Logistical Regression on Potential Diabetics

Sharjeel Ahmad[1], Daniel Shafi Batla[2], and Talha Ahmed[3]

[1]24100083
[2]24100003
[3]24100033

November 10, 2023

Instructor : Shaheena Bashir

1 Research Question

How are the **odds** of contracting diabetes related with several factors including, number of pregnancies/year, glucose concentration, BMI (Body Mass Index), blood pressure, insulin concentration, age etc?

2 Abstract

On 14 November 2019, the IDF (International Diabetes Foundation) released new figures that would that would highlight the alarming growth in the prevalence of diabetes globally (Press Release). Compared to 2017, an astonishing increase of *38 million* diabetics globally and Pakistan having reached the top 10 solely in terms of raw number of diabetics. Therefore, it is not difficult to deduce that the causes of diabetes are the crux to mitigating this world-wide phenomenon.

This report aims to build a model accurate enough to statistically infer whether a random person is diabetic (response-binary variable) based on certain diagnostic measurements (details in Introduction). This model will hopefully shed some light on some of the key factors and hence can help in strategizing counter-measures against it. R-Software is used for analyzing the explanatory variables, and a Logistic Regression technique with appropriate plots for interpretation later on.

3 Introduction

Our data comes from *Kaggle* database called, "**Pima Indians Diabetes Database**. It contains 768 observations and 9 columns (1 of them being *"Outcome"* i.e. response variable and the rest predictor variables). Below is the tabular representation of the data.

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31.0	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0.0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38.0	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1

Figure 1: First 14 rows of the dataset

3.1 Extraction

The first step in any data analysis is to import the data in the relevant software. Here, the raw data was presented to us in the *.csv* format which we decided to keep when extracting it onto **R**. Below shows the straightforward approach to importing the file "diabetes". The file was saved in the folder called CS. DTL

```
## Importing Data
my_data<-data.frame<-Diabetes_Dataset <- read.csv("diabetes.csv")
```

The data was converted into a data frame for manipulation and then renamed *my_data* for simplicity purposes

3.2 Variable Change

As you may have seen from the above table, the response variable is a collection of 0's and 1's i.e. binary. However, *R* has taken it as an integer similar to that of Glucose, Blood Pressure etc. Therefore, to stay in the true spirit of Logistic Regression, we will modify the data type and change it into a factor. Below show's the one-line code for doing exactly that and also the *R*-ouptut showing each independent, continuous predictor/response variable along with its type.

```
attach(my_data)
#Class of Outcome is integer, so convert to factor.
class(Outcome)

## [1] "integer"
```

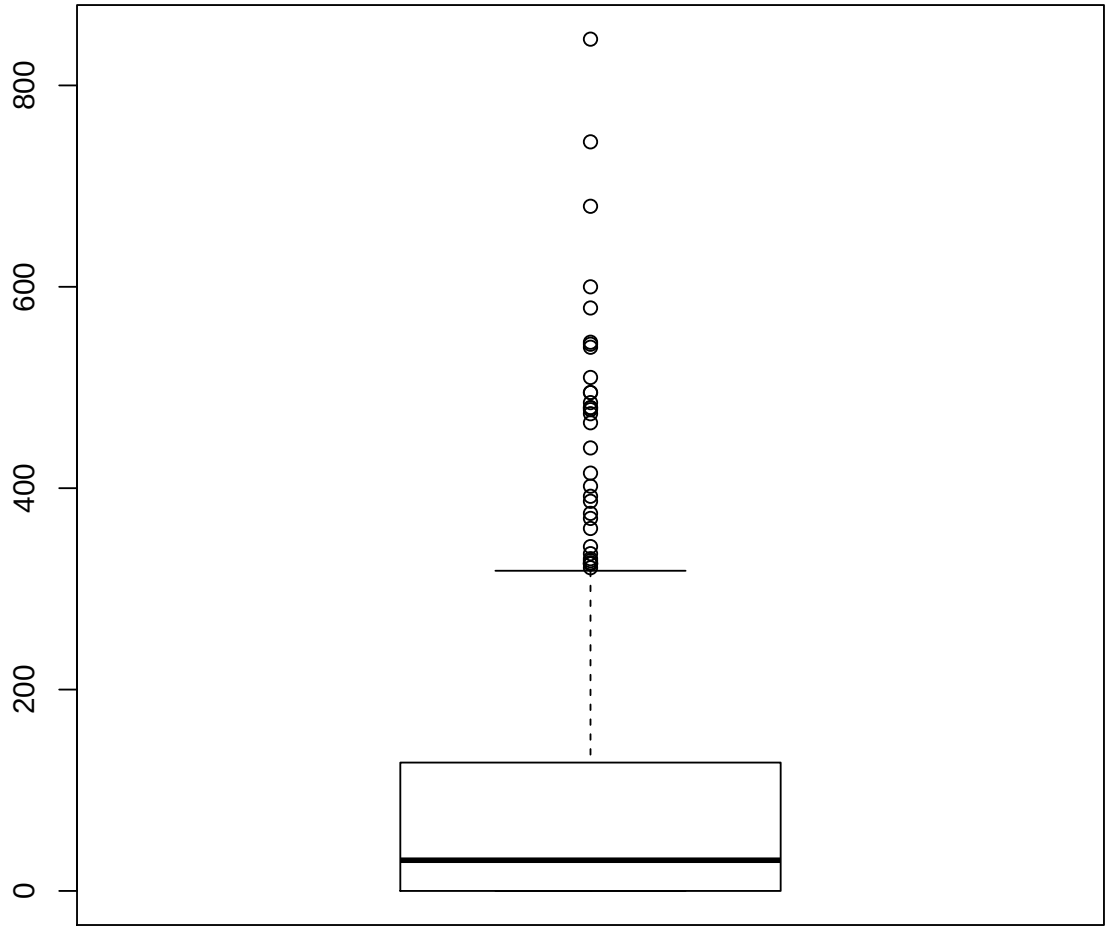
```r
Outcome <- as.factor(Outcome)
class(Outcome)

## [1] "factor"

str(Outcome)

##  Factor w/ 2 levels "0","1": 2 1 2 1 2 1 2 1 2 2 ...

## Each Predictor Variable and its type:
class(Pregnancies)

## [1] "integer"

class(Glucose)

## [1] "integer"

class(BloodPressure)

## [1] "integer"

class(SkinThickness)

## [1] "integer"

class(Insulin)

## [1] "integer"

class(BMI)

## [1] "numeric"

class(DiabetesPedigreeFunction)

## [1] "numeric"

class(Age)

## [1] "integer"
```

