Etude des relations de causalité

ZIDAN Loubna

2023-04-10

Table de matières:

1- Introduction:

Le diabète est une maladie chronique caractérisée par une hyperglycémie, c'est-à-dire une concentration élevée de glucose dans le sang. Selon l'Organisation mondiale de la santé (OMS), le nombre de personnes atteintes de diabète dans le monde est passé de 108 millions en 1980 à 422 millions en 2014. Cette augmentation rapide est principalement due à l'augmentation de l'obésité, de la sédentarité et du vieillissement de la population.

Comprendre les relations causales entre les différentes variables liées au diabète est crucial pour prévenir et traiter cette maladie. Les variables liées au diabète comprennent, entre autres, l'âge, le sexe, l'indice de masse corporelle (IMC) ou bien (BMI) en englais, la consommation de tabac et d'alcool, l'hypertension artérielle, l'hypercholestérolémie, la glycémie à jeun, la glycémie postprandiale, etc.

Dans ce projet, nous chercherons à étudier les relations causales entre ces différentes variables liées au diabète à l'aide de techniques statistiques avancées telles que l'analyse de régression. Nous espérons que les résultats de cette recherche permettront de mieux comprendre les facteurs qui contribuent au développement du diabète et de développer des stratégies efficaces de prévention et de traitement.

2- Données:

Les données que nous allons utiliser proviennent de l'ensemble de données Pima Indians Diabetes. Cet ensemble de données contient des informations sur 768 femmes d'origine indienne Pima âgées de 21 ans et plus. Les variables sont les suivantes :

- "Pregnancies" : le nombre de fois que la patiente a été enceinte
- "Glucose" : la concentration de glucose dans le sang de la patiente (mesurée en mg/dl)
- "BloodPressure" : la pression artérielle diastolique de la patiente (mesurée en mmHg)
- "SkinThickness" : l'épaisseur de la peau du triceps de la patiente (mesurée en mm)
- "Insulin" : la concentration d'insuline dans le sang de la patiente (mesurée en mU/l)

• “BMI” (Body Mass Index) : l’indice de masse corporelle de la patiente (calculé comme le poids en kilogrammes divisé par le carré de la taille en mètres)

• “DiabetesPedigreeFunction” : une mesure de la probabilité de développer le diabète en fonction de l’historique familial de la patiente.

• “Age” : l’âge de la patiente (mesuré en années)

• “Outcome” : une variable binaire qui indique si la patiente a été diagnostiquée avec le diabète (1) ou non (0).

3- Méthodes:

Étant donné que la variable de sortie (outcome) dans notre ensemble de données est une variable binaire qui indique si la patiente a été diagnostiquée avec le diabète (1) ou non (0), nous devons utiliser une régression logistique.

La régression logistique est une technique statistique qui permet de modéliser la relation entre une variable binaire dépendante et plusieurs variables indépendantes continues ou catégorielles. Dans notre cas, la variable de sortie est binaire (diabète ou non-diabète) et les variables indépendantes sont continues. Par conséquent, la régression logistique est plus appropriée pour notre analyse.

La régression logistique nous permettra de déterminer l’effet de chaque variable indépendante sur la probabilité de diabète chez les femmes Pima, ajustée pour les autres variables indépendantes dans le modèle.

Cependant, il est important de noter que la régression logistique est une méthode statistique qui présuppose certaines hypothèses, telles que :

1- L’hypothèse de non-multicollinéarité : cette hypothèse stipule que les variables indépendantes ne sont pas fortement corrélées les unes aux autres. Si les variables indépendantes sont fortement corrélées, cela peut rendre difficile la distinction de l’effet de chaque variable sur la variable dépendante.

2- L’hypothèse de normalité : cette hypothèse stipule que la distribution des résidus (différences entre les valeurs observées et les valeurs prédites) est normale. Nous pouvons vérifier cette hypothèse en traçant un histogramme des résidus et en vérifiant s’ils suivent une distribution normale.

En vérifiant ces hypothèses, nous pouvons nous assurer que notre modèle de régression logistique est approprié pour les données et que les résultats de l’analyse sont fiables.

