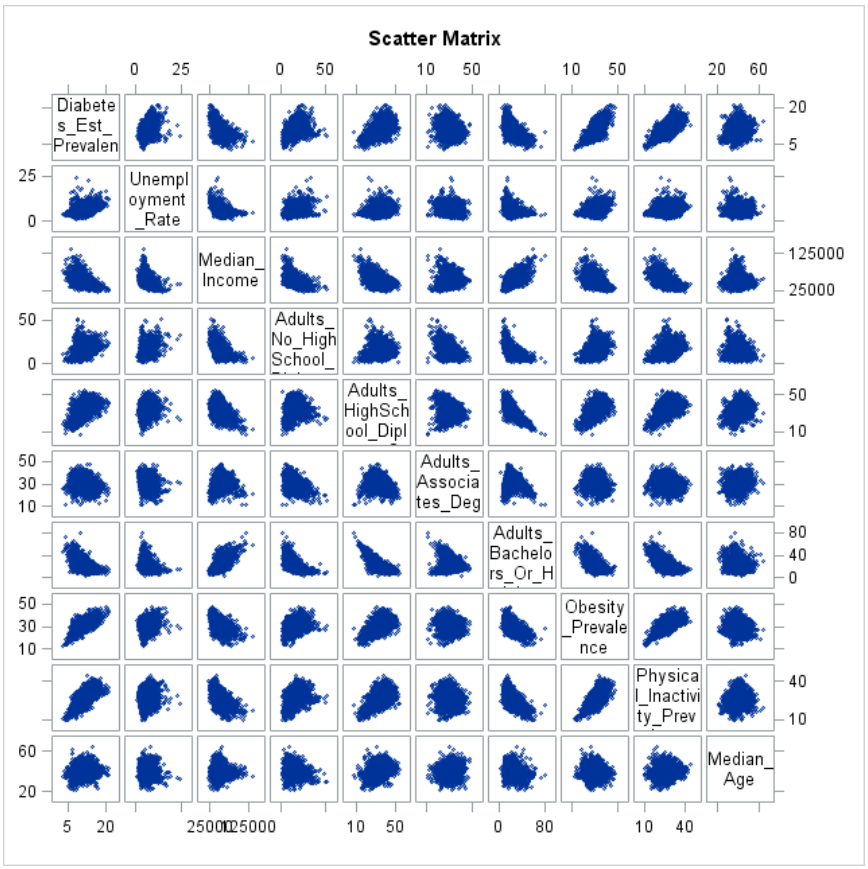
	Diagnosed_Percent	yhat
Diagnosed_Percent Diagnosed Diabetes Est. Percent	1.00000	0.85392 <.0001
yhat Predicted Value of new_y	0.85392 <.0001	1.00000

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

III. William Chirciu

3.1: Scatterplots



3.2: Descriptives

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Diabetes_Est_Prevalence	3139	11.37400	2.51661	35703	3.00000	21.10000
Unemployment_Rate	3139	6.24562	2.26862	19605	1.20000	24.00000
Median_Income	3139	49533	12884	155484533	22045	134609
Adults_No_HighSchool_Diploma	3139	14.18621	6.54159	44531	1.27900	51.47900
Adults_HighSchool_Diploma_Only	3139	34.57702	7.07204	108537	6.46400	54.64200
Adults_Associates_Degree	3139	30.43518	5.12375	95536	11.27800	47.42800
Adults_Bachelors_Or_Higher	3139	20.80159	9.13997	65296	2.98500	80.21000
Obesity_Prevalence	3139	31.47509	4.50874	98800	12.80000	47.80000
Physical_Inactivity_Prevalence	3139	26.82440	5.17697	84202	9.80000	44.40000
Median_Age	3139	40.71911	5.21521	127817	21.60000	64.50000
Age_Obesity	3139	1280	232.06008	4017782	449.28000	1949
Age_Inactivity	3139	1095	256.74501	3436117	347.30000	2078
Unemployment_Age	3139	254.17108	97.17935	797843	41.16000	769.44000
Unemployment_Obesity	3139	199.16990	86.37420	625194	42.84000	747.04000
Unemployment_Inactivity	3139	170.21292	78.19152	534298	30.84000	645.78000

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

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	B	1.59971	1.40545	1.14	0.2551	.	0
Unemployment_Rate	B	-0.74903	0.10425	-7.19	<.0001	0.00990	100.96040
Median_Income	B	-0.00001461	0.00000282	-5.18	<.0001	0.41967	2.38281
Adults_No_HighSchool_Diploma	B	0.01805	0.00576	3.13	0.0017	0.39038	2.56157
Adults_HighSchool_Diploma_Only	B	-0.03405	0.00508	-6.71	<.0001	0.42979	2.32671
Adults_Associates_Degree	B	-0.05627	0.00625	-9.00	<.0001	0.54019	1.85119
Adults_Bachelors_Or_Higher	0	0	.	.	.	.	.
Obesity_Prevalence	B	0.13758	0.06066	2.27	0.0234	0.00741	135.01449
Physical_Inactivity_Prevalence	B	0.23733	0.05664	4.19	<.0001	0.00644	155.19551
Median_Age	B	0.06818	0.03175	2.15	0.0318	0.02020	49.50034
Age_Obesity	B	-0.00028400	0.00137	-0.21	0.8354	0.00551	181.54662
Age_Inactivity	B	-0.00139	0.00130	-1.07	0.2838	0.00500	199.99357
Unemployment_Age	B	0.01457	0.00176	8.27	<.0001	0.01891	52.89040
Unemployment_Obesity	B	0.01350	0.00303	4.45	<.0001	0.00806	124.02340

3.4:Regression Results All Variables + Interaction Terms

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	B	3.21904	0.75164	4.28	<.0001	.	0
Unemployment_Rate	B	-0.74093	0.10390	-7.13	<.0001	0.00997	100.29707
Median_Income	B	-0.00001443	0.00000280	-5.15	<.0001	0.42559	2.34969
Adults_No_HighSchool_Diploma	B	0.01890	0.00572	3.30	0.0010	0.39542	2.52899
Adults_HighSchool_Diploma_Only	B	-0.03375	0.00505	-6.68	<.0001	0.43418	2.30319
Adults_Associates_Degree	B	-0.05521	0.00617	-8.95	<.0001	0.55484	1.80232
Adults_Bachelors_Or_Higher	0	0	.	.	.	.	.
Obesity_Prevalence	B	0.13122	0.01649	7.96	<.0001	0.10023	9.97675
Physical_Inactivity_Prevalence	B	0.17395	0.00742	23.43	<.0001	0.37505	2.66628
Median_Age	B	0.02683	0.01192	2.25	0.0245	0.14327	6.97980
Unemployment_Age	B	0.01430	0.00171	8.37	<.0001	0.02011	49.72832
Unemployment_Obesity	B	0.01282	0.00217	5.89	<.0001	0.01571	63.66323

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

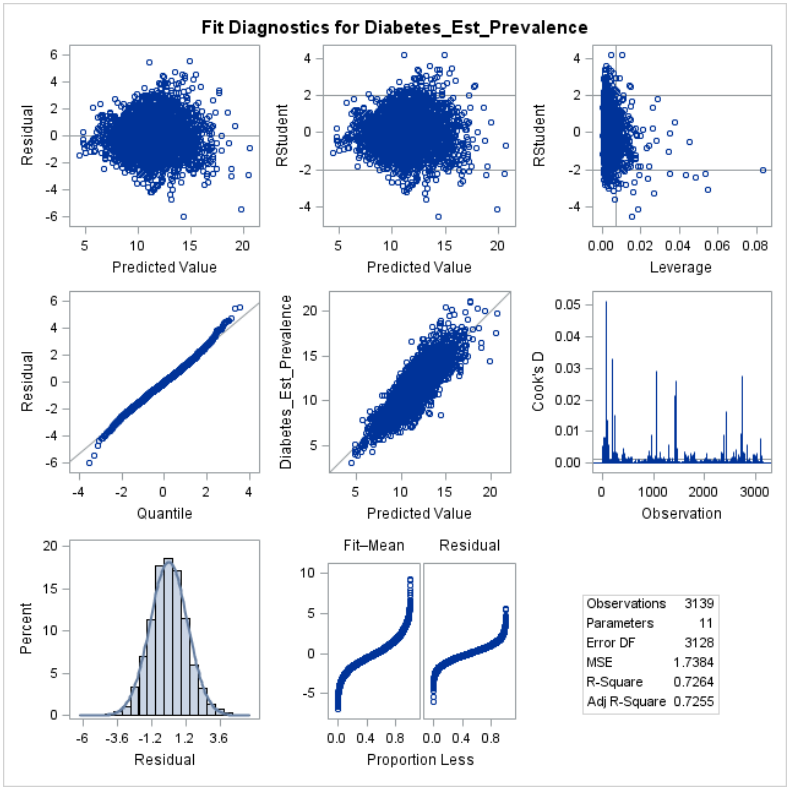
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	14436	1443.63203	830.45	<.0001
Error	3128	5437.63851	1.73838		
Corrected Total	3138	19874			

