Statistical Learning Lab

Assignment - 2

Logistic Regression Assignment

NAME : SUNNY KUMAR

ROLL NO : 22IM10040

Show the code snippets and the corresponding output for the following:

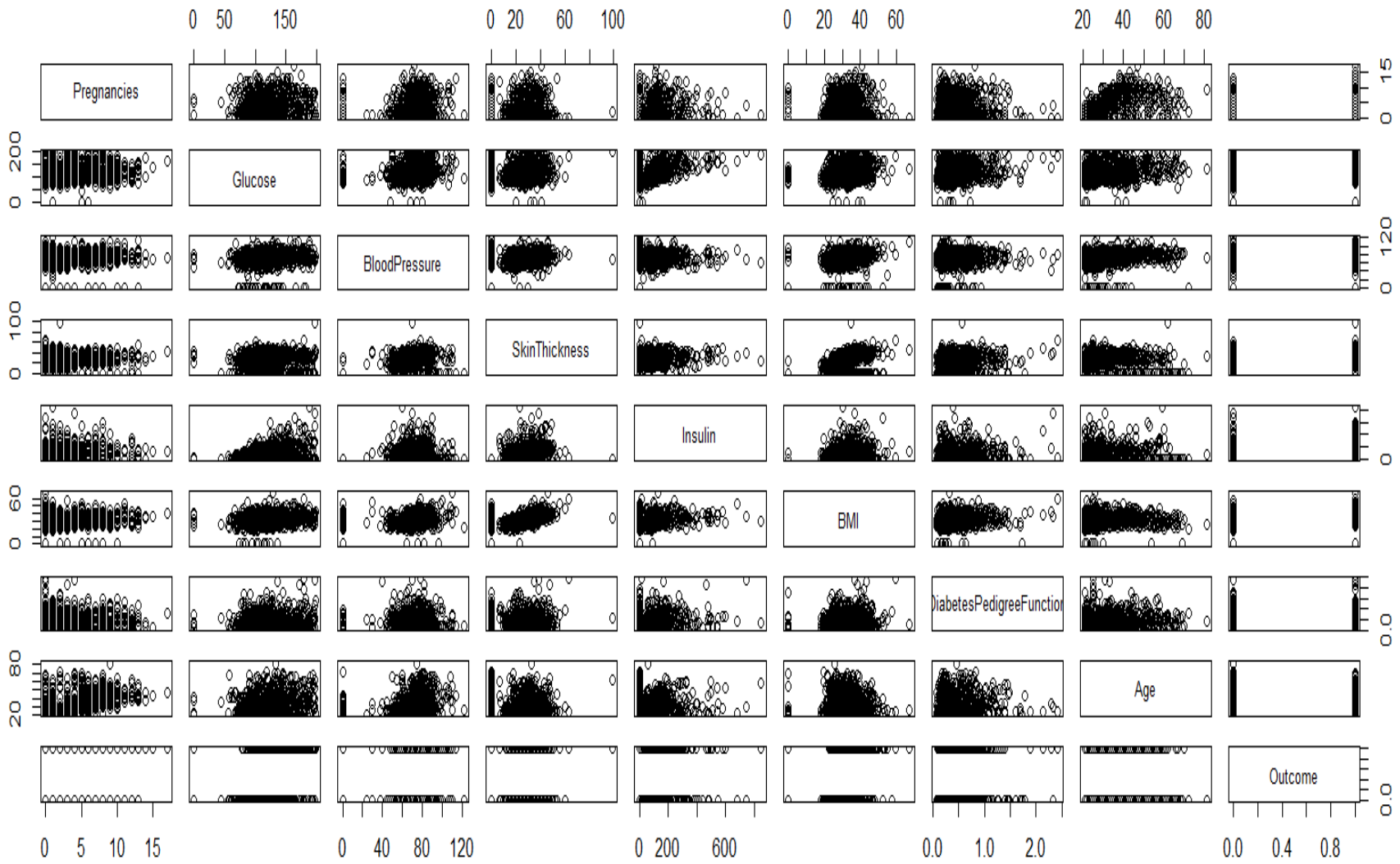
1. **Load the dataset "diabetes.csv". Display first few rows of the dataset**.