Nous allons effectuer les étapes suivantes :

a)Charger les données

b)Vérifier la qualité des données

c)Explorer les données

d)Prétraitement des données

e)Diviser les données en ensemble d’entraînement et ensemble de test

f)Créer le modèle de régression logistique

g)vérifier les hypotheses de modèle

h)Evaluer le modèle

a)Charger les données:

```
df <- read.csv("C:\\Users\\loubn\\Desktop\\M1\\Projet_causalité\\diabetes.csv")
```

b)Vérifier la qualité des données:

```
summary(df)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00
##     Outcome
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
```

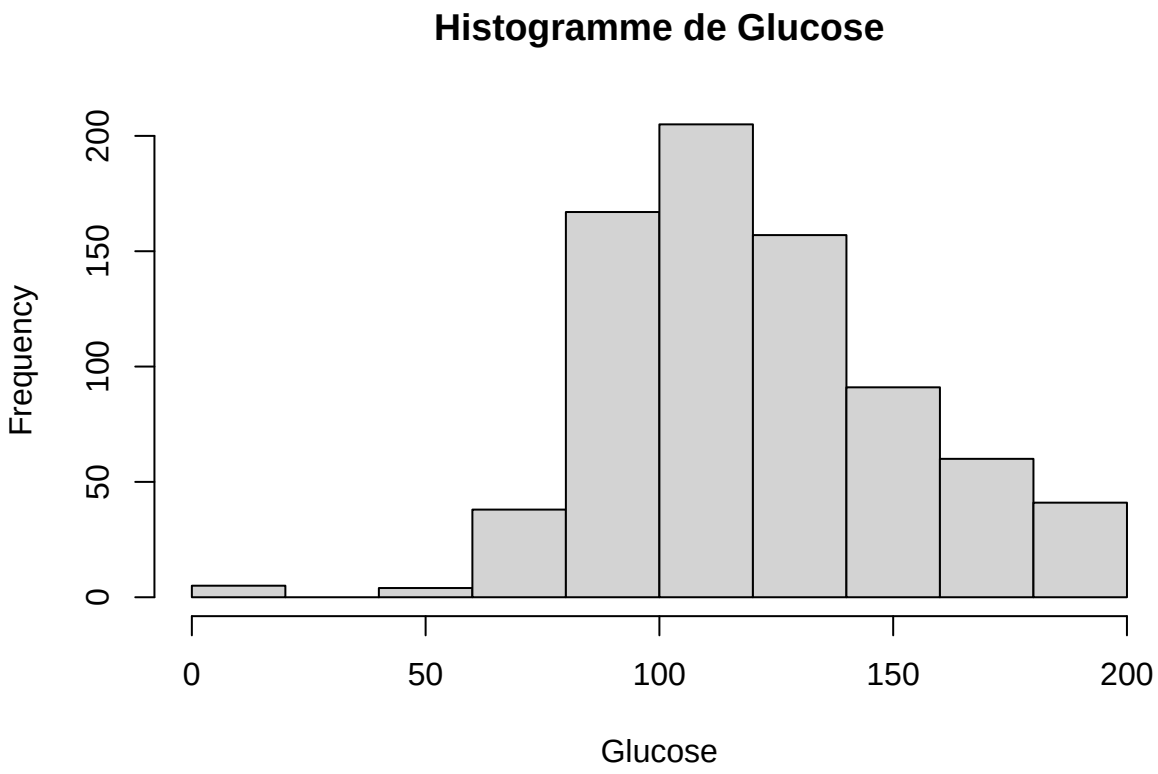
```
sapply(df, function(x) sum(is.na(x)))
```

```
##              Pregnancies                  Glucose            BloodPressure
##                        0                        0                        0
##            SkinThickness                  Insulin                      BMI
##                        0                        0                        0
## DiabetesPedigreeFunction                      Age                  Outcome
##                        0                        0                        0
```

