

Données manquantes - cas pratique sur les données Pima

Alexis Guyonvarch, Philippe Real

14/12/2019

Contents

Introduction	1
Exploration des données	2
Statistiques descriptives	2
Analyse exploratoire - mécanisme des données manquantes	4
Distributions marginales des variables	6
Imputations multiples	8
Imputations multiples avec les méthodes JM et FCS du package MICE	8
Imputations multiples avec la méthode MIPCA du package missMDA	9
Validation de la qualité des imputations	10
Analyse de sensibilité du modèle	12
Modélisation - Régression logistique	15
Listwise deletion	15
Mise en oeuvre avec les données imputées	15
Conclusion	16
ANNEXE	17
ACM - aide à l'interprétation	17
Imputation simple	17
Calcul du nombre d'imputations pour les fonctions MICE	18
Transformation des données	19
Graphiques “stripplot” - qualité des imputations multiples	20
Choix de modèle	20

Introduction

Le jeu de données **Pima** sur lequel nous allons nous concentrer pour le projet provient de l’Institut national du diabète, des maladies digestives et rénales des Etats-Unis. Il est issu des travaux de recherche datant de la fin des années 1980, de J.W. Smith et *al...* Les auteurs ont formaté et étudié ces données devenues classiques dans le domaine des statistiques. L’article initial, *Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus*, décrit ces données et les résultats de leur étude.

Le jeu de données comprend les résultats d'analyses biologiques et facteurs de prédisposition renseignés dans le cadre d'un suivi médical établi en vue de dépister le diabète des patients. Ces derniers sont des femmes adultes issues de la communauté des indiens Pima. Originaires du Mexique, les Indiens Pimas des États-Unis se sont installés en Arizona, il y a environ trente mille ans et sont restés génétiquement isolés des populations voisines pendant des millénaires. La prévalence du diabète de type 2 est particulièrement élevée au sein de la communauté établie aux Etats-Unis, probablement en raison de facteurs environnementaux (sédentarisation, surconsommation).

Le jeu de données, mis à disposition sur le portail *kaggle*, contient les variables suivantes :

- **npreg** : Nombre de grossesses
- **glu** : Concentration de glucose dans le sang
- **bp** : Pression sanguine (mm Hg)
- **skin** : Epaisseur de la peau (mm)
- **Insuline** : Taux d'insuline présent dans le sang (mu U/ml)
- **bmi** : Indice de masse corporelle (poids en kg/(taille en m)élevée au carré)
- **ped** : Antécédents familiaux de diabète sucrégénétique
- **age** : Age
- **type** : Variable réponse binaire (0/1)

Objectif

L'objectif de l'étude consiste en la modélisation de la pathologie, le diabète de type 2, dit non-insulino dépendant, en fonction des caractéristiques des individus. La variable cible étant qualitative binaire, nous procéderons à une régression logistique, de type binomial, de la variable **type** à partir des variables quantitatives à notre disposition dans le jeu de données.

Exploration des données

La première section de la partie **exploration des données** s'intéresse à la description des données. La deuxième section traite l'analyse exploratoire dans le but de déduire le mécanisme de données manquantes à l'oeuvre. La troisième section s'appuie sur la lecture des graphique des distributions marginales (les densités distinctes d'une variable X en fonction des données manquantes ou non d'une variable Y) pour confirmer ou infirmer nos hypothèses.

Statistiques descriptives

Le jeu comprend 768 observations et 10 colonnes. Chaque observation concerne une femme adulte pour laquelle les résultats d'analyse et facteurs de prédisposition au diabète sont renseignés en colonne.

La variable endogène concerne la pathologie "diabète" caractérisée dans le jeu de données par la variable binaire **type**, qui prend la valeur 0 si l'individu ne souffre pas diabète ou 1, si au contraire celui-ci est touché par la pathologie. Notons que l'échantillon qui comprend 35% d'individus diabétiques est donc relativement équilibré pour la modélisation.

Pour prédire le diabète chez l'individu, 8 variables quantitatives exogènes sont à disposition : **npreg**, **glu**, **skin**, **Insuline**, **bmi**, **ped**, **age**.

##	vars	n	mean	sd	median	trimmed	mad	min	max	range
##	npreg	1	768	3.85	3.37	3.00	3.46	2.97	0.00	17.00
##	glu	2	763	121.69	30.54	117.00	119.66	29.65	44.00	199.00
##	bp	3	733	72.41	12.38	72.00	72.29	11.86	24.00	122.00
##	skin	4	541	29.15	10.48	29.00	28.88	10.38	7.00	99.00

```

## Insuline      5 394 155.55 118.78 125.00 135.32 81.54 14.00 846.00 832.00
## bmi          6 757  32.46   6.92  32.30  32.11  6.82 18.20  67.10  48.90
## ped          7 768   0.47   0.33   0.37   0.42   0.25  0.08   2.42   2.34
## age          8 768  33.24  11.76  29.00  31.54 10.38 21.00  81.00  60.00
## type         9 768   0.35   0.48   0.00   0.31   0.00  0.00   1.00   1.00
##              skew kurtosis    se
## npreg        0.90     0.14 0.12
## glu          0.53    -0.29 1.11
## bp           0.13     0.89 0.46
## skin         0.69     2.88 0.45
## Insuline    2.15     6.23 5.98
## bmi          0.59     0.84 0.25
## ped          1.91     5.53 0.01
## age          1.13     0.62 0.42
## type         0.63    -1.60 0.02

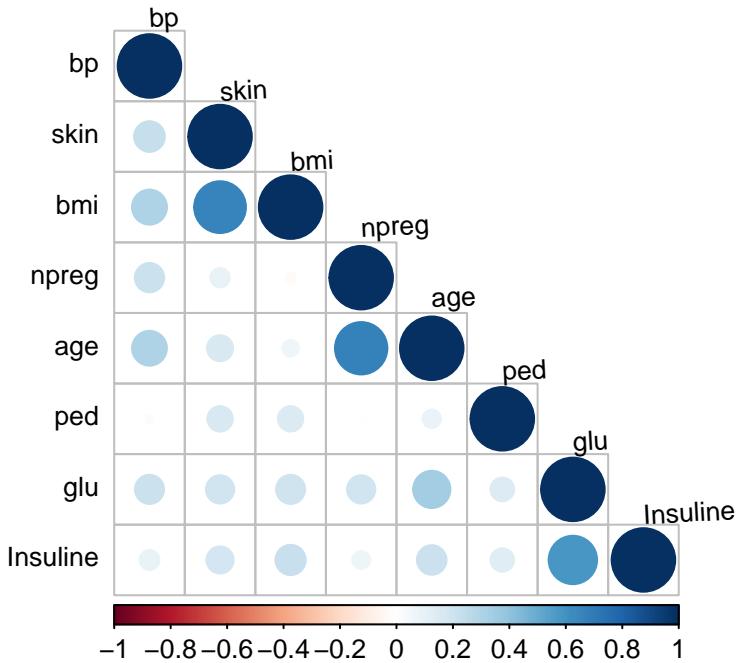
```

Corrélations

Les coefficients de corrélation (Pearson) sont positifs et supérieur à 0,5 :

- entre les variables **skin** et **bmi** (0,65);
- entre les variables **glu** et **Insulin** (0,58);
- entre les variables **npreg** et **age**(0,54).

De manière évidente, l'épaisseur de peau, conséquemment de masse grasse, est fortement liée à l'indice de masse corporelle. De même, la sécrétion d'insuline est directement reliée à la concentration de glucose dans le sang. Enfin, le nombre de grossesses est fonction de l'âge des femmes étudiées. Il faut donc s'attendre à la multicolinéarité dans la régression logistique que nous allons mettre en oeuvre.

