** **

**ECOLE NATIONALE**

**DES SCIENCES APPLIQUEES OUJDA**

****

**Rapport de Projet**

**Modélisation Statistique**

**--IDSCC4-----2021-2022--**

Préparé par ***:*** ***ALLA ISMAIL***

***LAMKADAM YASSIR***

Encadrant ***: Mme ELMEHDI Rachida***

***Résumé/Abstract***

La modélisation dans les mathématiques consiste à définir une expression mathématique qui décrit un phénomène réel en faisant introduire des variables (vecteurs au sens algébrique). C’est donc chercher pour ce phénomène un modèle adéquat, noté mathématiquement par une fonction qui est souvent une fonction inconnue et à déterminer. Par ailleurs, il existe des phénomènes dont lesquels il faut intervenir un certain ensemble de contraintes entre les variables qui doivent être prises en considération dans la détermination de la fonction.

Les modèles cherchés ont trois caractéristiques principales. Ils sont hypothétiques, modifiables et adéquats pour certains problèmes dans certaines situations ? Il n’y a donc pas le bon modèle d’une manière absolue.

La modélisation est un des objets de la statistique. De ce fait, la modélisation statistique consiste principalement dans la représentation des bases de données observées par des modèles théoriques qui décrivent au mieux ces données en utilisant des variables explicatives et des variables à expliquer et en tenant compte surtout de leur nature aléatoire. Elle sert entre autres à faire de l’inférence et à établir des prédictions afin d’en tirer des conclusions et d’en prendre des décisions.

Les méthodes de modélisation statistique sont diverses et variées mais les principales parmi elles sont le modèle linéaire (gaussien), le modèle linéaire généralisé, les modèles non linéaires, les modèles mixtes, les modèles pour données répétées, les modèles pour séries chronologiques, l’analyse discriminante et la classification, …

Dans ce projet, on va se focaliser sur le modèle linéaire (gaussien) : régression linéaire simple, régression linéaire multiple et ANOVA.

# 

# Listes des figures

[Figure 1:data summary](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106571)

[Figure 2:summary régression linéaire simple](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106572)

[Figure 3: commande plot régression linéaire simple](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106573)

[Figure 4: droite de régression linéaire simple](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106574)

[Figure 5: intervalle de confiance des paramètres du model](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106580)

[Figure 6:data summary](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106571)

[Figure 7: plot de la matrice de corrélation.](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106572)

[Figure 8:summary régression linéaire multiple](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106571)

[Figure 9: modéle après ré-estimation.](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106572)

[Figure 10: modéle après 2éme ré-estimation.](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106572)

[Figure 11 :test de Fisher](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106571)

FIGURE 12 : COEFICIENT DE DETERMINATION R2……………………………………………………………………………………………………………………………………………

[Figure 13: intervalle de confiance des paramètres du mode](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106572)

[Figure 14 :les données de test (my\_df).](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106571)

[Figure 15: le résultat de la prédiction ponctuelle](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106572)

[Figure 16: le résultat de la prédiction par intervalle](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106572)

[Figure 17 :data summary](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106571)

[Figure 18: anova](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106572)

Table des matières

[**I.** **Régression linéaire Simple**](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106306)

[1. Introduction et présentation de données :](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106307)

[2. Régression linéaire simple *et* l’estimation des paramètres:](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106308)

[3. Coefficient de détermination:](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106309)

[4. Tests d’hypothèse :](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106308)

[5. Intervalle de confiance :](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106309)

[6. intervalle de Prédiction  :](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106309)

[**II.** **Régression Linéaire Multiple :**](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106311)

[1. Introduction et présentation de données :](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106307)

[2. Matrice de corrélation:](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106308)

[3. régression linéaire multiple et l’estimation des paramètres:](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106309)

[4. Tests d’hypothèse :](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106308)

[5. Coefficient de détermination:](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106309)

[6. Intervalle de confiance :](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106309)

[7. intervalle de Prédiction  :](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106309)

[**III.** **analyse de variance anova :**](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106311)

[1. Introduction et présentation de données :](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106307)

[2. analyse de variance anova :](file:///C:\Users\pc\AppData\Local\Temp\modélisation%20statistique.docx#_Toc92106308)