3.3 Data Cleaning

For statistical analysis, we took box plots, summaries, of every column variable in the data set. We noticed that a lot of predictor variables were skewed towards zero(One shown as an

example) This skewness does not make physical sense for some predictors. The abundance of these zero values prevented us from obtaining accurate results. To overcome this we replaced the zero values for BMI, skin thickness, insulin, and blood pressure with the mean of the non-zero values. Since we are modeling our data onto a logistical regression, there was no need to normalize the predictor variables.

```
boxplot(Insulin)
```




```
summary(Insulin)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     0.0    30.5    79.8   127.2   846.0
```

Below shows the *R* code for replacing zeros with the mean of the non-zero values of "BloodPressure".

```
## data_temp is subset of my_data containing the columns BloodPressure and DPF
data_temp<-my_data[, c("BloodPressure", "DiabetesPedigreeFunction")]
#Replacing all zeros in BloodPressure column with NA
data_temp["BloodPressure"][data_temp["BloodPressure"] == 0] <- NA
##Using na.omit to remove the NA
data_temp<-na.omit((data_temp))
##Calculating the mean of BloodPressure with only non-zero values
x_1<-mean(data_temp$BloodPressure)
x_1

## [1] 72.40518

##Replacing all zeros in BloodPressure with the mean i.e. x_1
my_data["BloodPressure"][my_data["BloodPressure"] == 0] <- x_1
## Similar process for other variables:
##Insulin:

data_temp<-my_data[, c("Insulin", "DiabetesPedigreeFunction")]
data_temp_2<-data_temp
data_temp["Insulin"][data_temp["Insulin"] == 0] <- NA
data_temp<-na.omit((data_temp))
x_2<-mean(data_temp$Insulin)
x_2

## [1] 155.5482

my_data["Insulin"][my_data["Insulin"] == 0] <- x_2
##SkinThickness:

data_temp<-my_data[, c("SkinThickness", "DiabetesPedigreeFunction")]
data_temp_2<-data_temp
data_temp["SkinThickness"][data_temp["SkinThickness"] == 0] <- NA
data_temp<-na.omit((data_temp))
x_3<-mean(data_temp$SkinThickness)
x_3

## [1] 29.15342

my_data["SkinThickness"][my_data["SkinThickness"] == 0] <- x_3
##BMI:

data_temp<-my_data[, c("BMI", "DiabetesPedigreeFunction")]
data_temp_2<-data_temp
```

```r
data_temp["BMI"][data_temp["BMI"] == 0] <- NA
data_temp<-na.omit((data_temp))
x_4<-mean(data_temp$BMI)
x_4

## [1] 32.45746

my_data["BMI"][my_data["BMI"] == 0] <- x_4
##Glucose:

data_temp<-my_data[, c("Glucose", "DiabetesPedigreeFunction")]
data_temp_2<-data_temp
data_temp["Glucose"][data_temp["Glucose"] == 0] <- NA
data_temp<-na.omit((data_temp))
x_5<-mean(data_temp$Glucose)
x_5

## [1] 121.6868

my_data["Glucose"][my_data["Glucose"] == 0] <- x_5
```

