THE UNIVERSITY OF TEXAS AT ARLINGTON



MULTIPLE LINEAR REGRESSION PROJECT

BLOOD SUGAR LEVEL ANALYSIS FOR WOMEN WITH HYPERGLYCEMIA

APPLIED REGRESSION ANALYSIS IE 5318- 005

Under the Guidance of and Submitted to: Dr. Aera K Leboulluec

SUBMITTED BY:

Achuth Chandrasekhar Venkata Subramaniam- 1001487479

Anutham Eashwar Padmanabhan- 1001371738

Nobel Victor Bombacha - 1001583388

Salvius Stephen Kinny- 1001552262

Tejas Sutar Nandkumar- 1001558442

Submitted on: May 3rd, 2018, Thursday

PROJECT PROPOSAL

1. PROBLEM AND VARIABLE DESCRIPTION

Problem Description:

We aim to analyze the blood sugar levels in pregnant women diagnosed with Hyperglycemia, which is a restorative condition that causes sugar levels to ascend higher than common. It is popularly known as Type-II diabetes. If there should arise an occurrence of Type-II, the body does not utilize insulin consistently. i.e. protection for insulin starts. To start with, the pancreas makes additional insulin to modify the condition, however after some time, it can't keep up and create enough insulin to keep blood sugar at ordinary levels. This medical condition tragically is a long-lasting malady and has brought about numerous well-being difficulties and has caused health issues for decades. Our objective is to present valuable research insights on this medical condition.

Variable Description:

**Sugar level**-Response Variable (Y) in women diagnosed with Hyperglycemia with their Body Mass Index levels (BMI), Blood Pressure (BP), Thickness of Triceps Skin (ST) & Insulin Content as the independent variables namely X1, X2, X3 & X4 respectively. **Sugar level** is favored as the response variable as it aids in determining whether the woman is affected by Hyperglycemia.

Body Mass Index-(BMI)

BMI is the measure of weight (in kg's) to height (in m) proportion to choose if a person is underweight or obese. When a person's weight is higher compared to his height, he is considered obese. An obese individual has high sugar levels which causes hyperglycemia. This is the reason we consider it as one of our independent variables

Blood Pressure-(BP)

Blood pressure is the pressure of coursing blood on the dividers of blood vessels, as a rule alludes to the pressure in extensive veins of the fundamental dissemination. It is estimated when heart contracts (systole) and when heart grows or rests (diastolic). A person who has Type-II diabetes has large amounts of Blood Pressure and subsequently is one of our independent variables.

Thickness of Triceps Skin-(TOTS)

Skin weighs 11.02 lbs. and is approximately 610 µm thick and with an area of 22 sq. ft. High sugar level leads to decrease in Thickness of Triceps Skin and subsequently is one of our independent variables.

Insulin Content

Hormone produced in the pancreas by the islets and this regulates the amount of sugar in the blood. The lack of insulin causes a form of hyperglycemia and subsequently is one of our independent variables.

Data collection:

The data was collected from pregnant women affected by hyperglycemia of a local hospital with age ranging between 22 to 48. The number of observations is forty (n).

Source of data: https://gist.github.com/ktisha/c21e73a1bd1700294ef790c56c8aec1f

Multiple Linear Regression:

The data set obtained consists of Response (Y) & Predictor Variables (X1, X2, X3, X4) which is listed below:

Y- Blood Sugar Levels

X1-- Body Mass Index (BMI)

X2- Blood Pressure (BP)

X3--Thickness of Triceps Skin(ST)

X4-Insulin Content

2. DISCUSSION ON MATRIX SCATTER PLOT:

The matrix in Figure 1 depicts the relationship between Y and X1, X2, X3, X4 and the relationship between all the independent variables. In the Blood Sugar Level (Y) vs. Body Mass Index (X1) plot it is noted that there is no curvature. For plots: Blood Sugar Level (Y) vs. Blood Pressure (X2), Blood Sugar Level (Y) vs. Thickness of Triceps Skin (X3) and Blood Sugar Level (Y) vs. Insulin Content (X4), there is no curvature.

It is also noted that in the Blood Sugar Level (Y) vs. Body Mass Index (X1), there is upward linear trend, in Blood Sugar Level (Y) vs. Blood pressure (X2) we note scattering of data values with linear trend, in Blood Sugar Level (Y) vs. Thickness of Triceps Skin (X3) there is a downward trend and between Blood Sugar Level (Y) vs. Insulin content (X4) again there is a downward trend with scattering of data values.

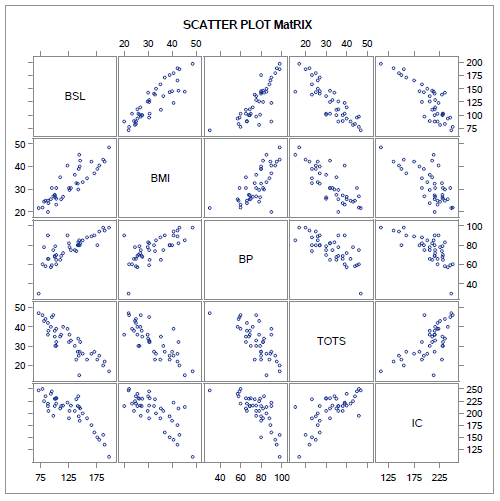
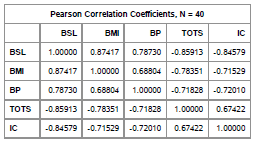


Figure 1: Matrix Scatter Plot

3. DISCUSSION ON PAIRWISE CORRELATION:



Tsable 1: Correlation Matrix

The correlation between the Blood Sugar Level (Y) and BMI (X1), Blood Pressure (X2), Skin Thickness (X3) and Insulin content (X4) are shown in the (Figure 1) matrix scatter plot and the (Table 1) correlation matrix.

Response vs. Predictor Variable (Y vs. X1): The correlation between Blood Sugar Level (Y) and Body Mass Index (X1) has a value of 0.87417 which shows a positive correlation with linear trend.

Response vs. Predictor Variable (Y vs. X2): The correlation between Blood Sugar Level (Y) and Blood Pressure (X2) has a value of 0.78730 which shows a positive correlation with linear trend.

Response vs. Predictor Variable (Y vs. X3): The correlation between Blood Sugar Level (Y) and Thickness of Triceps Skin (X3) has a value of -0.85913 that shows a negative correlation and negative linear trend.

Response vs. Predictor Variable (Y vs. X4): The correlation between Blood Sugar Level (Y) and Insulin content (X4) has a value of -0.84579 which shows a negative correlation and negative linear trend.

Predictor vs. predictor (X1Vs. X2): The correlation between body mass index(X1) and the Blood Pressure(X2) has a correlation value of 0.68804 this indicates a positive correlation.

Predictor vs. predictor (X1 Vs. X3): The correlation between body mass index (X1) and the Thickness of Triceps Skin(X3) has a correlation value of -0.78351 which indicates a negative correlation, multicollinearity may exist (as correlation value > cut-off value of 0.7).

Predictor vs. predictor (X1 Vs. X4): The correlation between Body Mass Index (X1) and Insulin content (X4) has a correlation value of -0.71529 this indicates a negative correlation, multicollinearity may exist (as correlation value > cut off value of 0.7).

Predictor vs. predictor (X2 Vs. X3): The correlation between blood pressure (X2) and Thickness of Triceps Skin (X3) has a correlation value of -0.71828 this indicates a negative correlation., multicollinearity may exist (as correlation value > cut off value of 0.7).

Predictor Vs. predictor (X2 Vs. X4): The correlation between blood pressure (X2) and Insulin content (X4) has a correlation value of -0.72010 this indicates a negative correlation, multicollinearity may exist (as correlation value > cut off value of 0.7).

Predictor Vs. predictor (X3 Vs. X4): The correlation between Thickness of Triceps Skin(X3) and the Insulin content (X4) has a correlation value of 0.67422 this indicates a positive correlation.

4. POTENTIAL COMPLICATIONS:

We analyze that there is no curvilinearity as suggested by the correlation matrix and scatter plots. It is observed that there are outliers, which may be influential in our analysis.

PRELIMINARY MULTIPLE LINEAR REGRESSION MODEL ANALYSIS

A preliminary multiple regression model analysis was performed on SAS software. The output tells us the impact of each predictor variable on the response variable. The fitted model is expressed as:

Yi = ẞo + ẞ1xi1 + ẞ2xi2 + ẞ3xi3 + ẞ4xi4 + Ԑi

Y (Blood Sugar Level) = ẞo + ẞ1\* (Body Mass Index) + ẞ2\* (Blood Pressure) + ẞ3\*(Thickness of Triceps Skin) + ẞ4\* (Insulin Content)

Where,

Dependent Variable Yi = Blood Sugar level (mg/dL)

Independent Variable 1 xi1 = Body Mass Index (kg/m2)

Independent Variable 2 xi2 = Blood Pressure in (mmHg)

Independent Variable 3 xi3 = Skin Thickness(μm)

Independent Variable 4 xi4 = Insulin level (mL/cc)

The results are as below after regressing Blood Sugar level on Body Mass Index, Blood Pressure, Thickness of Triceps Skin and Insulin Content. The results contain detail regarding parameter estimates and the analysis of variance.

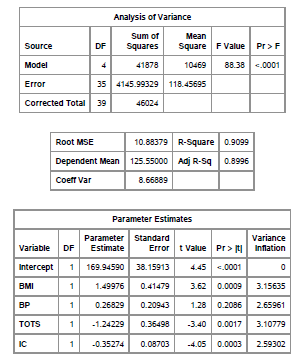


Table 2: Anova and Parameter Estimates for preliminary model

The fitted regression model is as follows based on SAS output (Table 2):

Ŷi = 169.9459 + 1.49976xi1 + 0.26829 xi2 - 1.24229xi3 - 0.35274 xi**4**

Blood Sugar level (Y) = 169.9459 + (1.49976\* Body Mass Index) + (0.26829 \* Blood Pressure) - (1.24229\* Thickness of Triceps Skin) - (0.35274 \* Insulin Content)

It is noted that when there is a unit change in Body Mass Index, blood sugar level increases by 169.9459 units when Blood pressure, thickness of triceps skin and insulin content are kept constant. Similarly, when there is a change in Blood pressure by one unit, blood sugar level increases by 0.26829 units when BMI, Thickness of triceps skin and Insulin content are kept constant. The Blood sugar level also decreases by 1.24229 units when Thickness of triceps skin changes by one unit and when BMI, Blood pressure and Insulin content are fixed. Also, when there is a unit change in Insulin Content, blood sugar level decreases by 0.35274 units when BMI, blood pressure and thickness of triceps skin are kept constant.

The parameter β0 = 169.9459 indicates y intercept in the regression plane. This means that the blood sugar level will have a value of 169.9459 units when BMI, blood pressure, thickness of triceps skin and insulin content are all having a value of zero units. However, it is not meaningful to have Blood sugar level to have a constant value of 169.9459 units when BMI, blood pressure, thickness of triceps skin and insulin content have zero units.

**MODEL ASSUMPTIONS**

Residual Analysis is to be performed to verify the preliminary model. The assumptions for the analysis are:

i) Linear Model-The current MLR model form is reasonable

ii) Constant Variance-The residuals have constant variance

iii) Normality-The residuals are normally distributed

iv) Correlation-The residuals are not correlated

v) There are no outliers

vi) The predictors are not highly correlated with each other

**MLR MODEL FORM IS REASONABLE:**

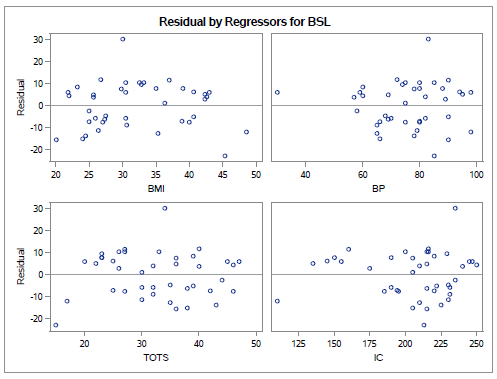


Figure 2: Residuals vs. Predictor plots for the preliminary model

From Figure 2, we can infer that there is no curvature in the Residuals vs. Predictor plots. Therefore, we draw an inference that transformation is not required and that the MLR model is reasonable.