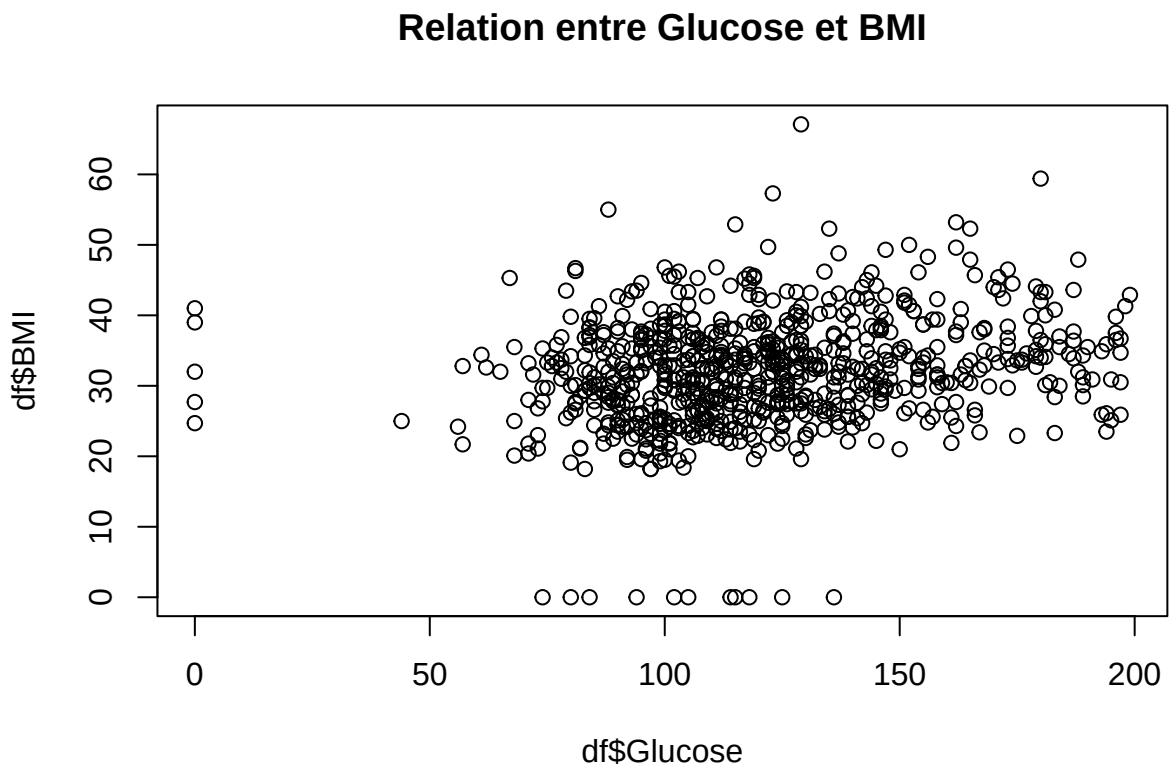
Bien que la variable de sortie soit binaire (diabète ou non-diabète) et que les variables indépendantes soient continues, nous pouvons remarquer que la qualité des données est bonne puisqu'il n'y a pas de valeurs manquantes.

c)Explorer les données:

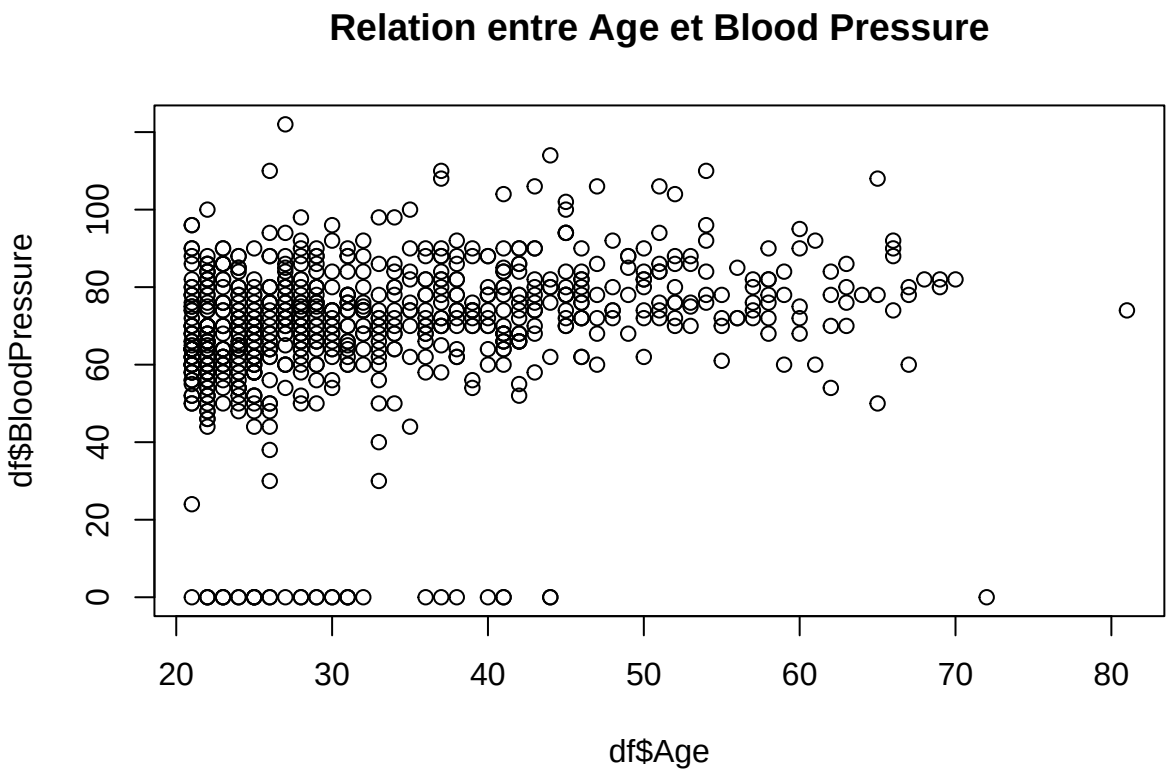
```
hist(df$Glucose, main="Histogramme de Glucose", xlab="Glucose")
```