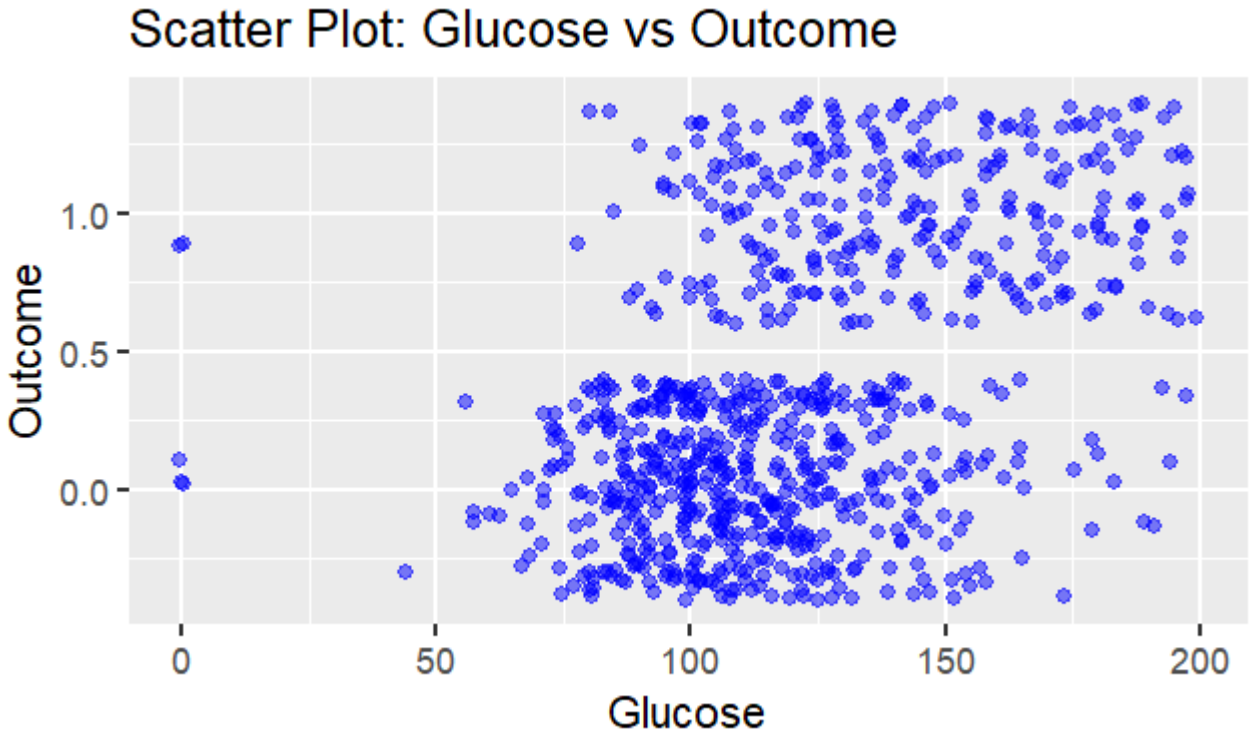
   Ans : Loaded the dataset using environment -> import dataset and imported dataset.