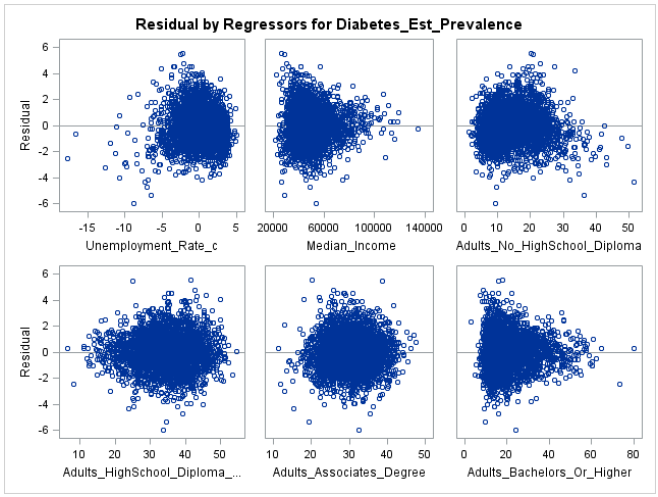
Root MSE	1.31847	R-Square	0.7264
Dependent Mean	11.37400	Adj R-Sq	0.7255
Coeff Var	11.59200		

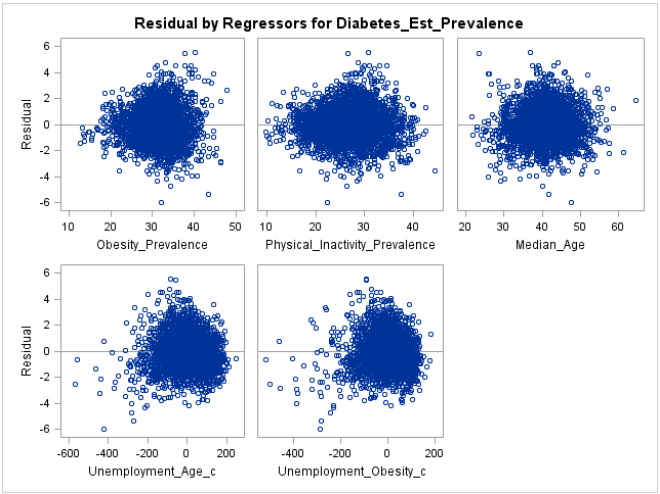
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	B	-1.40852	0.47316	-2.98	0.0029
Unemployment_Rate_c	B	0.74093	0.10390	7.13	<.0001
Median_Income	B	-0.00001443	0.00000280	-5.15	<.0001
Adults_No_HighSchool_Diploma	B	0.01890	0.00572	3.30	0.0010
Adults_HighSchool_Diploma_Only	B	-0.03375	0.00505	-6.68	<.0001
Adults_Associates_Degree	B	-0.05521	0.00617	-8.95	<.0001
Adults_Bachelors_Or_Higher	0	0	.	.	.
Obesity_Prevalence	B	0.21126	0.00793	26.64	<.0001
Physical_Inactivity_Prevalence	B	0.17395	0.00742	23.43	<.0001
Median_Age	B	0.11616	0.00528	22.01	<.0001
Unemployment_Age_c	B	-0.01430	0.00171	-8.37	<.0001
Unemployment_Obesity_c	B	-0.01282	0.00217	-5.89	<.0001

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

Train/Test Split

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	DIABETES
Random Number Seed	807231000
Sampling Rate	0.8
Sample Size	2507
Selection Probability	0.800192
Sampling Weight	0
Output Data Set	DIABETES_SPLIT

3.9 Adjusted R-Square Model Selection Top Results

Number in Model	Adjusted R-Square	R-Square	Variables in Model
10	0.7316	0.7327	Unemployment_Rate_c Median_Income Adults_No_HighSchool_Diploma Adults_HighSchool_Diploma_Only Adults_Bachelors_Or_Higher Obesity_Prevalence Physical_Inactivity_Prevalence Median_Age Unemployment_Age_c Unemployment_Obesity_c
10	0.7316	0.7327	Unemployment_Rate_c Median_Income Adults_No_HighSchool_Diploma Adults_Associates_Degree Adults_Bachelors_Or_Higher Obesity_Prevalence Physical_Inactivity_Prevalence Median_Age Unemployment_Age_c Unemployment_Obesity_c
10	0.7316	0.7327	Unemployment_Rate_c Median_Income Adults_No_HighSchool_Diploma Adults_HighSchool_Diploma_Only Adults_Associates_Degree Obesity_Prevalence Physical_Inactivity_Prevalence Median_Age Unemployment_Age_c Unemployment_Obesity_c
10	0.7316	0.7327	Unemployment_Rate_c Median_Income Adults_HighSchool_Diploma_Only Adults_Associates_Degree Adults_Bachelors_Or_Higher Obesity_Prevalence Physical_Inactivity_Prevalence Median_Age Unemployment_Age_c Unemployment_Obesity_c
9	0.7308	0.7317	Unemployment_Rate_c Median_Income Adults_HighSchool_Diploma_Only Adults_Associates_Degree Obesity_Prevalence Physical_Inactivity_Prevalence Median_Age Unemployment_Age_c Unemployment_Obesity_c
9	0.7304	0.7314	Unemployment_Rate_c Median_Income Adults_No_HighSchool_Diploma Adults_Bachelors_Or_Higher Obesity_Prevalence Physical_Inactivity_Prevalence Median_Age Unemployment_Age_c Unemployment_Obesity_c

3.10: Final Model Regression Results

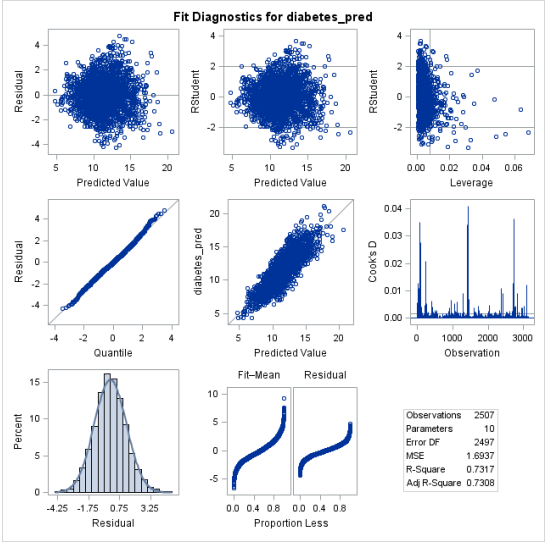
Number of Observations Read	3133
Number of Observations Used	2507
Number of Observations with Missing Values	626

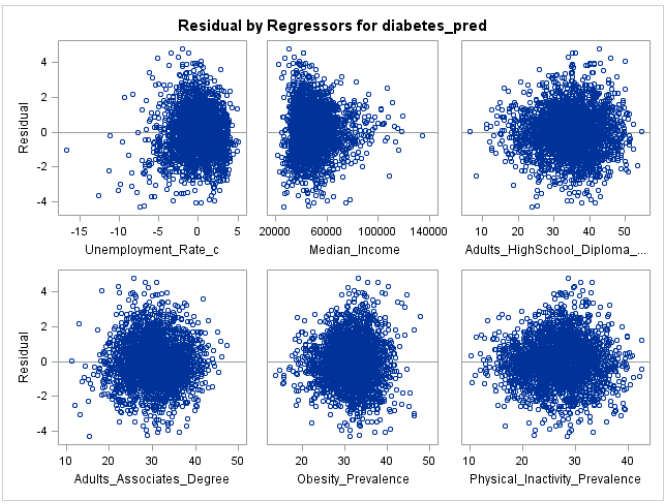
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	11535	1281.71119	756.75	<.0001
Error	2497	4229.16001	1.69370		
Corrected Total	2506	15765			