**THE RESIDUALS HAVE CONSTANT VARIANCE:**

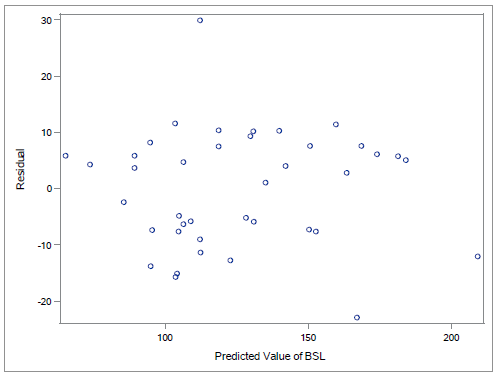


Figure 3: Residuals vs. predicted value plot for preliminary model

From Figure 3, we can see that random scatter of points is observed and there is no funnel shape. This concludes that the model has constant variance satisfying our second assumption.

**MODIFIED LEVENE TEST**: **TEST FOR NON-CONSTANT VARIANCE**

Modified Levene test is a test to check for constant variance. The data set was split into two groups. The first group, g1 has 19 observations. The second group, g2 has 21 observations. f-test is first performed to check if sample variances are equal and then t-test is performed to check if the error variances are constant. Below is the output obtained from the SAS software:

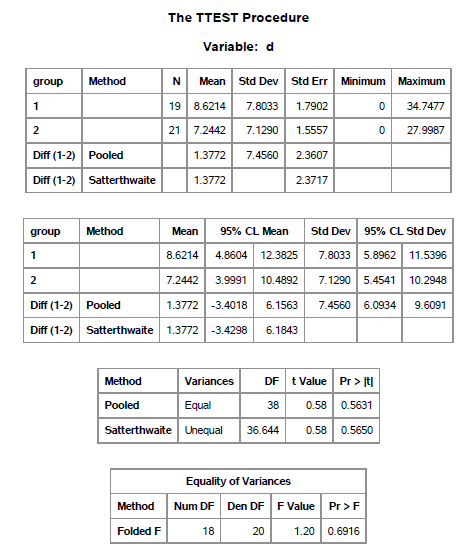


Table 3: Modified Levene Test for Preliminary Model

F- Test:

H0: σ1 = σ2 (Variances are equal)

versus

H1: σ1 ≠ σ2 (Variances are not equal)

Decision Rule: Reject H0 if p < α

p value from Table 3 is 0.6916 and significance level α = 0. 05.

Conclusion: We fail to reject H0 as p > α.

Interpretation: Variances of the two populations were equal and we can proceed with the t-test for

checking for constant error variances.

T- Test:

H0: Variance is Constant

versus

H1: Variance is not Constant

Decision Rule: Reject H0 if p < α.

P value from Table 3, p value is 0.5650 and significance level α = 0. 05.

Conclusion: We fail to reject H0 as p > α.

Interpretation: Therefore, we are 95% confident that the model has a constant variance and the test

results match with the residual analysis plot (Figure 3: Residuals versus Fitted Value) and it was concluded that the variance is constant. Hence, we have proceeded with the analysis.

**NORMALITY:**

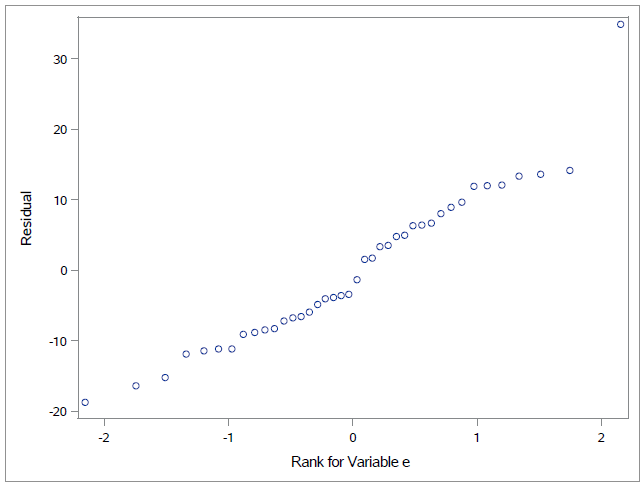


Figure 4: Normal probability plot for preliminary model

From figure 4, it is interpreted that the plot is pretty straight. We can conclude that the distribution of the error term is normal and we will conduct a normality test to verify the result.

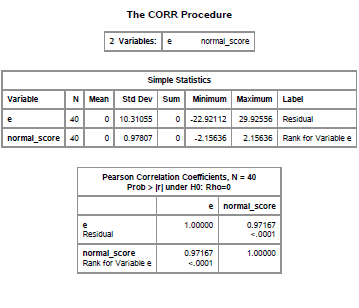


Table 4: Normality test for preliminary model

H0: Normality is satisfied

versus

H1: Normality is violated

Decision Rule: Reject H0 if ρ̂< c (α, n)

Sample Correlation (ρ̂) from SAS output (Table 4) is, ρ̂= 0.97167; α= 0.05; n=40; c (0.05, 40) = 0.9712.

**Conclusion**: We fail to reject H0 as ρ̂> c (α, n). Hence, we are 95% confident that the normality of the model is satisfied, agreeing with our earlier inference made using the normal probability plot. Therefore, the residuals are normally distributed.

**RESIDUALS ARE UNCORRELATED:**

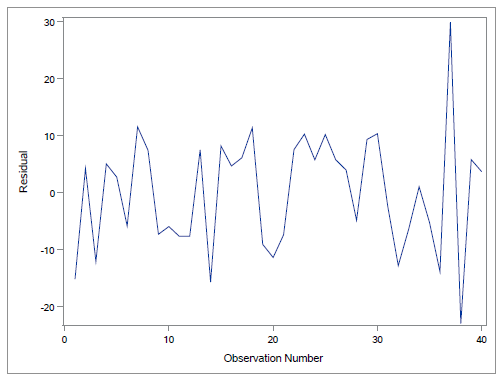


Figure 5: Time series plot

Time series plot is a graph which displays residual values against time. Here, the time series plot for residuals vs observation number shows the scattering of the residual values and it has no obvious trends

indicating random jaggedness.

**DIAGNOSTICS**

**Outliers, Leverages and Influences**

Outliers in our dataset are noted using two tests, Bonferroni Outlier Test for identifying Y outliers and test using leverage values for identifying X outliers. The total number of observations (n) used is 40 and parameter, p is equal to 5 (p=5, n=40).

**Bonferroni Outlier Test – to check for Y Outliers**

The Bonferroni Outlier test was used to check for any data points that’s are Y outlying or have extreme

Y values. The studentized deleted residual is represented by ti, p represents the number of parameters

and n represents the total observation points. The studentized residual, ti for each observation is

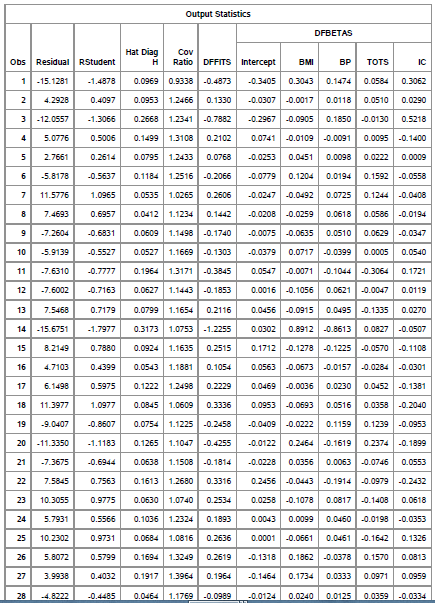
shown in the appendix.

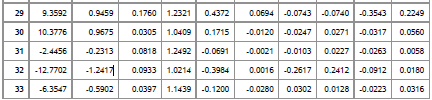
**Decision Rule**: If |ti| > t (1-α/2n; n-p-1), then observation ‘i’ is a Y outlier.

**Cut-off Value**: t (1-α/2n; n-p-1) = t (1 – 0.05/80; 40-5-1) = t (0.999375; 34) = 3.52

**Conclusion**: All values of studentized residuals, |ti| are less than 3.52. Hence, there are no Y –

outliers.





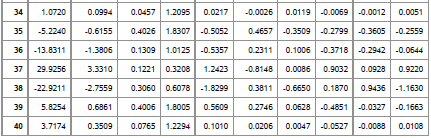


Table 5: Output statistics

The studentized deleted residual is represented as ti but here it is represented as RStudent, as seen in table 5. The RStudent value is calculated using SAS and it is compared with the cut-off value t (1-(α/2n), n - p - 1). We compare the cut-off value with the absolute value of RStudent to find the outliers

(n=40, p = 5, α = 0.05). The cut-off value is compared with the absolute value of RStudent from table 5, and noticed that, there is no value greater than the cut-off value. Therefore, we conclude there is no Y-outlier.

**Test for X Outliers**

Leverage values, hii is utilized to identify X outliers in data points; hii values for each value is

shown in table 5. We intend to find the x-outlier hence, we calculate the cut-off value given by ‘2p/n’, where p🡪 Number of predictors; n🡪 Number of observations

**Decision Rule**: If hii > 2𝑝/𝑛, then observation ‘i’ is an X outlier.

Cut-off Value: 2𝑝/𝑛=2∗5/40= 0.25

**Conclusion:** 5 data points have leverage value, hii > 0.25. The X – outliers are noted at:

Observations (3,14,35,38,39)

**Influential Test for Outliers**

We have identified outliers, and now we intend to verify if they have an influence on the model.

Computing DFFITS Cook’s distance will aid us.

**a. Influence on the Fitted Values – DFFITS**

DFFITS value for each observation is shown in Table 5.

**Decision Rule**: If |DFFITSi| > 2√(p/n), then outlier ‘i’ has an influence on the fitted values.

Cut-off value= 2√(p/n) = 2√ (5/40) = 0.707106781.

**Conclusion:** Observations 3 and 38 have |DFFITS| value greater than 0.707106781.

Hence, we infer that the above three X outlying values have an influence on the fitted values.

**b. Influence on Individual LSEs – DFBETAS**

DFBETAS value for each observation is shown in Table 5.

**Decision Rule**: If |DFBETASik| >2/√𝑛, outlier ‘i’ has an influence on the Least Squares Errors.

Cut-off value: 2/ √𝑛 = 2/√40= 0.316227.

By comparing the cut-off value with the table values of DFBETAS, we infer that:

b0: 35,38

b1: 14,35,38

b2: 14,39

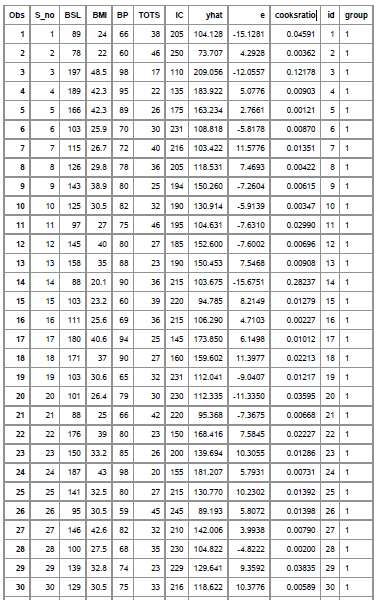
b3: 35,38

b4: 3,38

**c. Combined Influence on all the LSEs – Cook’s Distance**

The value of Cook’s Distance, Di for each observation is shown in the Table D Decision Rule: If Di > F (0.5; p, n-p), then outlier ‘i’ has a combined influence on all the LSEs Cut-off value: F(0.5; p, n-p) = F(0.5; 5, 35) = 0.88731

**Conclusion:** No observations were found to exceed the cut-off value of 0.88731 (from table 6). Therefore, it can be concluded that no observation is found to be higher than the cut-off value, there are no outliers that will influence the model.