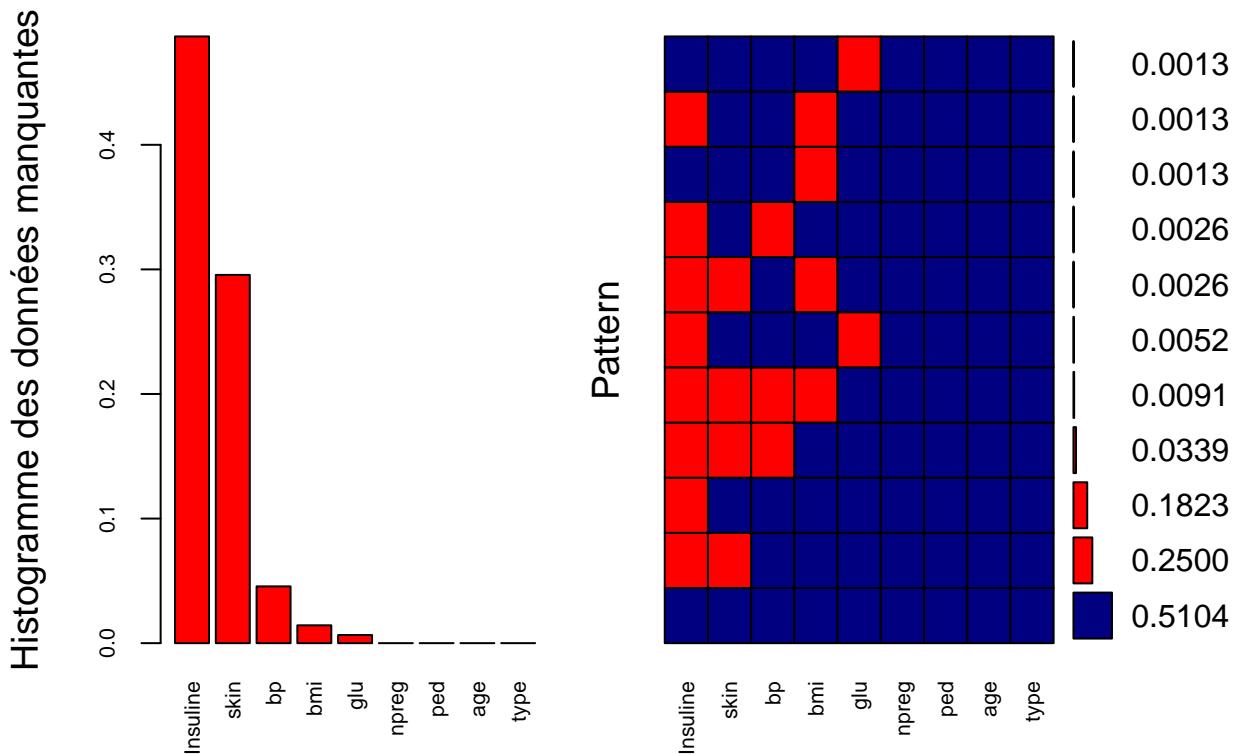


Analyse exploratoire - mécanisme des données manquantes

Le pourcentage de données manquantes est élevé dans le jeu de données, les informations sont exhaustives pour seulement 51% des individus. Le graphique confirme par ailleurs la prédominance de plusieurs combinaisons :

- les valeurs des variables **Insuline** et **skin** sont simultanément manquantes pour 29% des individus;
- les valeurs de la variable **Insuline** sont manquantes pour 18% des individus, toutes les autres variables étant connues;
- les valeurs des variables **Insuline**, **skin** et **bp** sont simultanément manquantes pour 4% des individus.

A elle seule, la variable **Insuline**, quand elle est manquante, regroupe 8 patterns dont 4 concernent également la variable **Skin**.



```

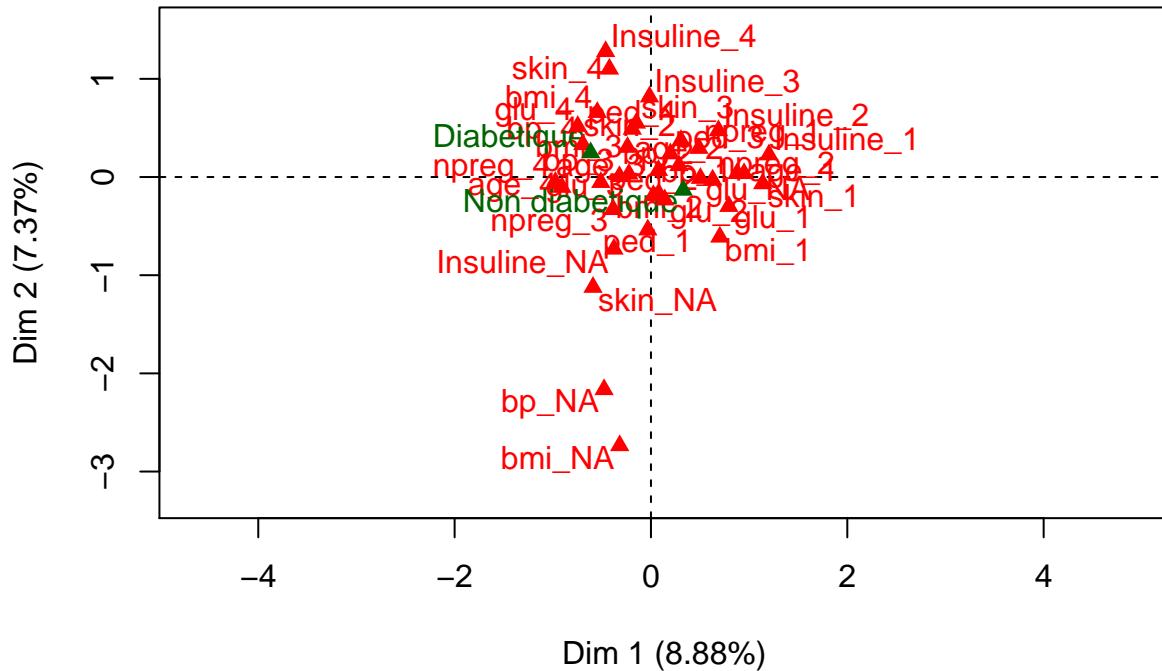
## 
##  Variables sorted by number of missings:
##    Variable      Count
##    Insuline 0.486979167
##    skin 0.295572917
##    bp 0.045572917
##    bmi 0.014322917
##    glu 0.006510417
##    npreg 0.000000000
##    ped 0.000000000
##    age 0.000000000
##    type 0.000000000

```

Analyse des correspondances multiples

L'analyse des correspondances multiples conforte le diagnostic précédent. Les valeurs manquantes des variables **Insuline** et **skin** d'une part, **bp** et **bmi** d'autre part, constituent des catégories liées (les statistiques d'aide à l'interprétation des axes sont en annexe). Ces combinaisons pourraient être davantage associées aux quartiles supérieurs des variables **age** et **npreg**. Nous pourrions ainsi postuler que la collecte des analyses médicales, imparfaite, le soit plus encore pour certains individus, notamment parmi les plus âgés se pliant moins volontiers au protocole de suivi, peut-être longitudinal. Pour la représentation, a été ajoutée comme variable qualitative supplémentaire la variable réponse. Nous pouvons d'ores et déjà observer que la pathologie semble toucher les individus dont les valeurs des variables **Insuline**, **ped**, **bmi**, **skin** et **glu** appartiennent aux derniers quartiles de leur distribution.