**Régression linéaire Simple**

# 

**Introduction :**

La régression linéaire simple est un [modèle de régression linéaire](https://en.wikipedia.org/wiki/Linear_regression) avec une seule [variable explicative](https://en.wikipedia.org/wiki/Covariate) . C'est-à-dire qu'il concerne des points d'échantillonnage à deux dimensions avec [une variable indépendante et une variable dépendante](https://en.wikipedia.org/wiki/Dependent_and_independent_variables) (conventionnellement, les coordonnées *x* et *y* dans un [système de coordonnées cartésiennes](https://en.wikipedia.org/wiki/Cartesian_coordinate_system) ) et trouve un fonction linéaire (une [ligne droite](https://en.wikipedia.org/wiki/Straight_line) non verticale) qui, aussi précisément que possible, prédit les valeurs des variables dépendantes en fonction de la variable indépendante.

**1-présentation de données :**

On charge la base de données à deux colonnes y et x, avec y la variable à expliquer, et x la variable explicative, les nombres des variables d’entrainement n= 768

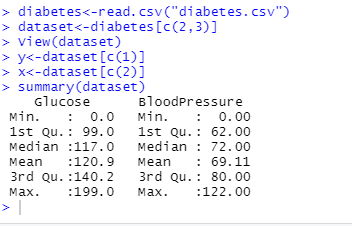


Figure  : data summary

La commande sammray nous permet d’avoir une vision générale sur notre base de données

Et d’afficher les différents indices statistiques (la moyenne, ….)

# 2-Régression linéaire simple et l’estimation des paramètres:

# 

Figure : summary régression linéaire simple

La régression linéaire R utilise la fonction lm() pour créer un modèle de régression à partir de la base de données train, et la stocke dans la variable dataset.lm1

La commande summary permet de donner un aperçu sur la régression sur la base de données.

**L’estimation des paramètres** :

L’intercepté (0) = 103.47629

La pente (1) = 0.25205

# Donc la droite de régression de notre modèle est : Y = 103.47629 + 0.25205\*X

# 3.Coefficient de détermination et coefficient de corrélation linéaire :

# Le coefficient de détermination exprime le rapport entre la variance expliquée par le modèle de régression et la variance totale. Il sert à évaluer la qualité de l’ajustement

# On ici le R2 = 0.022 << 1

# Donc le modèle n’ajuste pas bien le nuage de pointes

# Car la variance totale est très élevés

**4-** **Tests d’hypothèse au niveau de risque 5 %:**

**Significativité des paramètres (du modèle) :**

Test de student :

Pour 0:

l’hypothèse :

H0 : 0 = 0 VS H1 : 0 ! = 0

Nous avons Pr <2**e**-16 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 0  est très significative dans notre modèle.

Pour 1:

l’hypothèse : H0 : 1 = 0 VS H1 : 1! = 0

Nous avons Pr est largement inférieure à 0,05 donc on rejette l’hypothèse H0, alors 1  est très significative dans notre modèle.

