Jeremy Qin, Edouard Cayouette et Francisco Lopez

STT3510 - H23

Projet Biostatistique

Le projet constitue à établir la relation entre la probabilité qu'une mère donne naissance à un bébé de faible poids et les variables prédictives. Cette étude a été faite sur des données comportant 189 observations provenant d'un centre médical d'Amérique du Nord au cours des années 1980.

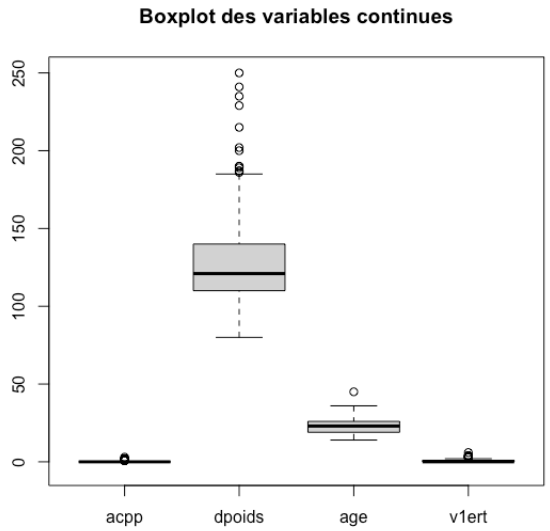
Un nouveau né est dit de faible poids s'il ne pèse pas plus de 2,5 kg. Nous avons aussi accès à plusieurs variables prédictives de la mère, et nous donnerons une description de celles-ci. Tout d'abord, les variables prédictives sont: le nombre d'accouchements prématurés précédents (**acpp**), le poids de la mère en livres lors de la dernière période menstruelle (**dpoids**), la présence d'irritabilité utérine (**iuter**) , les antécédents d'hypertension (**htension**), le status tabagique pendant la grossesse (**fumeur**), l'âge de la mère (**age**), la race de la mère (**race**) et le nombre de visites chez le médecin au cours du premier trimestre (**v1ert**).

Ayant ces informations en main, nous pouvons regarder plus en détails chacune d'entre elles. En premier lieu, nous pouvons déterminer la moyenne et l'écart type des variables continues.

```
> mean
      acpp      dpoids          age       v1ert
 0.1957672 129.8148148  23.2380952   0.7936508
```

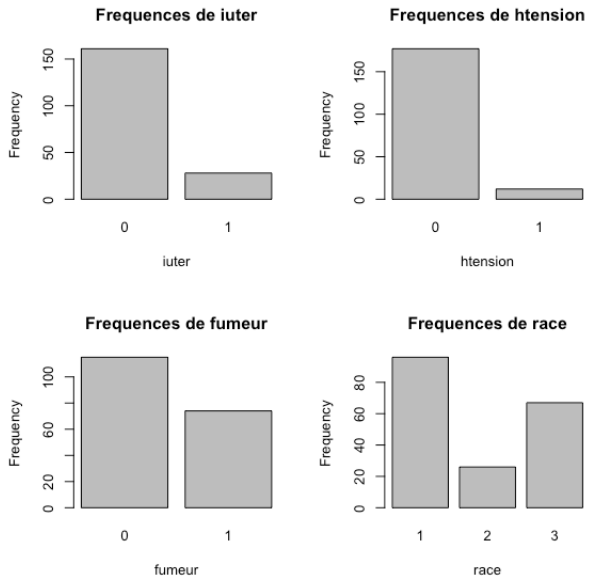
```
> sd
      acpp     dpoids         age      v1ert
 0.4933419 30.5793804  5.2986779  1.0592861
```

Pour mieux visualiser les chiffres, nous pouvons les représenter dans des boxplots afin de nous donner une idée de comment les valeurs sont distribuées. Nous avons donc le graphique suivant.



Ainsi, nous pouvons voir que la distribution des valeurs de **acpp** et **v1ert** ne sont pas très dispersées et sont concentrées sur de très petites valeurs. Toutefois, on voit que pour la variable **dpoids**, il y a présence de quelques valeurs très loin de la moyenne et la présence d'une valeur similaire pour la variable **age**. Cela pourrait être signe de faire attention à ces variables lors de la modélisation.

Par la suite, nous pouvons faire la même chose pour les variables catégorielles, mais au lieu de déterminer la moyenne et l'écart type, nous nous intéressons aux fréquences de celles-ci. Nous obtenons les graphiques suivants.



Avec ces fréquences, nous pouvons remarquer que très peu de mères ont une irritabilité utérine et des antécédents d'hypertension. La majorité des mères ne fumaient pas pendant leurs grossesses et la majorité des mères étaient des personnes blanches. Avec une bonne vue maintenant des données, nous pouvons passer à la modélisation. Pour faire cela, nous nous tournons vers les modèles de régression logistique. Le premier modèle que l'on considère est le modèle additif. En faisant le test d'adéquation, on obtient une p-valeur de 0.122.

```
glm(formula = faible.poids ~ acpp + dpoids + age + v1ert + iuter +
    htension + fumeur + race, family = binomial(logit), data = poidsNaissance.dat)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 201.28  on 179  degrees of freedom
AIC: 221.28

> 1 - pchisq(201.28, 179)
[1] 0.1216379
```