MCA factor map

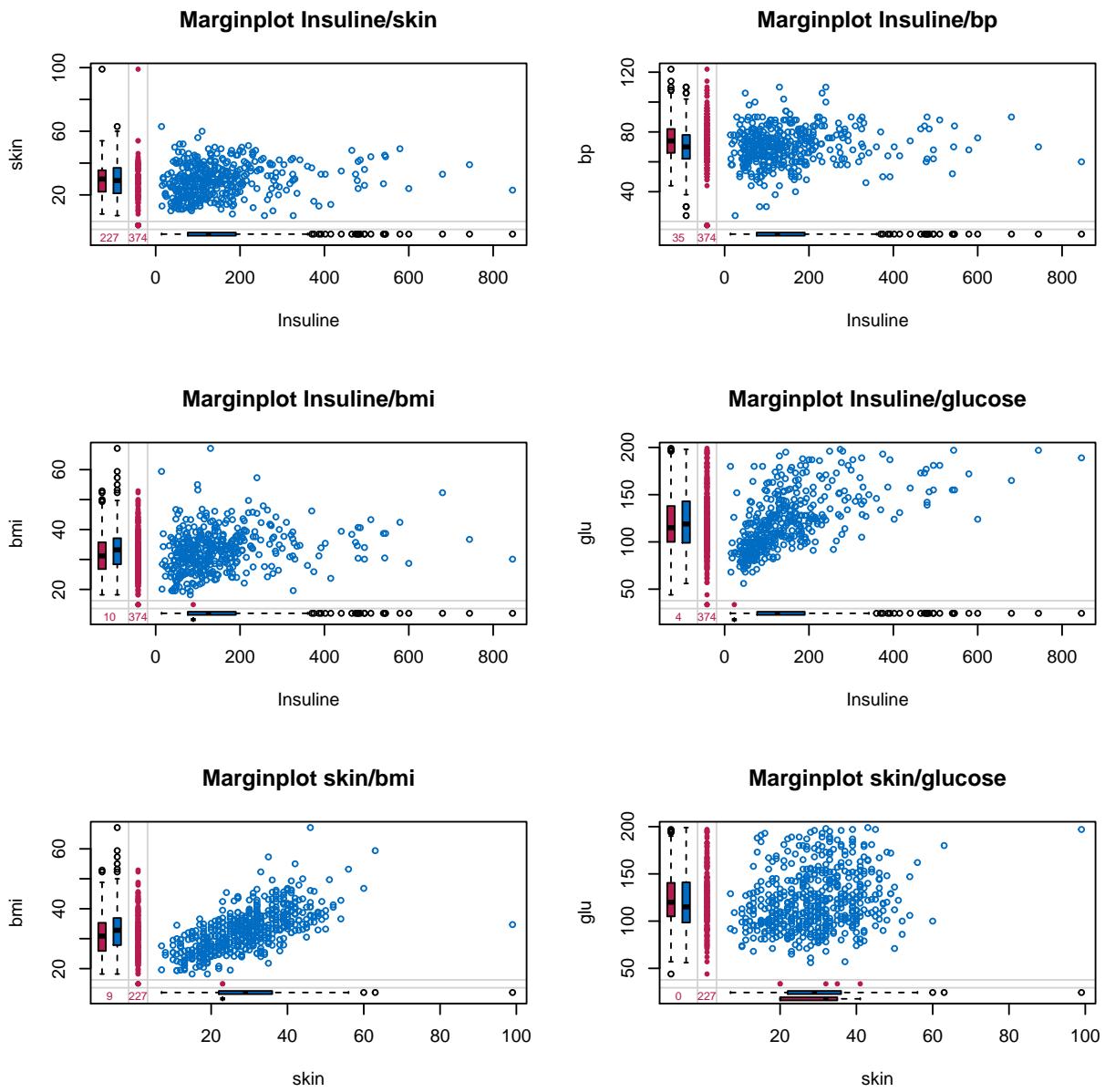


Distributions marginales des variables

Nous procémons dans cette section à l'exploration des distributions croisées entre 2 variables en fonction de la disponibilité ou non des valeurs pour celles-ci. Le but est de comparer les distributions, figurées par des "boîtes à moustache", d'une variable X en fonction de l'absence ou la disponibilité des valeurs pour une variable Y, et inversement. Nous commençons l'analyse en nous intéressant aux variables présentant les parts de données manquantes les plus élevées, **Insuline** (48,7%) et **skin** (29,6%) et, ce faisant, en comparant leurs distributions en fonction de la disponibilité des valeurs pour les variables **glu** (taux de glucose), **bp** (pression sanguine) et **bmi** (indice de masse corporelle). Nous nous intéressons ensuite aux variables **age** et **npreg** en fonction de la disponibilité des valeurs des variables **Insuline** et **skin**, et ce dans le but de confirmer ou d'infirmer l'hypothèse émise à l'issue de l'analyse en correspondances multiples.

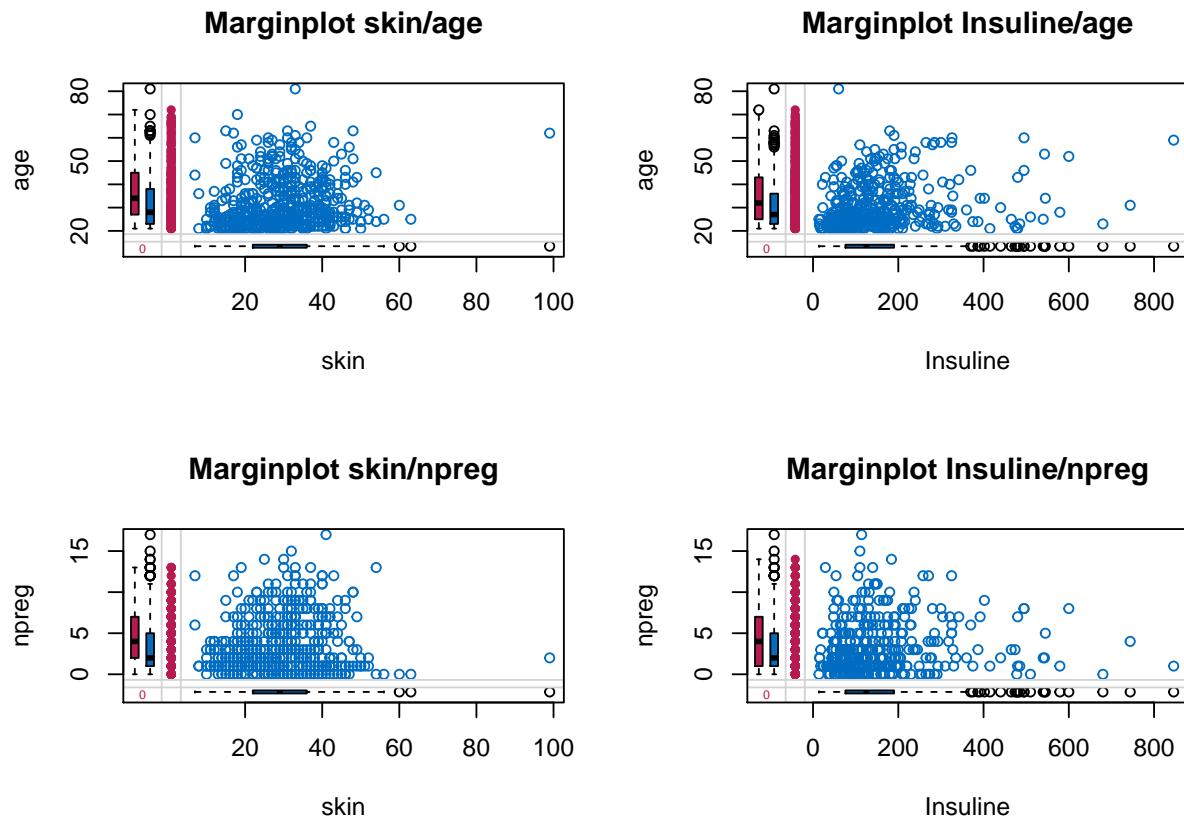
i) Distributions croisées variables **Insuline**, **skin**, **glu**, **bp**, **bmi**

Les graphiques suivants illustrent la relative similarité des distributions des variables **glu**, **bp**, **bmi** que les valeurs de la variable **Insuline** ou **skin** soient renseignées ou manquantes. Nous pouvons noter en outre qu'en cas de données manquantes pour **Insuline**, la valeur des variables **Skin** et **bp** sont elles aussi absentes. Comme indiqué plus haut, nous pouvons postuler que certains relevés d'analyses sont plus complexes et requièrent un personnel spécialisé. Cette automatique du lien entre valeurs manquantes n'est en revanche pas vérifié pour les variables **glu** et **bmi**, bien que celles-ci ne contiennent qu'une seule valeur disponible concomitamment à une valeur manquante pour la variable **Insuline**. L'examen des différents axes des abscisses plaide pour l'absence de dispositif particulier, soit des modèles MCAR pour les différentes variables étudiées ici.



i) Distributions croisées avec les variables `age` et `npreg`

Les graphiques ci-après s'intéressent aux distributions des variables `age` et `npreg` en fonction de la disponibilité des valeurs pour les variables **Insuline** et **skin**. Comme indiqué plus haut, nous pouvons postuler un pattern de type "MAR" entre ces variables. En effet, la différence de distribution, figurée par les 2 boîtes à moustache, rouge et bleue, est patente pour les variables `age` et `npreg` en fonction de l'absence ou de la disponibilité des valeurs pour les variables. Nous avons déjà avancé l'hypothèse d'un lien pour ces deux variables induits par l'absence de collecte pour certains individus. Nous pouvons ajouter, à la suite de ce qui a été postulé, que ces difficultés dans la collecte des analyses concerneraient en premier chef les individus parmi les plus âgés au sein de la population de l'échantillon, le lien avec le nombre de grossesses provenant quant à lui de la corrélation avec la variable `age` identifiée ci-avant.



Au final, les boîtes à moustache rouges, celles figurant les distributions des données manquantes, sont globalement proches des boîtes à moustache bleues, celles figurant les distributions des données observées. Cela traduirait des dispositifs MCAR pour les variables `glu`, `bp`, `bmi`, `Insuline` et `skin`. S'agissant de ces deux dernières variables, les différentes analyses effectuées au paragraphe précédent semblent corroborer la coexistence de dispositif MAR avec la variable `age`, et dans une moindre mesure à la variable `npreg`. Ce lien avait déjà été mis en évidence lors de l'analyse des correspondances multiples.

Imputations multiples

Dans cette partie, nous recourons à plusieurs méthodes d'imputation multiple avant d'effectuer les régressions. La première section concernera la mise en oeuvre des deux méthodes d'imputation multiple fournies avec la librairie `Mice`. Dans la deuxième section, seront “implémentées” les méthodes d'imputation du package `missMDA`. Sera discutée dans une troisième section la qualité des imputations avec ces différentes méthodes.

Deux essais d'imputation “simple” est fourni en annexe pour illustration.

Imputations multiples avec les méthodes JM et FCS du package MICE

Nous mettons en oeuvre 2 méthodes adaptées à l'imputation de variables de type numérique :

- la méthode Joint Modeling (JM),
- la méthode Fully Conditionnal Specification (FCS).

i) Imputation multiple - méthode Joint Modeling

En annexe sont détaillées les analyses nous permettant de déterminer le seuil minimal d'imputations visant la convergence de l'algorithme. Ici, le seuil retenu est de 100 jeux de données.

```
imp.mi.jm <- mice(data.miss, m=100, method="norm", seed=111119, print=F)
```

i) Imputation multiple - méthode Fully Conditionnal Specification

Pour la méthode FCS, le nombre d'imputations est identique (100 jeux de données).