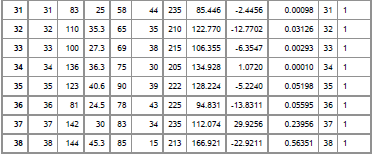




Table 6: Cook’s distance for preliminary model

**Variance Inflation Factor**

From Table 7 below, we observe that maximum Variance Inflation Factor (VIFmax = 3.15635 which is less than 5 (cut-off value). The average VIF value is 2.87919 which is also less than 5. Hence it is concluded that, we do not have a serious multicollinearity problem.

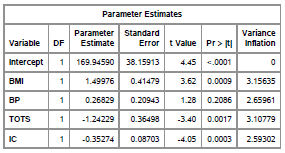


Table 7: VIF table for preliminary model

**DISCUSSION ON REMEDIAL MEASURES**

There is a need for appropriate justification to eliminate outliers from our dataset such as experimental or typing error. We have 5 X-outlying values DFBETAS, DFFITS and Cook’s Distance were put to use to verify the influence. We also observe that our dataset has a constant variance and that the normality is satisfied, they did not have an effect on the model assumptions. Further, the Variance Inflation Factor for each predictor variable is less than the cut-off value of 5. Therefore, there is no need to remove outliers and we can proceed with our analysis with the preliminary model. The preliminary model is given by:

Blood Sugar level (Y) = 169.9459 + (1.49976\* Body Mass Index) + (0.26829 \* Blood Pressure) - (1.24229\* Thickness of Triceps Skin) - (0.35274 \* Insulin Content)

**DISCUSSION OF REGRESSION OUTPUT**

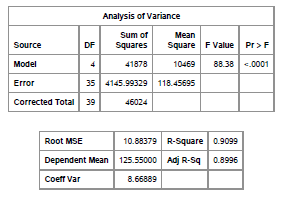


Table 8: Analysis of variance for preliminary model

**Regression Sum of Squares:** SSR explains the extent to which the regression model presents the data. From Table 8, SSR is noted as 41878. The degree of freedom for SSR is 4.

**Error Sum of Squares:** SSE presents the amount of deviation of predicted values from the mean value of the response variable. A smaller SSE means a good fit of the model to the data. From Table 8, SSE is noted as 4145.99329 and the degree of freedom associated with SSE is 35.

**Total Sum of Squares:** SSTO measures the deviation of the individual response variable with the overall mean of the response variables. From Table 8 SSTO is noted as 46024 and the degrees of freedom associated with SSTO is 39.

**Mean Regression Sum of Squares:** MSR represents the fraction between SSR and degrees of freedom associated with it. From Table 8 MSR is noted to be 10469.

**Mean Error Sum of Squares:** MSE represents the fraction between SSE and degrees of freedom associated with it. From Table 8, MSE is noted to be 118.45695.

**Coefficient of Determination**: R2 represents the extent to which the regression model fits the data. It can be computed by the fraction between MSR and MSE. From the ANOVA table 8, the value of R2 is noted to be 0.9099. This implies that, 90.99% of the variability is accounted by the predictors represented in our model.

**Significance of Predictors**

We compare the values of R2 and adjusted R2 to draw inferences. From Table 8, value of R2 is 0.9099 and value of adjusted R2 is 0.8996. The difference between the two the tells us by how much the predictors are insignificant in our model, difference between R2 and adjusted R2 is close to 1% and therefore, we infer that 1% of predictor is insignificant. We also draw inferences from our model by comparing the p-value of each predictor with the significance level (α = 0.05). If p-value is found to be lesser than the significance level, then we can conclude that the predictor is significant in the model. Table 9 is used to lookup and compare the values.

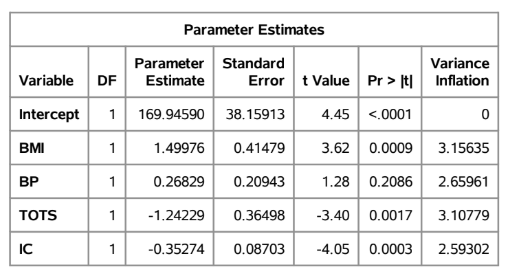


Table 9: Test of significance for preliminary model

p-value of Body Mass Index(BMI) is 0.0009 and this value is lesser than the significance level of 0.05 and we infer that x1 is a significant predictor.

p-value of Blood Pressure(BP) is 0.2086 higher than 0.05, nd we infer that x2 is not a significant predictor.

p-value of Thickness of Tricep Skin is 0.0017 and which is lesser than level of significance,i.e 0.05 and we infer that x3

is a significant predictor.

p-value of Insulin is 0.0003, that is lesser than 0.05 and we infer that x4 is a significant predictor.

**Test for checking the Significance of model**

H0: Yi = β0 + εi (Reduced Model)

Versus

H1: Yi = β0 + β1xi1 + β2xi2 + β3xi3 + β4xi4 + εi (Full Model)

Decision Rule: Reject H0 if p-value < α

From Table 8, p-value is approximately equal to 0.0001 and significance level, α = 0.05

Conclusion:We reject H0 since p-value is less than α.

Hence, we are 95% confident that full model is significant.

**III. EXPLORATION OF INTERACTION TERMS**

PARTIAL REGRESSION PLOTS

In order to get better understanding of unexplained variability in the model we add interaction terms to the model i.e. multiplying the predictor variables in pairs. From our model we get six such interaction terms which is added to the model only on the basis of the interaction pair having a linear trend. Interaction terms aid in lowering the correlation between the predictor variables, thereby, reducing the issue of serious multicollinearity in the chosen model. The below plots represent residual of model vs interaction terms.

Partial regression plots help in representing the consequences of adding a new variable to an existing model. Body Mass Index(BMI), Blood Pressure(BP), Thickness of Triceps Skin(TOTS), Insulin Content(IC) are the predictor variables with the response variable being Blood Sugar Levels(BSL).

Interaction terms:

Body Mass Index(X1) vs Blood Pressure(X2)

Body Mass Index(X1) vs Thickness of Triceps(X3)

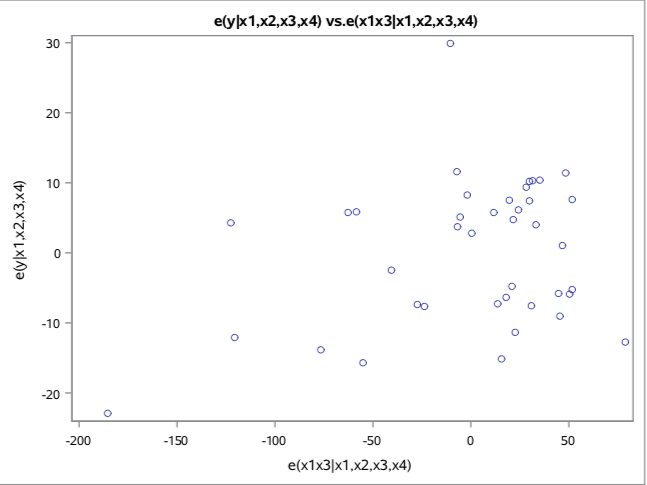
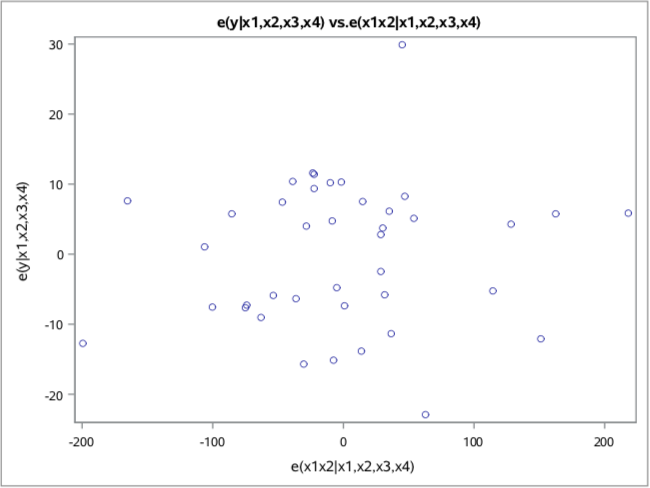
Body Mass Index(X1) vs Insulin Content(X4)

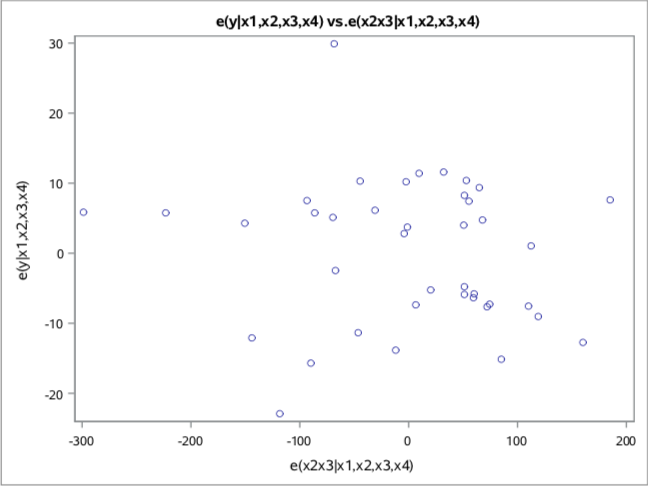
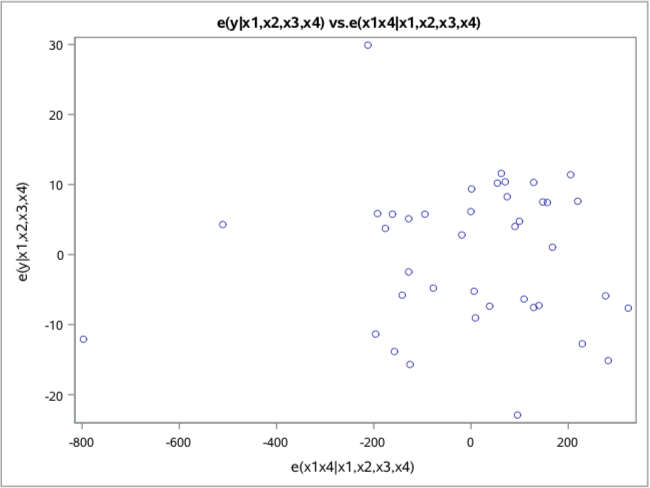
Blood Pressure(X2) vs Thickness of Triceps(X3)

Blood Pressure(X2) vs Insulin Content(X4)

Thickness of Triceps(X3) vs Insulin Content(X4)