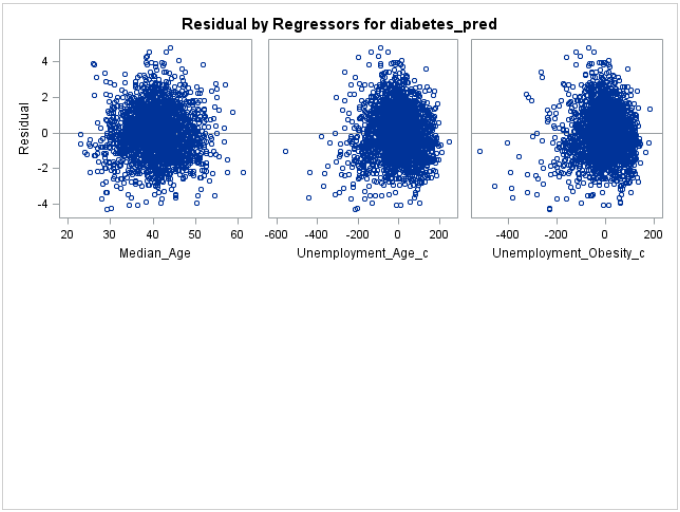
Root MSE	1.30142	R-Square	0.7317
Dependent Mean	11.36881	Adj R-Sq	0.7308
Coeff Var	11.44730		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-0.45459	0.42860	-1.06	0.2890	0
Unemployment_Rate_c	1	0.58497	0.12231	4.78	<.0001	0.52281
Median_Income	1	-0.00001793	0.00000273	-6.56	<.0001	-0.09348
Adults_HighSchool_Diploma_Only	1	-0.03957	0.00535	-7.40	<.0001	-0.11296
Adults_Associates_Degree	1	-0.06875	0.00586	-11.73	<.0001	-0.14130
Obesity_Prevalence	1	0.21275	0.00882	24.12	<.0001	0.38078
Physical_Inactivity_Prevalence	1	0.17862	0.00795	22.46	<.0001	0.36892
Median_Age	1	0.11424	0.00579	19.73	<.0001	0.23698
Unemployment_Age_c	1	-0.01250	0.00198	-6.31	<.0001	-0.45479
Unemployment_Obesity_c	1	-0.01091	0.00247	-4.43	<.0001	-0.32187

3.11: Final Model Diagnostics/Residuals







3.12: Final Model Error

Validation statistics for model

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	626	1.30504	1.03168

3.13: Test Set R-Square

2 Variables:	Diabetes_Est_Prevalence yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Diabetes_Est_Prevalence	626	11.36262	2.49738	7113	3.00000	20.00000	
yhat	626	11.30471	2.16505	7077	4.28860	20.49000	Predicted Value of diabetes_pred

Pearson Correlation Coefficients, N = 626
Prob > |r| under H0: Rho=0

	Diabetes_Est_Prevalence	yhat
Diabetes_Est_Prevalence	1.00000	0.85272 <.0001
yhat Predicted Value of diabetes_pred	0.85272 <.0001	1.00000

3.14: Two Predictions Results

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	11.1072	0.0451	11.0188	11.1955	8.5537	13.6607	.
2	.	11.1968	0.0456	11.1073	11.2863	8.6432	13.7503	.

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.

IV. Ramkumar Perumal

V. Outputs

4.1

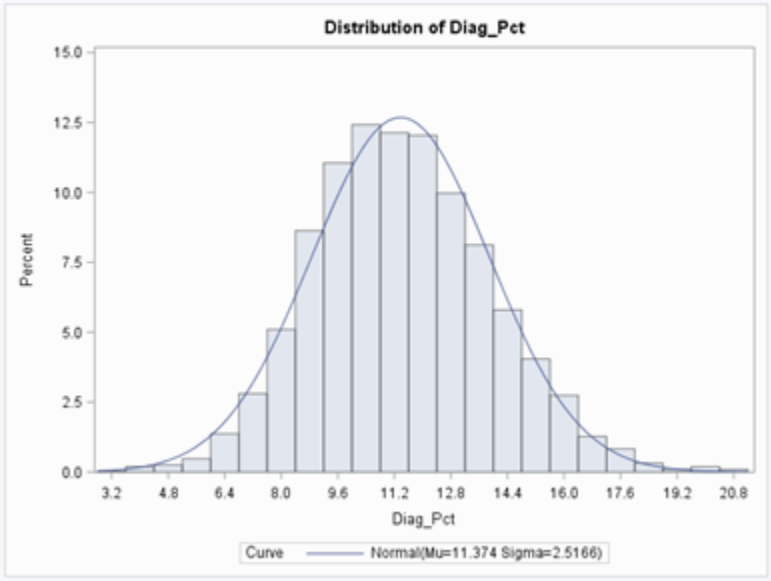
Diabetes Prevalence Dataset

Obs	State	County	Diag_Pct	Unemp_Rate	Income	lt_HighSchool	only_HighSchool	Col_Degree	Bach_Degree	Obesity	Ph_Inactivity	Age	region	d_Northeast	d_Midwest	d_South
1	Alabama	Autauga County	12.4	5.9	54487	12.417	34.331	28.66	24.593	36.4	30.9	37.9	South	0	0	1
2	Alabama	Baldwin County	11.1	6.1	56480	9.972	28.692	31.788	29.547	29.3	24.5	41.8	South	0	0	1
3	Alabama	Barbour County	18.2	10.6	32884	26.236	34.927	25.969	12.868	44.2	30.6	38.3	South	0	0	1
4	Alabama	Bibb County	14.6	7.2	43079	19.302	41.816	26.883	12	38.4	37.5	40	South	0	0	1
5	Alabama	Blount County	14.4	6.1	47213	19.969	32.942	34.039	13.05	35.8	29	40.2	South	0	0	1
6	Alabama	Bullock County	19.5	8.8	34278	33.438	34.147	22.16	10.255	40.2	29.8	38.5	South	0	0	1
7	Alabama	Butler County	17.5	8.6	35409	18.94	39.852	25.127	16.08	36	33.9	40.5	South	0	0	1
8	Alabama	Calhoun County	16.3	8	41778	17.663	32.072	32.612	17.653	36.1	32.5	38.8	South	0	0	1

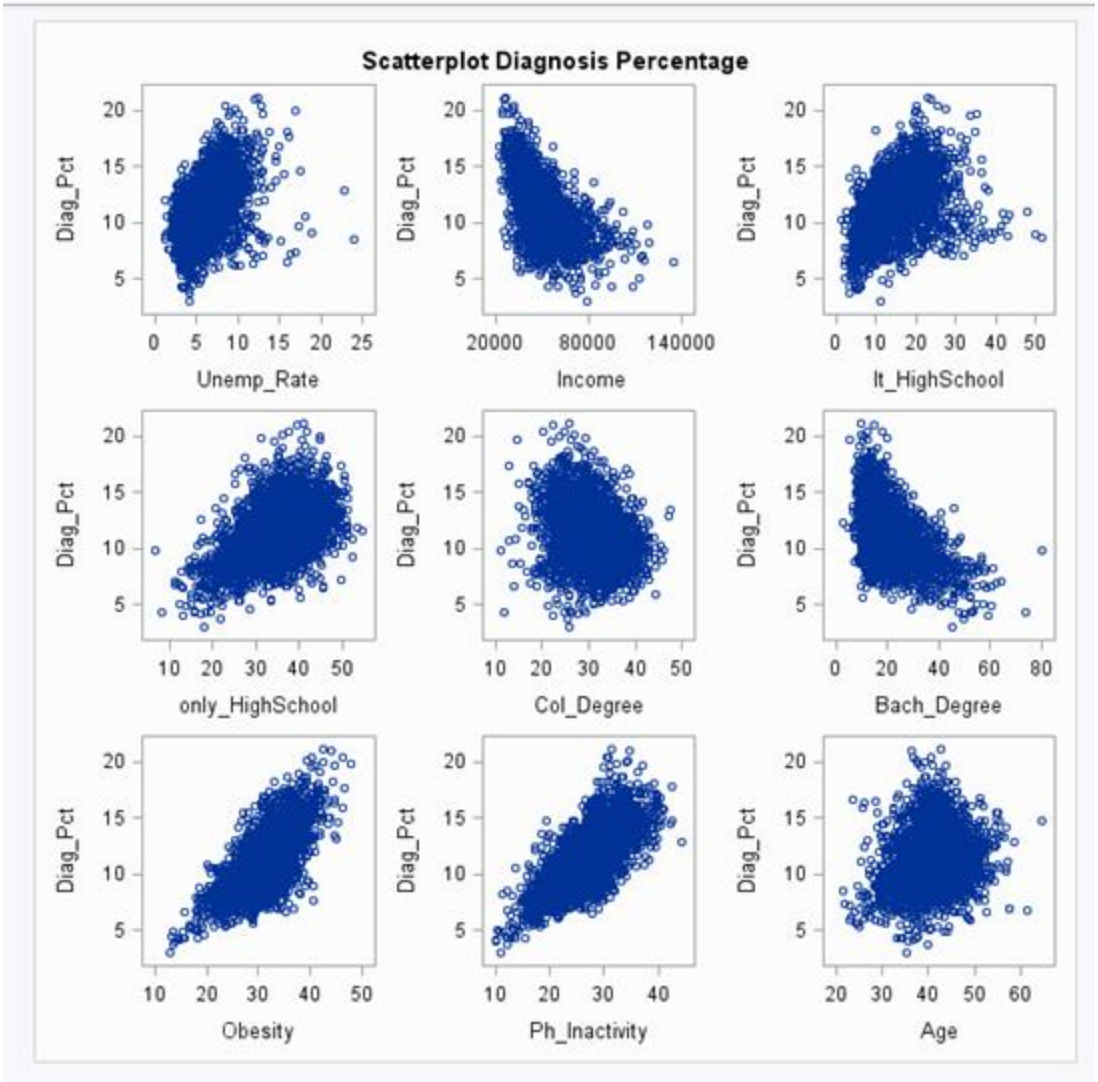
4.2

Explore the Distribution for Diagnosis percentage

The UNIVARIATE Procedure



4.3



4.4

Pearson Correlation Coefficients, N = 3139
Prob > |r| under H0: Rho=0

	Diag_Pct	Unemp_Rate	Income	lt_HighSchool	only_HighSchool	Col_Degree	Bach_Degree	Obesity	Ph_Inactivity	Age	d_Northeast	d_Midwest	d_South
Diag_Pct	1.00000	0.45995 <.0001	-0.57082 <.0001	0.47694 <.0001	0.50158 <.0001	-0.29975 <.0001	-0.56141 <.0001	0.68145 <.0001	0.72708 <.0001	0.19103 <.0001	-0.14587 <.0001	-0.17788 <.0001	0.53225 <.0001