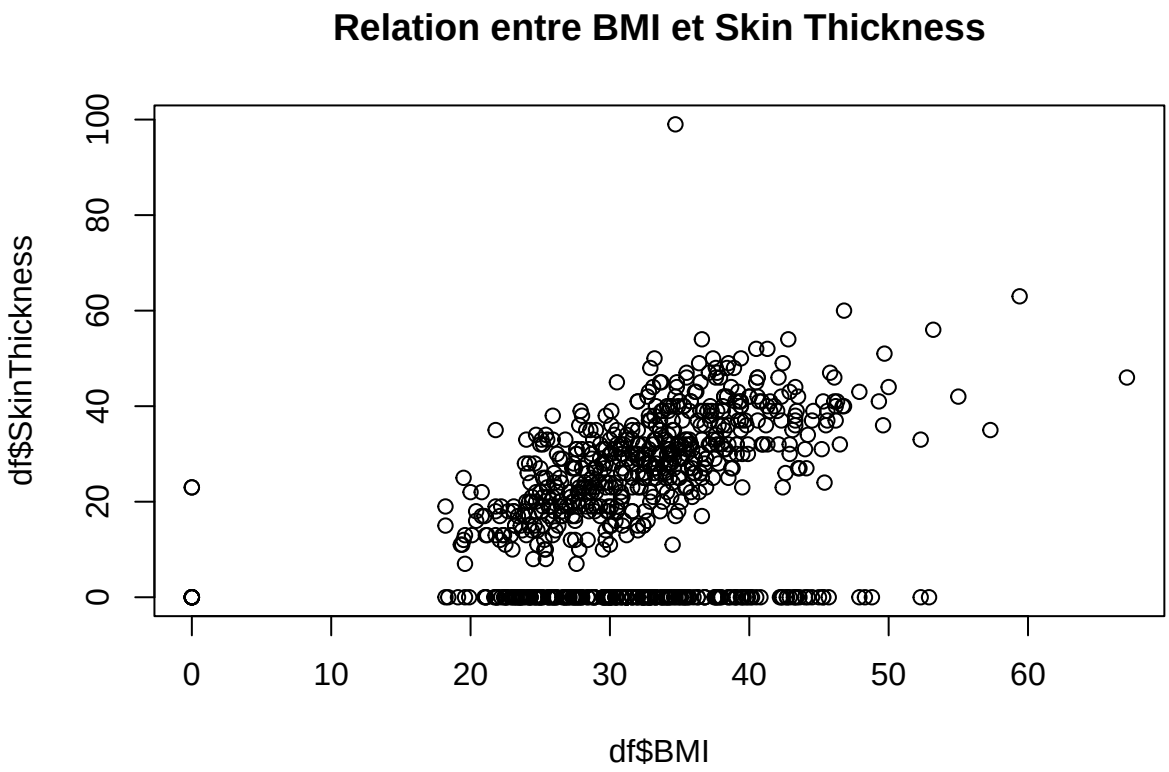

```
plot(df$Glucose, df$BMI, main = "Relation entre Glucose et BMI")
```




```r
plot(df$Age, df$BloodPressure, main = "Relation entre Age et Blood Pressure")
```




```r
plot(df$BMI, df$SkinThickness, main = "Relation entre BMI et Skin Thickness")
```



-D'après l'histogramme du glucose, il est observé que la concentration de glucose dans le sang de la patiente présente une fréquence élevée entre 100 et 150 mg/dl, ce qui suggère qu'une grande partie des patientes ont une concentration de glucose dans cette plage

-Lorsque nous examinons la relation entre le glucose et l'BMI, nous constatons que la corrélation est nulle, car le nuage de points tend à former une ligne droite horizontale.