4 Method

Since the outcome variable is measured in a binary format i.e., the presence or absence of diabetes in the subjects, we use a logistical Regression. A logistical regression model assumes that aside from the dichotomous nature of the outcome, the predictor variables have a low degree of multicolinearity. The latter assumption will be confirmed in the analysis section. The general equation for the regression is:

$$\log(\frac{p}{1-p}) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon; \epsilon \sim Bin(1, p) \tag{1}$$

The following terms in the equation are the key to interpreting the logistic regression.

$$\frac{p}{1-p} \tag{2}$$

$$\log(\frac{p}{1-p}) \tag{3}$$

(2) is interpreted as the odds ratio and and acts as a function to the probability, while on the other hand (3) is the *log* odds. The odds signify the likelihood of a particular event occurring, relative to another. In the equation shown β_o refers to the intercept term, and all other $\beta's$, all the way till β_n refer to the coefficients of the independent variables being used in the regression. A positive β coefficient indicates a positive rate of change in the *log*

odds of the response variable. The magnitude of the beta coefficients determines the extent of contribution from the independent variable which are listed as X_1, X_2 and so on. The beta coefficient of a predictor variable refers to the marginal change in the *log* odds of the outcome when the variable in question is increased by one unit, keeping all other variables fixed. More than often, due to complexity of imagining "*log* odds", we are interested in the following quantity:

$$e^{\beta_n}$$

This quantity represents the marginal change in the odds ratio of the response variable when the predictor variable X_n changes by 1 unit.
For our statistical analysis we theorize the logistical model to be:

$$\log(\frac{p}{1-p}) = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 \epsilon; \epsilon \sim Bin(1, p) \quad (4)$$

$$p = \text{probability of Outcome} = 1 \text{ i.e. person has diabetes}$$

$$X_1 = Pregnancies, X_2 = Glucose, X_3 = BloodPressure, X_4 = SkinThickness$$

$$X_5 = Insulin, X6 = BMI, X_7 = DiabetesPedigreeFunction(DPF), X_8 = Age$$

Note: ϵ follows a binomial distribution because the response variable is dichotomous.

To determine the effectiveness of our model we have R output the summary of our logistic regression, Chi-squared deviance test, Variance Inflation Factor (VIF), We will also be analyzing the QQ plots and multicollinearity of the variables. Moreover, we will be looking at the predicted probabilities, confidence interval of the odds ratio and Pearson's correlation for analysis of our regression.

5 Analysis

Before we move on to the logistic regression analysis, its important that we prove why are we able to apply logistic regression. From Fig.1, one can trivially see that the response variable is dichotomous so one assumption for the logit regression is fulfilled. The second, and perhaps the most important is the test of independence of each variable i.e. there exist no serious concern of multi-collinearity which might otherwise, inflate the standard errors and reliabilty of each predictor variable.
For this, we thought it was best to do it with two methods. And if the interpretation of the two methods coincide, we can be pretty sure about the result. The two methods in question are already mentioned in "**Method**" section:

- Pearson's Correlation
- VIF (discussed later once logistic regression is done)

Finding the pearson's correlation is pretty straightforward. Below shows the *R*-code + output.

```r
cor(my_data[,-9])
```

```
##                           Pregnancies   Glucose BloodPressure SkinThickness
## Pregnancies                1.00000000 0.1279115   0.208522309    0.08298907
## Glucose                    0.12791147 1.0000000   0.218366918    0.19299109
## BloodPressure              0.20852231 0.2183669   1.000000000    0.19281584
## SkinThickness              0.08298907 0.1929911   0.192815844    1.00000000
## Insulin                    0.05602701 0.4201571   0.072516882    0.15813897
## BMI                        0.02156505 0.2309412   0.281267706    0.54239773
## DiabetesPedigreeFunction -0.03352267 0.1370597  -0.002763364    0.10096644
## Age                        0.54434123 0.2665335   0.324594939    0.12787247
##                              Insulin        BMI DiabetesPedigreeFunction
## Pregnancies                0.05602701 0.02156505             -0.033522673
## Glucose                    0.42015709 0.23094124              0.137059710
## BloodPressure              0.07251688 0.28126771             -0.002763364
## SkinThickness              0.15813897 0.54239773              0.100966445
## Insulin                    1.00000000 0.16658610              0.098633942
## BMI                        0.16658610 1.00000000              0.153399971
## DiabetesPedigreeFunction   0.09863394 0.15339997              1.000000000
## Age                        0.13673386 0.02551918              0.033561312
##                                   Age
## Pregnancies                0.54434123
## Glucose                    0.26653352
## BloodPressure              0.32459494
## SkinThickness              0.12787247
## Insulin                    0.13673386
## BMI                        0.02551918
## DiabetesPedigreeFunction   0.03356131
## Age                        1.00000000
```

From the above *R*-output, one can see every value is less than 0.8, hence there is no serious concern of multi-collinearity.