4.5

Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

Bach_Degree =	100 * Intercept - 1.69E-6 * Unemp_Rate - 488E-12 * Income - 1 * lt_HighSchool - 1 * only_HighSchool - 1 * Col_Degree + 4.23E-6 * Obesity - 5.3E-6 * Ph_Inactivity + 1.38E-6 * Age + 5. * d_South

4.6

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	-2.99125	0.45257	-6.61	<.0001	.	0
Unemp_Rate	1	0.25260	0.01174	21.51	<.0001	0.67883	1.47313
Income	1	-0.00001615	0.00000262	-6.16	<.0001	0.42311	2.36345
lt_HighSchool	1	-0.01099	0.00555	-1.98	0.0476	0.36594	2.73267
only_HighSchool	1	-0.00562	0.00484	-1.16	0.2458	0.41144	2.43047
Col_Degree	1	-0.02756	0.00622	-4.43	<.0001	0.47434	2.10817
Obesity	1	0.20298	0.00773	26.27	<.0001	0.39693	2.51937
Ph_Inactivity	1	0.11036	0.00738	14.96	<.0001	0.33035	3.02710
Age	1	0.11204	0.00494	22.69	<.0001	0.72657	1.37633
d_Northeast	1	0.31134	0.11605	2.68	0.0073	0.55586	1.79903
d_Midwest	1	0.30456	0.08407	3.62	0.0003	0.30558	3.27243
d_South	1	1.63730	0.08963	18.27	<.0001	0.24199	4.13240

4.7

Linear regression full model

The REG Procedure
Model: MODEL1
Dependent Variable: Diag_Pct

Number of Observations Read	2982
Number of Observations Used	2982

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	14027	1275.18559	1132.29	<.0001
Error	2970	3344.81416	1.12620		
Corrected Total	2981	17372			

Root MSE	1.06123	R-Square	0.8075
Dependent Mean	11.35255	Adj R-Sq	0.8067
Coeff Var	9.34791		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	-3.30488	0.40430	-8.17	<.0001	0	0
Unemp_Rate	1	0.29437	0.01132	26.01	<.0001	0.25419	1.47290
Income	1	-0.00001168	0.00000238	-4.90	<.0001	-0.06099	2.39301
lt_HighSchool	1	-0.00780	0.00522	-1.49	0.1353	-0.02016	2.80826
only_HighSchool	1	-0.00455	0.00438	-1.04	0.2992	-0.01320	2.49308
Col_Degree	1	-0.02424	0.00557	-4.35	<.0001	-0.05066	2.09286
Obesity	1	0.19116	0.00703	27.18	<.0001	0.34841	2.53532
Ph_Inactivity	1	0.12001	0.00660	18.19	<.0001	0.25708	3.08132
Age	1	0.10865	0.00455	23.86	<.0001	0.22653	1.39084
d_Northeast	1	0.19966	0.10470	1.91	0.0566	0.02107	1.88353
d_Midwest	1	0.19852	0.07683	2.58	0.0098	0.03905	3.52211
d_South	1	1.50303	0.08196	18.34	<.0001	0.31023	4.41457

4.8

Train Set

Obs	Selected	State	County	Diag_Pct	Unemp_Rate	Income	H_HighSchool	only_HighSchool	Col_Degree	Bach_Degree	Obesity	Ph_Inactivity	Age	region	d_Northeast	d_Midwest	d_South	new_y
1	0	Alabama	Autauga County	12.4	5.9	54467	12.417	34.331	28.66	24.550	36.4	38.9	37.9	South	0	0	1	
2	0	Alabama	Baldwin County	11.1	6.1	55460	9.572	28.652	31.788	29.547	29.3	24.5	41.8	South	0	0	1	
3	1	Alabama	Barbour County	18.2	10.6	32664	26.236	34.507	25.969	12.968	44.2	30.6	38.3	South	0	0	1	18.2
4	1	Alabama	Bibb County	14.6	7.2	43079	15.332	41.816	26.883	12	39.4	37.5	40	South	0	0	1	14.6
5	0	Alabama	Blount County	14.4	6.1	47213	10.993	32.942	34.839	13.05	35.8	29	40.2	South	0	0	1	
6	1	Alabama	Bullock County	17.5	8.6	35409	19.54	39.852	25.127	16.08	38	33.9	40.5	South	0	0	1	17.5
7	0	Alabama	Calhoun County	16.3	8	41779	17.683	32.072	32.612	17.653	36.1	32.5	39.8	South	0	0	1	
8	0	Alabama	Chambers County	16	6.7	39630	19.737	39.411	26.368	12.484	38	32.6	43.1	South	0	0	1	
9	1	Alabama	Cherokee County	12.9	5.8	41456	18.717	35.958	31.383	13.962	35.5	35.9	45.2	South	0	0	1	12.9
10	0	Alabama	Chilton County	14.1	6.3	44188	19.562	41.254	23.589	14.855	36	29.8	38.3	South	0	0	1	
11	1	Alabama	Choctaw County	16.6	9	32691	21.167	37.257	29.639	11.96	39.5	33.1	44.6	South	0	0	1	16.6
12	1	Alabama	Clarke County	16.8	12.2	34061	18.958	45.118	23.787	12.137	38.4	29.4	41.1	South	0	0	1	16.8
13	1	Alabama	Clay County	14.4	7.2	38512	20.386	34.575	28.975	11.062	39.1	31.6	43.6	South	0	0	1	14.4
14	1	Alabama	Cleburne County	12.7	6.6	43403	20.815	37.97	24.688	11.028	34.6	32.2	41.3	South	0	0	1	12.7
15	0	Alabama	Coffee County	15.5	6.7	48632	14.864	29.344	32.126	23.667	34.2	29.8	38.2	South	0	0	1	

4.9

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	8301.34595	922.37177	832.07	<.0001
Error	1780	1973.18715	1.10853		
Corrected Total	1789	10275			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-3.64969	0.44013	76.22488	68.76	<.0001
Unemp_Rate	0.28902	0.01424	456.90881	412.17	<.0001
Income	-0.00000956	0.00000260	15.00917	13.54	0.0002
Col_Degree	-0.02113	0.00609	13.36374	12.06	0.0005
Obesity	0.19267	0.00871	542.30930	489.21	<.0001
Ph_Inactivity	0.11524	0.00804	227.63526	205.35	<.0001
Age	0.10787	0.00532	455.75688	411.14	<.0001
d_Northeast	0.35559	0.13085	8.18730	7.39	0.0066
d_Midwest	0.25602	0.09674	7.76327	7.00	0.0082
d_South	1.47598	0.10318	226.81913	204.61	<.0001

Bounds on condition number: 4.2718, 185.6

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	only_HighSchool	10	0.0000	0.8081	10.1236	0.12	0.7252
2	lt_HighSchool	9	0.0002	0.8080	9.9091	1.79	0.1815

4.10

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	Ph_Inactivity		1	0.5641	0.5641	2254.19	2313.64	<.0001
2	Unemp_Rate		2	0.1142	0.6783	1197.59	634.47	<.0001
3	Obesity		3	0.0431	0.7214	800.306	276.15	<.0001
4	d_South		4	0.0345	0.7558	482.913	251.94	<.0001
5	Age		5	0.0477	0.8036	42.7113	433.29	<.0001
6	Col_Degree		6	0.0023	0.8059	23.1295	21.39	<.0001
7	Income		7	0.0011	0.8070	15.2065	9.88	0.0017

4.11

Linear regression on selected models - Test

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	2982
Number of Observations Used	1790
Number of Observations with Missing Values	1192

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	8301.34595	922.37177	832.07	<.0001
Error	1780	1973.18715	1.10853		
Corrected Total	1789	10275			