-La relation entre l'âge et la tension artérielle indique clairement que la variable de la tension artérielle n'a pas de corrélation avec l'âge des patientes.

-Enfin, en ce qui concerne la relation entre le BMI et l'épaisseur de la peau du triceps, nous pouvons observer une corrélation positive entre les deux variables. Autrement dit, plus l'une augmente, plus l'autre augmente en même temps. Cependant, cette augmentation ne s'applique pas aux patientes ayant une épaisseur de peau du triceps égale à zéro.

#Nous pouvons également utiliser des corrélations pour explorer les relations entre les variables.

```
cor(df)
```

```
##                           Pregnancies    Glucose BloodPressure SkinThickness
## Pregnancies                1.00000000 0.12945867    0.14128198   -0.08167177
## Glucose                    0.12945867 1.00000000    0.15258959    0.05732789
## BloodPressure              0.14128198 0.15258959    1.00000000    0.20737054
## SkinThickness             -0.08167177 0.05732789    0.20737054    1.00000000
## Insulin                   -0.07353461 0.33135711    0.08893338    0.43678257
## BMI                        0.01768309 0.22107107    0.28180529    0.39257320
## DiabetesPedigreeFunction -0.03352267 0.13733730    0.04126495    0.18392757
## Age                        0.54434123 0.26351432    0.23952795   -0.11397026
```

```
## Outcome                   0.22189815 0.46658140    0.06506836    0.07475223
##                              Insulin        BMI DiabetesPedigreeFunction
## Pregnancies              -0.07353461 0.01768309              -0.03352267
## Glucose                   0.33135711 0.22107107               0.13733730
## BloodPressure             0.08893338 0.28180529               0.04126495
## SkinThickness             0.43678257 0.39257320               0.18392757
## Insulin                   1.00000000 0.19785906               0.18507093
## BMI                       0.19785906 1.00000000               0.14064695
## DiabetesPedigreeFunction  0.18507093 0.14064695               1.00000000
## Age                      -0.04216295 0.03624187               0.03356131
## Outcome                   0.13054795 0.29269466               0.17384407
##                                  Age    Outcome
## Pregnancies               0.54434123 0.22189815
## Glucose                   0.26351432 0.46658140
## BloodPressure             0.23952795 0.06506836
## SkinThickness            -0.11397026 0.07475223
## Insulin                  -0.04216295 0.13054795
## BMI                       0.03624187 0.29269466
## DiabetesPedigreeFunction  0.03356131 0.17384407
## Age                       1.00000000 0.23835598
## Outcome                   0.23835598 1.00000000
```

D'après cette table de corrélation entre les variables, nous pouvons constater qu'il y a une forte corrélation (54%) entre l'âge et le nombre de grossesses ("Pregnancies"). De plus, il y a une corrélation moins forte (46%) entre le résultat ("Outcome") et le glucose, qui est la variable la plus corrélée avec "Outcome".

d)Prétraitement des données:

```
# Convertir la variable Outcome en facteur
df$Outcome <- as.factor(df$Outcome)

# Standardiser les variables numériques
num_cols <- c("Pregnancies", "Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI", "DiabetesPe
df[num_cols] <- scale(df[num_cols])
```

e)Diviser les données en ensemble d'entraînement et ensemble de test:

```
set.seed(123)
training.samples <- df$Outcome %>%
  createDataPartition(p = 0.8, list = FALSE)
train.data <- df[training.samples, ]
test.data <- df[-training.samples, ]
```

f)Créer le modèle de régression logistique:

```
model <- glm(Outcome ~ Pregnancies + Glucose + BloodPressure + SkinThickness + Insulin + BMI + Diabetes
```

g)Vérifier les hypotheses du modèle:

1- Vérifier l'hypothèse de non-multicollinéarité:

```
vif(model)
```

```
##              Pregnancies                  Glucose            BloodPressure
##                 1.440356                 1.204816                 1.180210
##            SkinThickness                  Insulin                      BMI
##                 1.535105                 1.419226                 1.234531
## DiabetesPedigreeFunction                      Age
##                 1.032761                 1.541349
```

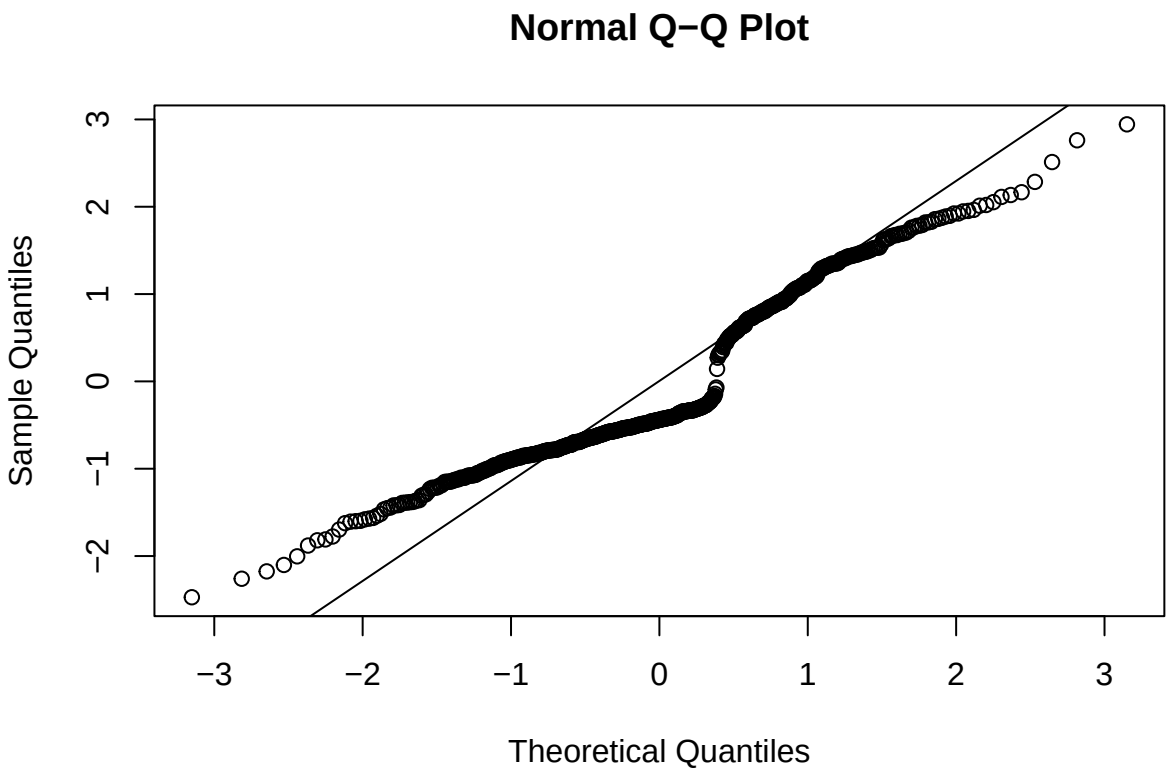
Lorsque la VIF (variance inflation factor) d'une variable est supérieure à 5, il est recommandé de la retirer du modèle. Dans notre cas, aucune de nos variables n'a une VIF supérieure à 5, ce qui indique qu'il n'y a pas de problème de multicolinéarité.

2- Vérifier la normalité des résidus

```
# Faire une prédiction sur les données de test
pred <- predict(model, newdata = test.data, type = "response")

# Calculer les résidus
residus <- residuals(model, type = "deviance")

# Tracer un graphique en Q-Q des résidus
qqnorm(residus)
qqline(residus)
```



Si la ligne de régression théorique et les résidus diffèrent considérablement, cela peut indiquer que les résidus ne suivent pas une distribution normale. Cependant, dans notre cas, nous pouvons observer que la ligne de régression théorique et les résidus suivent la même direction et sont proches, ce qui indique que les résidus suivent une distribution normale.