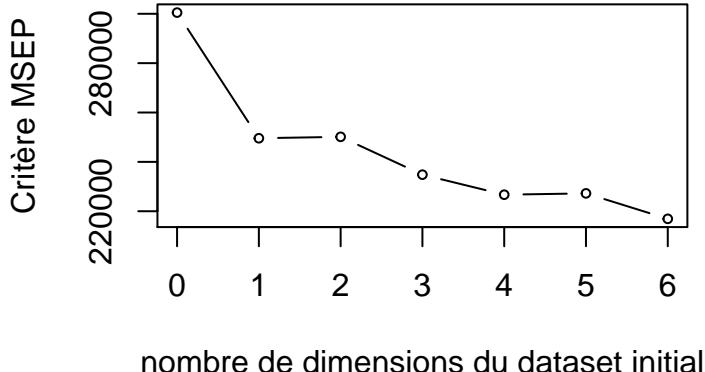
```
imp.mi.fcs <- mice(data.miss, m=100, seed=111119, print = F)
```

Imputations multiples avec la méthode MIPCA du package missMDA

i) Choix du nombre de dimensions

La fonction MIPCA du package missMDA s'appuie sur une imputation multiple par réduction des dimensions au moyen d'un analyse en composantes principales. Nous procérons au cours d'une première étape à l'estimation du nombre de dimensions à l'imputation avec la méthode de validation croisée "kfold". Selon le critère "MSEP", le nombre de dimensions à retenir de 6.

```
par(mfrow=c(1,1))
## 1ère étape : le nombre de dimensions axes à choisir
nb.kfold<- estim_ncpPCA(data.miss, ncp.min = 0, ncp.max = 6,
                           method.cv = "Kfold", nbsim = 100, verbose = F)
plot(names(nb.kfold$criterion), nb.kfold$criterion,type="b", ylab="Critère MSEP",
      xlab="nombre de dimensions du dataset initial", cex=.6)
```



```
#Choix du nombre de dimensions
ncp.res <-nb.kfold$ncp
```

i) Imputation multiple - MIPCA

Les imputations avec la fonction MIPCA sont opérées avec en paramètre d'entrée le nombre de dimension calculée à l'étape précédente.

```
## Multiple Imputation method "bayes"
imp.mipca <- MIPCA(data.miss, ncp=ncp.res, verbose=F, method.mi="Bayes")
```

Validation de la qualité des imputations

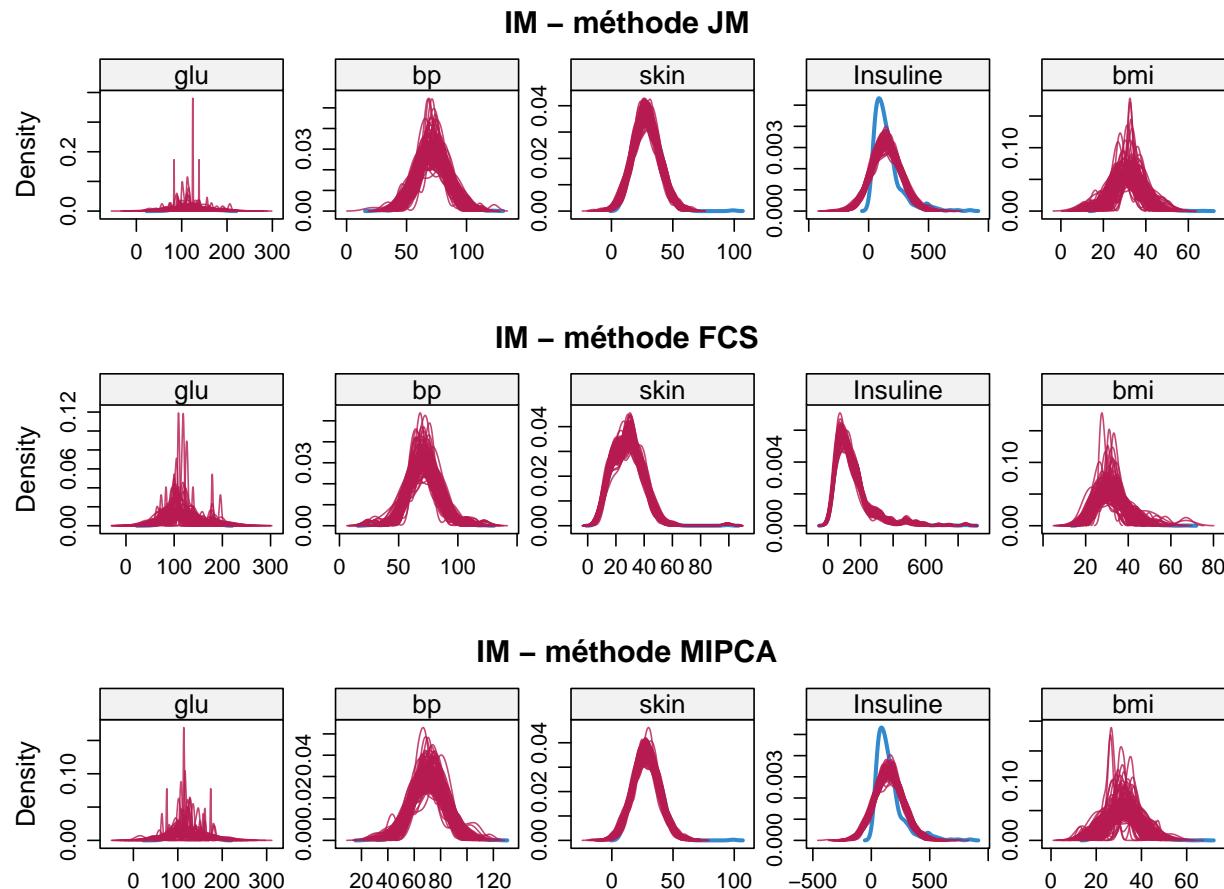
Il s'agit ici de vérifier l'adéquation des données imputées au modèle et hypothèses émises de deux manières :

- une première validation consiste en l'examen comparé des densités des valeurs imputées et observées pour les variables **Insuline**, **skin**, **bp**, **bmi** et **glu**;
- l'autre méthode dite de “sur imputation” consiste à retirer une valeur observée pour procéder ensuite à l'imputation de la valeur devenue manquante dans le jeu de données. Cette étape est reconduite pour toutes les valeurs observées des variables manquantes pour comparer ensuite la qualité des imputations.

Analyse des distributions

Avec les méthodes JM (MICE) et MIPCA (missMDA), les distributions des valeurs imputées de la variable **Insuline** ont des profils assez divergents de la densité des valeurs observées. Les moyennes et les variances des valeurs imputées sont plus élevées. Cela pourrait se justifier dans le cas du dispositif “MAR” présumé pour cette variable. En revanche, une partie des valeurs imputées sont négatives et sont donc aberrantes.

Avec la méthode FCS, les densités des variables imputées semblent en revanche en adéquation avec les densités des variables observées.

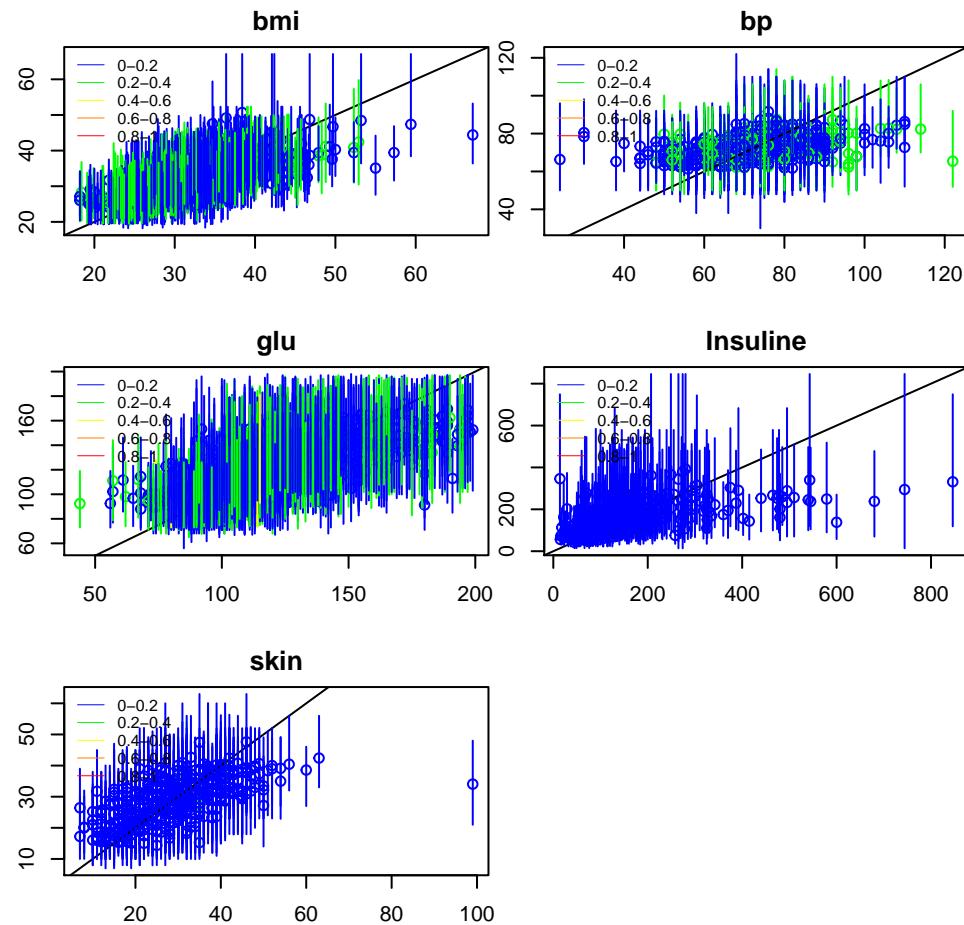


“Overimputation”

Comme explicité plus haut, il s’agit ici d’imputer, pour la variables avec données manquantes, chaque valeurs observée avec la méthode retenue et comparer ensuite les valeurs obtenues aux valeurs observées. La bissectrice correspond à une prédiction parfaite des valeurs obervées (en abscisse) et prédictes (en ordonnées). La qualité de l’imputation est en outre mesurée au moyen d’intervalles de confiance figurés pour chaque valeur observée par les segments verticaux du graphique.