Root MSE	1.05287	R-Square	0.8080
Dependent Mean	11.39693	Adj R-Sq	0.8070
Coeff Var	9.23818		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-3.64969	0.44013	-8.29	<.0001	0
Unemp_Rate	1	0.28902	0.01424	20.30	<.0001	0.25102
Income	1	-0.00000956	0.00000260	-3.68	0.0002	-0.05181
Col_Degree	1	-0.02113	0.00609	-3.47	0.0005	-0.04418
Obesity	1	0.19267	0.00871	22.12	<.0001	0.35266
Ph_Inactivity	1	0.11524	0.00804	14.33	<.0001	0.24944
Age	1	0.10787	0.00532	20.28	<.0001	0.22740
d_Northeast	1	0.35559	0.13085	2.72	0.0066	0.03852
d_Midwest	1	0.25602	0.09674	2.65	0.0082	0.05071
d_South	1	1.47598	0.10318	14.30	<.0001	0.30709

4.12

Linear regression on selected models - Test

The REG Procedure
Model: MODEL2
Dependent Variable: new_y

Number of Observations Read	2982
Number of Observations Used	1790
Number of Observations with Missing Values	1192

Analysis of Variance

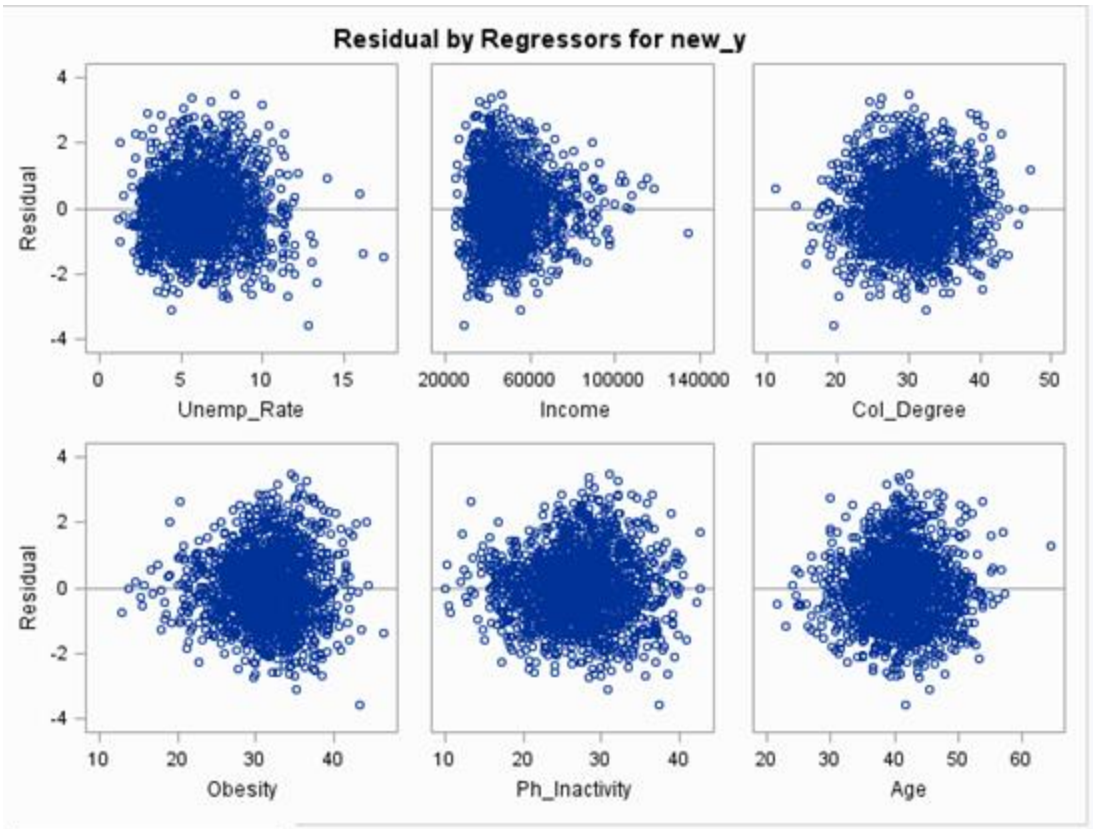
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	8291.03901	1184.43414	1064.11	<.0001
Error	1782	1983.49409	1.11307		
Corrected Total	1789	10275			

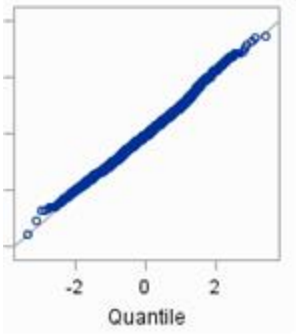
Root MSE	1.05502	R-Square	0.8070
Dependent Mean	11.39693	Adj R-Sq	0.8062
Coeff Var	9.25708		

Parameter Estimates

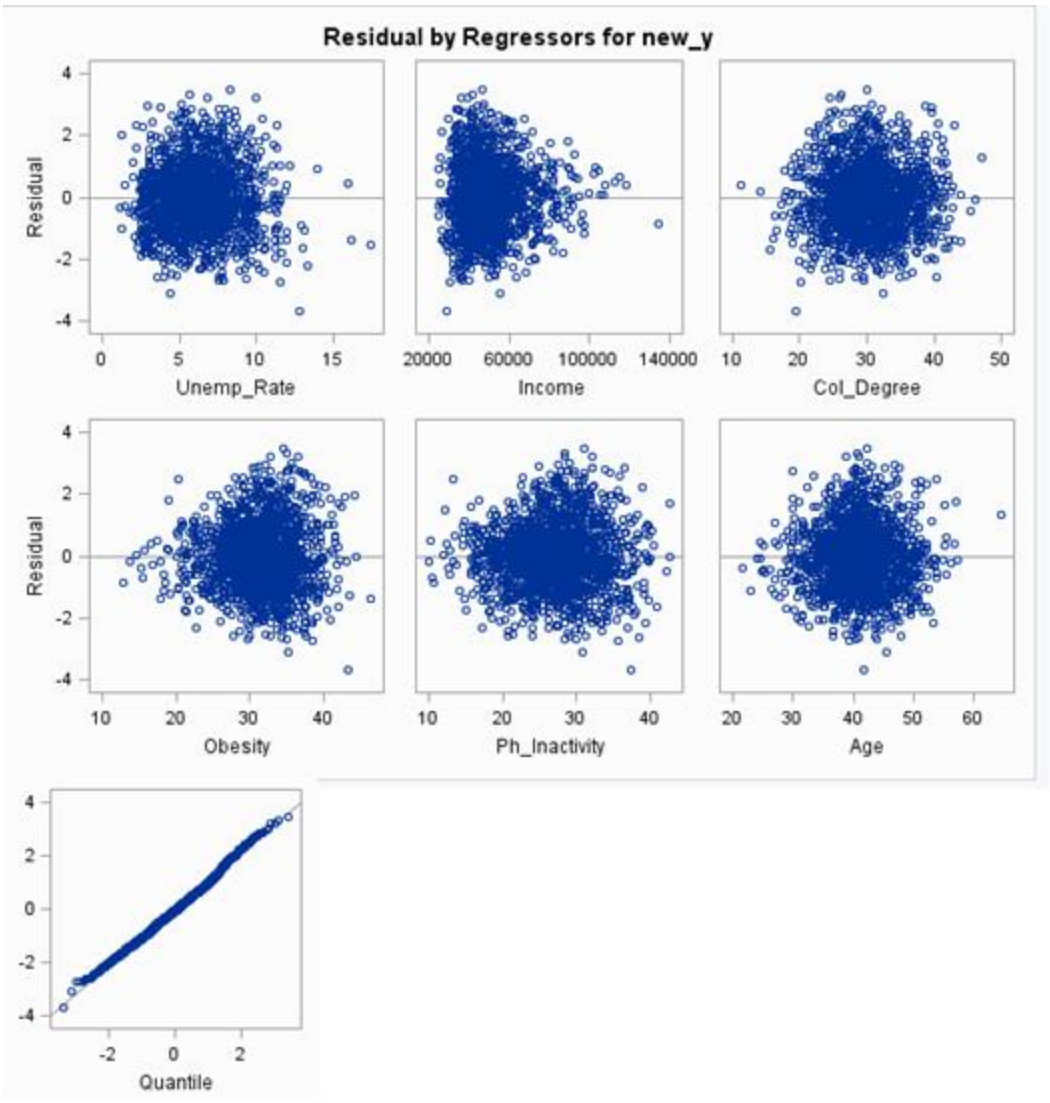
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-3.65871	0.43523	-8.41	<.0001	0
Unemp_Rate	1	0.28301	0.01395	20.29	<.0001	0.24580
Income	1	-0.00000801	0.00000255	-3.14	0.0017	-0.04342
Col_Degree	1	-0.02695	0.00561	-4.80	<.0001	-0.05635
Obesity	1	0.19868	0.00823	24.14	<.0001	0.36368
Ph_Inactivity	1	0.11878	0.00797	14.90	<.0001	0.25711
Age	1	0.11003	0.00528	20.83	<.0001	0.23195
d_South	1	1.23042	0.06077	20.25	<.0001	0.25600