5.1 Logistic Regression Summary

Before moving on to the analysis, it's imperative we see the result of the regression *R* outputs for us. Below shows the logistic regression output:

```r
logistical<- glm(Outcome~.,data=my_data,family="binomial")
summary(logistical)
```

```
##
## Call:
## glm(formula = Outcome ~ ., family = "binomial", data = my_data)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6703  -0.7181  -0.3946   0.7134   2.3769
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -9.0967954  0.8125684 -11.195  < 2e-16 ***
## Pregnancies               0.1250180  0.0323845   3.860 0.000113 ***
## Glucose                   0.0373503  0.0038787   9.630  < 2e-16 ***
## BloodPressure            -0.0088036  0.0085614  -1.028 0.303810
## SkinThickness             0.0034830  0.0131395   0.265 0.790952
## Insulin                  -0.0007875  0.0011736  -0.671 0.502194
## BMI                       0.0931107  0.0178394   5.219  1.8e-07 ***
## DiabetesPedigreeFunction  0.8660799  0.2963413   2.923 0.003471 **
## Age                       0.0131406  0.0095104   1.382 0.167061
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 993.48  on 767  degrees of freedom
## Residual deviance: 713.30  on 759  degrees of freedom
## AIC: 731.3
##
## Number of Fisher Scoring iterations: 5
```

With reference to above output and (4), we have the following logistic regression equation (correct to 3 dp):

$$\log(\frac{p}{1-p}) = -9.097 + 0.125X_1 + 0.037X_2 - 0.009X_3 + 0.003X_4 + -0.008X_5 + 0.093X_6 + 0.866X_7 + 0.013X_8 \quad (5)$$

Running the logistical regression , the summary tells us that the following variables that accurately show consistency with our model and show that our regression is a good predictor of the odds since they have a low p score. From the table we can see that Pregnancies, Diabetes Pedigree function (based on the family history of having diabetes), Glucose level, and BMI have a low p score which means that they are good predictors and consistently dictate the odds of having diabetes. On the other hand, Age, Skin Thickness, Insulin, and, Blood pressure are not very good predictors since they have a higher p score, although all of them have a positive trend. From this result, one can estimate to some extent, that diabetes is more of a genetic problem than a physical problem. However, on the other hand, for Insulin level and Blood pressure to have low relevance, we must reason that the inconsistencies in data points and missing values of these measurements, which forced us to replace them with their means (Recall **3.3**) greatly hamper the regression model for these explanatory variables. Maybe, for insulin one can speculate that subjects taking insulin dose will have a normal amount of insulin in their blood but will still have a

positive diagnosis of diabetes and hence the low relevance. With respect to age, the reason why it has a high p-score might be due to the lack of distinction between type 1 and type 2 diabetes in the data collection. Type 1 diabetes is typically observed after the age of 20 while type 2 is usually a risk for people above the age of 45. This may explain the relative irrelevancy of age in this report.
All the regression coefficients represent the change in the log odds when any one of the parameters is increased by one unit whilst keeping every other component constant (Recall **4**). The values are small as these represent the change in the log of odds rather than the odds themselves, which are calculated by taking the exponential of these coefficients. The negative values mean that the increase in one unit of these variables will decrease the odds of having diabetes. As a concise finish to the logistic regression summary, below is the R output showing the odds of diabetes with respect to one specific predictor variable.

```
##Log Odds of each variable
coef(logistical)
```

```
##               (Intercept)               Pregnancies                   Glucose
##             -9.0967953822             0.1250179779             0.0373502939
##             BloodPressure           SkinThickness                   Insulin
##             -0.0088036314             0.0034829585            -0.0007875195
##                       BMI DiabetesPedigreeFunction                       Age
##              0.0931106991             0.8660798697             0.0131406001
```

```
##exponent of each value gives the Odd Ratio (OR)
exp(coef(logistical))
```

```
##               (Intercept)               Pregnancies                   Glucose
##              0.0001120242             1.1331688248             1.0380565821
##             BloodPressure           SkinThickness                   Insulin
##              0.9912350070             1.0034890310             0.9992127905
##                       BMI DiabetesPedigreeFunction                       Age
##              1.0975832300             2.3775721684             1.0132273172
```

```
##Doing the following will give us the increase/decrease of Odds in terms of percentag
(exp(coef(logistical))-1) * 100
```

```
##               (Intercept)               Pregnancies                   Glucose
##             -99.98879758              13.31688248               3.80565821
##             BloodPressure           SkinThickness                   Insulin
##              -0.87649930               0.34890310              -0.07872095
##                       BMI DiabetesPedigreeFunction                       Age
##               9.75832300             137.75721684               1.32273172
```