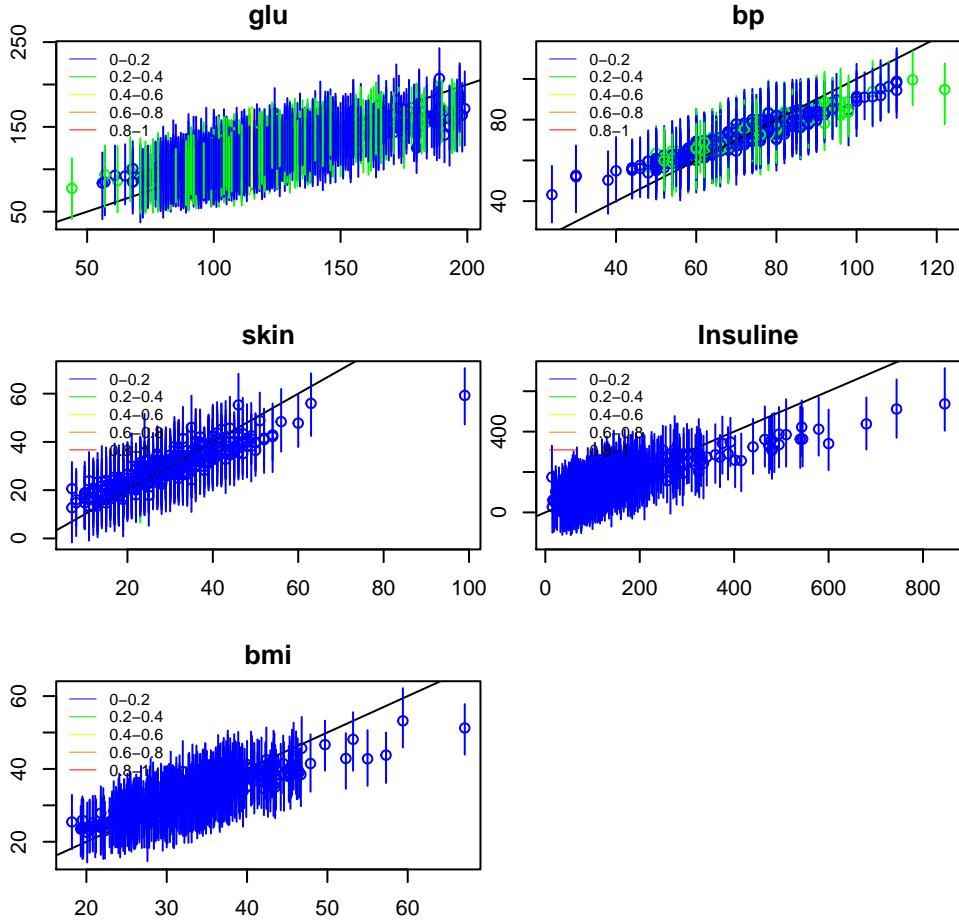
L’overimputation de la fonction mice semble montrer une imputation biaisée pour les plus fortes valeurs des variables **Insuline** et **skin**.

i) FCS - mice



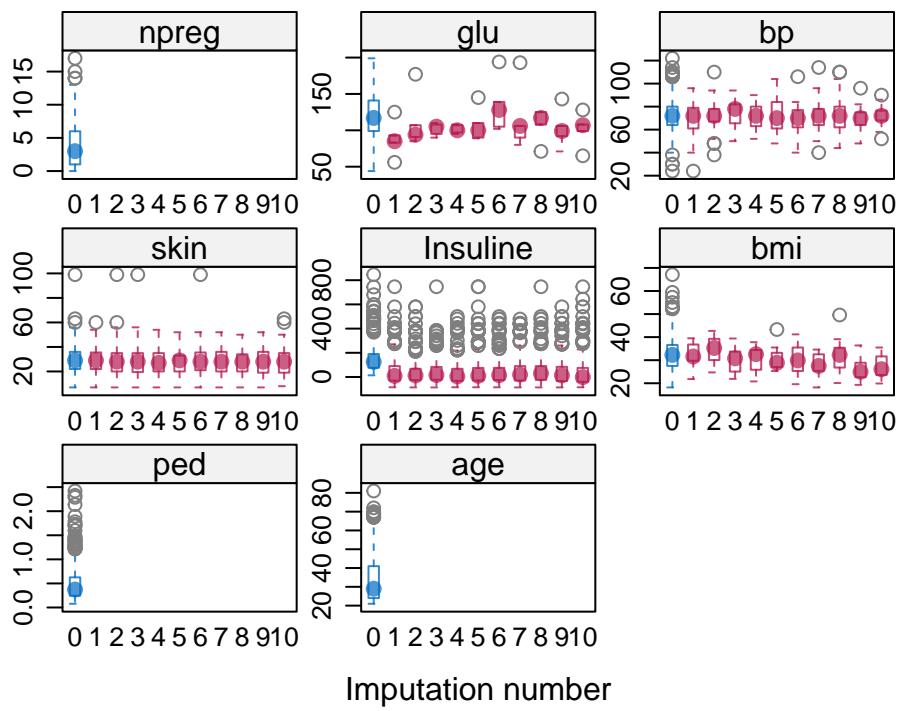
Au regard des graphiques, l’imputation la méthode Bayes de la librairie missMDA semble plus satisfaisante, notamment pour la variable **skin**. Les imputations de variable **Insuline** présentent à nouveau un biais pour les valeurs les plus élevées, bien que celui-ci soit moins fort que dans le cas précédent.

ii) Méthode bayes - missMDA

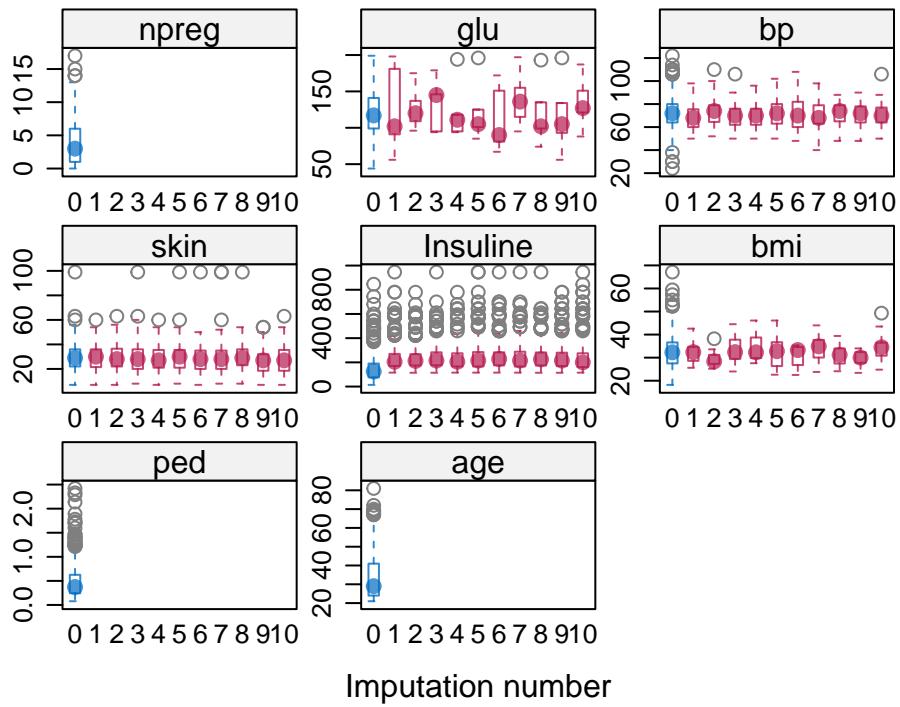


Analyse de sensibilité du modèle

L'analyse de sensibilité consiste à tester la robustesse du modèle d'imputation en ajoutant une perturbation à l'une des variables explicatives, l'imputation s'appliquant à la volée. L'objectif est de vérifier les hypothèses émises quant aux dispositifs de données manquantes, en l'espèce les modèles MCAR conjecturés plus haut entre variables avec données manquantes. Ici, les imputations sont visiblement robustes à une modification des valeurs des variables **Insuline** avec une perturbation, tour à tour, négative, puis positive d'une valeur de 100 (valeur approximative de l'écart type des données observées).