4.13


Residual by Regressors for new_y
Residual
Unemp_Rate
Income
Col_Degree
Obesity
Ph_Inactivity
Age


Quantile

4.14


Residual by Regressors for new_y
Residual
Unemp_Rate
Income
Col_Degree
Obesity
Ph_Inactivity
Age
Quantile

4.15

Validation Model1 - Test Set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	2982
Number of Observations Used	1790
Number of Observations with Missing Values	1192

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	8301.34595	922.37177	832.07	<.0001
Error	1780	1973.18715	1.10853		
Corrected Total	1789	10275			

Root MSE	1.05287	R-Square	0.8080
Dependent Mean	11.39693	Adj R-Sq	0.8070
Coeff Var	9.23818		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.64969	0.44013	-8.29	<.0001
Unemp_Rate	1	0.28902	0.01424	20.30	<.0001
Income	1	-0.00000956	0.00000260	-3.68	0.0002
Col_Degree	1	-0.02113	0.00609	-3.47	0.0005
Obesity	1	0.19267	0.00871	22.12	<.0001
Ph_Inactivity	1	0.11524	0.00804	14.33	<.0001
Age	1	0.10787	0.00532	20.28	<.0001
d_Northeast	1	0.35559	0.13085	2.72	0.0066
d_Midwest	1	0.25602	0.09674	2.65	0.0082
d_South	1	1.47598	0.10318	14.30	<.0001

4.16

Validation Model2 - Test Set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	2982
Number of Observations Used	1790
Number of Observations with Missing Values	1192

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	8291.03901	1184.43414	1064.11	<.0001
Error	1782	1983.49409	1.11307		
Corrected Total	1789	10275			

Root MSE	1.05502	R-Square	0.8070
Dependent Mean	11.39693	Adj R-Sq	0.8062
Coeff Var	9.25708		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.65871	0.43523	-8.41	<.0001
Unemp_Rate	1	0.28301	0.01395	20.29	<.0001
Income	1	-0.00000801	0.00000255	-3.14	0.0017
Col_Degree	1	-0.02695	0.00561	-4.80	<.0001
Obesity	1	0.19868	0.00823	24.14	<.0001
Ph_Inactivity	1	0.11878	0.00797	14.90	<.0001
Age	1	0.11003	0.00528	20.83	<.0001
d_South	1	1.23042	0.06077	20.25	<.0001

4.17

Validation statistics for Model1

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	1192	1.07739	0.86452

Validation statistics for Model1

The CORR Procedure

2 Variables:	Diag_Pct yhat

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Diag_Pct	1192	11.28591	2.43962	13453	3.80000	19.70000	
yhat	1192	11.27231	2.14230	13437	4.02337	18.64746	Predicted Value of new_y

Pearson Correlation Coefficients, N = 1192 Prob > |r| under H0: Rho=0		
	Diag_Pct	yhat
Diag_Pct	1.00000	0.89741 <.0001
yhat Predicted Value of new_y	0.89741 <.0001	1.00000

4.18

Validation statistics for Model2

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	1192	1.07650	0.86248

Validation statistics for Model2

The CORR Procedure

2 Variables:	Diag_Pct yhat

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Diag_Pct	1192	11.28591	2.43962	13453	3.80000	19.70000	
yhat	1192	11.27495	2.13702	13440	4.14473	18.76007	Predicted Value of new_y

Pearson Correlation Coefficients, N = 1192
Prob > |r| under H0: Rho=0

	Diag_Pct	yhat
Diag_Pct	1.00000	0.89764 <.0001
yhat Predicted Value of new_y	0.89764 <.0001	1.00000

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
Published: November 7, 2017

VI. Charles Saporito

5.1

Breakdown of Prevalence

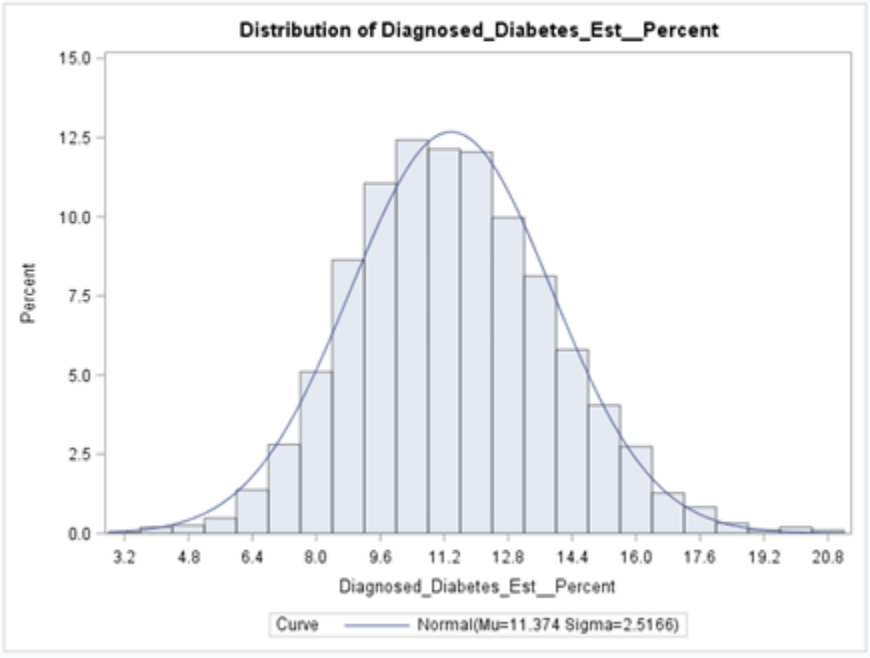
The MEANS Procedure

Analysis Variable : Diagnosed_Diabetes_Est__Percent							
Mean	Std Dev	Std Error	Lower 95% CL for Mean	Upper 95% CL for Mean	25th Pctl	50th Pctl	75th Pctl
11.3740045	2.5166089	0.0449180	11.2859329	11.4620760	9.6000000	11.3000000	13.0000000