Alors on va désigner les données la droit de régression linéaire simple

# 

Figure : commande plot régression linéaire simple

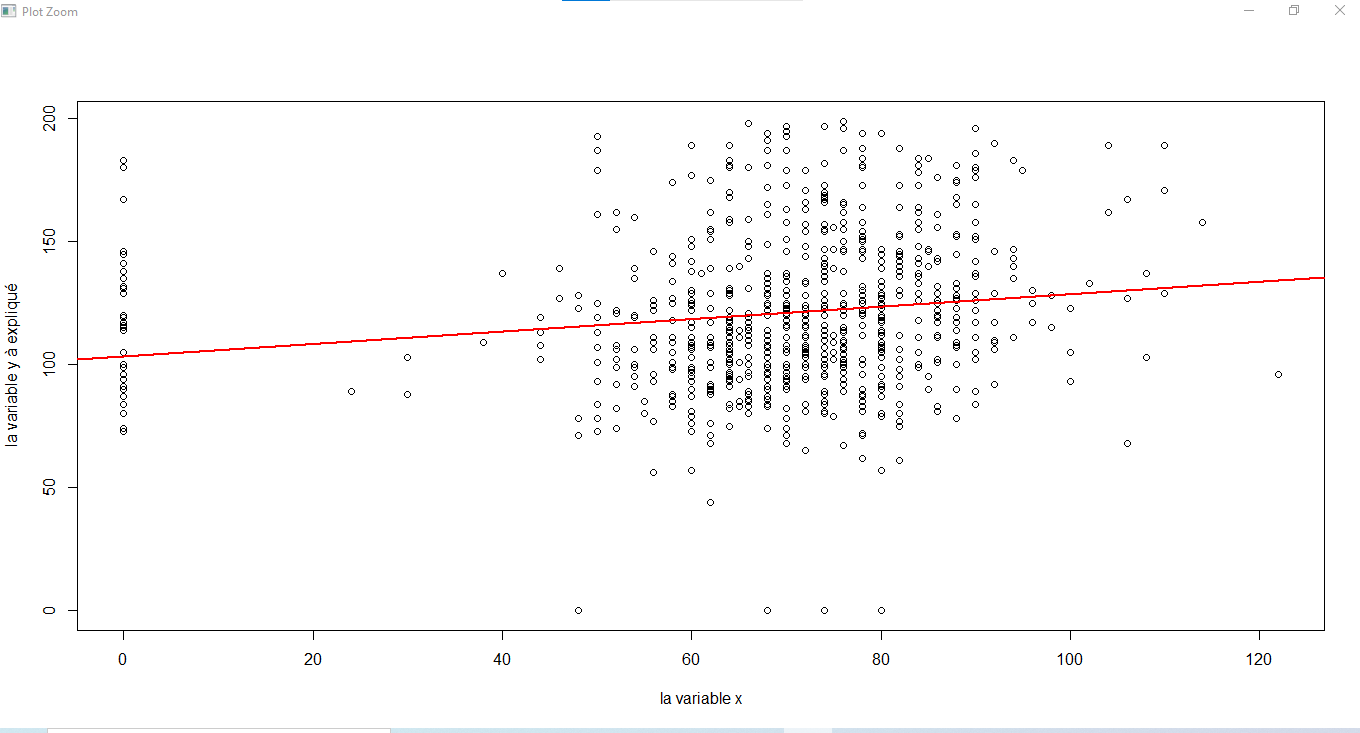


Figure : droite de régression linéaire simple

**5-Intervalle de confiance :**

L'intervalle de confiance est un indicateur mathématique qui permet de chiffrer la zone d'incertitude, lors d'une enquête ou d'un sondage portant sur un échantillon de population.

On effectue la détermination de l’intervalle de confiance de notre modèle par la commande ‘confint(dataset.lm1)’ Et on obtient :

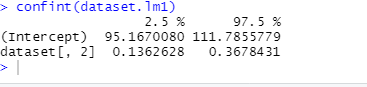


Figure5: intervalle de confiance des paramètres du model

# Ici l’intervalle de confiance de beta0 c’est [95.167 ; 111.785] avec 95%

# de confiance

# Ici l’intervalle de confiance de beta1 c’est [0.1362628 ; 0.3678431] avec 95%de confiance

# 6-intervalle de Prédiction :

# soit Xi\* = 95 donc Y͠ i\* = 103.47629 + 0.25205\*95 = 9854.1923

# donc l’intervalle de prédiction est : Yi € [Y͠ i\* - σ1 \* ttab ; Y͠ i\* + σ1 \* ttab]

# soit

# hi\*=0.6533 e

# σ1 2= σ2 2\*(1+ hi\*)

# ttab=2.228

# resudial standard errer =31.62 🡺 σ1 =(31.62\*31.62\*(1+0.6533))⌃(1/2)

# σ1=40.65

# intervalle=[9854.1923- 40.65\*2.228; 9854.1923+40.65\*2.228]

# intervalle de prédiction de Y\* est [9763.62 ; 9944.76]

# plus l’intervalle de prédiction est large et plus la prédiction perd de précision

Régression Linéaire Multiple

# Introduction :

L'analyse de régression multiple est utilisée lorsque nous voulons prédire la valeur d'une variable en fonction de la valeur de deux ou plusieurs autres variables. La variable que nous voulons prédire est appelée variable dépendante et la variable que nous utilisons pour prédire la variable dépendante est appelée variable indépendante. Dans ce cas, la variable dépendante est le et la variable indépendante est toute autre variable de l’ensemble de données.