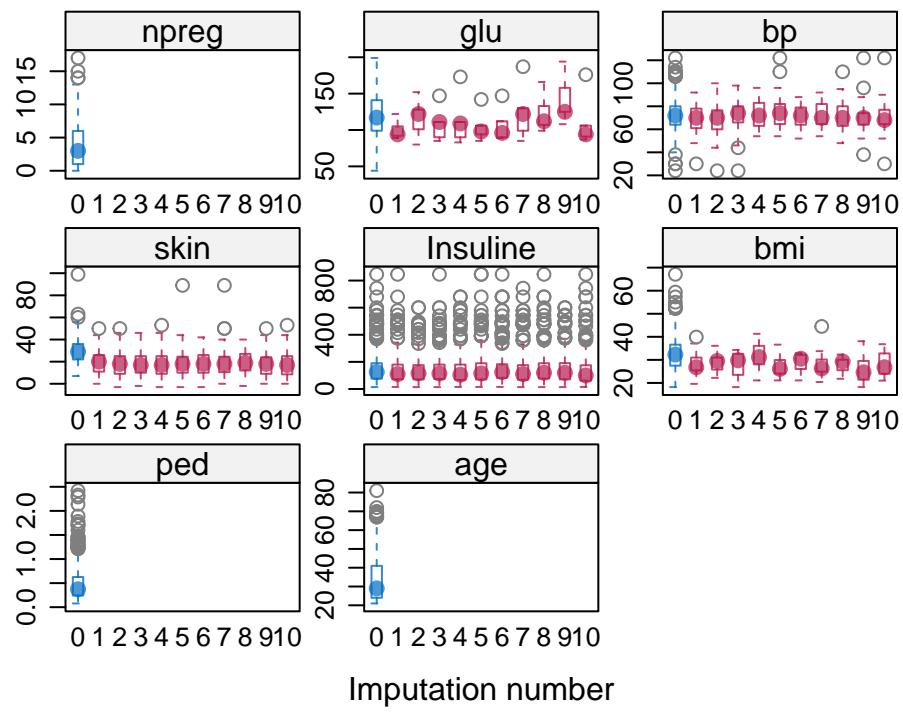


Les graphiques ci-dessus montrent qu'en retranchant 100 à chaque valeur imputée de la variable Insulin, les autres imputations ne semblent pas modifiées pour autant.

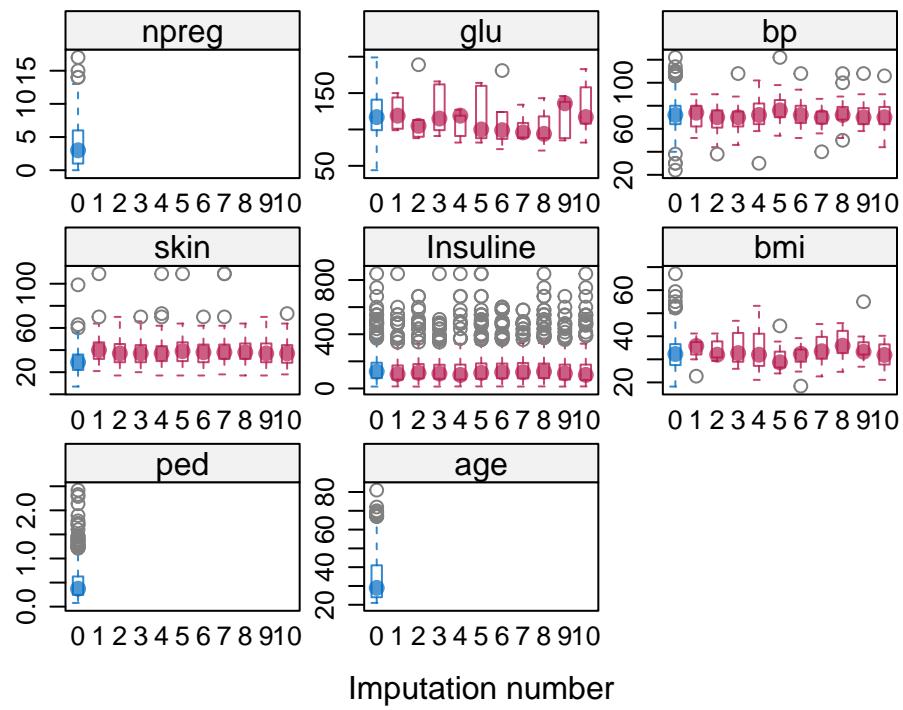


A nouveau, les graphiques ci-dessus montrent qu'en augmentant de 100 chaque valeur imputée, l'impact sur les imputations des autres variables semblent nulles.

Nous observons maintenant la robustesse de l'imputation à une modification des valeurs des variables `skin` avec cette fois-ci une perturbation d'une valeur de 10 (écart type des données observées).



A nouveau, il semble que les perturbations introduites pour la variable `skin` n'affectent pas les autres imputations.



Modélisation - Régression logistique

Listwise deletion

Nous mettons en oeuvre au cours de cette étape une régression logistique, avec suppression des observations comportant des données manquantes, ce qui correspond à la méthode par défaut de la régression logistique (méthode dite des cas concrets ou “listwise deletion”). Au seuil de 5%, seules les coefficients des variables `glu`, `bmi` et `ped`, sont significatifs. Au seuil de 10%, la variable `age` est également significative.

La régression logistique nous donne les estimateurs suivants:

```
##  
## Call:  
## glm(formula = type ~ ., family = binomial(link = "logit"), data = data.miss)  
##  
## Deviance Residuals:  
##      Min       1Q   Median       3Q      Max  
## -2.7823  -0.6603  -0.3642   0.6409   2.5612  
##  
## Coefficients:  
##             Estimate Std. Error z value Pr(>|z|)  
## (Intercept) -1.004e+01  1.218e+00 -8.246 < 2e-16 ***  
## npreg        8.216e-02  5.543e-02  1.482  0.13825  
## glu         3.827e-02  5.768e-03  6.635 3.24e-11 ***  
## bp          -1.420e-03  1.183e-02 -0.120  0.90446  
## skin        1.122e-02  1.708e-02  0.657  0.51128  
## Insuline    -8.253e-04  1.306e-03 -0.632  0.52757  
## bmi         7.054e-02  2.734e-02  2.580  0.00989 **  
## ped         1.141e+00  4.274e-01  2.669  0.00760 **  
## age         3.395e-02  1.838e-02  1.847  0.06474 .  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
## Null deviance: 498.10  on 391  degrees of freedom  
## Residual deviance: 344.02  on 383  degrees of freedom  
##  (376 observations deleted due to missingness)  
## AIC: 362.02  
##  
## Number of Fisher Scoring iterations: 5
```

Mise en oeuvre avec les données imputées

i) MICE - Méthode FCS

Comme attendu au vu des dispositifs MCAR conjecturés plus haut entre variables avec données manquantes, les estimateurs des coefficients de la régression sont proches de ceux obtenus par régression avec la méthode des cas concrets et surtout, le sens des relations est préservé (négatifs pour les variables `bp` et `Insuline`). Nous remarquons par ailleurs que l'hypothèse MAR entre les variables `Insuline` et `skin` d'une part, `age` et `npreg` d'autre part, semble confortée à l'issue de la régression : les plus fortes variations en valeurs absolue des coefficients concernent ces variables. Avec les imputations des variables `Insuline` et `skin`, les facteurs `age` et `npreg` sont devenus moins influents sur la prédiction de la pathologie car probablement liés par un dispositif MAR à `Insuline` et `skin`.

```

##           estimate   std.error   statistic      df   p.value
## (Intercept) -9.2086572638 0.853109344 -10.7942286 693.1621 0.000000e+00
## npreg         0.1216436843 0.032719907  3.7177270 747.5964 2.160431e-04
## glu          0.0386630008 0.004751491  8.1370257 460.3735 3.774758e-15
## bp            -0.0093517896 0.008691120 -1.0760166 703.0560 2.822889e-01
## skin          0.0061127999 0.014450455  0.4230178 328.8118 6.725588e-01
## Insuline     -0.0009845081 0.001392793 -0.7068587 191.6328 4.805129e-01
## bmi          0.0908537893 0.021103634  4.3051253 544.8683 1.979227e-05
## ped           0.8750919350 0.298909413  2.9276158 750.7090 3.519007e-03
## age          0.0137964406 0.009776780  1.4111437 733.6865 1.586261e-01

```

i) missMDA - Méthode bayes

Les estimateurs des coefficients de la régression mis en oeuvre sur les jeux de données imputées avec la méthode MIPCA sont très proches de ceux calculés auparavant. A nouveau, nous sommes confortés dans nos hypothèses par la convergence des résultats des deux estimations.

```

##           estimate   std.error   statistic      df   p.value
## (Intercept) -9.0985129278 0.839858264 -10.8333910 709.8511 0.000000e+00
## npreg         0.1214328916 0.032718754  3.7114155 749.6304 2.213595e-04
## glu          0.0373612599 0.004448996  8.3976837 572.9198 4.440892e-16
## bp            -0.0086069827 0.008620462 -0.9984363 720.0231 3.184033e-01
## skin          0.0104329248 0.013711571  0.7608847 384.3380 4.471925e-01
## Insuline     -0.0004133463 0.001246719 -0.3315472 245.9021 7.405136e-01
## bmi          0.0849034450 0.020287806  4.1849496 569.4388 3.303349e-05
## ped           0.8612289287 0.298072436  2.8893276 753.3926 3.971505e-03
## age          0.0133021608 0.009745180  1.3649989 735.4218 1.726708e-01