5.2

Distribution of Prevalence

The UNIVARIATE Procedure

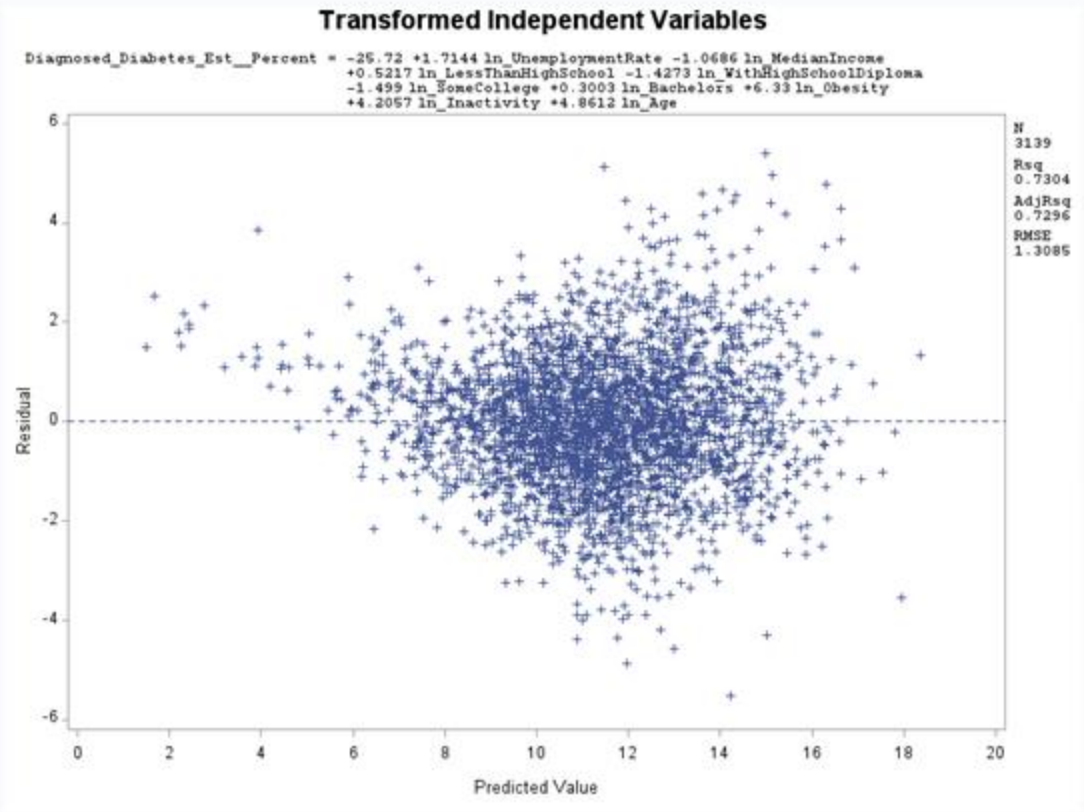


5.3

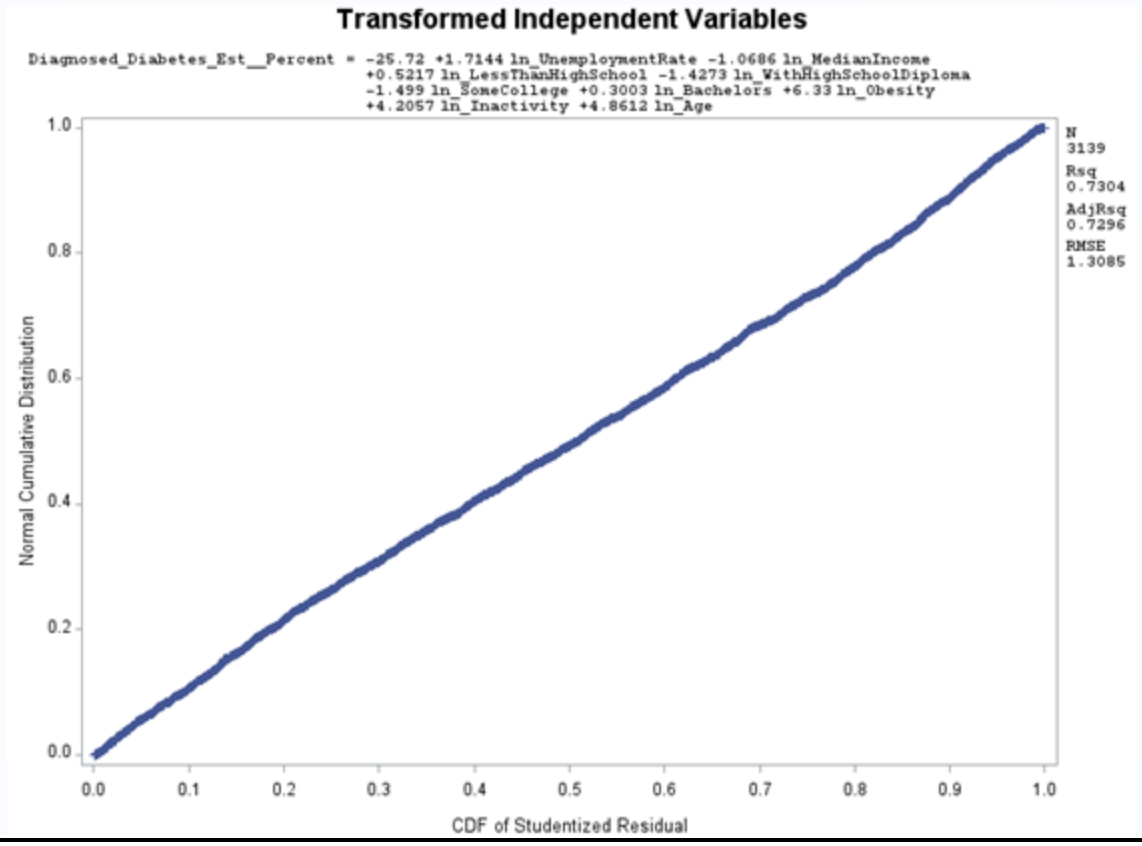
Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-25.71991	2.50966	-10.25	<.0001	0
ln_UnemploymentRate	1	1.71438	0.07605	22.54	<.0001	1.42510
ln_MedianIncome	1	-1.06861	0.14768	-7.24	<.0001	2.35022
ln_LessThanHighSchool	1	0.52173	0.10380	5.03	<.0001	4.37230
ln_WithHighSchoolDiploma	1	-1.42726	0.21207	-6.73	<.0001	4.26881
ln_SomeCollege	1	-1.49899	0.17305	-8.66	<.0001	1.69596
ln_Bachelors	1	0.30027	0.14663	2.05	0.0407	6.30528
ln_Obesity	1	6.32995	0.23993	26.38	<.0001	2.47622
ln_Inactivity	1	4.20570	0.19956	21.08	<.0001	3.06996
ln_Age	1	4.86121	0.20850	23.31	<.0001	1.38705

VIF shown in the above output

5.4


Transformed Independent Variables
Diagnosed_Diabetes_Est__Percent = -25.72 +1.7144 ln_UnemploymentRate -1.0686 ln_MedianIncome
+0.5217 ln_LessThanHighSchool -1.4273 ln_WithHighSchoolDiploma
-1.499 ln_SomeCollege +0.3003 ln_Bachelors +6.33 ln_Obesity
+4.2057 ln_Inactivity +4.8612 ln_Age
N
3139
Rsq
0.7304
AdjRsq
0.7296
RMSE
1.3085
Residual
Predicted Value
6
4
2
0
-2
-4
-6
0
2
4
6
8
10
12
14
16
18
20

The REG Procedure



5.5

Transformed Independent Variables

The REG Procedure
Model: MODEL1
Dependent Variable: Diagnosed_Diabetes_Est__Percent

Number of Observations Read	3139
Number of Observations Used	3139

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	14516	1612.91621	941.97	<.0001
Error	3129	5357.71285	1.71228		
Corrected Total	3138	19874			

Root MSE	1.30854	R-Square	0.7304
Dependent Mean	11.37400	Adj R-Sq	0.7296
Coeff Var	11.50465		

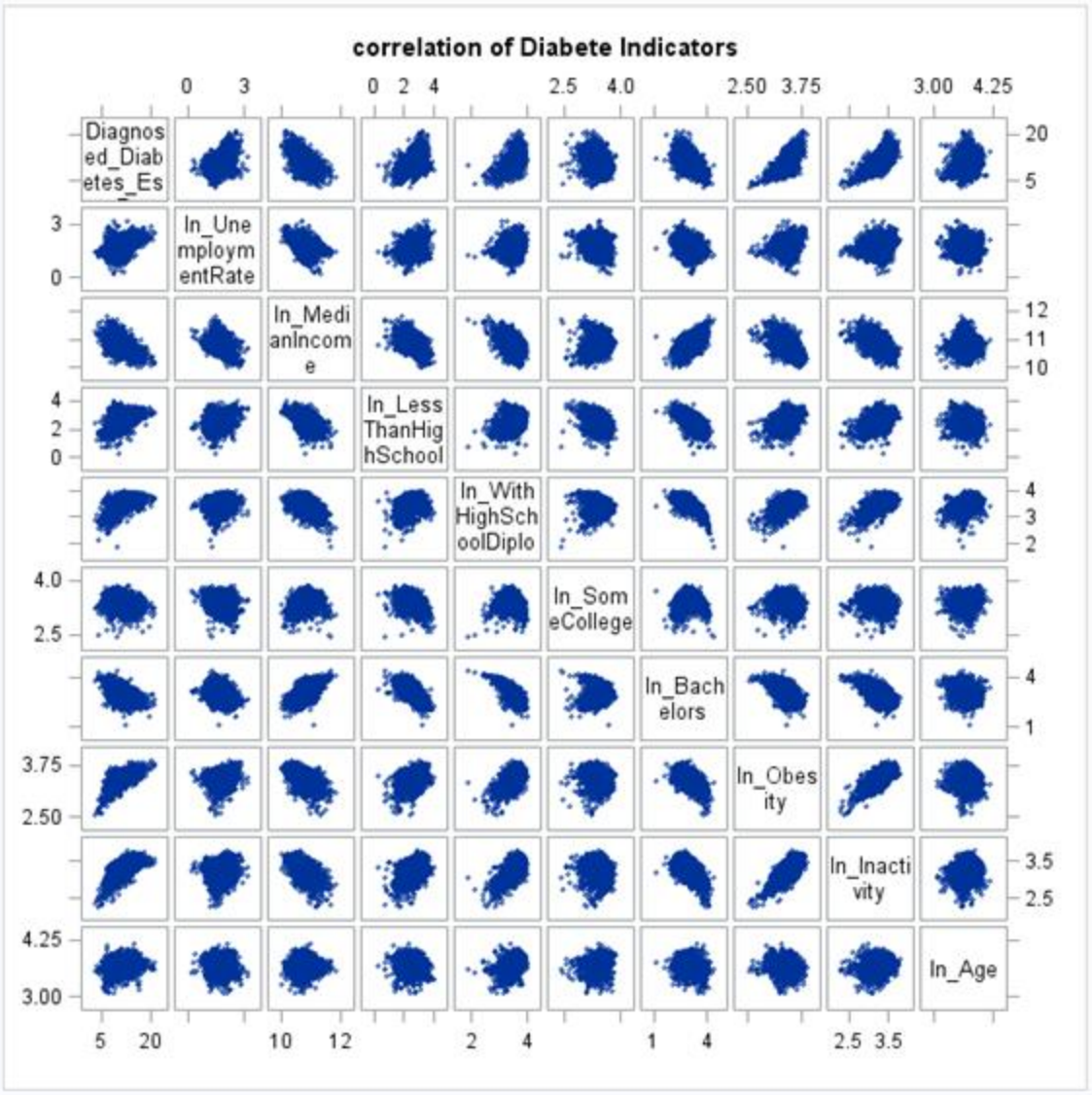

correlation of Diabete Indicators
Diagnosed_Diabetes_Es
ln_UnemploymentRate
ln_MedianIncome
ln_LessThanHighSchool
ln_WithHighSchoolDiplo
ln_SomeCollege
ln_Bachelors
ln_Obesity
ln_Inactivity
ln_Age