**1-Présentation des données :**

Les ensembles de données se composent de plusieurs variables prédictives médicales (indépendantes) et d'une variable quantité de Glucose (dépendante). Les variables indépendantes incluent le nombre de grossesses que la patiente a eues, son IMC, son taux d'insuline, son âge, etc.

On charge la base de données ‘diabètes’ qui contient 6 variables continues et 768 exemples pour l’entrainement

On utilise la commande summary(diabètes) pour donner une aperçu sur la base de données

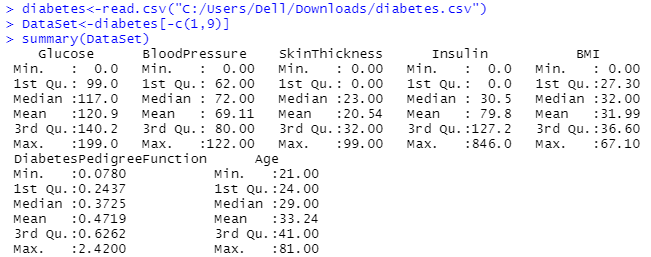


Figure 6 : summary data

La commande summray nous permet d’avoir une vision générale sur notre base de données

Et d’afficher les différents indices statistiques (la moyenne, ….)

**2-Matrice de corrélation :**

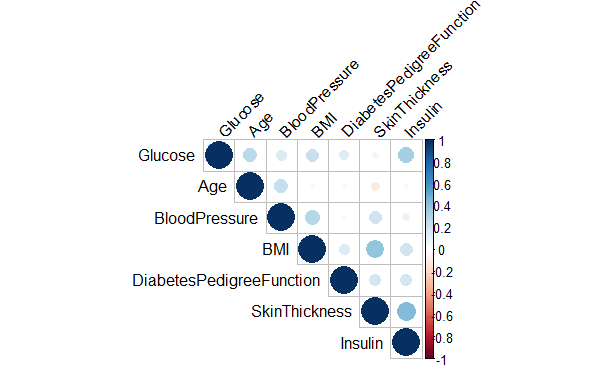
****

Figure 7: plot de la matrice de corrélation.

**Remarque :**

D’aprés le plot de la matrice de corrélation on remarque que toutes les variables sont indépendantes

**3-Régression linéaire multiple :**

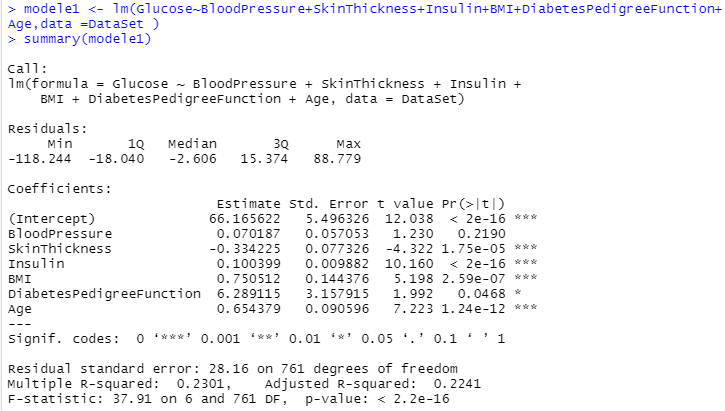
****

Figure 8: summary régression linéaire multiple

La régression linéaire R utilise la fonction lm() pour créer un modèle de régression à partir de la base de données train, et la stocke dans la variable modele1

La commande summary permet de donner un aperçu sur la régression sur la base de données.

**L’estimation des paramètres**

0 = 66.165622

1 = 0.070187

2 = -0.334225

3 = 0.100399

4 = 0.750512

5 = 6.289115

6 = 0.654379

# Donc , le modèle de la régression linéaire multiple s’écrit sous forme matricielle

# est : Y=X\* (avec X c’est la matrice des données).

**4-** **Tests d’hypothèse au niveau de risque 5 %:**

**Significativité des paramètres :**

***Test de student :***

H0: 0=0 VS H1: 0 ≠ 0

Nous avons Pr < 2e-16 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 0 est très significative dans notre modèle.

H0: 1=0 VS H1: 1 ≠ 0

Nous avons 0.2190 est superieur à 0,05 donc on accepte l’hypothèse H0, alors on élemine 1 et on ré-estime le modele

H0: 2=0 VS H1: 2 ≠ 0

Nous avons 1.75e-05 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 2 est très significative dans notre modèle

H0: 3=0 VS H1: 3 ≠ 0

Nous avons Pr < 2e-16 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 3 est très significative dans notre modèle.

