Logistic Regression

Aleksandra Kalisz
Postgraduate Diploma in
Science in Data Analytics
National College of Ireland
Dublin, Ireland
21118876@student.ncirl.ie

***Abstract*— The aim of the assignment is to estimate a binary logistic regression model to facilitate diabetes diagnosis based on blood results. The dataset is split into a suitable training and test dataset, and the models are evaluated on the test dataset using a confusion matrix. The final model is then tested on the rest of the data. The assignment concludes with the calculation of the probability that these prediabetic cases are diagnosed as diabetic.**

Introduction

This assignment focuses on analyzing a dataset of blood sample details of diabetic patients collected in an Iraqi University Hospital in 2020 to estimate a binary logistic regression model that can facilitate diabetes diagnosis based on blood results. The analysis includes exploratory data analysis, splitting the dataset into a suitable training and test dataset, and evaluating the models on the test dataset using a confusion matrix. The objective of this project is to evaluate the use of binary logistic regression modelling techniques to predict the likelihood of prediabetic patients developing diabetes based on their blood sample.

Binary Logistics

Binary logistic regression is a statistical method used to model binary outcomes. The algorithm estimates the probability of the binary outcome based on the values of the independent variables. The model generates coefficients for each independent variable, indicating the strength and direction of the relationship between predictor and outcome. The model is fit by maximizing the likelihood function and evaluated using measures such as the confusion matrix. The resulting model can be used to predict the probability of the binary outcome for new observations.[i]

$$p = 1 / (1 + exp(-(b0 + b1x1 + b2x2 + ... + bk*xk)))$$

where:

p is the predicted probability of the binary outcome (e.g., the probability of a patient developing diabetes).

exp - is the exponential function.

b0 - is the intercept (the constant term in the model). b1, b2, **bk -** are the coefficients for each independent variable (predictor)

x1, x2, ..., xk. x1, x2, ..., xk are the values of the independent variables (predictors) for the observation being predicted.

The Data

In this assignment, we analyze a dataset containing blood sample details of diabetic patients collected in an Iraqi University Hospital in 2020. The dataset includes 12 relevant variables which describe:

Gender -Male/Female.
AGE -Patient age.
Urea - A diamine, chief nitrogenous waste product in humans.
Cr - Creatinine Ratio, a parameter to assess kidney function.
HbA1c -Average blood glucose (sugar) Levels
Chol - Cholesterol, a parameter to assess liver function.
TG - Triglycerides a type of fat in the blood used to transport energy.
HDL -High-density lipoprotein, the “good” cholesterol.
LDL - Low-density lipoprotein, the “bad” cholesterol.
VLDL - Very-low-density lipoprotein cholesterol.
BMI - Body-Mass-Index.
CLASS - N/Y/P – class of diabetes.

```
'data.frame':   1000 obs. of  14 variables:
 $ ID        : int  502 420 680 634 721 759 636 788 82 132 ...
 $ No_Pation: int  17975 47975 87656 34224 34225 34230 34231 34232 46815 34
 $ Gender    : chr  "F" "F" "F" "F" ...
 $ AGE       : int  50 50 50 45 50 32 31 33 30 45 ...
 $ Urea      : num  4.7 4.7 4.7 2.3 2 3.6 4.4 3.3 3 4.6 ...
 $ Cr        : int  46 46 46 24 50 28 55 53 42 54 ...
 $ HbA1c     : num  4.9 4.9 4.9 4 4 4 4.2 4 4.1 5.1 ...
 $ Chol      : num  4.2 4.2 4.2 2.9 3.6 3.8 3.6 4 4.9 4.2 ...
 $ TG        : num  0.9 0.9 0.9 1 1.3 2 0.7 1.1 1.3 1.7 ...
 $ HDL       : num  2.4 2.4 2.4 1 0.9 2.4 1.7 0.9 1.2 1.2 ...
 $ LDL       : num  1.4 1.4 1.4 1.5 2.1 3.8 1.6 2.7 3.2 2.2 ...
 $ VLDL      : num  0.5 0.5 0.5 0.4 0.6 1 0.3 1 0.5 0.8 ...
 $ BMI       : num  24 24 24 21 24 24 23 21 22 23 ...
 $ CLASS     : chr  "N" "N" "N" "N" ...
```