h)Evaluer le modèle:

```
summary(model)
```

```
##
## Call:
## glm(formula = Outcome ~ Pregnancies + Glucose + BloodPressure +
##     SkinThickness + Insulin + BMI + DiabetesPedigreeFunction +
##     Age, family = "binomial", data = train.data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4719  -0.7674  -0.4402   0.7776   2.9436
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -0.850841   0.105914  -8.033 9.49e-16 ***
## Pregnancies               0.336385   0.120759   2.786  0.00534 **
## Glucose                   1.094443   0.129596   8.445  < 2e-16 ***
## BloodPressure            -0.287765   0.107555  -2.676  0.00746 **
```

```
## SkinThickness            -0.009735   0.121630  -0.080  0.93621
## Insulin                  -0.082021   0.110226  -0.744  0.45681
## BMI                       0.636012   0.130450   4.876 1.09e-06 ***
## DiabetesPedigreeFunction  0.309976   0.112273   2.761  0.00576 **
## Age                       0.181526   0.120799   1.503  0.13291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 796.05  on 614  degrees of freedom
## Residual deviance: 598.41  on 606  degrees of freedom
## AIC: 616.41
##
## Number of Fisher Scoring iterations: 5
```

Les résultats de la régression logistique indiquent que certaines variables sont significativement associées à l'issue du diabète. Plus spécifiquement, on peut observer :

-La variable "Pregnancies" a un coefficient positif (0,336) et un p-value significative (0,005), ce qui suggère que le nombre de grossesses antérieures est positivement associé à l'issue du diabète.

-La variable "Glucose" a un coefficient positif élevé (1,094) et une p-value très significative (< 2e-16), ce qui suggère qu'un taux de glucose élevé est fortement associé à un risque accru de diabète.

-La variable "BloodPressure" a un coefficient négatif (-0,288) et une p-value significative (0,007), ce qui suggère que des niveaux élevés de pression artérielle peuvent être associés à un risque plus faible de diabète.

-La variable "BMI" a un coefficient positif (0,636) et une p-value très significative (1,09e-06), ce qui suggère que des niveaux élevés d'IMC sont fortement associés à un risque accru de diabète.

-Les autres variables (SkinThickness, Insulin, DiabetesPedigreeFunction et Age) n'ont pas de p-value significative, ce qui suggère qu'elles ne sont pas significativement associées à l'issue du diabète dans ce modèle.

-La déviance nulle était de 796.05 et la déviance résiduelle est de 598.41, ce qui indique que le modèle est assez bon pour ajuster les données. L'AIC (criterion d'information d'Akaike) est de 616.41, ce qui est un bon ajustement.

4-Interpréter les résultats:

Nous pouvons dire que la régression logistique a montré que certaines variables étaient significativement associées à l'issue du diabète, comme le nombre de grossesses antérieures, le taux de glucose, la pression artérielle et l'IMC.

Il est également important de noter que les autres variables (SkinThickness, Insulin, DiabetesPedigreeFunction et Age) n'ont pas été trouvées significativement associées à l'issue du diabète dans ce modèle. Cela peut suggérer que ces variables ne sont pas des facteurs de risque importants pour le développement du diabète dans cet échantillon de données.

Il est également intéressant de noter que la déviance résiduelle est inférieure à la déviance nulle, ce qui indique que le modèle est capable de bien ajuster les données. En outre, l'AIC est relativement faible, ce qui suggère que le modèle est un bon ajustement pour les données.

Cependant, il est important de prendre en compte les autres facteurs qui peuvent influencer l'issue du diabète, tels que les antécédents familiaux et les habitudes de vie.

5- Conclusion:

Dans l'ensemble, notre analyse suggère que les variables "Glucose", "BMI" et "Age" peuvent être des facteurs de risque significatifs pour le diabète, En utilisant la régression logistique avec les variables "Pregnancies", "Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI", "DiabetesPedigreeFunction", et "Age" comme prédicteurs, nous avons trouvé que Glucose, BMI, DiabetesPedigreeFunction et le nombre de grossesses (Pregnancies) étaient significativement associés à la présence de diabète (Outcome) chez les patients. Les coefficients de régression pour ces variables étaient positifs, ce qui indique que des niveaux plus élevés de ces variables étaient associés à une probabilité plus élevée de diabète.

Cependant, il est important de noter que l'analyse est basée sur des données observées, et ne prouve pas nécessairement une relation de causalité. D'autres facteurs, tels que les facteurs génétiques ou environnementaux, pourraient également influencer le développement du diabète. Par conséquent, des études complémentaires pourraient être nécessaires pour confirmer ces résultats et mieux comprendre les mécanismes sous-jacents.