PARTIAL REGRESSION PLOTS FOR INTERACTION TERMS BEFORE





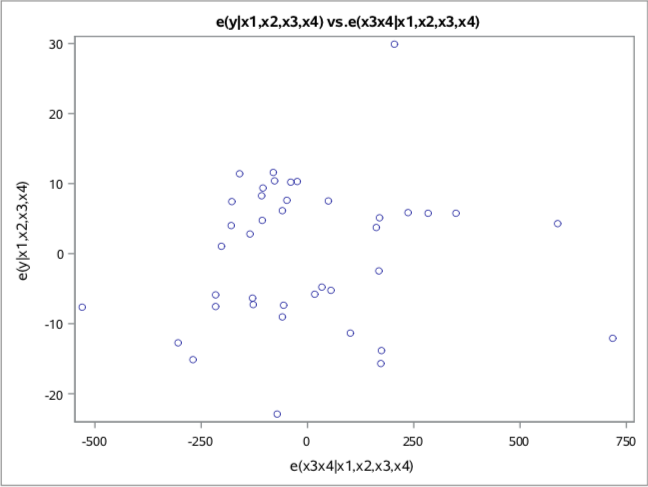
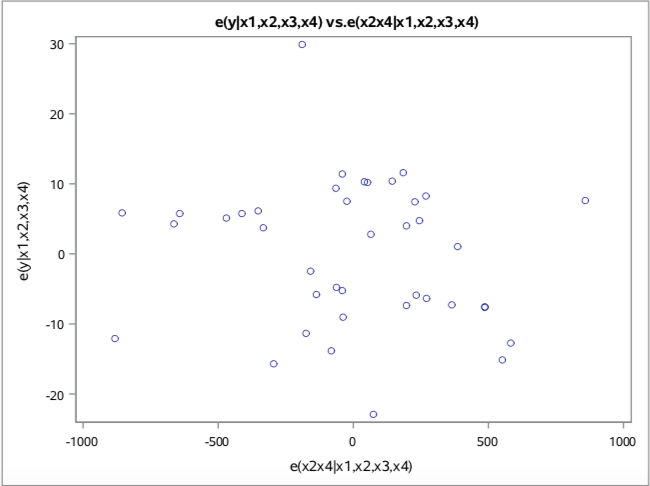
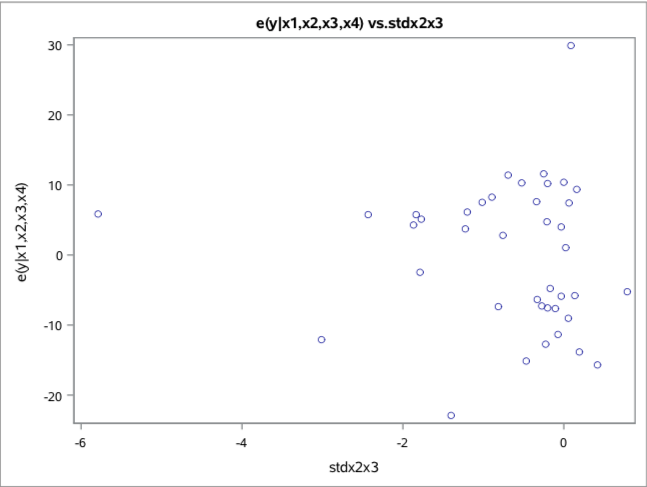
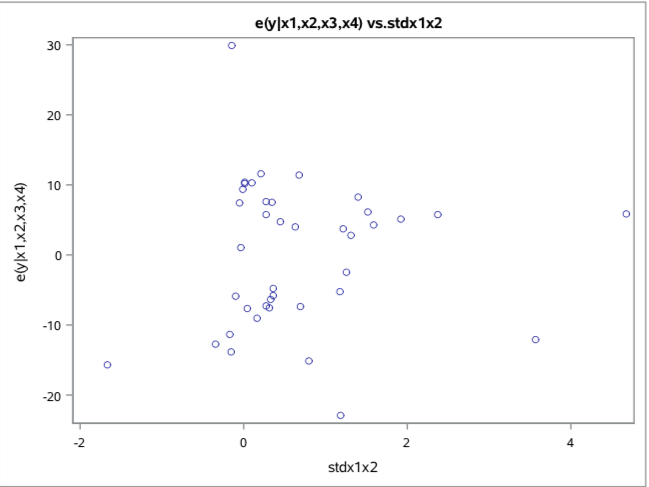


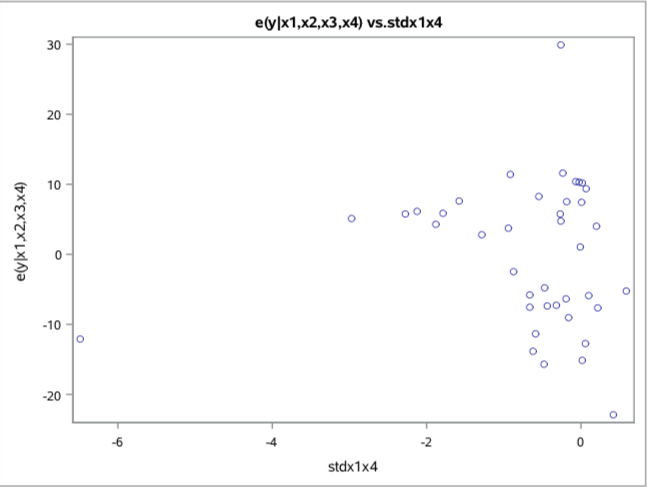
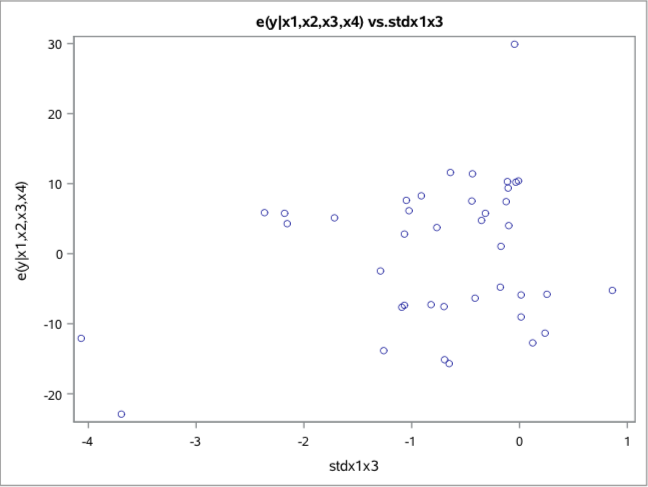
Figure 6: Partial regression plots of interaction terms before standardization.

From Figure 6 we can conclude that the interaction terms X1 X3 and X3 X4 have a linear trend and their addition will have a significant effect on the model. We also standardize the variables for confirmation.

PARTIAL REGRESSION PLOTS AFTER STANDARDIZATION

Prior to plotting the graphs, the six interaction terms were standardized (to a mean of zero and standard deviation of one, to distribute the terms normally). This aids in reducing variability of model. This is done to reduce multicollinearity. Residuals of Blood Sugar Level are plotted against the standardized interaction terms and we search for linear trend.





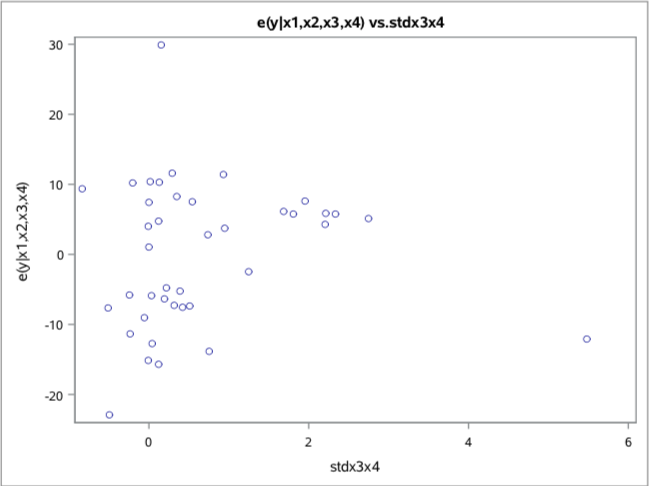
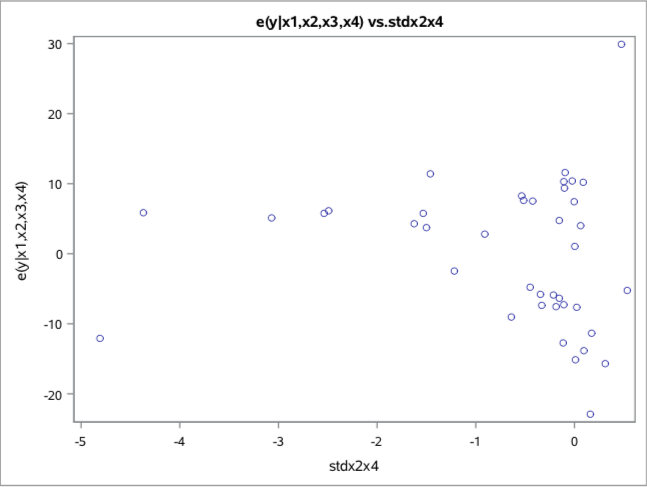


Figure 7: Residual Vs. Standardized Interaction terms plots

From Figure 7 we infer that there is some linear trend in std X1 X3 and std X3 ­X4. Therefore, we add these two terms to our model in order to reduce variability of Blood Sugar Levels. The other plots mentioned above are not taken into consideration because they all show random cloud formations when their standardized terms were plotted against the residual of full model.

CORRELATION OF THE INVOLVED INTERACTION TERMS

Before Standardization

Correlation helps in explaining the extent to which the variables are related to each other. The below matrix represents the correlation between the interaction terms and predictor variables. We do this to check for multicollinearity in the model. The correlation for the interaction terms is generally higher because of direct multiplication. If correlation is greater than 0.7 it implies that there is multicollinearity between the terms. To overcome this issue, we do standardization. The unstandardized interaction terms from Pearson Correlation Coefficients are shown below.

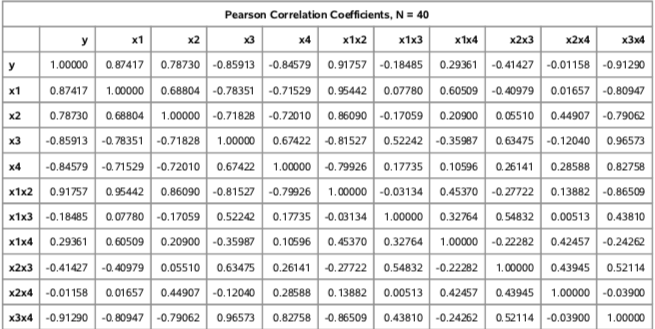


Table 10: Correlation matrix before standardization

After Standardization

The standardized interaction terms are a product of the standardized predictors. The Pearson Correlation Coefficients after standardization of interaction terms are shown below in the Table 10 . To reduce multicollinearity, the standardization is done which indirectly reduces the VIF values that gives us an accurate prediction of the fitted values for Blood Sugar Level.

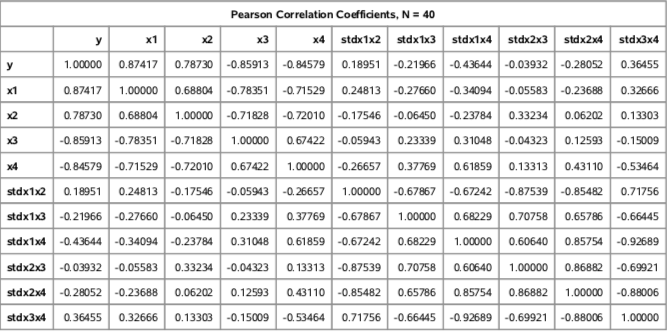


Table 11: Correlation matrix after standardization

Discussion

From Table 9 and Table 10, we understand that after standardization there is considerably less correlation between the interaction after standardizing the interaction terms. Therefore, addition of the standardized interaction terms into our preliminary model, helps in reducing the issue of multicollinearity. We continue the analysis to get the best models after adding the standardized interaction terms.

**IV. MODEL SEARCH**

Here we aim to find lack of fitness, outliers and select the best possible model. The Significance Level of 0.10 is selected for the model selection. We deploy the below mentioned methods for model search:

Backward Deletion

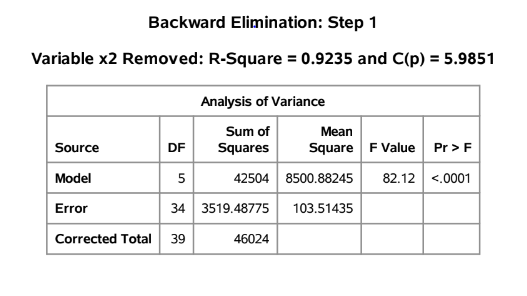
Best Subset Selection

Stepwise elimination Method

We select models if the significance level is less than 0.1 and there should be no multicollinearity.

Backward Deletion

The Backward deletion method uses the full model with all potential predictor variables. This method regresses the dependent variable on the full set of predictor variables. p-values of various predictors are computed and the predictor with the largest p-value is deleted. The elimination process is carried one step at a time and the regression is run over and again. The regression is performed till the predictor variables have p-value less than 0.10, performing these steps helps us in obtaining the best model.