Figure 1. Data Set

Cleaning and transformations of the Diabetes dataset involves removing the data entries with the "P" class, which represents prediabetic patients. These entries are excluded from the analysis as the objective is to predict whether prediabetic patients will develop diabetes. The unnecessary variables ID and No_Pation are also removed from the dataset. Next, the code checks for missing values in the remaining variables to ensure the quality of the data. After that, the Gender and CLASS variables are transformed from character strings to binary variables for modelling purposes. The "F" in the Gender variable is recoded to 0 to represent females, while the remaining values (i.e., "M") are recoded to 1 to represent

males. The CLASS variable is recoded to 0 to represent patients without diabetes and 1 to represent patients with diabetes. Finally, the AGE and Cr variables are transformed from character strings to numeric values to enable numerical analysis. A final check of the data is performed to ensure that all variables are in the appropriate format.

```
'data.frame':   947 obs. of  12 variables:
 $ Gender: num  0 0 0 0 0 0 0 0 0 0 ...
 $ AGE   : num  50 50 50 45 50 32 31 33 30 45 ...
 $ Urea  : num  4.7 4.7 4.7 2.3 2 3.6 4.4 3.3 3 4.6 ...
 $ Cr    : num  46 46 46 24 50 28 55 53 42 54 ...
 $ HbA1c : num  4.9 4.9 4.9 4 4 4 4.2 4 4.1 5.1 ...
 $ Chol  : num  4.2 4.2 4.2 2.9 3.6 3.8 3.6 4 4.9 4.2 .
 $ TG    : num  0.9 0.9 0.9 1 1.3 2 0.7 1.1 1.3 1.7 ...
 $ HDL   : num  2.4 2.4 2.4 1 0.9 2.4 1.7 0.9 1.2 1.2 .
 $ LDL   : num  1.4 1.4 1.4 1.5 2.1 3.8 1.6 2.7 3.2 2.2
 $ VLDL  : num  0.5 0.5 0.5 0.4 0.6 1 0.3 1 0.5 0.8 ...
 $ BMI   : num  24 24 24 21 24 24 23 21 22 23 ...
 $ CLASS : num  1 1 1 1 1 1 1 1 1 1 ...
```

Figure 2. Cleaned Data Set

After cleaning data histograms with density function are created for the original variables in the dataset, including AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, and BMI. The purpose is to check the normality of the data and determine if any transformations are needed. The function histogram is defined to create the histograms with normal distribution curves. After creating the histograms, the data is transformed to improve normality for the variables that are skewed or have a non-normal distribution.

The transformations include:

```
# Transform Urea variable using log transformation
diabetes$Urea <- log(diabetes$Urea)

# Transform HbA1c variable using square root transformation
diabetes$HbA1c <- sqrt(diabetes$HbA1c)

# Transform LDL variable using reciprocal transformation
diabetes$LDL <- 1/diabetes$LDL

# Transform VLDL variable using natural logarithm transformati
diabetes$VLDL <- log(diabetes$VLDL)

# Transform BMI variable using square transformation
diabetes$BMI <- log(diabetes$BMI)
```

Figure 3. Data Transformations

Finally, the histograms are created again for the transformed variables to verify normality.

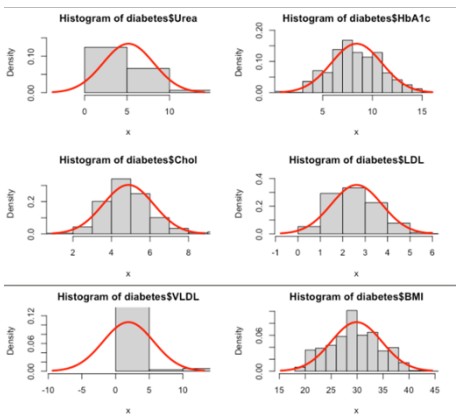

Figure 4. Histograms of Transformed Data

Exploratory Data Analysis

To gain insight into the relationships between the variables and the target variable we used various visualization techniques like histograms, scatterplots, and correlation matrix to have good understanding of the data.