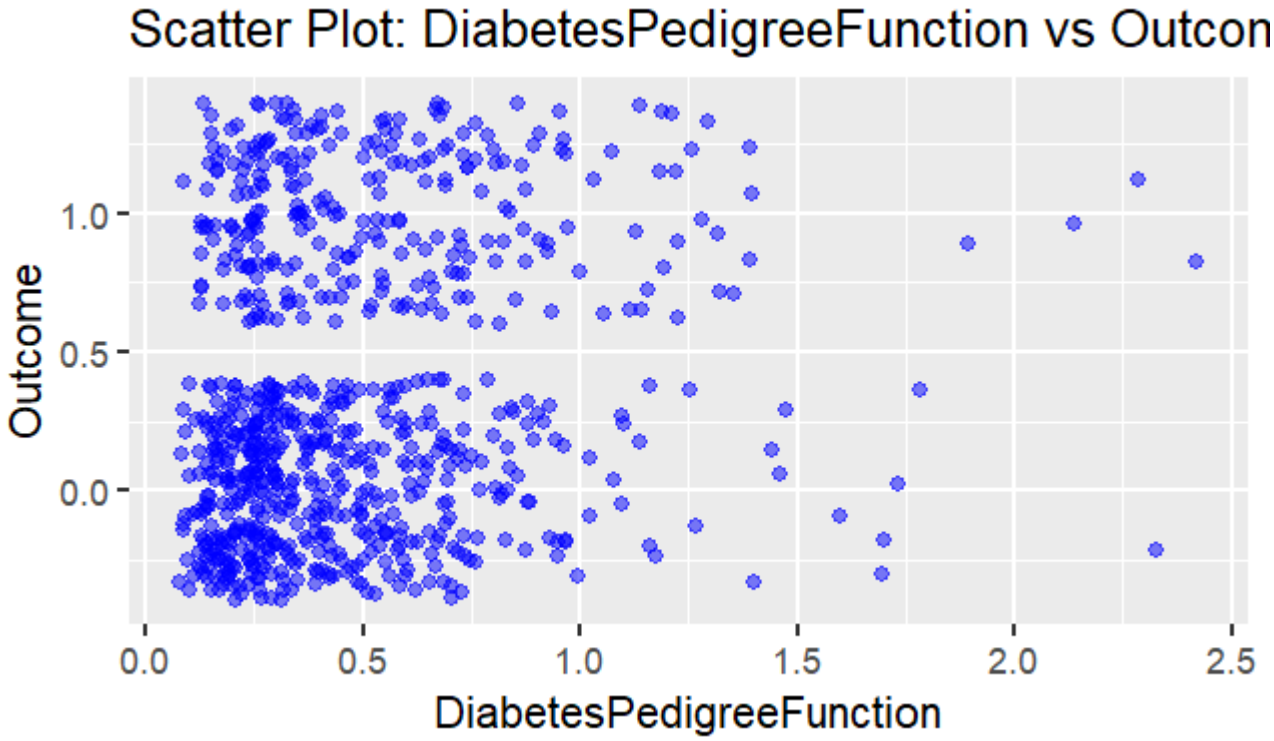
```
> View(diabetes)
> df <- diabetes
> View(df)
> head(df)
  Pregnancies Glucose BloodPressure SkinThickness Insulin  BMI DiabetesPedigreeFunction Age Outcome
1           6     148            72            35       0 33.6                    0.627  50       1
2           1      85            66            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
```

2. **Perform preliminary analysis to show how the variables are related to each other. Use scatter plot, box plot etc. to visualize how different variables impact the "Outcome" variable.**

   Ans : Scatter plot of among different variables is given below:

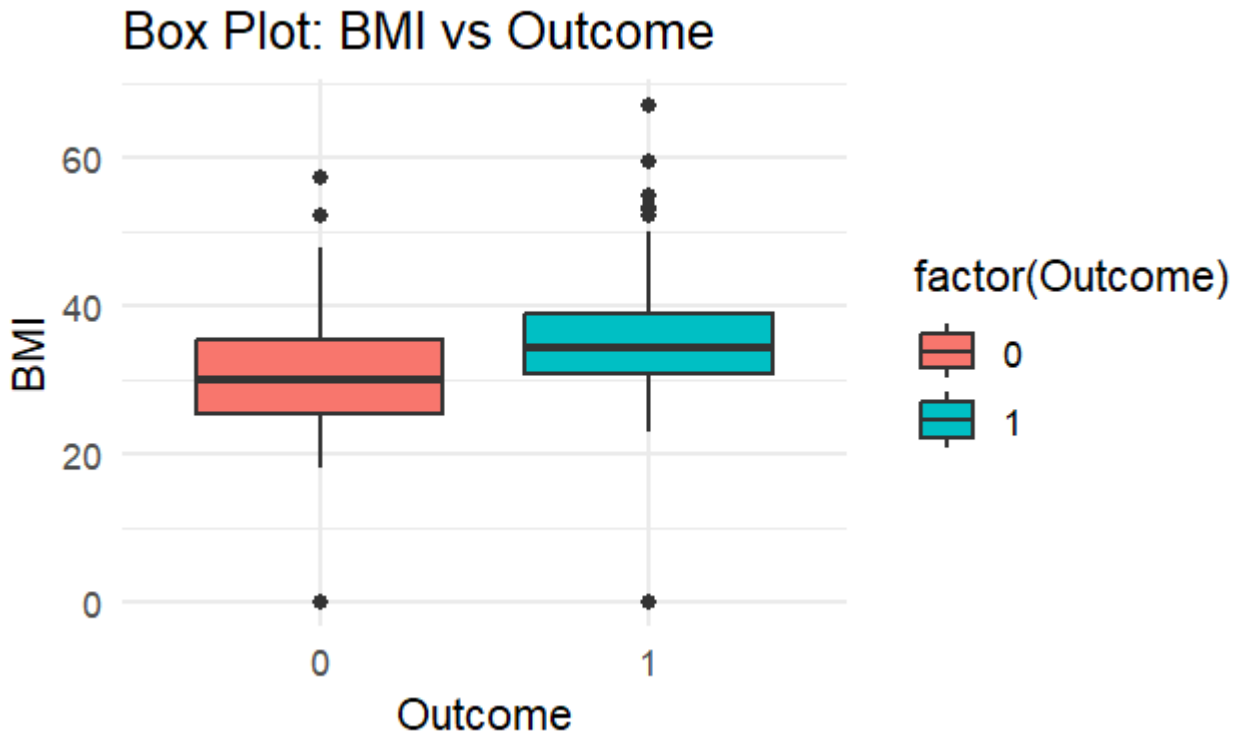

Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunctio
Age
Outcome



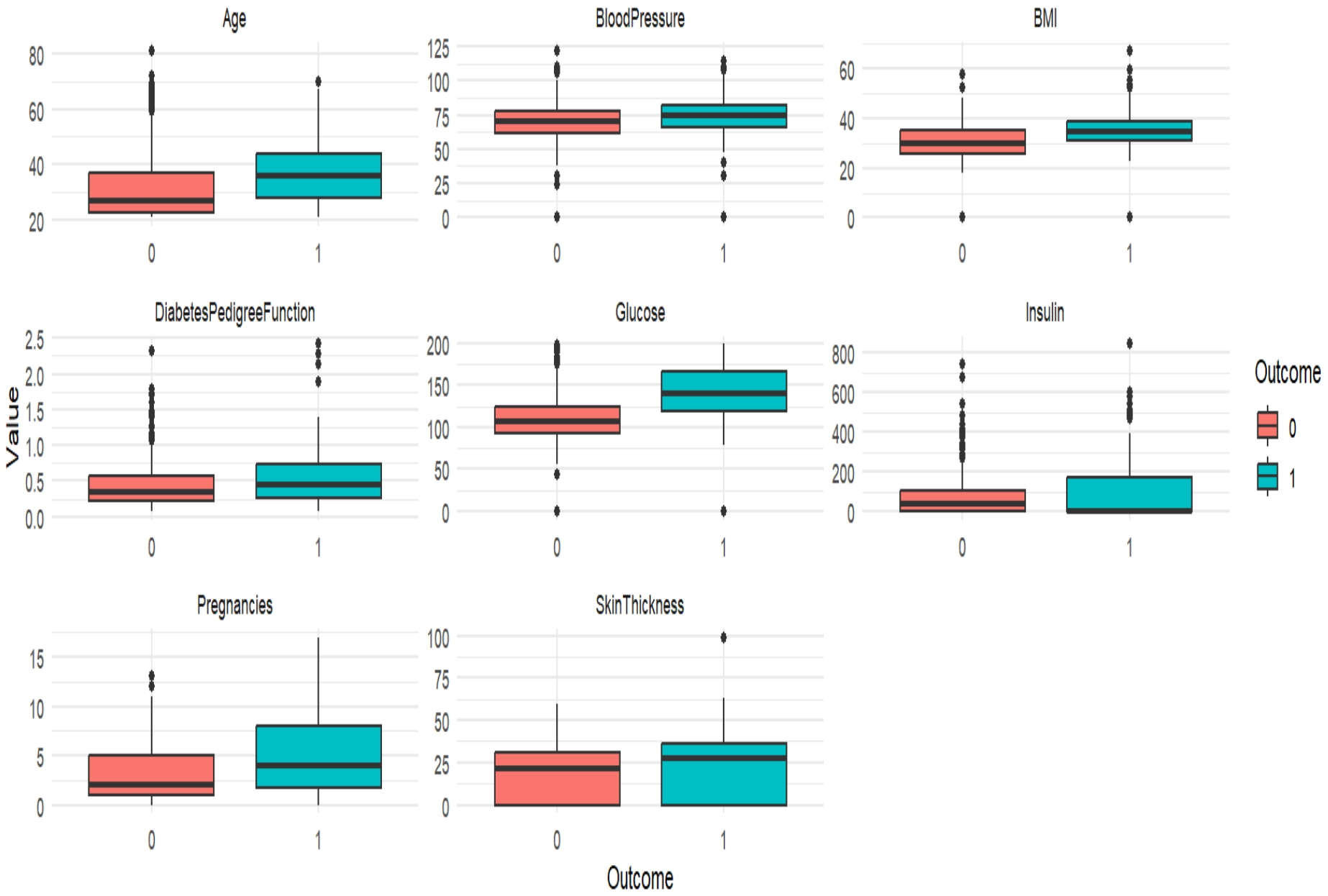


From the above two scatter plots we, cannot say anything that some particular column is impacting the outcome variable.

For lower boxplot : *ggplot(diabetes, aes(x = factor(Outcome), y = BMI, fill = factor(Outcome))) +*
geom_boxplot() +
labs(title = "Box Plot: BMI vs Outcome", x = "Outcome", y = "BMI") +
theme_minimal()



My Inference : The median BMI for individuals with Outcome = 1 (diabetic) is higher than for individuals with Outcome = 0 (non-diabetic). This suggests that diabetics tend to have a higher BMI on average. BMI appears to have a relationship with the Outcome variable. Higher BMI values are more associated with diabetes (Outcome = 1). Similarly, All the plots are shown below.


Age
BloodPressure
BMI
DiabetesPedigreeFunction
Glucose
Insulin
Pregnancies
SkinThickness
Value
Outcome
0
1

3. **Randomly sample 80% of the data as training data and rest as test data. Fit a Logistic Regression model with all the predictors on training data. From the summary which factors seem to be significant? Explain how the predictors impact the log-odds ratio of diagnosed with diabetes (Outcome)**

```
trn <- sample(dim(df)[1], 615) # 80% of 768
trn
df_train <- df[trn,]
df_test <- df[-trn,]
head(df_test)
df_test <- subset(df_test,select = -`Outcome`)
head(df_test)
dim(df_train)
dim(df_test)
#fit logistic regression model
glm.fits = glm(Outcome ~. , family = binomial , data = df_train)