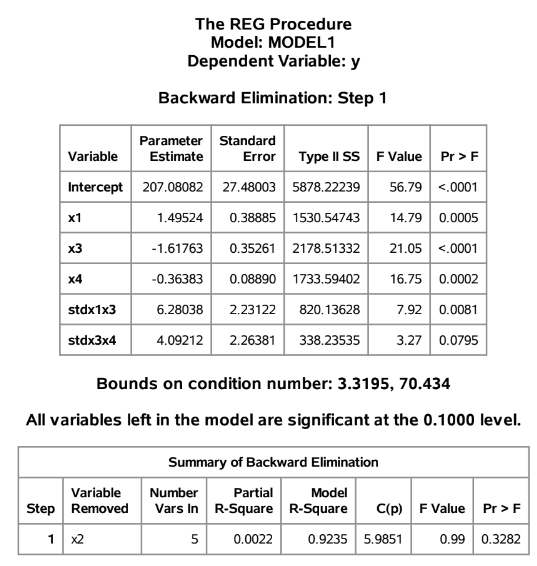


Table 12: Backward Deletion Result

From Table 12, the best model has the predictor variable Body Mass Index(X1), Blood Pressure(X2), Thickness of Triceps(X3), Insulin Content(X4), stdX1X3, stdX3X4, the removed predictor is X2.

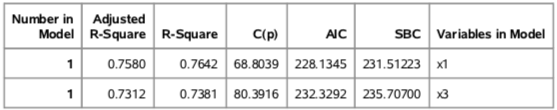
Using the Backwards Deletion method, we get the following as the best model:

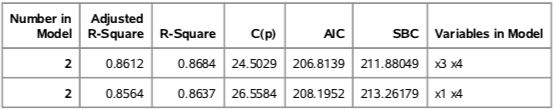
Blood Sugar Level(Yi) = Body Mass Index(X1) + Blood Pressure(X2) + Thickness of Triceps(X3) + Insulin Content(X4) + stdX1X3 + stdX3X4

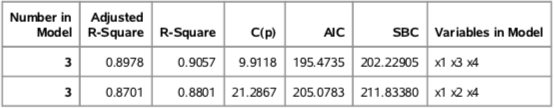
Best Subset Method

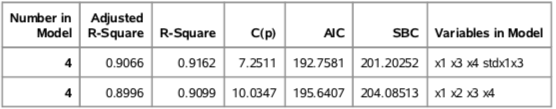
Best Subset Method is a procedure which is programmed to compute regression coefficients and predict responses such that the variance is low. Here we take into account the following variables Body Mass Index(X1), Blood Pressure(X2), Thickness of Triceps(X3), Insulin Content(X4), stdX1X3, stdX3X4. SAS helps to identify the best models. This method begins with the full model consideration, predictor is segregated to all possible subsets, post which the method regresses the Y on all available subsets.

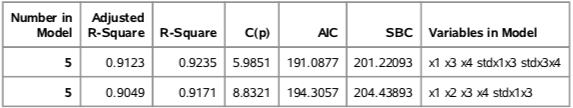
The best model is selected on the basis of High R2, Low Sum of the Squares Error, High Adjusted R2 and Low Mean Sum of Squares. The Bias model is found from the variation in Mallows Cp with the number of predictors, p. Other attributes for the best model include Low Mallows Cp (approximately equal to number of predictors), low Alkaline Information Criterion(AIC) and Schwarz’s Bayesian Criterion(SBC) are considered.











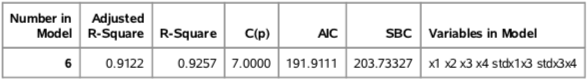


Table 13: Best Subset Selection

After careful consideration based on the values from Table 13, we select the following models that satisfy all the criteria for selection.

The first model has R2=0.9235 and adjusted R2= 0.9123 with Cp=5.9851 as shown in Table 13. We note that the first model has a low Cp value and a high R2 and adjusted R2 values.

The second model of subset regression is also based on R2 and Adjusted R2 values and has 6 variables (BMI, BP, Thickness of Skin, Insulin Content, stdx1x3, stdx3x4). The variables in this model have R2= 0.9257 and Adjusted R2= 0.9122 and Cp=7.0000 as shown in Table 13. Both the models have less AIC and SBC values while they have R2 and adjusted R2 values.

**Model 1**: Blood Sugar Level(Yi) = Body Mass Index(X1) +Thickness of Triceps(X3) + Insulin Content(X4) + stdX1X3 + stdX3X4 ­(Model Number 5 from Table 13)

**Model 2**: Blood Sugar Level(Yi) = Body Mass Index(X1) + Blood Pressure(X2) +Thickness of Triceps(X3) + Insulin Content(X4) + stdX1X3 + stdX3X4(Model Number 6 from Table 13)

Stepwise Regression Method

The Step Wise Regression Method starts with no predictors in the model, the predictors are added or deleted on the basis of the p-values to get the best model. P-value of the predictor must be less than the cut-off level of significance 0.10. The variables with a higher p-value than 0.10 are removed from the model. Finally, we get a model in which all variables have p-values below the level of significance of 0.10.

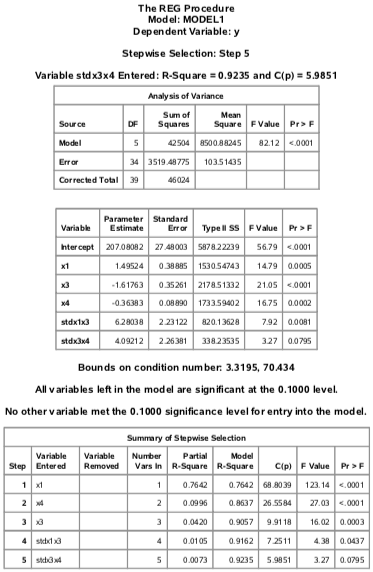


Table 14: Stepwise Regression Model

From Table 14, The best model has the following predictors

Blood Sugar Level(Yi) = Body Mass Index(X1) + Thickness of Triceps(X3) + Insulin Content(X4) + stdX1X3 + stdX3X4.

Potential Best Models

The three model search methods have helped us find two potential best models which are given by:

Model 1:

The predictors in this model are Total Body Mass Index(X1), Thickness of Triceps(X3), Insulin Content(X4), stdX1X3, stdX3X4.

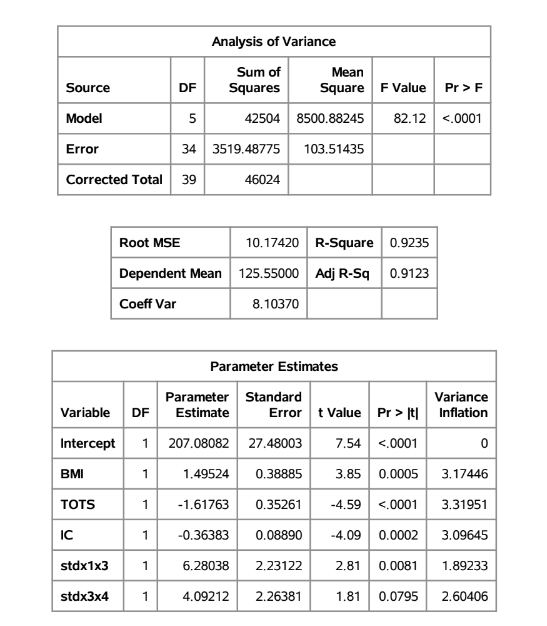


Table 15: ANOVA Table for Model 1

Blood Sugar Level = 207.08082 + 1.49524\*Body Mass Index – 1.61763\*Thickness of Triceps Skin – 0.36383\*Insulin Content + 6.28038\* stdX1X3 + 4.09212\* stdX3X4

This model has R2 of 0.9235, and adjusted R2 of 0.9123, the difference between the two being 0.0112, which is very minimal, also the maximum VIF value is 3.31951, which is lesser than 5 indicating that there is no issue of serious multicollinearity. All predictors in the model are significant at 0.10 level of significance.

Model 2:

The predictors in this model are Total Body Mass Index(X1), Blood Pressure(X2), Thickness of Triceps(X3), Insulin Content(X4), stdX1X3, stdX3X4.

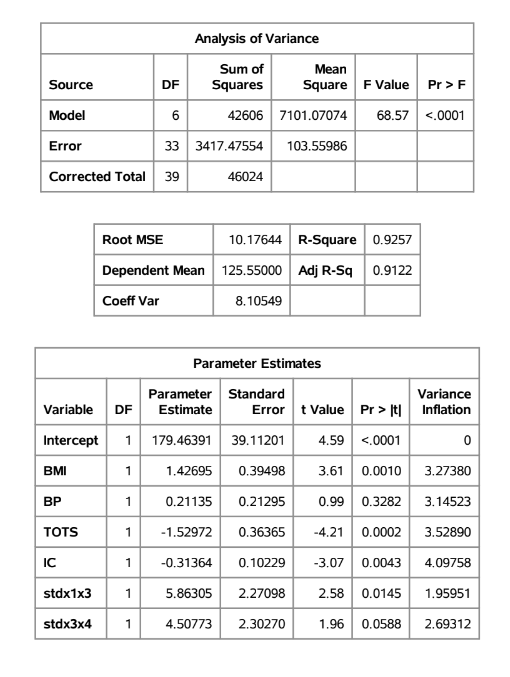


Table 16: ANOVA Table for Model 2

Blood Sugar Level = 179.46391 + 1.42695\*Body Mass Index + 0.21135\*Blood Pressure – 1.52972\*Thickness of Triceps Skin – 0.31364\*Insulin Content + 5.86305\* stdX1X3 + 4.50773\* stdX3X4

This model has R2 of 0.9257, and adjusted R2 of 0.9122, the difference between the two being 0.0135, which is very minimal, also the maximum VIF value is 4.09758, which is lesser than 5 indicating that there is no issue of serious multicollinearity. All predictors in the model are significant at 0.10 level of significance.

**V. MODEL SELECTION**

Model selection interprets the model assumption and model search, by which an optimum model is selected in the predictor variables are not influenced by explanatory variables. The model assumption for all potential best model are considered in the model selection process which includes factors such as outliers, influence of the outliers and multicollinearity. From the model search module two models are selected and compared, following are the selected models

Model 1 Blood Sugar Level = 207.08082 + 1.49524\*Body Mass Index – 1.61763\*Thickness of Triceps Skin – 0.36383\*Insulin Content + 6.28038\* stdX1X3 + 4.09212\* stdX3X4

Model 2: Blood Sugar Level = 179.46391 + 1.42695\*Body Mass Index + 0.21135\*Blood Pressure – 1.52972\*Thickness of Triceps Skin – 0.31364\*Insulin Content + 5.86305\* stdX1X3 + 4.50773\* stdX3X4

**MODEL 1 – VERIFICATION OF ASSUMPTIONS**

**Residual Analysis  
1. Model Form**

The graphs are plotted with residual versus Body Mass Index (X1) Thickness of Triceps Skin (X3) Insulin Content (X4) stdx1x3 & stdx3x4to see the combined patterns for collinearity.

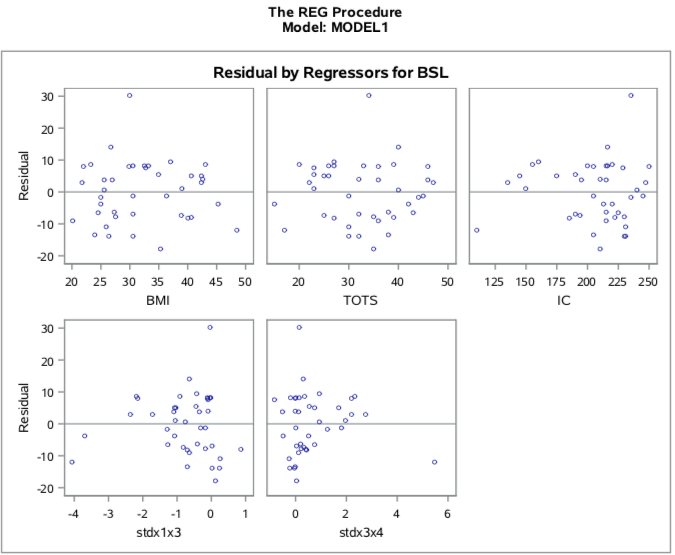


Figure 8: Residual Vs Predictors and Interaction terms of the Model1

From the figure 8, the residuals versus the independent predictors, Body Mass Index (X1) Thickness of Triceps Skin (X3) Insulin Content (X4) does not show a curvature. It was also observed that the standardized interaction terms stdx1x3 & stdx3x4 are show a linear trend. Therefore, it was concluded that the MLR Model form 1 is reasonable and that stdx1x3 & stdx3x4 can be added to the model.