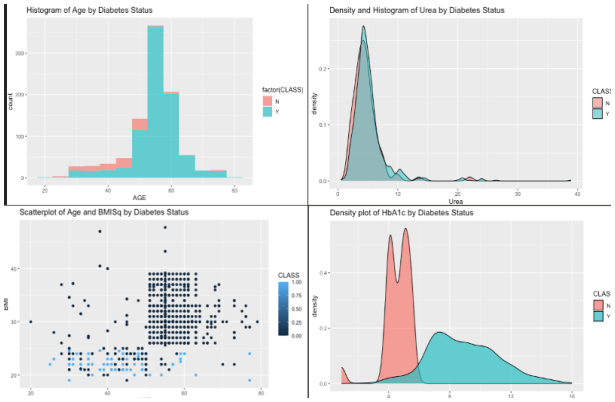

Figure 5.Relationship Between Different Variables Plots

The bar chart indicates that the number of patients examined for diabetes was roughly the same for both diabetic and non-diabetic patients. However, the age distribution of the patients differed between the two groups. Non-diabetic patients were more frequently recorded in the age range of 45-65, while diabetic patients were also often recorded in the age range of 60 and above. Furthermore, the chart indicates that there were very few people in the age range of 30-40 who were tested positive for diabetes. These findings suggest that age is a significant factor in the prevalence of diabetes, and that early testing for diabetes may not be a priority for individuals in their 30s and 40s.

The plot of Urea by Diabetes status shows that the distribution of Urea levels is similar for both groups, with a peak density around 0.3. However, there is a slightly higher density for individuals with diabetes compared to those without diabetes. The plot suggests that Urea levels may not be a good predictor for differentiating between diabetes and non-diabetes, as the distribution is relatively similar for both groups.

The scatter plot "Age vs BMI by Diabetes Status" indicates that patients without diabetes have a relatively stable BMI ranging from 20 to 25 across all ages. However, for patients with diabetes, the BMI increases with age, with a peak of nearly 40 for age group 50-65. Additionally, the plot shows that there are some patients younger than 40 years old with a BMI above 35 units, which may indicate early-onset obesity and an increased risk of developing diabetes. Overall, the plot highlights the importance of monitoring BMI as a risk factor for diabetes and the need for early intervention in patients with elevated BMI levels.

Based on the density plot of Blood Sugar (HbA1c), it can be observed that the non-diabetic group (represented in red) has the highest density of HbA1c values between 3.8 and 4.6. Additionally, the density of HbA1c values between 5 and 13 for non-diabetics reaches almost 0.2. This suggests that

there is a clear difference in the distribution of HbA1c values between the diabetic and non-diabetic groups. The plot highlights the importance of HbA1c as a diagnostic marker for diabetes and emphasizes the need for regular monitoring of blood sugar levels to maintain good health.

To have an even better understanding of the data we created corelation matrix:

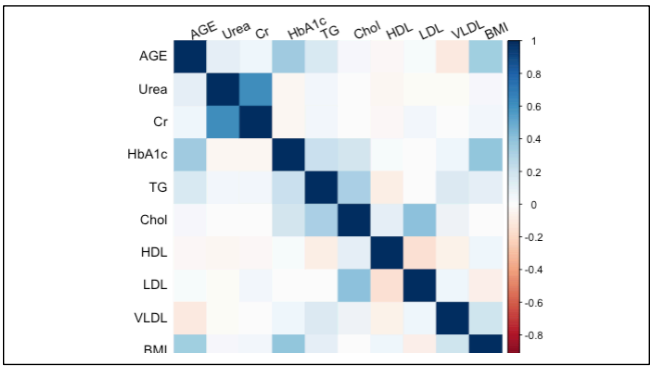

Figure 6. Correlation Matrix

A correlation matrix shows the pair- wise correlation coefficients between different variables. The coefficient ranges from -1 to 1, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and 1 indicates a perfect positive correlation. In our matrix we can see high correlation between Cr and Urea, Blood sugar and Age, Age and BMI, Chol and TG, LDL and Chol. This means that if one variable increases the other variable more likely to increase as well.