Note: Ignoring the coefficient of (Intercept), some of the odds are the following:

- If pregnancies increase by 1/year, then the odds of having diabetes increase by $\approx 13\%$

- If blood pressure increases by 1 mm Hg, then the odds of having diabetes decreases by $\approx 0.011\%$
- If the DPF scores of likelihood increases by 1, the odds of diabetes increases by $\approx 38\%$
- and so on...

5.2 Variance Inflation Factor(VIF)

Now that we have run the regression summary, we are fully equipped to check the VIF of each predictor variable. If the interpretation of this test matches with that of the Pearson's Correlation, we can say with full confidence that the predictor variables are independent of each other with no serious concern of multi-collinearity. Refer to the *R*-output below.

```
> vif(logistical)
            Pregnancies                 Glucose           BloodPressure
               1.417677                1.231559                1.217035
          SkinThickness                 Insulin                     BMI
               1.362541                1.190011                1.507551
DiabetesPedigreeFunction                     Age
               1.012094                1.534286
```

Figure 2: VIF of each predictor variable

Looking at the output above, one can trivially see that every value is less than 2.5 hence there is no significant case of multi-collinearity and therefore our second assumption for applying logistic regression is approved.

5.3 Predicted Probabilities and Plots

Before, we move on to plotting the predicted probabilities for our regression model, let's mathematically examine what is actually the predicted probability (denoted by p). Let's refer to our regression equation in (5). Notice, if we solve for p, we get the following result:

$$p = \frac{exp(-9.907 + 0.125X_1 + ... + 0.013X_8)}{[1 + exp(-9.907 + 0.125X_1 + ... + 0.013X_8)]} \tag{6}$$

There's an important point to note here. p, from looking (6) can also be interpreted as:

$$p = Pr(Outcome = 1|X_1 = x_1, X_2 = x_2, ..., X_8 = x_8) = \frac{exp(-9.907 + 0.125X_1 + ... + 0.013X_8)}{[1 + exp(-9.907 + 0.125X_1 + ... + 0.013X_8)]} \tag{7}$$

i.e. the conditional probability that the person has diabetes given a certain level of values for each predictor variable. In our case, these are the values in one single row of our dataset. Now that we have approached this mathematically, let's look at it graphically. Below shows the R-code of only one of the plots (the rest will be similar). We wil analyze only few of the plots and the interpretation of the rest of the plots will be of similar nature.
So looking at the first plot (Fig .3), we can see clearly that as the number of pregnancies/year increases the probability of Outcome = 1 i.e. the probability of getting

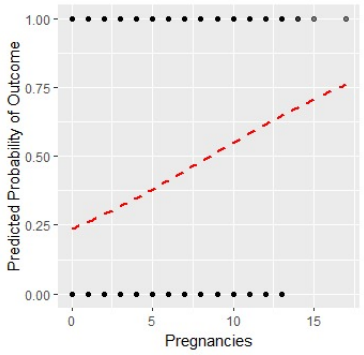


Figure 3: Predicted Probabilities of Outcome = 1 vs Pregnancies

diabetes increases as well. Looking at this another way, the outcome is more likely to be 0 if the person has less number of pregnancies/year. Similar interpretation can be done for the rest (plots in the appendix at the end).
We can also look at this another way by using the "*predict*" function: Refer to the code below:

```
## Using predict function to display the least 6 p values.
head(sort(predict(logistical, type = "response")))

##        590        618        681         98         91        462
## 0.01166695 0.01185643 0.01266520 0.01377073 0.01461340 0.01544460

## Printing the rows specified above.
my_data[590,]

##     Pregnancies Glucose BloodPressure SkinThickness  Insulin  BMI
## 590           0      73      72.40518      29.15342 155.5482 21.1
##     DiabetesPedigreeFunction Age Outcome
## 590                    0.342  25       0

my_data[618,]

##     Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 618           2      68            62            13      15 20.1
##     DiabetesPedigreeFunction Age Outcome
## 618                    0.257  23       0

my_data[681,]

##     Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 681           2      56            56            28      45 24.2
##     DiabetesPedigreeFunction Age Outcome
## 681                    0.332  22       0
```

```
my_data[98,]

##    Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI
## 98           1      71            48            18      76 20.4
##    DiabetesPedigreeFunction Age Outcome
## 98                    0.323  22       0

my_data[91,]

##    Pregnancies Glucose BloodPressure SkinThickness  Insulin  BMI
## 91           1      80            55      29.15342 155.5482 19.1
##    DiabetesPedigreeFunction Age Outcome
## 91                    0.258  21       0

my_data[462,]

##     Pregnancies Glucose BloodPressure SkinThickness  Insulin  BMI
## 462           1      71            62      29.15342 155.5482 21.8
##     DiabetesPedigreeFunction Age Outcome
## 462                    0.416  26       0
```

Looking at the rows printed, we can see that the max value of pregnancies/year is 2 (which is quite less) and the Outcome is 0 in all such cases. This is a direct visualization of the plot above and the predict function.