**2. Constant Variance**

The following is the graph plotted between residuals and fitted values.

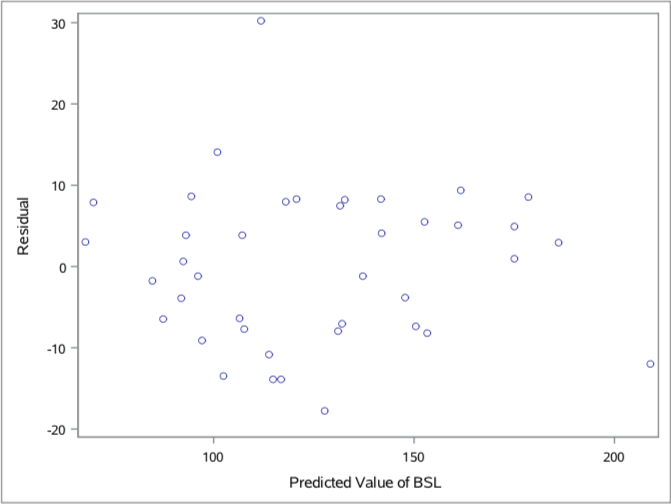


Figure 9: Residual Vs Fitted Values-Model 1

From the above Figure 9, the residual plot between Residual Vs Fitted values of Blood Sugar Level has a random point cloud. There is no funnel shape formation in the graph. Therefore, it was concluded that the Model 1 has a constant variance.

**3. Normal Probability plot for Residuals**

The graph between residual values and normal scores was plotted in SAS as shown below

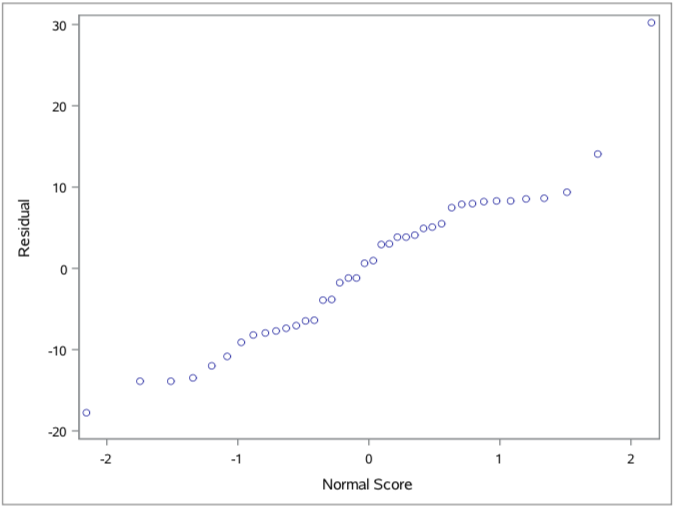


Figure 10: Normal Probability Plot – Model 1

From the Figure 10, the plot is seen to be pretty straight, but it has a slight S-Shaped shorter tail at the top. Since there is a straight line with a slight S shape, it was concluded that Normality of Model 1 is satisfied.

**4.Time Series Plot**

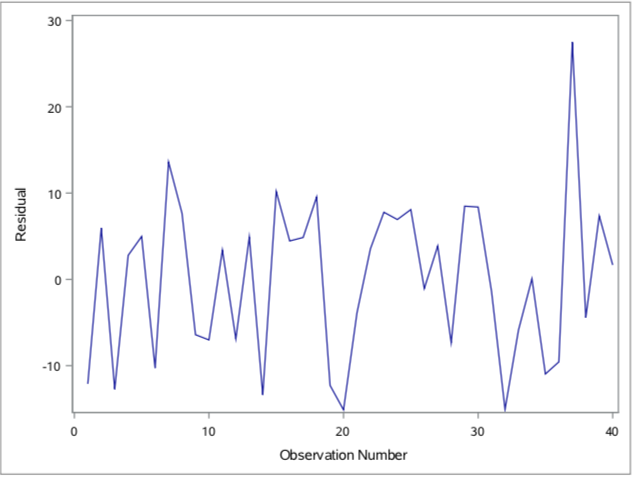
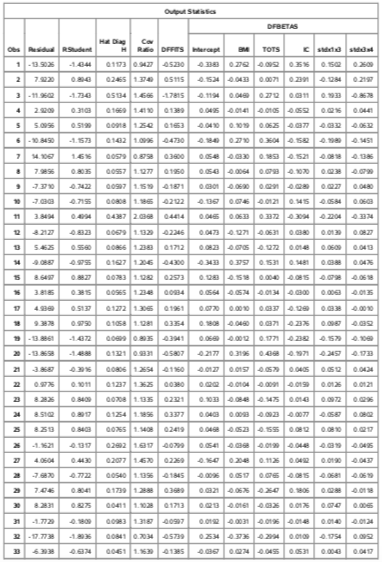


Figure 11: Time series plot of Model 1

The time series graph was plotted for model 1 and we can see that there is a random jaggedness throughout the graph without any trend being followed therefore we can say that the MLR Model 1 is reasonable.

**Diagnostics Check**



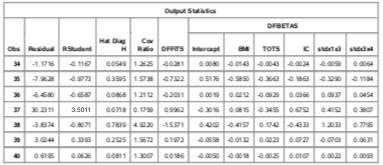


Table 17: Output Statistics

**Test for Y-Outliers-Bonferroni Outlier Test**

The Bonferroni Outlier test was used to check for any data points that’s are Y outlying or have extreme

Y values. The studentized deleted residual is represented by ti, p represents the number of parameters

and n represents the total observation points. The studentized residual, ti for each observation is

shown in the appendix.

**Decision Rule**: If |ti| > t (1-α/2n; n-p-1), then observation ‘i’ is a Y outlier.

**Cut-off Value**: t (1-α/2n; n-p-1) = t (1 – 0.05/80; 40-5-1) = t (0.999375; 34) = 3.52

**Conclusion**: All values of studentized residuals, |ti| are less than 3.52. Hence, there are no Y –

outliers.

The studentized deleted residual is represented as ti but here it is represented as RStudent, as seen in table 5. The RStudent value is calculated using SAS and it is compared with the cut-off value t (1-(α/2n), n - p - 1). We compare the cut-off value with the absolute value of RStudent to find the outliers

(n=40, p = 5, α = 0.05). The cut-off value is compared with the absolute value of RStudent from table 5, and noticed that, there is no value greater than the cut-off value. Therefore, we conclude there is no Y-outlier.

**Test for X Outliers**

Leverage values, hii is utilized to identify X outliers in data points; hii values for each value is

shown in table 5. We intend to find the x-outlier hence, we calculate the cut-off value given by ‘2p/n’, where p🡪 Number of predictors; n🡪 Number of observations

**Decision Rule**: If hii > 2𝑝/𝑛, then observation ‘i’ is an X outlier.

Cut-off Value: 2𝑝/𝑛=2∗5/40= 0.25

**Conclusion:** 6 data points have leverage value, hii > 0.25. The X – outliers are noted at:

Observations (3,11,26,35,38,39)

**Influential Test for Outliers**

We have identified outliers, and now we intend to verify if they have an influence on the model.

Computing DFFITS Cook’s distance will aid us.

**a. Influence on the Fitted Values – DFFITS**

DFFITS value for each observation is shown in Table 5.

**Decision Rule**: If |DFFITSi| > 2√(p/n), then outlier ‘i’ has an influence on the fitted values.

Cut-off value= 2√(p/n) = 2√ (5/40) = 0.707106781.

**Conclusion:** Observations 3,35,38 have |DFFITS| values greater than 0.707106781.

Hence, we infer that the above X outlying value have an influence on the fitted values.

**b. Influence on Individual LSEs – DFBETAS**

DFBETAS value for each observation is shown in Table 5.

**Decision Rule**: If |DFBETASik| >2/√𝑛, outlier ‘i’ has an influence on the Least Squares Errors.

Cut-off value: 2/ √𝑛 = 2/√40= 0.316227.

By comparing the cut-off value with the table values of DFBETAS, we infer that:

b0: 35,38

b1: 35,38

b2: 11,35

b3: 38

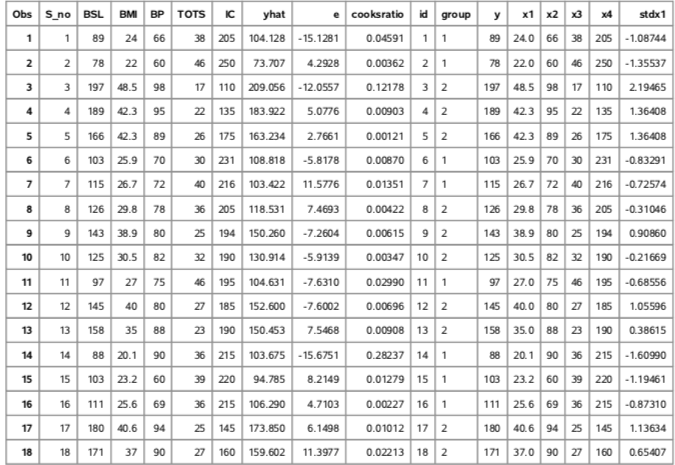
b4: 35,38

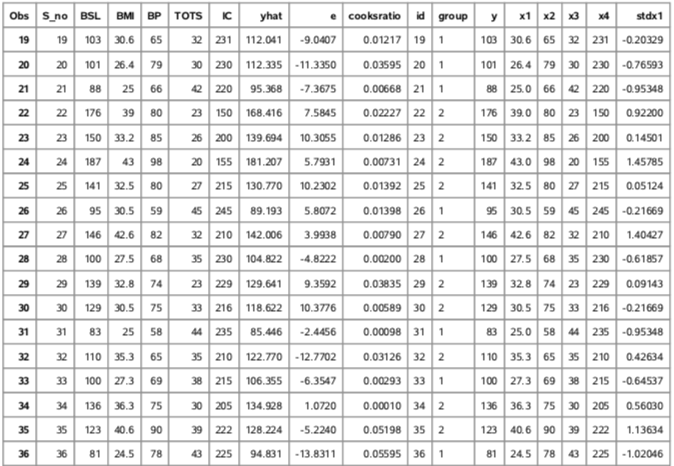
b5: 3,11,38

**c. Combined Influence on all the LSEs – Cook’s Distance**

The value of Cook’s Distance, Di for each observation is shown in the Table D Decision Rule: If Di > F (0.5; p, n-p), then outlier ‘i’ has a combined influence on all the LSEs Cut-off value: F (0.5; p, n-p) = F(0.5; 5, 35) = 0.88731

**Conclusion:** No observations were found to exceed the cut-off value of 0.88731 (from table 18). Therefore, it can be concluded that no observation is found to be higher than the cut-off value, there are no outliers that will influence the model.





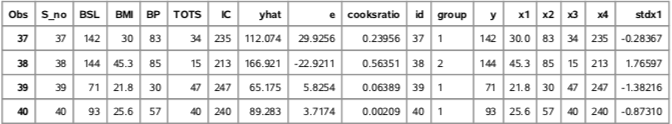


Table 18: Cook’s distance for preliminary model

**MODEL 2 – VERIFICATION OF ASSUMPTIONS**

**Residual Analysis  
1. Model Form**

The graphs are plotted with residual versus Body Mass Index (X1) Blood Pressure (X2) Thickness of Triceps Skin (X3) Insulin Content (X4) stdx1x3 & stdx3x4 to see the combined patterns for collinearity.