Looking at matrix we would like to investigate the Creatinine levels (Cr). To do that we created a scatter plot of its levels by diabetes status:

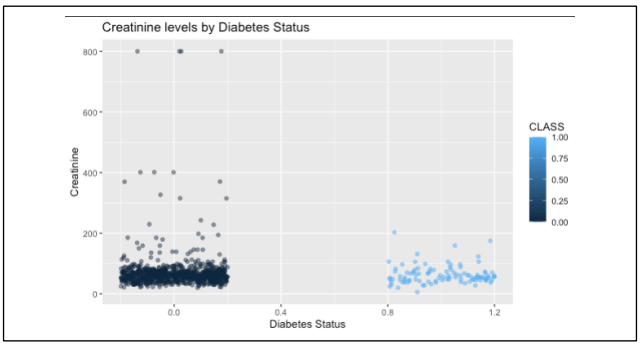

Figure 8. Creatinine Levels

From this plot we can see that there are about five outliers at scale of 800. An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. Outliers can be caused by measurement or recording errors, natural deviations in populations, or due to the existence of some underlying causal factors. In the case of the 'Cr' variable in the diabetes dataset, if the value of 800 is due to an error in measurement or recording, then it should be removed from the analysis. However, if it is a genuine value and there are no other errors, it should not be removed as it is an important part of the data and could provide important insights into the population being studied.

After little bit of investigation on the internet, it shows that average Creatinine levels for men and women are in the range of 0.7 to 2.0 in these circumstances I decided to remove outliers in range of 800. [ii]

After saving the new dataset we produced the boxplot of all the variables:

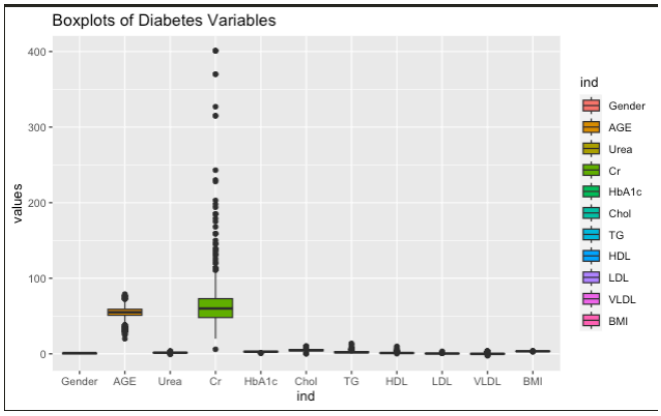

Figure 9. Boxplot of All Variables

In the above boxplot we can see median of Cr and AGE is higher than the rest of the variables, outliers are still present, but we decided in the dataset as they could be investigated further.

Splitting the Data

The next step is to split the data. We used function to split the data randomly.

```
set.seed(123)
index <- createDataPartition(diabetes$CLASS, p = 0.7, list = FALSE)
train <- diabetes[index,]
test <- diabetes[-index,]
```

Figure 7. Splitting the Data

Fitting the Model

Using binary logistic regression model to predict diabetes diagnosis using the training data. The glm() function is used to fit the model, with the CLASS variable as the response variable and the other variables as predictors. The family argument is set to "binomial" to specify a binary logistic regression model.

```
# fitting a binary logistic regression model to predict diabetes diagnosis.
reg_model <- glm(CLASS ~ AGE + Urea + Cr + HbA1c + Chol + TG + HDL + LDL + VLDL + BMI,
          data = train, family = binomial)

summary(reg_model)
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
Call:
glm(formula = CLASS ~ AGE + Urea + Cr + HbA1c + Chol + TG + HDL +
    LDL + VLDL + BMI, family = binomial, data = train)

Deviance Residuals:
      Min        1Q    Median        3Q       Max
-2.64787  -0.00691  -0.00065  -0.00006   2.22142

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 114.193780  23.234444   4.915 8.89e-07 ***
AGE          -0.009424   0.039527  -0.238 0.811565
Urea          0.735105   1.131511   0.650 0.515908
Cr           -0.019115   0.016924  -1.129 0.258713
HbA1c        -7.461304   1.690059  -4.415 1.01e-05 ***
Chol         -0.848583   0.270227  -3.140 0.001688 **
TG           -2.026225   0.595871  -3.400 0.000673 ***
HDL          -0.313335   0.555762  -0.564 0.572894
LDL           0.250976   1.314607   0.191 0.848593
VLDL         -0.465244   0.612252  -0.760 0.447321
BMI         -28.237478   6.311228  -4.474 7.67e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 424.905  on 660  degrees of freedom
Residual deviance:  64.831  on 650  degrees of freedom
AIC: 86.831

Number of Fisher Scoring iterations: 10
```