5.4 Deviance Chisq Test

Before carrying out any regression analysis, it's key to know which values are essential to the data and which one's, if not included, wont affect the overall result that much. We have carried out this process using the deviance "Chisq" test. Below shows the relevant *R*-code and output.

```
anova(logistical, test = "Chisq")

## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Outcome
##
## Terms added sequentially (first to last)
##
##
##                          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                       767     993.48
```

```
## Pregnancies                1   37.274       766     956.21 1.026e-09 ***
## Glucose                    1  186.545       765     769.67 < 2.2e-16 ***
## BloodPressure              1    0.977       764     768.69 0.3228385
## SkinThickness              1   13.911       763     754.78 0.0001917 ***
## Insulin                    1    0.001       762     754.78 0.9799580
## BMI                        1   30.522       761     724.25 3.302e-08 ***
## DiabetesPedigreeFunction   1    9.057       760     715.20 0.0026171 **
## Age                        1    1.898       759     713.30 0.1683534
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the similar threshold as we did for the p-values in the regression summary, we can deduce the following: Pregnancies, Glucose, SkinThickness, BMI and DBF are all relevant and important to include in the dataset. While the rest i.e. Age, BloodPressure and Insulin are not relevant. However, like we said in the regression summary, any result for Insulin should not be regarded as accurate due to its original skewed distribution towards the value 0 (Refer to the boxplot in the appendix at the end). One last thing to note is that this test should not be confused with what we concluded from our regression summary. The two have different purposes. The former is used for checking relevance of the predictor variable to the dataset as a whole while the latter checks the effect of the predictor variable on the response variable.

5.5 Confidence Interval of Odd Ratios (OR)

Another key aspect of our statistical analysis is the confidence interval of the Odd Ratios. It gives an expected range for the true odds ratio of the predictor variable to fall within. Depending on the width of the confidence interval, the reliability of the odds of outcome can increase/decrease as well. Recalling the odd ratio's calculated in **5.1**, the following output shows the confidence interval of those OR's.

```
exp(confint(logistical))

## Waiting for profiling to be done...

##                                 2.5 %       97.5 %
## (Intercept)              2.161215e-05 0.0005246437
## Pregnancies              1.064179e+00 1.2084527436
## Glucose                  1.030408e+00 1.0462189651
## BloodPressure            9.746822e-01 1.0080147789
## SkinThickness            9.783014e-01 1.0299601851
## Insulin                  9.969466e-01 1.0015742577
## BMI                      1.060471e+00 1.1374135659
## DiabetesPedigreeFunction 1.336557e+00 4.2748122128
## Age                      9.944216e-01 1.0322876023
```

```r
## Let's have a look at the OR's again for better understanding
exp(coef(logistical))
```