```

Conclusion

Partant de l'analyse des données et d'hypothèses sur les dispositifs des données manquantes, nous avons mis en oeuvre trois méthodes d'imputation multiples (Joint Modeling, Fully Conditionnal Specification avec mice et bayes avec missMDA). Après une première validation, les 2 méthodes retenues (Fully Conditionnal Specification avec mice et bayes avec missMDA) ont servi à modéliser la variable réponse "diabète". La convergence des résultats obtenus avec ces 2 méthodes se conjugue à une diminution importante de la dispersion des estimateurs. Les sorties, densité et stripplot, montrent cependant des imputations davantage en adéquation des données observées pour la méthode FCS du package mice que nous préférerions ici. A partir des jeux de données imputées avec cette méthode, nous avons donc conservé en annexe une étape de choix de modèle en recourant à la fonction pool.compare disponible dans la librairie mice. Au final, le modèle sélectionné pour le jeu de donnée Pima complété avec la méthode FCS du package mice serait le suivant :

- bp
- skin
- Insuline
- ped
- age

ANNEXE

ACM - aide à l'interprétation

Comme indiqué plus haut, le 1er axe oppose les 1ers quartiles des variables `age`, `bp`, `bmi`, `Insuline`, `glu` aux derniers quartiles de ces mêmes variables. Le 2e axe oppose quant à lui les données manquantes des variables `skin`, `Insuline`, `bp` et `glu` aux derniers quartiles de celles-ci.

```
##           Dim 1      Dim 2
## npreg_1    2.976678752 1.287719e+00
## npreg_2    3.555211267 9.617845e-03
## npreg_3    1.320248123 1.162086e+00
## npreg_4    8.107735121 2.940545e-02
## glu_1      6.113929291 1.042705e+00
## glu_2      0.193629966 6.365538e-01
## glu_3      0.484214082 1.090445e-02
## glu_4      5.260890885 3.038805e+00
## glu_NA     0.064128908 4.912006e-05
## bp_1       7.826151478 1.671527e-02
## bp_2       0.053625303 5.173938e-02
## bp_3       0.959797695 3.967660e-04
## bp_4       4.017515891 1.107315e+00
## bp_NA      0.403623129 9.988081e+00
## skin_1     9.834298653 4.134077e-02
## skin_2     0.595321518 1.069025e+00
## skin_3     0.149145248 2.471205e+00
## skin_4     1.186292464 9.662344e+00
## skin_NA    3.999681239 1.742392e+01
## Insuline_1 7.236251399 3.106887e-01
## Insuline_2 2.368248354 1.349721e+00
## Insuline_3 0.001109442 4.024570e+00
## Insuline_4 1.044546763 9.666726e+00
## Insuline_NA 2.683510986 1.225034e+01
## bmi_1      4.806765252 4.384369e+00
## bmi_2      0.069284847 3.616544e-01
## bmi_3      0.545116554 1.061349e+00
## bmi_4      2.843338826 4.972664e+00
## bmi_NA     0.056339505 5.012836e+00
## ped_1      0.010052366 3.384407e+00
## ped_2      0.005719703 4.108318e-01
## ped_3      0.380560649 6.584077e-01
## ped_4      0.350775707 2.786276e+00
## age_1      9.969485336 1.749011e-02
## age_2      0.727165108 1.488025e-01
## age_3      2.640331935 3.591831e-02
## age_4      7.159278255 1.130164e-01
```

Imputation simple

Dans un premier temps on traite les valeurs manquantes par imputation simple avec le package MICE :

- par le biais de la méthode PMM (predictive mean matching),

- puis au moyen d'une régression linéaire - non bayésienne - stochastique.

Ces imputations n'ont pas été conservées en raison de la sensibilité à la spécification du modèle s'agissant de la méthode paramétrique (régression linéaire - non bayésienne - stochastique) et du biais souvent généré par la méthode semi paramétrique (predictive mean matching).

```
#Predictive mean matching
imp.si.pmm <- mice(data.miss, m=1, seed = 111119, print = F)

#régression stochastique avec bootstrap
imp.si.norm <- mice(data.miss, method = "norm.nob", m = 1, maxit = 1, print = F)
```

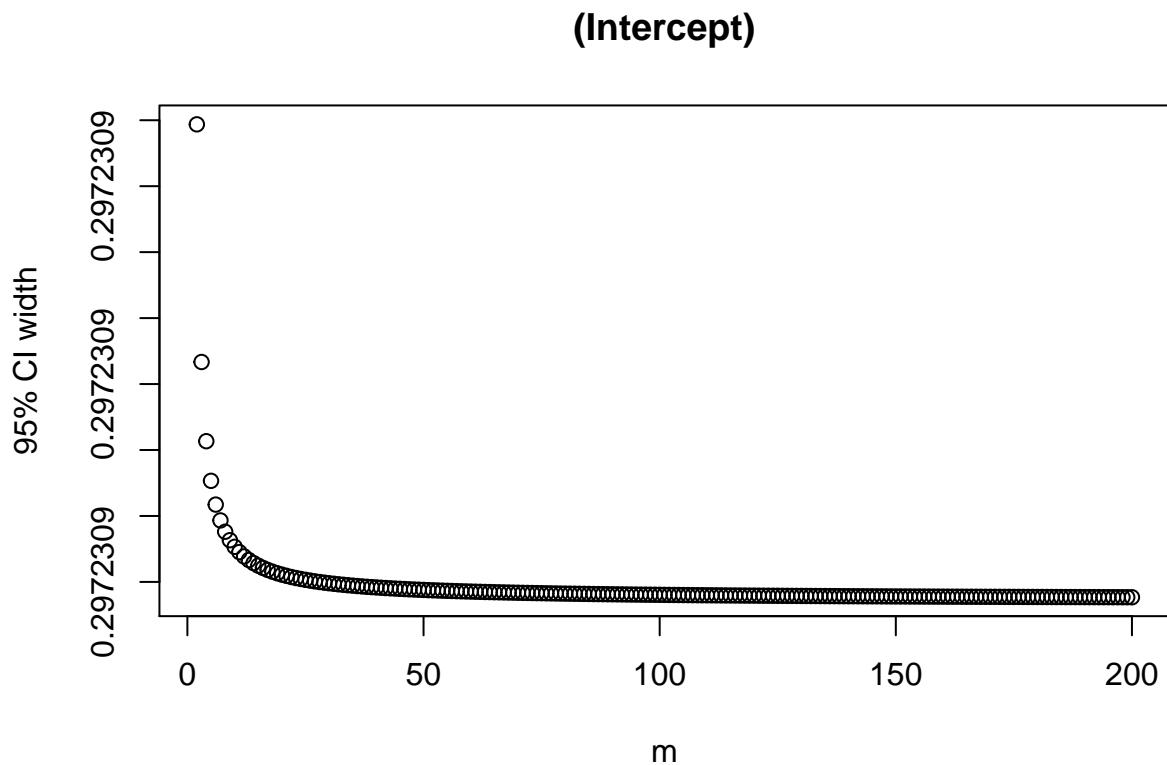
Calcul du nombre d'imputations pour les fonctions MICE

Les graphiques nous permettent de déduire le seuil d'imputations à prévoir pour s'assurer de la convergence des estimateurs mais éviter dans le même temps un temps de computation machine trop grand.