Figure 10. All Variables Logistic Regression Model

The output provides estimates of the coefficients for each predictor variable, along with their standard errors, z-values, and corresponding p-values (Pr(>|z|)). These values can be used to determine the significance of each predictor variable in the model. The intercept of the model is estimated to be 114.193780, and it is significant (p-value < 0.001). The variables HbA1c, Chol, TG, BMI are also significant predictors of diabetes, as indicated by their small p-values. The deviance residuals are also provided, which can be used to assess the goodness of fit of the model. The null deviance is the deviance when only the intercept is included in the model, while the residual deviance is the deviance of the fitted model. The difference between these two values can be used to calculate the percentage of deviance explained by the model. In this case, the model explains approximately 84.73% of the deviance in the data. Finally, the AIC (Akaike Information Criterion) value is provided, which can be used to compare different models. Lower AIC values indicate better fitting models. In this case, the AIC value is 86.831, which suggests a good fit for the model.

Knowing our predictor variables, we can create the model only with these variables which are significant.

```
```{r}
# fit a binary logistic regression model to predict diabetes diagnosis with the significant var
new_model <- glm(formula = CLASS ~ HbA1c + BMI  + Chol + TG, family = binomial, data = train)
summary(new_model)
```

```
Call:
glm(formula = CLASS ~ HbA1c + BMI + Chol + TG, family = binomial,
    data = train)

Deviance Residuals:
      Min        1Q    Median        3Q       Max
-2.62849  -0.00578  -0.00066  -0.00006   1.78253

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 111.8587    21.7416   5.145 2.68e-07 ***
HbA1c        -8.0395     1.6882  -4.762 1.91e-06 ***
BMI         -27.5140     5.6620  -4.859 1.18e-06 ***
Chol         -0.7469     0.2346  -3.184 0.001453 **
TG           -1.9918     0.5313  -3.749 0.000177 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 424.905  on 660  degrees of freedom
Residual deviance:  67.199  on 656  degrees of freedom
AIC: 77.199

Number of Fisher Scoring iterations: 10
```

Figure 11.Figure 11. New Logistic Regression Model

In this model, all the independent variables are statistically significant at the 0.05 level, except for the age variable, which was not included in this model. The model's deviance residuals indicate that the model fits the data well. A lower residual deviance suggests a better fit of the model. The AIC value (Akaike Information Criterion) is a measure of the model's goodness of fit, with lower values indicating a better fit. Overall, this logistic regression model suggests that HbA1c, BMI, Chol, and TG are significant predictors of the presence of diabetes in this dataset.

Generating Predictive Values

O test data we generate predicted values for the test data using the fitted new_model and create confusion matrix.

```
pred_class
   N   Y
0 240   4
1   5  33
```

Figure 12. Confusion Matrix

From the confusion matrix, we can see that there are a total of 282 test cases. Out of these, 240 cases were predicted as "N" (no diabetes) and 33 cases were predicted as "Y" (diabetes). The model correctly predicted 240 "N" cases and 5 "Y" cases. However, it misclassified 4 actual "N" cases as "Y" and 33 actual "Y" cases as "N".

The next step is to calculate the accuracy, sensitivity, and specificity of the model on the test data using the confusion matrix. The a**ccuracy** is calculated as the **proportion** of correctly predicted observations, the **precision** is calculated as the proportion of true positives (predicted diabetic cases among actual diabetic cases), and the Recall is calculated as the proportion of true negatives (predicted non-diabetic cases among actual non-diabetic cases).