H0: 4=0 VS H1: 4 ≠ 0

Nous avons 2.59e-07 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 4 est très significative dans notre modèle.

H0: 5=0 VS H1: 5 ≠ 0

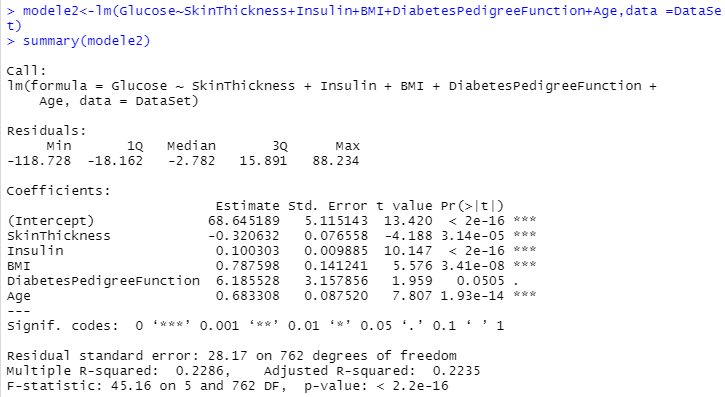
Nous avons 0.0468 est inférieure à 0,05 donc on rejette l’hypothèse H0, alors 5 est significative dans notre modèle.

H0: 6=0 VS H1: 6 ≠ 0

Nous avons 1.24e-12 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 6 est très significative dans notre modèle.

***Ré-estimation du modèle :***

Voici les résultats après l’élimination de 1

 Figure 9: modéle après ré-estimation.

***Test de student (après ré-estimation) :***

H0: 0=0 VS H1: 0 ≠ 0

Nous avons Pr < 2e-16 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 0 est très significative dans notre modèle.

H0: 2=0 VS H1: 2 ≠ 0

Nous avons 3.14e-05 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 2 est très significative dans notre modèle

H0: 3=0 VS H1: 3 ≠ 0

Nous avons Pr < 2e-16 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 3 est très significative dans notre modèle.

H0: 4=0 VS H1: 4 ≠ 0

Nous avons 3.41e-08 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 4 est très significative dans notre modèle.

H0: 5=0 VS H1: 5 ≠ 0

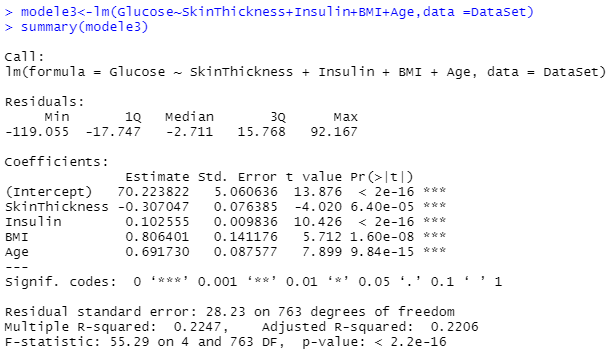
Nous avons 0.0505 est superieur à 0,05 donc on accepte l’hypothèse H0, alors on élemine 5 et on ré-estime le modele.

H0: 6=0 VS H1: 6 ≠ 0

Nous avons 1.93e-14 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 6 est très significative dans notre modèle.

***Ré-estimation du moldele :***

Voici les résultats après l’élimination de 5



*Figure 10: modéle après 2éme ré-estimation.*

***Test de student (après ré-estimation) :***

H0: 0=0 VS H1: 0 ≠ 0

Nous avons Pr < 2e-16 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 0 est très significative dans notre modèle.

H0: 2=0 VS H1: 2 ≠ 0

Nous avons 6.40e-05 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 2 est très significative dans notre modèle

H0: 3=0 VS H1: 3 ≠ 0

Nous avons Pr < 2e-16 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 3 est très significative dans notre modèle.

H0: 4=0 VS H1: 4 ≠ 0

Nous avons 1.60e-08 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 4 est très significative dans notre modèle.