MICE JM

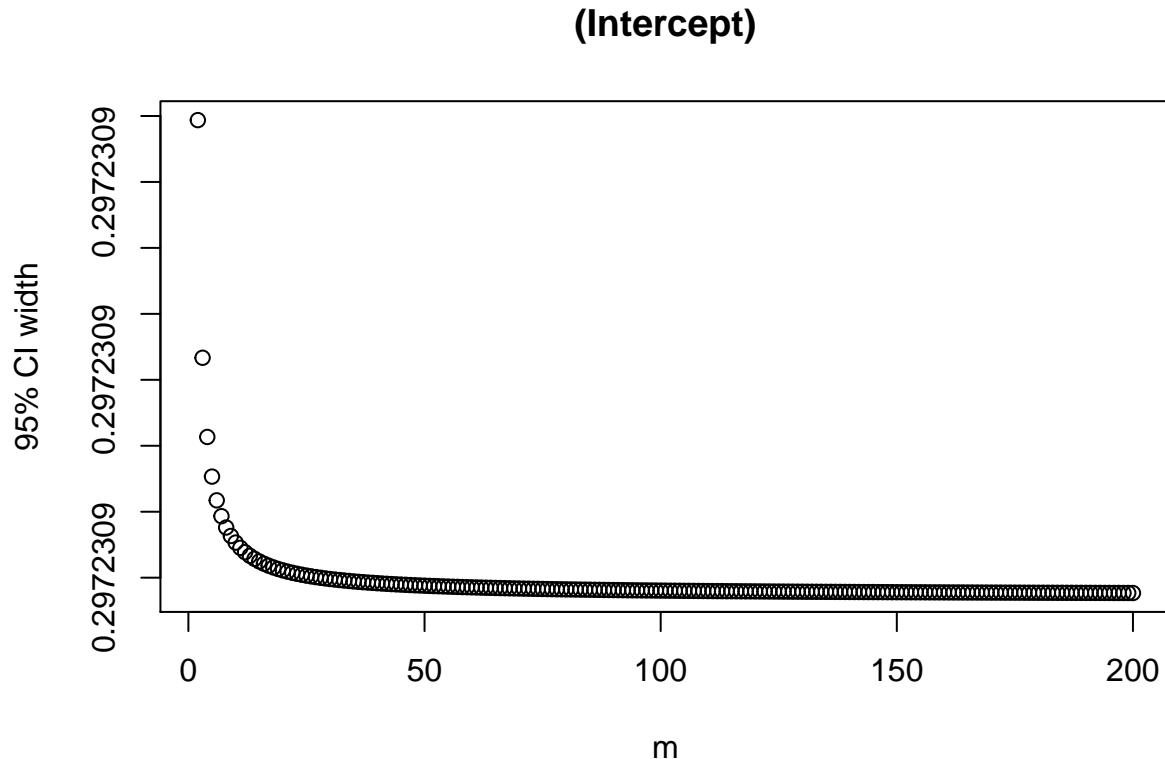
La convergence semble être atteinte entre 50 et 100 imputations.

```
mice.jm.conv <- mice(data.miss, m=200,method='norm',seed=221219, print=F)
res.mice.jm.conv<-with(mice.jm.conv,glm(type~1, data=data.miss, family = "binomial"))
plot(res.mice.jm.conv)
```



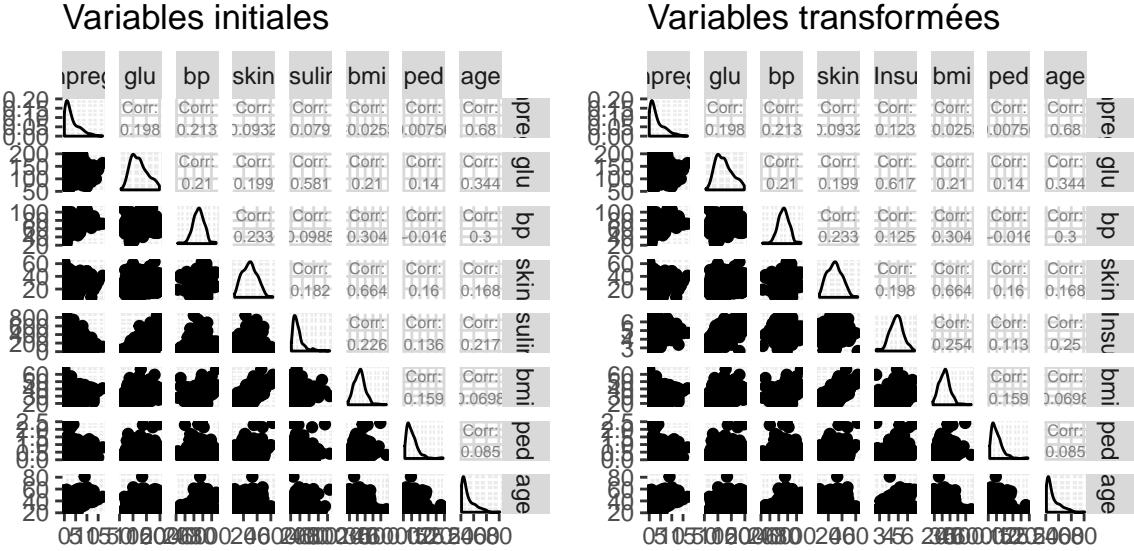
MICE FCS

```
imp.mice.fcs.conv <- mice(data.miss, m=200, seed=221219, print=F)
res.imp.mice.fcs.conv<-with(imp.mice.fcs.conv,glm(type~1,data=data.miss,family = "binomial"))
plot(res.imp.mice.fcs.conv)
```



Transformation des données

Préalablement aux étapes d'imputation et de modélisation, nous avons procédé à une transformation des variables visant à renforcer la linéarité des liens entre elles. *In fine*, après plusieurs essais (logarithme, logistic, racine carrée), seule la transformation logarithmique de la variable insuline est conservée. Les différences de résultats, notamment les imputations multiples, étant complexes à interpréter, nous avons privilégié le jeu de donnée non transformé pour le projet. Les résultats sont en annexe pour rendre compte de l'ensemble de notre démarche.



Graphiques “stripplot” - qualité des imputations multiples

Les graphiques “stripplot” qui consomment trop de taille mémoire et disque ne sont donc pas intégrés au rapport. Mais les sorties Les sorties montrent néanmoins que les imputations figurent bien dans les plages de valeurs attendues. Avec la méthode FCS de mice, les imputations semblent toutefois plus en adéquation des données observées.

```
#Imputations multiples
#FCS
#stripplot(imp.mi.fcs, pch = 20, cex = 1.2)
#stripplot(imp.mipca.prelim, pch = 20, cex = 1.2)
```

Choix de modèle

Pour spécifier notre modèle, nous utilisons ici les possibilités offertes par la fonction pool.compare du package mice en nous référant à la documentation disponible sur la page <https://stefvanbuuren.name/mice/reference/pool.html>.

Nous acceptons au cours de cette 1ère étape l'utilité de la variable **Insuline** qui comportait beaucoup de données manquantes (52%).

```
#Test des rapports de vraisemblance
diabete.insuline.glm <- with(data=imp.mi.fcs ,
                                exp=glm(type ~ npreg + glu + bp + skin +
                                         Insuline + bmi + ped + age
                                         ,family=binomial(link="logit")))

diabete.noinsuline.glm <- with(data=imp.mi.fcs ,exp=glm(type ~ npreg + glu + bp + skin
                                         + bmi + ped + age
                                         ,family=binomial(link="logit")))

pool.compare(diabete.insuline.glm, diabete.noinsuline.glm, method = "likelihood")$pvalue
```

Au regard des différentes comparaisons et du test final, nous conservons avec la méthode FCS du package mice les modèles excluant les variables :

```
+bmi +glu +npreg

#age
diabete.insuline.noage.glm1 <- with(data=imp.mi.fcs ,
                                         exp=glm(type ~ npreg + glu + bp + skin +
                                                 Insuline + bmi + ped,
                                                 family=binomial(link="logit")))
stat_likelihood.glm1.age<- pool.compare(diabete.insuline.glm,
                                         diabete.insuline.noage.glm1, method = "likelihood")

#skin
diabete.insuline.noskin.glm1 <- with(data=imp.mi.fcs ,
                                         exp=glm(type ~ npreg + glu + bp +
                                                 Insuline + bmi + ped + age,
                                                 family=binomial(link="logit")))
stat_likelihood.glm1.skin<- pool.compare(diabete.insuline.glm,
                                         diabete.insuline.noskin.glm1, method = "likelihood")

#bp
diabete.insuline.nobp.glm1 <- with(data=imp.mi.fcs ,
                                         exp=glm(type ~ npreg + glu + skin +
                                                 Insuline + bmi + ped + age,
                                                 family=binomial(link="logit")))
stat_likelihood.glm1.bp<- pool.compare(diabete.insuline.glm,
                                         diabete.insuline.nobp.glm1, method = "likelihood")

#bmi
diabete.insuline.nobmi.glm1 <- with(data=imp.mi.fcs ,
                                         exp=glm(type ~ npreg + glu + bp + skin +
                                                 Insuline + ped + age,
                                                 family=binomial(link="logit")))
stat_likelihood.glm1.bmi<- pool.compare(diabete.insuline.glm,
                                         diabete.insuline.nobmi.glm1, method = "likelihood")

#glu
diabete.insuline.noglu.glm1 <- with(data=imp.mi.fcs ,
                                         exp=glm(type ~ npreg + bp + skin +
                                                 Insuline + bmi + ped + age,
                                                 family=binomial(link="logit")))
stat_likelihood.glm1.glu<- pool.compare(diabete.insuline.glm,
                                         diabete.insuline.noglu.glm1, method = "likelihood")

#npreg
diabete.insuline.nopreg.glm1 <- with(data=imp.mi.fcs ,
                                         exp=glm(type ~ bp + skin + glu+ Insuline
                                                 + bmi + ped + age,
                                                 family=binomial(link="logit")))
stat_likelihood.glm1.npreg<- pool.compare(diabete.insuline.glm,
                                         diabete.insuline.nopreg.glm1, method = "likelihood")

#Au final, on teste le modèle sans les variables bmi, glu et preg

diabete.insuline.glm3<- with(data=imp.mi.fcs ,
                                         exp=glm(type ~ bp + skin + Insuline + ped + age,
                                                 family=binomial(link="logit")))

stat_likelihood.glm1.final<- pool.compare(diabete.insuline.glm,
```

```
diabete.insuline.glm3, method = "likelihood")
##### SORTIE MODELE FINAL SELECTIONNE#####
summary(pool(diabete.insuline.glm3))
```