summary(glm.fits)
```

Interpretations: From the summary we can say that Columns : " Pregnancies , Glucose , BMI , DiabetesPedigreeFunction, Age" are significant , because their P value is less than 0.05.

```
Call:
glm(formula = Outcome ~ ., family = binomial, data = df_train)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)              -8.792745   0.808817 -10.871  < 2e-16 ***
Pregnancies               0.090120   0.035607   2.531  0.01137 *
Glucose                   0.034376   0.004109   8.365  < 2e-16 ***
BloodPressure            -0.011485   0.005930  -1.937  0.05280 .
SkinThickness             0.004239   0.007832   0.541  0.58830
Insulin                  -0.001433   0.001022  -1.402  0.16079
BMI                       0.094775   0.016864   5.620 1.91e-08 ***
DiabetesPedigreeFunction  0.895416   0.346096   2.587  0.00968 **
Age                       0.024800   0.010622   2.335  0.01955 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 805.53  on 614  degrees of freedom
Residual deviance: 577.82  on 606  degrees of freedom
AIC: 595.82

Number of Fisher Scoring iterations: 5
```

Variable-Specific Interpretation(Log Odds Ratio):

Pregnancies (0.090120, p = 0.01137):

- A one-unit increase in the number of pregnancies increases the log-odds of diabetes diagnosis by 0.090120. This is a statistically significant predictor.

Glucose (0.034376, p < 2e-16)*:

- A one-unit increase in glucose levels significantly increases the log-odds of diabetes diagnosis by 0.034376. This is a highly significant predictor.