```
##              (Intercept)              Pregnancies                  Glucose
##            0.0001120242            1.1331688248            1.0380565821
##            BloodPressure           SkinThickness                  Insulin
##            0.9912350070            1.0034890310            0.9992127905
##                      BMI DiabetesPedigreeFunction                      Age
##            1.0975832300            2.3775721684            1.0132273172
```

The exponential of the confidence intervals shows that the 97.5% of the observations have an increase of less than the value stated under the 97.5% column and 2.5% of the observation have an increase of the value than the value under 2.5% column of the confidence intervals. This means the 95% of the observations have an increase ranging from the value under the 2.5% column to the value under the 97.5% column. Moreover, the interval of the confidence interval can be trivially seen for e.g, the confidence interval for Pregnancies is $[1.064, 1.208]$. In addition, we can see that every OR lies well within their respective interval hence we retain our conclusions about OR's we did in **5.1**.

6 Conclusion

The Logistical Regression Model in the project predicts the likelihood of having diabetes given certain predictor variables. According to the data genetic factors like the Diabetes Pedigree Function have a significant impact to the risk of being diabetic. This is supported by the findings of various research papers including the one published by S.S. Rich where he concluded that both type 1 and type 2 diabetes has a mode of inheritance (1319). The high p-value for some predictor variables may be due to the outliers present within the data, these outliers skewed the