```
Confusion Matrix:
 240 5 4 33

Accuracy:  0.968
Precission:  0.868
Recall:  0.984
Score F1: 0.861
```

Figure 13. Classification

Based on these metrics, the model seems to be performing well with a high accuracy, precision, recall, and F1 score.

Testing the Model

Testing the final model on the CLASS=='P' cases by generating predicted probabilities for those cases and calculating the average probability of being diagnosed as diabetic. The predict() function is used again to generate predicted probabilities for the CLASS=='P' cases in the original diabetes data frame. The mean() function is used to calculate the average probability of being diagnosed as diabetic among these cases.
```

The cat() function is used to print the average probability to the console, with the round() function used to round the value to three decimal places.

```
# testing the model on the CLASS=='P' cases

predictions1 <- predict(new_model, new_data1 = diabetes[diabetes$CLASS == "P", ], type = "response")
prob_diabetic_p <- mean(predictions1 > 0.5)

cat("Probability of 'P' cases being diagnosed as diabetic: ", round(prob_diabetic_p, 3))
```

```
Probability of 'P' cases being diagnosed as diabetic:  0.103
```

Figure 14. Model Testing

Based on the logistic regression model, the probability of cases with CLASS == "P" being diagnosed as diabetic is 0.103, or approximately 10.3%.

Checking the Assumptions

Finally, checking whether there is multicollinearity among the predictor variables, whether the relationship between the predictor variables and the outcome variable is linear, and whether there are outliers or influential observations that may affect the model we have already checked.

Checking Multicollinearity

```
   HbA1c      BMI     Chol       TG
2.024993 2.658775 1.272345 1.586585
```

Figure 15. Multicollinearity

Based on the variance inflation factor (VIF) values, it appears that there is no significant multicollinearity among the independent variables. The maximum VIF value is 2.66, which is below the threshold of 5 indicating high multicollinearity. Therefore, we can conclude that the independent variables are not strongly correlated with each other and can be used in the logistic regression model without any issues of multicollinearity.

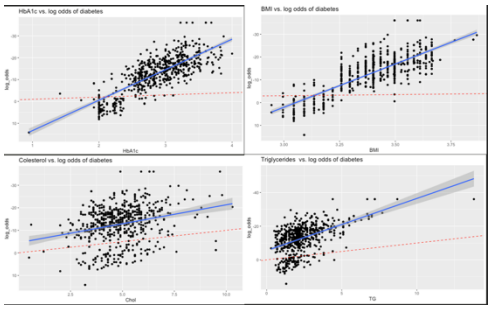

Figure 16. Linearity Assumptions

Checking relationship between the predictor variables and the outcome variable linearity.

Looking at the graphs we can see that all four predictors appear to have a roughly linear relationship with the odds of diabetes. This suggests that the linearity assumption for logistic regression is likely to hold for this model.

Conclusion

The results of the logistic regression model showed that HbA1c, Cr, TG and BMI were significant predictors of diabetes. Furthermore, the model achieved an accuracy rate of 96.8% on the test dataset, indicating good model performance. The probability of prediabetic patients being diagnosed with diabetes was calculated to be 10.3%. The findings suggest that age is a significant factor in the prevalence of diabetes, and early testing for diabetes may not be a priority for individuals in their 30s and 40s. Additionally, monitoring BMI is important as a risk factor for diabetes, and there is a need for early intervention in patients with elevated BMI levels.

Table of Figures

Bibliography

[i] "Binary Logistic Regression - Statistics Solutions." *Statistics Solutions*, 10 Sept. 2012, www.statisticssolutions.com/binary-logistic-regression/. Accessed 6 May 2023.

[ii] Sallsten, Gerd, and Lars Barregard. "Variability of Creatinine in Healthy Individuals." *International Journal of Environmental Research and Public Health*, vol. 18, no. 6, 7 May 2023, p. 3166, https://doi.org/10.3390/ijerph18063166.