BloodPressure (-0.011485, p = 0.05280):

- A one-unit increase in blood pressure slightly decreases the log-odds of diabetes diagnosis by 0.011485. However, this predictor is only marginally significant (. indicates a p-value close to 0.05).

SkinThickness (0.004239, p = 0.58830):

- This predictor has a small positive coefficient but is not statistically significant ($p > 0.05$), meaning its impact on the Outcome is uncertain.

Insulin (-0.001433, p = 0.16079):

- Insulin has a negligible negative coefficient and is not statistically significant.

BMI (0.094775, p = 1.91e-08)*:

- A one-unit increase in BMI significantly increases the log-odds of diabetes diagnosis by 0.094775. This is a strong and highly significant predictor.

DiabetesPedigreeFunction (0.895416, p = 0.00968):

- A one-unit increase in this metric increases the log-odds of diabetes diagnosis by 0.895416. This variable is statistically significant.

Age (0.024800, p = 0.01955):

- A one-year increase in age increases the log-odds of diabetes diagnosis by 0.024800. Age is statistically significant.

4. From the model fitted in problem 3, derive confusion matrix, accuracy, and F1-score on test data.

```
> pred <- predict(glm.fits , df_test , type = "response")
> pred_class <- ifelse(pred>=0.5 , 1 , 0)
> #Create confusion matrix
> table(df[-trn,]$Outcome , pred_class)
   pred_class
     0  1
  0 95 13
  1 23 22
```

Ans : Accuracy = (95+22)/(95+22+13+23) = 0.76470 . So, Accuracy = 76.47 %.

```
> #Test Accuracy
> mean(pred_class == df[-trn,]$Outcome)
[1] 0.7647059
```

- **Precision**: $\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$
- **Recall** (Sensitivity): $\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$
- **F1-Score**: $F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

```
tp <- conf_matrix[2, 2]  # True Positives
fp <- conf_matrix[1, 2]  # False Positives
fn <- conf_matrix[2, 1]  # False Negatives

# Precision
precision <- tp / (tp + fp)

# Recall
recall <- tp / (tp + fn)
# F1-Score
f1_score <- 2 * (precision * recall) / (precision + recall)

cat("Precision:", precision, "\n")
cat("Recall:", recall, "\n")
cat("F1-Score:", f1_score, "\n")
```

```
> f1_score <- 2 * (precision * recall) / (precision + recall)
> cat("Precision:", precision, "\n")
Precision: 0.6285714
> cat("Recall:", recall, "\n")
Recall: 0.4888889
> cat("F1-Score:", f1_score, "\n")
F1-Score: 0.55
```

So, we get the F1-score as 0.55.

5. Let's call the model fitted in problem 3 M1. Now choose predictors "Pregnancies", "Glucose" and "BMI" and fit a model (M2). Compare the deviances among these two models and perform hypothesis test to show whether M1 is significantly more informative than M2.

```
> M2 = glm(Outcome ~ Pregnancies + Glucose + BMI  , family = binomial , data = df_train)
> summary(M2)

Call:
glm(formula = Outcome ~ Pregnancies + Glucose + BMI, family = binomial,
    data = df_train)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.224589   0.706358 -11.644  < 2e-16 ***
Pregnancies  0.126192   0.030218   4.176 2.97e-05 ***
Glucose      0.033798   0.003646   9.270  < 2e-16 ***
BMI          0.088609   0.015329   5.781 7.45e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 805.53  on 614  degrees of freedom
Residual deviance: 594.86  on 611  degrees of freedom
AIC: 602.86

Number of Fisher Scoring iterations: 5
```

```
> deviance(M1) - deviance(M2)
[1] -17.03798
> dof = 8 - 3
> qchisq(0.95,dof)
[1] 11.0705
> chisquare = deviance(M1) - deviance(M2)
```

Interpretation:

From the above picture we can say that , deviance is less than critical value(-17.03 < 11.07) . So , fail to reject the null hypothesis and hypothesis testing. So, We can say that Our simple model (M2) with three columns **“Pregnancies”, “Glucose” and “BMI”** Is better than Full model (M1) using hypothesis testing. Or say M2 is significantly more informative than M1.

Description of the study:

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988, November). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care* (p. 261). American Medical Informatics Association.