data both towards the zero value as well as out of the range (as seen in the box-plots for Insulin). All in all in order to obtain a more accurate regression model, the data set should consist of the much more diverse genetic pool as well fully recorded data for all the predictor variables under examination to avoid averaging the NA values.

7 References

Learning, UCI Machine. "Pima Indians Diabetes Database." Kaggle, 6 Oct. 2016, https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

Press Release. "Latest Figures Show over 19 Million People Now Living with Diabetes in Pakistan as the Numbers Continue to Rise." Brandsynario, 14 Nov. 2019, https://www.brandsynario.com/latest-figures-show-over-19-million-people-now-living-with-diabetes-in-pakistan-as-the-numbers-continue-to-rise/.

Rich, Stephen S. "Mapping genes in diabetes: genetic epidemiological perspective." Diabetes 39.11 (1990): 1315-1319.

Senaviratna, N. A. M. R., and T. M. J. A. Cooray. "Diagnosing multicollinearity of logistic regression model." Asian Journal of Probability and Statistics 5.2 (2019): 1-9.

8 Appendix

8.1 Predicted Probabilities

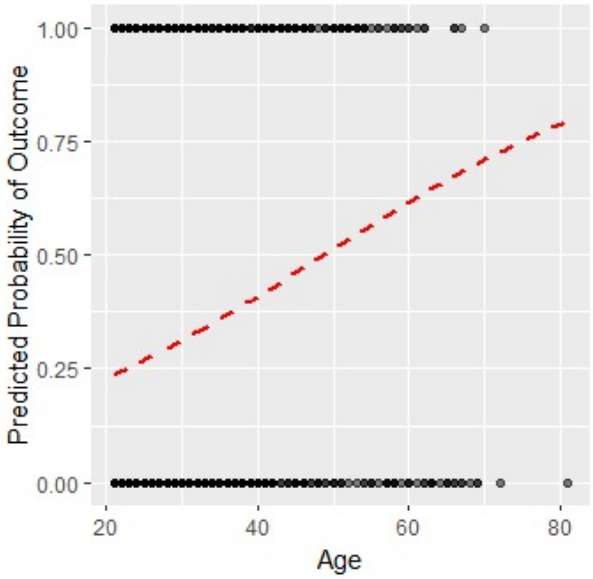


Figure 4: Predicted Probability of Age

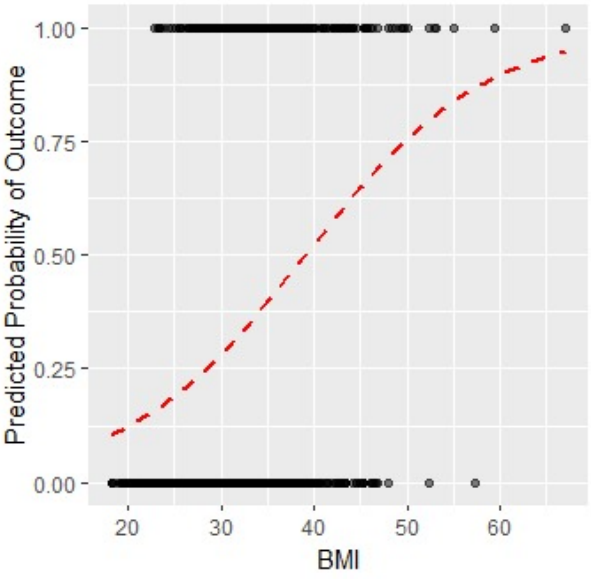


Figure 5: Predicted Probability of BMI

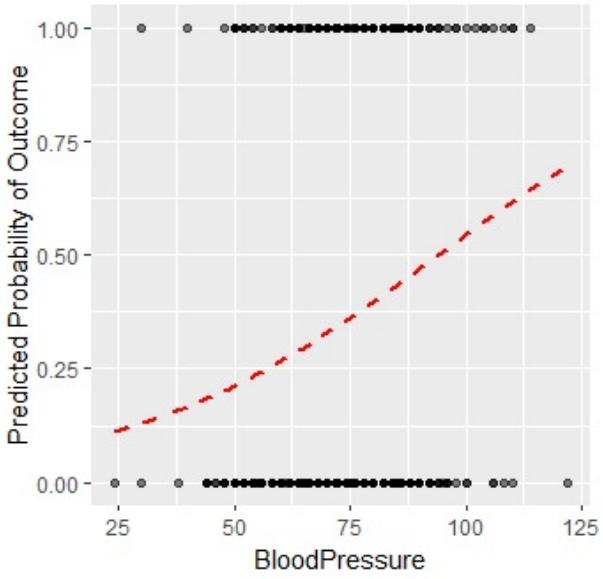


Figure 6: Predicted Probability of Blood Pressure

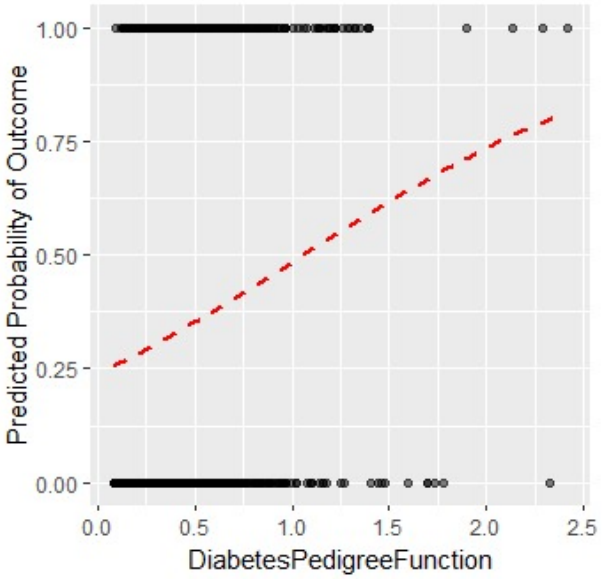


Figure 7: Predicted Probability of Pedigree Function

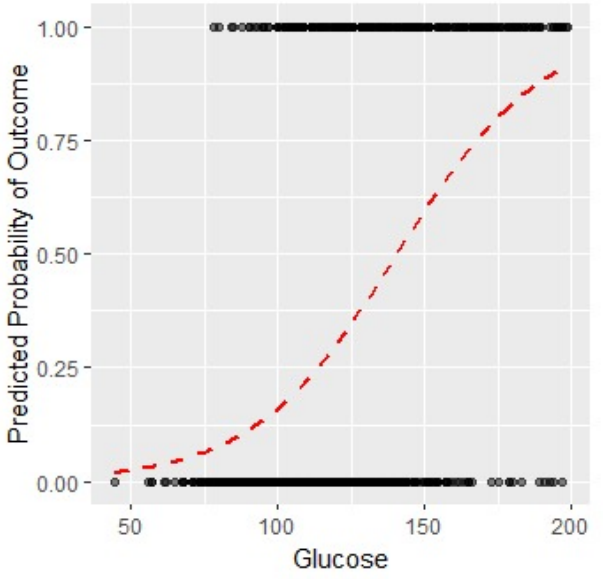


Figure 8: Predicted Probability of Glucose

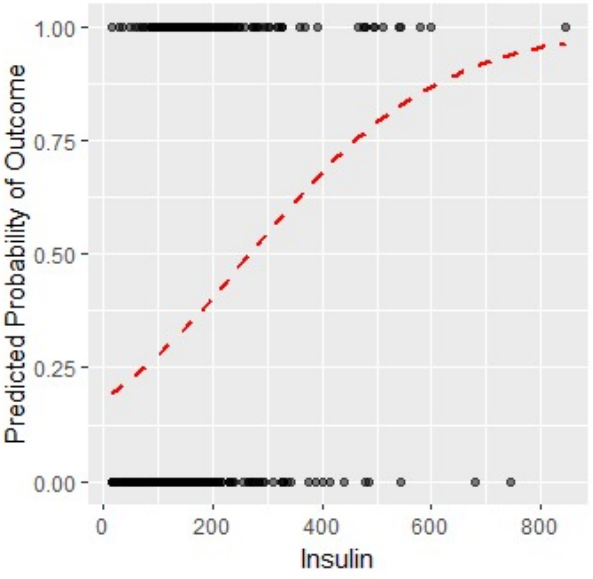


Figure 9: Predicted Probability of Insulin

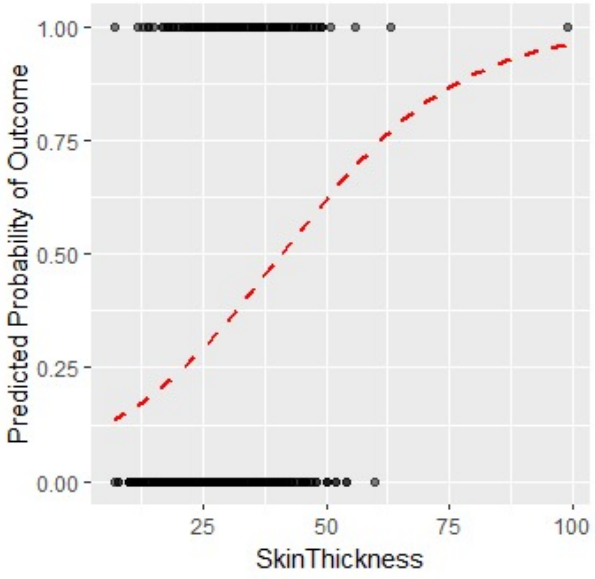


Figure 10: Predicted Probability of Skin Thickness