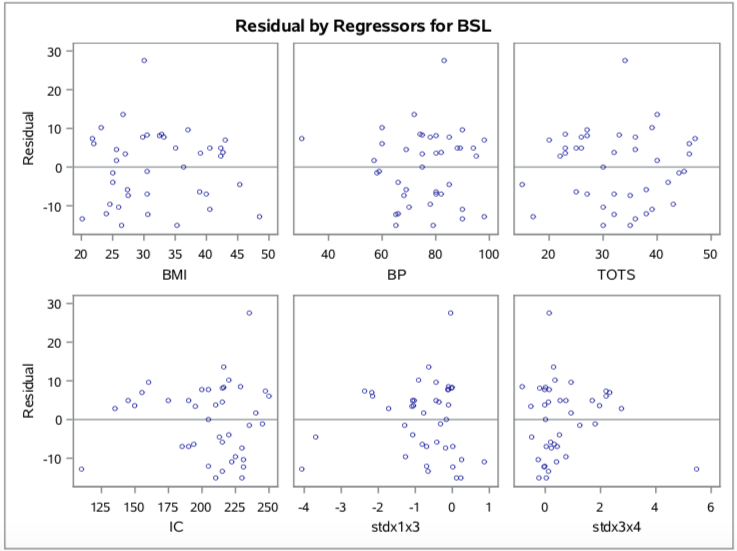


Figure 12: Residual Vs Predictors and Interaction terms of the Model2

From the figure 12, the residuals versus the independent predictors, Body Mass Index (X1) Blood Pressure (X2) Thickness of Triceps Skin (X3) Insulin Content (X4) does not show a curvature. It was also observed that the standardized interaction terms stdx1x3 & stdx3x4 are showing a linear trend. Therefore, it was concluded that the MLR Model form 2 is reasonable and that stdx1x3 & stdx3x4 can be added to the model.

**2. Constant Variance**

The following is the graph plotted between residuals and fitted values.

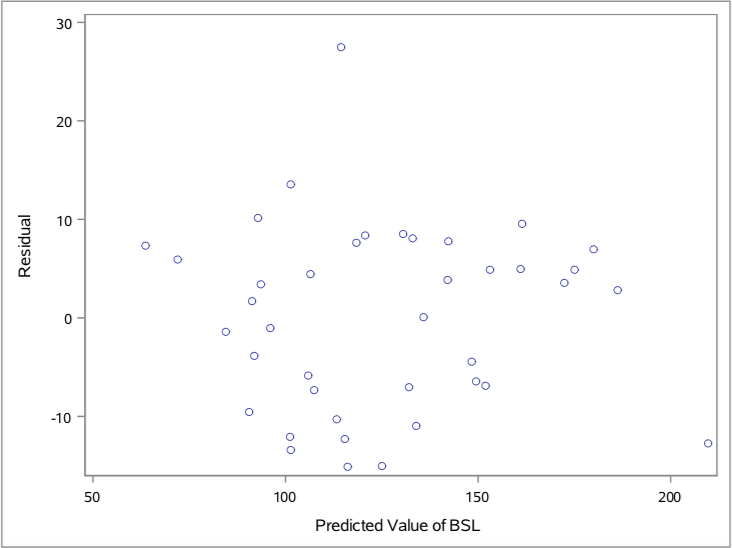


Figure 13: Residual Vs Fitted Values-Model 2

From the above Figure 13, the residual plot between Residual Vs Fitted values of Blood Sugar Level has a random point cloud. There is no funnel shape formation in the graph. Therefore, it was concluded that the Model 2 has a constant variance.

**3. Normal Probability plot for Residuals**

The graph between residual values and normal scores was plotted in SAS as shown below

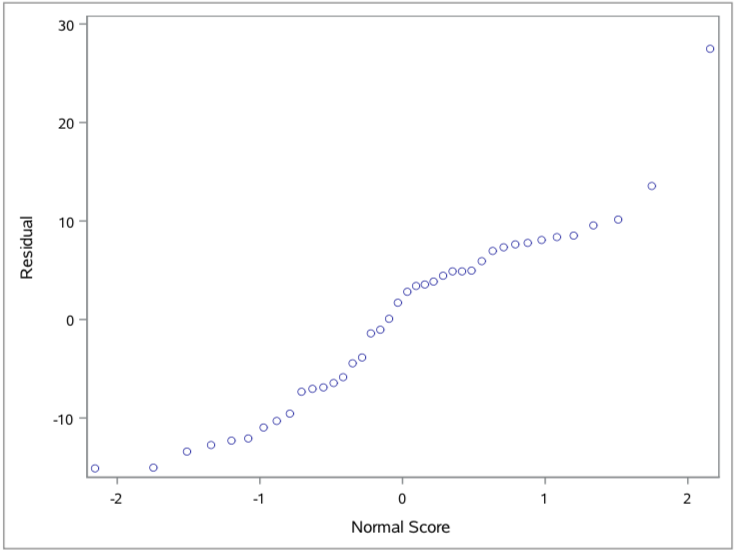


Figure 14: Normal Probability Plot – Model 2

From the Figure 14, the plot exhibits a pretty straight line but has a shorter tail Since the plot is predominantly a straight line we conclude that Normality of Model 2 is satisfied.

**4.Time Series Plot**

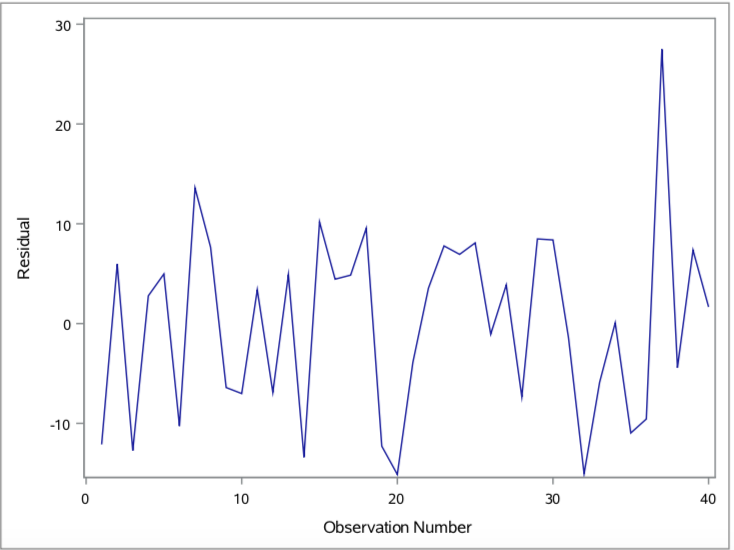
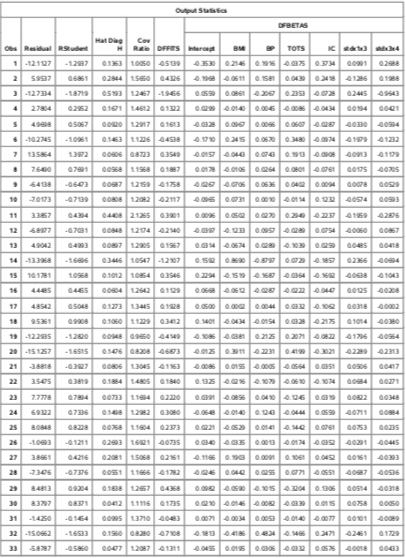


Figure 15: Time series plot of Model 2

The time series graph was plotted for model 1 and we can see that there is a random jaggedness throughout the graph without any trend being followed therefore we can say that the MLR Model 2 is reasonable.

**Diagnostics Check**



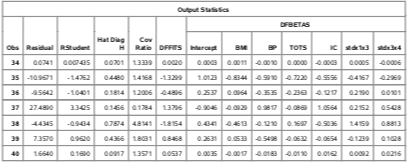


Table 19: Output Statistics

**Test for Y-Outliers-Bonferroni Outlier Test**

The Bonferroni Outlier test was used to check for any data points that’s are Y outlying or have extreme

Y values. The studentized deleted residual is represented by ti, p represents the number of parameters

and n represents the total observation points. The studentized residual, ti for each observation is

shown in the appendix.

**Decision Rule**: If |ti| > t (1-α/2n; n-p-1), then observation ‘i’ is a Y outlier.

**Cut-off Value**: t (1-α/2n; n-p-1) = t (1 – 0.05/80; 40-6-1) = t (0.999375; 33) = 3.53

**Conclusion**: All values of studentized residuals, |ti| are less than 3.53. Hence, there are no Y –

outliers.

The studentized deleted residual is represented as ti but here it is represented as RStudent, as seen in table 5. The RStudent value is calculated using SAS and it is compared with the cut-off value t (1-(α/2n), n - p - 1). We compare the cut-off value with the absolute value of RStudent to find the outliers

(n=40, p = 5, α = 0.05). The cut-off value is compared with the absolute value of RStudent from table 5, and noticed that, there is no value greater than the cut-off value. Therefore, we conclude there is no Y-outlier.

**Test for X Outliers**

Leverage values, hii is utilized to identify X outliers in data points; hii values for each value is

shown in table 5. We intend to find the x-outlier hence, we calculate the cut-off value given by ‘2p/n’, where p🡪 Number of predictors; n🡪 Number of observations

**Decision Rule**: If hii > 2𝑝/𝑛, then observation ‘i’ is an X outlier.

Cut-off Value: 2𝑝/𝑛=2∗6/40= 0.3

**Conclusion:** 6 data points have leverage value, hii > 0.3. The X – outliers are noted at:

Observations (3,11,14,35,38,39)

**Influential Test For Outliers**

We have identified outliers, and now we intend to verify if they have an influence on the model.

Computing DFFITS Cook’s distance will aid us.

**a. Influence on the Fitted Values – DFFITS**

DFFITS value for each observation is shown in Table 5.

**Decision Rule**: If |DFFITSi| > 2√(p/n), then outlier ‘i’ has an influence on the fitted values.

Cut-off value= 2√(p/n) = 2√ (6/40) = 0.7745966692.

**Conclusion:** Observations 3,14,35,38,39 have |DFFITS| values greater than 0.7745966692.

Hence, we infer that the above X outlying value have an influence on the fitted values.

**b. Influence on Individual LSEs – DFBETAS**

DFBETAS value for each observation is shown in Table 5.

**Decision Rule**: If |DFBETASik| >2/√𝑛, outlier ‘i’ has an influence on the Least Squares Errors.

Cut-off value: 2/ √𝑛 = 2/√40= 0.316227.