Figure 5.6

Regression Model to Determine Influential Points and Outliers

The REG Procedure
Model: MODEL1
Dependent Variable: Diagnosed_Diabetes_Est__Percent

Number of Observations Read	3139
Number of Observations Used	3139

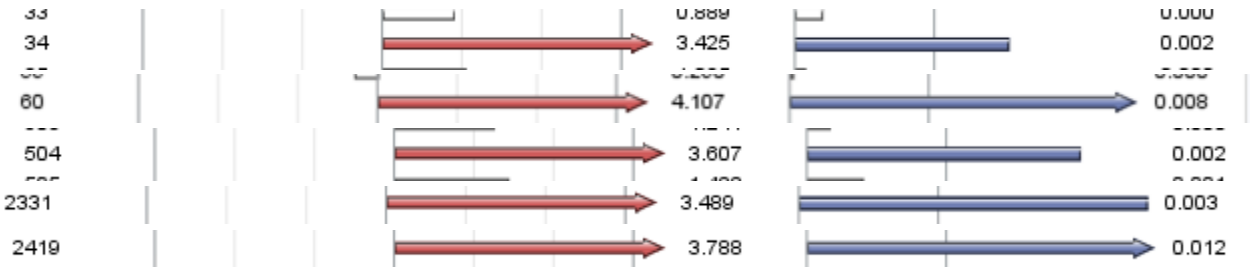
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	14516	1612.91621	941.97	<.0001
Error	3129	5357.71285	1.71228		
Corrected Total	3138	19874			

Root MSE	1.30854	R-Square	0.7304
Dependent Mean	11.37400	Adj R-Sq	0.7296
Coeff Var	11.50465		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-25.71991	2.50966	-10.25	<.0001
ln_UnemploymentRate	1	1.71438	0.07605	22.54	<.0001
ln_MedianIncome	1	-1.06861	0.14768	-7.24	<.0001
ln_LessThanHighSchool	1	0.52173	0.10380	5.03	<.0001
ln_WithHighSchoolDiploma	1	-1.42726	0.21207	-6.73	<.0001
ln_SomeCollege	1	-1.49899	0.17305	-8.66	<.0001
ln_Bachelors	1	0.30027	0.14663	2.05	0.0407
ln_Obesity	1	6.32995	0.23993	26.38	<.0001
ln_Inactivity	1	4.20570	0.19956	21.08	<.0001
ln_Age	1	4.86121	0.20850	23.31	<.0001



Without Outliers in Data

The REG Procedure
Model: MODEL1
Dependent Variable: Diagnosed_Diabetes_Est__Percent

Number of Observations Read	3136
Number of Observations Used	3136

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	14412	1601.33678	948.63	<.0001
Error	3126	5276.83677	1.68805		
Corrected Total	3135	19689			

Root MSE	1.29925	R-Square	0.7320
Dependent Mean	11.36668	Adj R-Sq	0.7312
Coeff Var	11.43033		

Figure 5.7

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-26.63580	2.49581	-10.67	<.0001	0
ln_UnemploymentRate	1	1.71280	0.07552	22.68	<.0001	0.25056
ln_MedianIncome	1	-1.00799	0.14696	-6.86	<.0001	-0.09738
ln_LessThanHighSchool	1	0.53148	0.10308	5.16	<.0001	0.09977
ln_WithHighSchoolDiploma	1	-1.40325	0.21072	-6.66	<.0001	-0.12744
ln_SomeCollege	1	-1.49085	0.17197	-8.67	<.0001	-0.10450
ln_Bachelors	1	0.30116	0.14571	2.07	0.0388	0.04807
ln_Obesity	1	6.27847	0.23836	26.34	<.0001	0.38360
ln_Inactivity	1	4.22710	0.19820	21.33	<.0001	0.34598
ln_Age	1	4.92275	0.20746	23.73	<.0001	0.25858

Interaction Terms

The GLM Procedure

Dependent Variable: Diagnosed_Diabetes_Est__Percent

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	2601.135261	867.045087	322.48	<.0001
Error	780	2097.176052	2.688687		
Corrected Total	783	4698.311314			

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-180.4828506	32.52451863	-5.55	<.0001
ln_MedianIncome	15.5770145	2.98922412	5.21	<.0001
ln_Inactivity	67.2152526	9.99546371	6.72	<.0001
ln_Median*ln_Inactiv	-5.5642249	0.92116196	-6.04	<.0001

Interaction Variable Multicolinearity

The CORR Procedure

3 Variables:	Diagnosed_Diabetes_Est__Percent ln_MedianIncome ln_Inactivity

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Diagnosed_Diabetes_Est__Percent	784	11.22717	2.44957	8802	3.80000	19.70000
ln_MedianIncome	784	10.78357	0.24208	8454	10.10479	11.67874
ln_Inactivity	784	3.26912	0.20313	2563	2.33214	3.74005

Pearson Correlation Coefficients, N = 784 Prob > |r| under H0: Rho=0			
	Diagnosed_Diabetes_Est__Percent	ln_MedianIncome	ln_Inactivity
Diagnosed_Diabetes_Est__Percent	1.00000	-0.55752 <.0001	0.70327 <.0001
ln_MedianIncome	-0.55752 <.0001	1.00000	-0.56323 <.0001
ln_Inactivity	0.70327 <.0001	-0.56323 <.0001	1.00000

Test and Train Sets for Diabetes Prevalence

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	PREVALENCE
Random Number Seed	71
Sampling Rate	0.75
Sample Size	2352
Selection Probability	0.75
Sampling Weight	0
Output Data Set	XV_ALL

Model Selection

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	3136
Number of Observations Used	2352
Number of Observations with Missing Values	784

Stepwise Selection: Step 8

Variable ln_LessThanHighSchool Entered: R-Square = 0.7474 and C(p) = 8.1129

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	11189	1398.67633	866.77	<.0001
Error	2343	3780.80074	1.61366		
Corrected Total	2351	14970			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-22.06033	2.56530	119.33289	73.95	<.0001
ln_UnemploymentRate	1.78706	0.08496	713.93282	442.43	<.0001
ln_MedianIncome	-1.04023	0.16342	65.38534	40.52	<.0001
ln_LessThanHighSchool	0.33685	0.08877	23.23352	14.40	0.0002
ln_WithHighSchoolDiploma	-1.73759	0.17364	161.58084	100.13	<.0001
ln_SomeCollege	-1.75322	0.17925	154.37457	95.67	<.0001
ln_Obesity	6.11404	0.26997	827.62848	512.89	<.0001
ln_Inactivity	4.38427	0.22405	617.92290	382.93	<.0001
ln_Age	4.70253	0.23759	632.13722	391.74	<.0001

Model Selection

The REG Procedure
Model: MODEL2
Dependent Variable: new_y

C(p) Selection Method

Number of Observations Read	3136
Number of Observations Used	2352
Number of Observations with Missing Values	784

Number in Model	C(p)	R-Square	Variables in Model
8	8.1129	0.7474	ln_UnemploymentRate ln_MedianIncome ln_LessThanHighSchool ln_WithHighSchoolDiploma ln_SomeCollege ln_Obesity ln_Inactivity ln_Age
9	10.0000	0.7475	ln_UnemploymentRate ln_MedianIncome ln_LessThanHighSchool ln_WithHighSchoolDiploma ln_SomeCollege ln_Bachelors ln_Obesity ln_Inactivity ln_Age

Validation of Model

The REG Procedure
Model: MODEL1
Dependent Variable: Diagnosed_Diabetes_Est__Percent

Number of Observations Read	3136
Number of Observations Used	3136

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	14412	1601.33678	948.63	<.0001
Error	3126	5276.83677	1.68805		
Corrected Total	3135	19689			

Root MSE	1.29925	R-Square	0.7320
Dependent Mean	11.36668	Adj R-Sq	0.7312
Coeff Var	11.43033		

5.9 references

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