H0: 6=0 VS H1: 6 ≠ 0

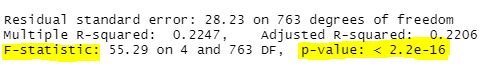
Nous avons 9.84e-15 est très inférieure à 0,05 donc on rejette l’hypothèse H0, alors 6 est très significative dans notre modèle.

**Significativité de modèle :**

***Test de Fisher:***

Le test de significativité global du modèle. Il consiste à tester si tous les coefficients du modèle, à l’exception de l’intercepte, sont nuls. Les hypothèses sont :

H0: 2 =3 =4 =6 = 0 VS H1: ∃j tel que j ≠ 0



*Figure 11: test de Fisher*

On a la p-value est largement inferieur a 0.05 donc on peut rejette fortement H0

Donc notre modèle finale s’écrit sous forme matricielle :

***Y=\*X (***avec X c’est la matrice des données***).***

# 5.Coefficient de détermination:

# Le coefficient de détermination exprime le rapport entre la variance expliquée par le modèle de régression et la variance totale. Il sert à évaluer la qualité de l’ajustement

# On ici le R2 = 0.22 = 22%

# 

# Figure 12: coefficient de détermination R2

# On ici le R2 = 0.022 << 1

# Donc le modèle n’ajuste pas bien le nuage de pointes

# Car la variance totale est très élevés

**6-Intervalle de confiance  :**

L'intervalle de confiance est un indicateur mathématique qui permet de chiffrer la zone d'incertitude, lors d'une enquête ou d'un sondage portant sur un échantillon de population.

On effectue la détermination de l’intervalle de confiance de notre modèle par la commande ‘***confint(modele3)***’

Et on obtient :

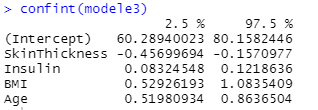


Figure 13: intervalle de confiance des paramètres du model

Donc

0 ∈ [60.2894 **;** 80.1582]

2 ∈ [-0.4569 **;** -0.1570]

3 ∈ [0.0832 **;** 0.1218]

4∈ [0.5292 **;** 1.0835]

6∈ [0.5198 **;** 0.8636]

**7- Prédiction  :**

Dans cette partie on va faire un prédiction ponctuelle et une prédiction par intervalle

Voici les données pour lesquels on va prédire le nouveau résultat



Figure 14: les données de test (my\_df).

**Prédiction ponctuelle :**

On a obtenue le résultat suivant avec la commande **‘ predict() ’**



Figure 15: le résultat de la prédiction ponctuelle .

Donc Ŷi\*=124.41

**Prédiction par intervalle :**

De même avec la commande **‘ predict() ’** on obtient le résultat suivant



*Figure 16: le résultat de la prédiction par intervalle .*

Donc :

Yi\*∈ [ 68.941 **;** 179.878]

**L’analyse de variance et de covariance**

**Introduction :**

L’analyse de variance, appelée tout simplement ANOVA a pour objectif d’expliquer une variable  
aléatoire quantitative en fonction de variables explicatives qualitatives appelées souvent facteurs  
qui peuvent avoir plusieurs modalités (niveaux). C’est une analyse d’une partition ou une  
classification des données en un certain ensemble de groupes ou classes appelés généralement les  
niveaux d’un facteur. Elle consiste dans la comparaison des moyennes empiriques de pour les  
différents niveaux pris par les facteurs.

**1-Présentation des données :**

Soit le facteur « région » de 4 niveaux (Northeast , northwest , southwest , southeast) de tailles respectives 324 , 325, 325 et 364 tel que la taille totale n=1338

On utilise la commande summary(datasetq) pour donner une aperçu sur la base de données

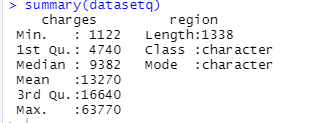


Figure 17: summary data

La commande summray nous permet d’avoir une vision générale sur notre base de données

Et d’afficher les différents indices statistiques (la moyenne, ….)

**2- analyse de variance anova :**

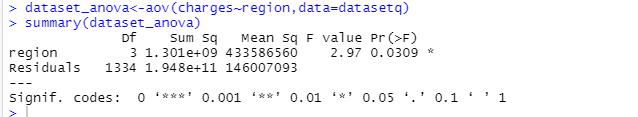


Figure 18: anova

On a F(3,1334)=2.97 avec p=0.0309<0.05