***Oral Glucose Tolerance Test (OGTT) Linear Regression******Analysis***

Elombe Calvert MS, MBBS

*Using Multiple Linear Regression to Test the Association between Oral Glucose Tolerance Test (OGTT) Blood Glucose Levels and Age, BMI, Systolic Blood Pressure, Diastolic Blood Pressure, Income to Poverty Ratio, Total Cholesterol, Triglycerides, Apolipoprotein, Glycosylated Hemoglobin, and Plasma Insulin Levels.*

**Question:**

Does there exist an association between blood glucose levels from the Oral Glucose Tolerance Test (OGTT) used to diagnose diabetes and Age, BMI, Systolic Blood Pressure, Diastolic Blood Pressure, Income to Poverty Ratio, Total Cholesterol, Triglycerides, Apolipoprotein, Glycosylated Hemoglobin, and Plasma Insulin Levels?

**Introduction:**

Diabetes Mellitus type 2 is a chronic condition in which the body's pancreas does not produce enough insulin or does not use insulin effectively due to acquired insulin resistance. This results in high blood sugar levels, which can damage the body's organs and lead to a range of complications such peripheral neuropathy, diabetic retinopathy, cardiovascular disease, Nephropathy among others.

Risk factors for diabetes type 2 include being overweight or obese, having high blood pressure, having high levels of lipids (fats) in the blood, being older, having a low income to poverty ratio, having a high BMI (a measure of body fat), and having high levels of glycosylated hemoglobin (a measure of long-term blood sugar control).

The gold standard for diagnosing diabetes is the Oral Glucose Tolerance Test (OGTT). In this test, a person is given a sweet drink containing a specific amount of glucose, and their blood sugar levels are measured at regular intervals over the next two hours.

Thus, in this study, I will use multiple linear regression to investigate the relationship between blood glucose levels from the OGTT test, and several potential predictor variables including blood pressure, lipids, age, income to poverty ratio, Apolipoprotein B, BMI, plasma insulin levels, and glycosylated hemoglobin. By carefully analyzing the strength and direction of the association between these variables, we will be able to determine whether they are significant predictors of OGTT blood glucose levels which can be used in the diagnosis of diabetes. Through this analysis, I aim to provide valuable insights that can improve our understanding of the factors that contribute to diabetes.

**Data Processing**

In this study, patients without OGTT levels were excluded from the analysis. To handle missing values, the median was used to impute continuous variables due to its robustness against non-normality and outliers. This method was chosen over simply dropping missing values from the dataset. In clinical practice, a single blood pressure measurement is not considered very useful, so the decision was made to average the four systolic and diastolic blood pressure measurements, resulting in two variables. All predictor variables were scaled to a common scale, but the outcome variable (OGTT levels) was left unscaled so that the fitted values of the model could be easily interpreted.

**Data Analysis**

[1] Regression Assumptions

In determining correlation among variables, a collinearity matrix between the variables showed only two reasonably correlated variables, which were total Cholesterol and Apolipoprotein B = 0.86, and Glycosylated Hemoglobin and OGTT levels = 0.62. In addition, from the VIF bar plot, only two predictors showed significant multicollinearity, Apolipoprotein and Total cholesterol, which is consistent with findings from our correlation matrix. Since the literature has little evidence to support Apolipoprotein B as a good predictor/ risk factor for diabetes, it will be dropped from the model. From the bar plot of the residual values from the model, the residuals are approximately normally distributed with a slight right skew. The linearity assumption was checked by looking if there exists any pattern seen in the fitted values which is indicated by the shape of the red line in the plot of residuals versus fitted values below. The red line is not horizontal but has a non-linear shape, thus indicating a non-linear relationship between the outcome variable and the predictors. From the Scale-Location plot below, the residuals display Heteroscedasticity which is non-constant variance in the residuals with differing fitted values.

[2] Variable selection

The base model which only includes the intercept and the full model which has all predictor variables excluding `Apolipoprotein B` was defined. This was done to reduce the model complexity and allow for more interpretable coefficients. To perform feature selection and construct the most robust model, iterative feature selection was used via stepAIC, backward selection. Thus, I started with a full model containing all the predictor variables, then insignificant predictors are iteratively removed leaving only variables that explain much of the variance in the target variable. This was ideal in this study where the aim is to assess the relationship between predictors and the outcome variable by finding only the features with a significant relationship with the outcome of interest. The features selected through this process as being important to outcome prediction are Age (RIDAGEYR), BMI (BMXBMI), Triglycerides (LBXTR), Glycosylated hemoglobin (LBXGH), Plasma insulin levels (LBXIN) and Systolic blood pressure (BPXSY) which are all significantly associated with the outcome, OGTT levels. The final model contains only 6 out of 10 predictor variables, all of which are statistically significant.

[3] Evaluate model fit

In assessing how well the model fits to the data, the **coefficient of determination**, also known as **R-squared** will be evaluated. In the final model, the R-squared was **0.426**. Thus, the predictor variables in the final model, RIDAGEYR, BMXBMI, LBXTR, LBXGH, LBXIN and BPXSY was able to explain ~43% of the variance that exists in the outcome variable LBXGLT.

[4] Alternative Model

In making the alternative model, of the 6-predictor variables in the final model, only two variables directly relate to blood glucose levels physiologically in the human body. Thus, a model was be constructed and evaluated based solely on Serum Plasma Insulin and Glycosylated Hemoglobin. The metric that was used to compare both the final model and the alternative model is the **Akaike information criterion (AIC),** where a lower AIC value indicates a better-fitting model. The AIC of the alternative model is **23545** whereas the AIC of the final model is **23443.8**. Thus, the final model has a lower AIC which means that it is better able to explain the data than the alternative model. The final model is the better model.

[5] Hypothesis Testing

In hypothesis testing for coefficients, the null hypothesis was stated as, H0: All the coefficients in the regression model are equal to zero and the alternate hypothesis, H1: At least one of the coefficients is significantly different from zero. Having an F-statistic of 290 with a significant p-value of 4.32596622148542e-278, we can reject the H0 and accept H1 that at least one coefficient is significantly different from zero. In essence, at least one of our predictor variables in our final model has a significant association with our outcome of interest.

In **conclusion**, multiple linear regression as a statistic method was used to determine significant associations between predictor variables Age, BMI, Triglycerides, Glycosylated Hemoglobin, Plasma insulin levels, Systolic blood pressure and our outcome variable OGTT blood glucose levels.