By comparing the cut-off value with the table values of DFBETAS, we infer that:

b0: 35,38

b1: 14,35,38

b2: 14,35,39

b3: 35

b4: 35,38

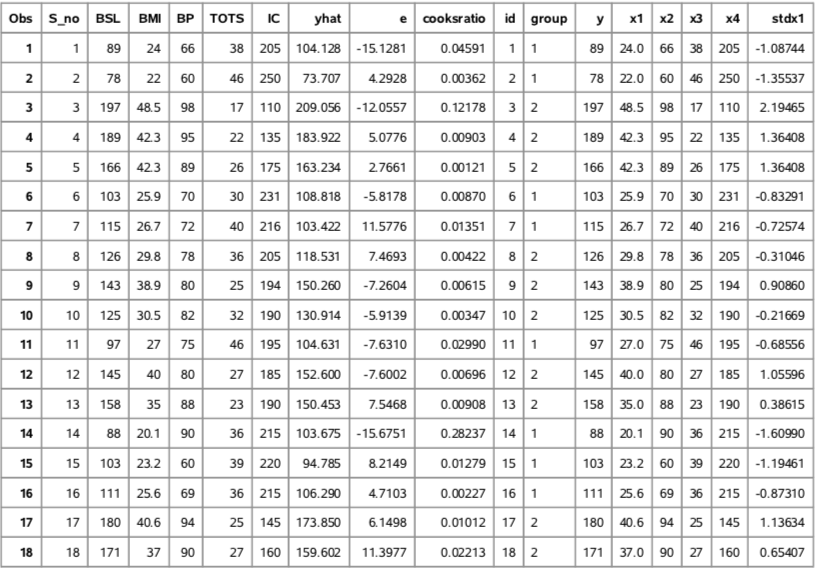
b5: 35,38

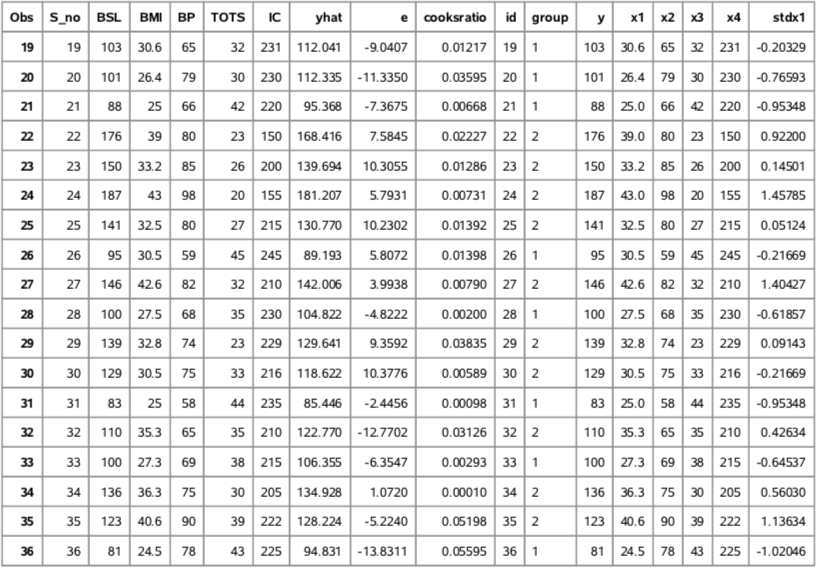
b6: 3,38

**c. Combined Influence on all the LSEs – Cook’s Distance**

The value of Cook’s Distance, Di for each observation is shown in the Table D Decision Rule: If Di > F (0.5; p, n-p), then outlier ‘i’ has a combined influence on all the LSEs Cut-off value: F(0.5; p, n-p) = F(0.5; 6, 34) = 0.91

**Conclusion:** No observations were found to exceed the cut-off value of 0.91 (from table 6). Therefore, it can be concluded that no observation is found to be higher than the cut-off value, there are no outliers that will influence the model.





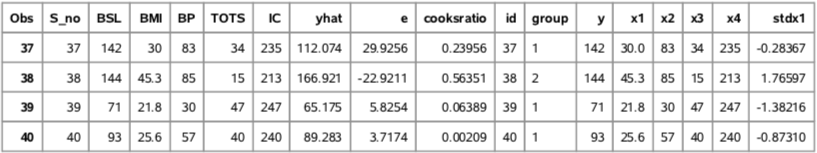


Table 20:

**Final Model Selection**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **PARAMETER** | **CONDITION** |  | **MODEL 1** |  | **MODEL 2** | **PREFERRED MODEL** |
| **Predictor Variables** | Less | 5 |  | 6 |  | Model 1 |
| **R2** | High | 0.9235 | | 0.9257 | | Model 2 |
| **Adjusted R2** | High | 0.9123 | | 0.9122 | | Model 1 |
| **MSE** | Low | 103.51435 | | 103.55986 | | Model 1 |
| **SSE** | Low | 3519.48775 | | 3417.47554 | | Model 2 |
| **Normality** | Normality Should be Satisfied | Normality Satisfied | | Normality Satisfied | | Model 1 |
| **Constant Variance** | Constant Variance | Constant Variance | | Constant Variance | | Equal |
| **Mallows Cp** | Approximately  Equal to number of parameters | 5.9851 | | 7.0000 | | Model 1 |
| **Multicollinearity** | Average VIF should  be less than 5.0 | The Average  VIF Factor is 2.34. | | The Average  VIF Factor is 3.11. | | Model 1 |

Table 21: Model Selection Criteria

The Table 21, draws comparisons between the two selected models. The R2 value of Model 2 is marginally higher than Model1, we can imply that Model 1 explains almost the same variability in Blood Sugar Levels as Model 2 and Adjusted R2 value of Model 1 is marginally higher than Model 2. Model 1 has a lower MSE and SSE value when compared to Model 2 . When checked for multicollinearity, it was observed that the VIF values of the predictors in the Model 1 are better than that of Model 2, Model 1 has an average VIF value of 2.34 while Model 2 has an average value of 3.11. Additionally, The Mallow’s Cp value is lesser in the case of model 1 and also close to the number of parameters. We also note that model 1 has all the predictors with p-values much less than 0.10, thus we can say that we are 90% confident that the regression model for model 1 is significant. Considering all the factors necessary we conclude that Model 1 is our best model and has been chosen for further analysis

Selected Model: Blood Sugar Level (Y)= 207.08082 + 1.49524\*Body Mass Index – 1.61763\*Thickness of Triceps Skin – 0.36383\*Insulin Content + 6.28038\* stdX1X3 + 4.09212\* stdX3X4

VI FINAL MULTIPLE LINEAR REGRESSION MODEL

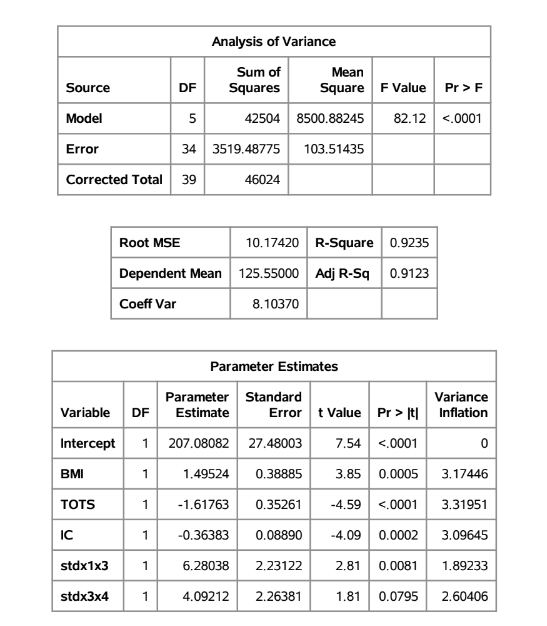


Table 22: Final MLR Model

The final model that we have selected is:

Blood Sugar Level (Y)= 207.08082 + 1.49524\*Body Mass Index – 1.61763\*Thickness of Triceps Skin – 0.36383\*Insulin Content + 6.28038\* stdX1X3 + 4.09212\* stdX3X4

We regress Blood Sugar Level on the predictors, Body Mass Index, Thickness of Triceps Skin, Insulin Content, stdX1X3 and stdX3X4 we obtained the above SAS Output as shown in Table 19.

We infer that our chosen MLR Model is good as a result of the linear relation between Blood Sugar Levels and the predictors. R2 square value for the chosen model is 0.9235, which implies that the model explains 92.35% of the variations in Blood Sugar Levels. We also make a note of the p – values for each of the predictors and each of them is less than 0.10, which suggests that we have a confidence of 90% for the model. We observe the VIF values are all less than five indicating that the model does not have the issue of serious multicollinearity.

INFERENCES ON THE PARAMETERS

We intend to perform inference tests and compute the intervals for our models at the level of significance α = 0.05, the number of parameters in our model is given by p=6.

Bonferroni Joint Confidence Interval for the parameters**:**

This helps us get the result of confidence intervals where all predictors lie simultaneously following other calculations.

α = 0.05, n=40, p=6, g (predictor variables) =5

The Bonferroni coefficient B= t (1-α/2g; n-p) =t (1-0.05/2(5);40-6) = t (0.995,34) = 2.728

1. CI for β1 = b1 ± B× s{b1} = 1.49524 ± (2.728) \* 0.38885 = (0.4344, 2.5560)

2. CI for β2 = b2 ± B× s{b2} = -1.61763 ± (2.728) \* 0.35261 = (-2.5795, -0.6557)

3. CI for β3 = b3 ± B× s{b3} = -0.36383 ± (2.728) \* 0.08890 = (-0.6063, -0.1213)

4. CI for β4 = b4 ± B× s{b4} = 6.28038 ± (2.728) \* 2.23122 = (0.1936, 12.3671)

5. CI for β5 = b5 ± B× s{b5} = 4.09212 ± (2.728) \* 2.26381 = (-2.0835, 10.2677)

**Conclusion:** We are 95% confident that β1 lies in (0.4344, 2.5560), β2 lies in (-2.5795, -0.6557),

β3 lies in (-0.6063, -0.1213), β4 lies in (0.1936, 12.3671), β5 lies in (-2.0835, 10.2677) simultaneously.

CONFIDENCE INTERVAL, PREDICTION INTERVAL AND CONFIDENCE REGION FOR MEAN RESPONSE AT Xh

We select the values for predictors as xh at 10th observation xT = (1, 30.5, 32, 190, 0.01544, 0.03393)

Confidence Interval for Mean Response

95 % Confidence Interval of mean response at X10­:

t (1-α/2; n-p) = t (1 -0.05/2; 40-6) = t (0.975;34) = 2.032

95% Confidence Interval for mean response (Blood Sugar Level):

= 130.914± 2.032\*2.892802635= 130.914±5.878174954 = (125.035825, 136.792175)

**Conclusion**: We are 95% confident that the Confidence Interval of mean Blood Sugar Level when the independent variables are (1, 30.5, 32, 190, 0.01544, 0.03393) will lie between 125.035825mg/dL and 136.792175 mg/dL.

*6.1.2.2 Confidence Region for The Entire Regression Surface:*

Working-Hotelling 1- α CR with Limits:

Where, = 6\* = 6 \* 2.38= 14.28

95% Confidence Region: 130.914± 3.7788\*= = (119.98499,141.84300)

**Conclusion**: We are 95% confident that the entire regression surface for the mean response (Blood Sugar Level) when the independent variables are (1, 30.5, 32, 190, 0.01544, 0.03393) will lie between boundary values 119.98499 mg/dL and 141.84300 mg/dL.

6.1.2.3 Prediction Interval for Mean Response:

95% Prediction Interval for xh =

95% Prediction Interval for xh =

= 130.914 ± 2.032\*29.43195= 130.914± 59.8057= (71.1082, 190.7197)

**Conclusion**: We are 95% confident that the actual values of Blood Sugar Levels when the independent variables are (1, 30.5, 32, 190, 0.01544, 0.03393), will lie between 71.1082 mg/dL and 190.7197 mg/dL.

VII. FINAL DISCUSSION:

Through this project we aimed to analyze the blood sugar levels in pregnant women diagnosed with Hyperglycemia (a diabetic condition) . We put to use general multi linear regression equation Yi = β0 + β1xi1 + ⋯ + βkxik + εi to estimate the blood sugar levels for the preliminary model. During the process, we checked the model assumptions by regression of blood sugar levels on the predictors Body Mass Index(BMI), BloodPressure(BP), Thickness of Triceps Skin(TOTS) and Insulin Content and assumptions. We noted that the tests for linearity, constant error variance and normality were satisfied for the preliminary model. We also observed that the model did not have serious multicollinearity problem. After this we checked for x & y outliers using leverage and Bonferroni outlier test and influences were checked. We observed no influential outliers.We proceeded with this model and found that adding the interaction term with the predictors explains variability better, thereby we added interaction terms and after further analysis we selected two models by model search procedures that is best subset, backward deletion and stepwise regression out of which the best two models were decided for further analysis.After comparison we chose Model I as our best model and we found that Model I has constant variance, no influential outlier and better coefficient of determination.

FUTURE ANALYSIS:

In order to obtain accurate results/outputs we consider the addition of variables that cause an impact. As we have blood sugar levels as our decision variable, we need to consider variables such as cholesterol levels,carbohydrate contents,fat percentage,protein level,white blood cells and red blood cells count ,haemoglobin concentration in blood and pancreas functionality for improved accuracy in blood sugar levels.

VIII.REFERENCES

Applied Linear Statistical Models by Neter, Kutner, Nachtsheim, Wasserman.

Multiple linear regression notes and Handouts by Dr. Aera Kim Leboulluec

<https://gist.github.com/ktisha/c21e73a1bd1700294ef790c56c8aec1f>

IX.GANTT CHART



X. APPENDIX