**Question**

The data was collected and made available by National Institute of Diabetes and Digestive and Kidney Diseases as part of the Pima Indians Diabetes Database. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here belong to the Pima Indian heritage (subgroup of Native Americans), and are females of ages 21 and above. The task is to investigate the factors associated with diabetes (outcome; 1 = Presence of diabetes, 0 =absence of diabetes)

**INTRODUCTION**

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. A study was conducted by the National Institute of Diabetes and Digestive and Kidney Diseases of females of ages 21 and above from a subgroup of Native Americans, known as the Pima Indian heritage, to provide data as part of the Pima Indians Diabetes Database.

The study had a specific objective which was to test for the factors associated with diabetes such as pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, Diabetes pedigree function and age. If the outcome is 1 then there was presence of diabetes but if the outcome was 0, then there was absence of diabetes in the person’s system.

The following hypothesis was formulated in order to achieve the above objective:

H0: There is no association between having diabetes and risk factors like pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, Diabetes pedigree function and age.

VS

H1: There is an association between having diabetes and those risk factors.

**Methods Used**

**Data**

The data is comprised of 8 predictors as described above with the response variable of whether or not someone is diabetic. The data was provided by the Lecturer.

**Model**

Logistic regression is the method used because the outcome variable is categorical. The data had 8 dependent variables each of which were continuous variables. The binary logistic regression was used, since the outcome was binary, yes/no to whether someone had diabetes.

**Libraries used**

Outliers- Used to check and eliminate outliers in the, basically clean the data.

The outlier function, used to find values with large differences from the sample mean, which can be outliers

Caret- this is a classification and regression training library. This package contains functions that streamline the model training process for regression and classification problems.

CreateDataPartition function was used. Which creates a series of test and training partitions.

Car- creates variance inflation and variance inflation factor for the generalized linear model.

**Analysis**

The model used according to the variables was binary logistic regression linear model. The steps in the analysis are:

**Cleaning the data**

The any (is.na ()) function was used to check for null values in the data.

The outlier function from the Outliers library was used to determine the outliers far from the mean in the data.

Hist function was added to visualize the distribution of the predictive variables.

The last process of this process was replacing the outlier values with the medians of their respective predictor variables.

**Partitioning the data**

Using the createDataPartiton function the data was split into test and training data. The training data is the used to come up with the model.

**Choosing a model**

After data preparation models with various combinations of the data were fitted and the best model chosen based on various determining factors.

Residuals and the VIF were also obtained from this process.

**Testing Assumptions**

First assumption for normality of the model was tested using the Shapiro test on the residuals.

Second assumption for constant variance was tested using the bptest

Last assumption of autocorrelation was done using the dwtest from the car library.

**Goodness of fit**

The HLtest function computes the classical Hosmer-Lemeshow (1980) goodness of fit test for a binomial glm object in logistic regression.

The final part of the analysis was determining the odds ratio of the model using the exp () function.

**Dividing the data into training and testing set**

The data was divided into two sets: Training and testing sets. We did this in order to improve the accuracy of the model. The data was divided into an 80 20 ratio whereby 80 was the training set and 20 was the testing set. The training set was used in developing our final model.

**Results/Interpretation**

Our final formula is

Log(p/(1-p) = -9.27321 + 0.15213(Pregnancies)+ 0.03810(Glucose) + 0.08779(BMI)+ 0.90231(DiabetesPedigreeFunction)

Our final formula is

Log(p/(1-p) = -9.27321 + 0.15213(Pregnancies)+ 0.03810(Glucose) + 0.08779(BMI)+ 0.90231(DiabetesPedigreeFunction)

The predicted value **(probability of diabetes)** for pregnancy is

Log(p/(1-p)) = -9.27321 + 0.15213(1) + 0.03810(0) + 0.08779(0) + 0.90231(0)

Log(p/(1-p)) = -9.12108

**exp (-9.12108) / (1+ exp (-9.12108)) = 0.000193**

The predicted value **(probability of diabetes)** for glucose level is

Log(p/(1-p)) = -9.27321 + 0.15213(0) + 0.03810(1) + 0.08779(0) + 0.90231(0)

Log(p/(1-p)) = -9.19911

**exp (-9.19911) / (1+ exp (-9.19911)) =0.000101**

The predicted value **(probability of diabetes)** for BMI is

Log(p/(1-p)) = -9.27321 + 0.15213(0) + 0.03810(0) + 0.08779(1) + 0.90231(0)

Log(p/(1-p)) = - 9.18542

**exp (-9.18542) / (1+ exp (-9.18542)) =0.000103**

The predicted value **(probability of diabetes)** for diabetes pedigree function is

Log(p/(1-p)) = -9.27321 + 0.15213(0) + 0.03810(0) + 0.08779(0) + 0.90231 (1)

Log(p/(1-p)) = -8.37079

**exp (-8.37079) / (1+ exp (-8.37079)) =0.000231**

**Checking accuracy of our the P-value of pregnancies, glucose, BMI and DiabetesPedigreeFunction were less than the level of significance. This meant that the predictors were probably an excellent addition to the model.**

**Conclusion:**

Using the coefficient estimates from our model, we can conclude that when a female is pregnant, has increasing glucose level, insulin, BMI, Diabetes Pedigree Function and skin thickness will likely have diabetes i.e.:

With a coefficient estimate of glucose = 0.039808 shows the probability of being diabetes increases with increase in glucose and the odds of having diabetes are higher with an increase in glucose level. An increased BMI might also indicate a risk of developing diabetes and normally, there is a high risk of developing diabetes as the age of the person increases (given other factors).

This shows that the factors do affect diabetes.

With accuracy of 0.7756098 of our model, indicates that 78% of the time our model classified the patients in a high-risk category when they actually had a high risk of getting diabetes. This shows our model is good. Assumptions of logistic regression hold ie:

* Little or no multicollinearity. For our model there is no multicollinearity between the independent variables
* Autocorrelation. This assumption holds as the p-value= 0.3747 is greater than the alpha level of 0.005 and with D = 1.9746 shows positive autocorrelation.
* Normality. For logistic regression, since it does not require a linear relationship between dependent and independent variables, normality do not need to be normally distributed. With p-values less than alpha level, the normality assumption does not hold.

**References:**

Machine learning Essentials: Practical Guide in R by Alboukadeal KASSAMBARA

Applied Logistic Regression by David W. Hosmer Jr., Stanley Lemeshow, And Rodney X. Sturdivant

Links;

<https://rpubs.com/soodrk/578110>

<https://www.kaggle.com/logeshk/pima-indians-diabetes-logistic-regression>

<https://rstudio-pubs-static.s3.amazonaws.com/534347_d1d5858091154928a404b40f67de2cb1.html>

<https://youtu.be/XycruVLySDg>