Le deuxième modèle que l'on considère est le modèle avec toutes les interactions En faisant le test d'adéquation, on obtient une p-valeur de 0.

```
glm(formula = faible.poids ~ dpoids + acpp + age + v1ert + iuter +
    htension + fumeur + race + dpoids:acpp + age:acpp + v1ert:acpp +
    iuter:acpp + htension:acpp + fumeur:acpp + race:acpp + age:dpoids +
    v1ert:dpoids + iuter:dpoids + htension:dpoids + fumeur:dpoids +
    race:dpoids + v1ert:age + iuter:age + htension:age + fumeur:age +
    race:age + iuter:v1ert + htension:v1ert + fumeur:v1ert +
    race:v1ert + htension:iuter + fumeur:iuter + race:iuter +
    fumeur:htension + race:htension + race:fumeur, family = binomial(logit),
    data = poidsNaissance.dat)

    Null deviance:  234.67  on 188  degrees of freedom
Residual deviance: 3171.84  on 145  degrees of freedom
AIC: 3259.8

> 1-pchisq(3171.84, 145)
[1] 0
```

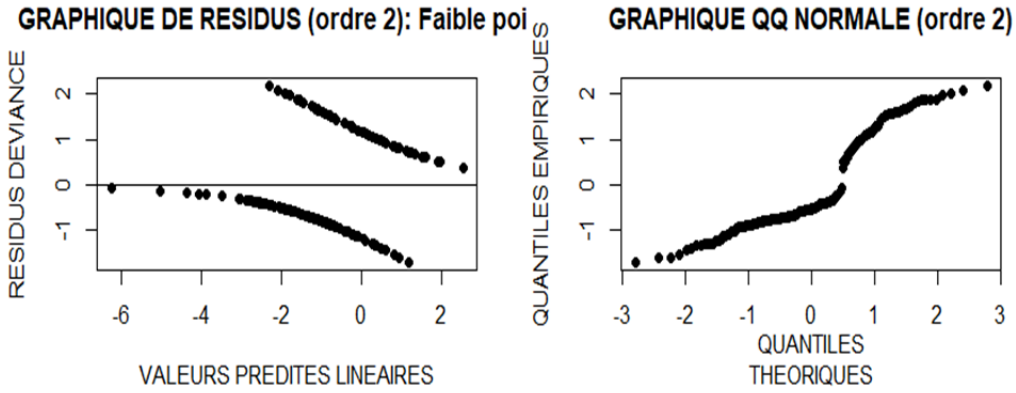
Le troisième que l'on considère est le modèle avec des interactions d'ordre 2. En faisant le test d'adéquation, on obtient une p-valeur de 0.193.

```
glm(formula = faible.poids ~ acpp + dpoids + age + v1ert + iuter +
    htension + fumeur + race + I(acpp^2) + I(dpoids^2) + I(age^2) +
    I(v1ert^2), family = binomial(logit), data = poidsNaissance.dat)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 190.99  on 175  degrees of freedom
AIC: 218.99

> 1-pchisq(190.99, 175)
[1] 0.1934218
```

On conclut que le modèle additif et le modèle avec interactions d'ordre 2 sont adéquats. En faisant une analyse de variance (ANOVA) entre les deux modèles, on obtient une p-valeur de 0.0357. Donc, le modèle réduit (dans ce cas c'est le modèle additif) n'est pas significatif comparé au modèle plus large.

```
Model 1: faible.poids ~ acpp + dpoids + age + v1ert + iuter + htension +
    fumeur + race
Model 2: faible.poids ~ acpp + dpoids + age + v1ert + iuter + htension +
    fumeur + race + I(acpp^2) + I(dpoids^2) + I(age^2) + I(v1ert^2)
  Resid. Df Resid. Dev Df Deviance
1       179     201.28
2       175     190.99  4   10.298
> 1-pchisq(10.298, 4)
[1] 0.03569622
```

Le graphique des résidus suivant présentent des patrons typiques de données non-groupée. C'est une bonne indication que nous avons un bon modèle. De plus, on voit qu'il n'y a pas de valeurs aberrantes.



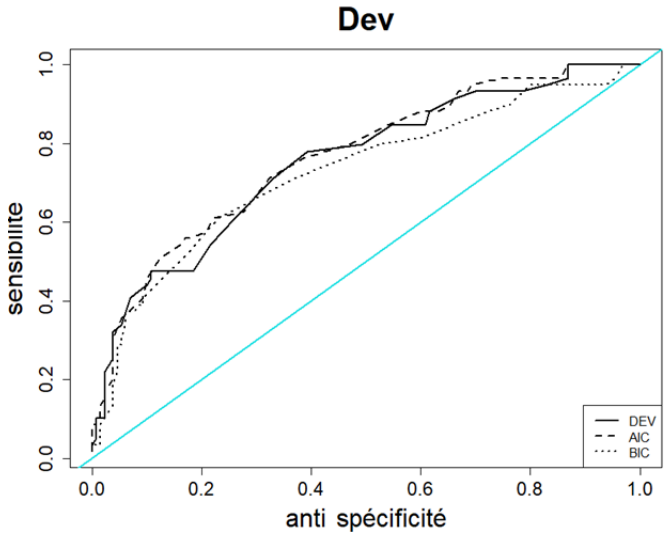
À partir du modèle avec interaction d'ordre 2, on va faire faire la sélection en reverse afin de faire un choix de modèle final. Après plusieurs itérations on arrive avec les modèles suivants :

Déviance : accp + htension + fumeur + race + acpp^2 + dpoids^2

AIC : accp + iuter + htension + fumeur + race + acpp^2 + dpoids^2

BIC : accp + htension + acpp^2 + dpoids^2

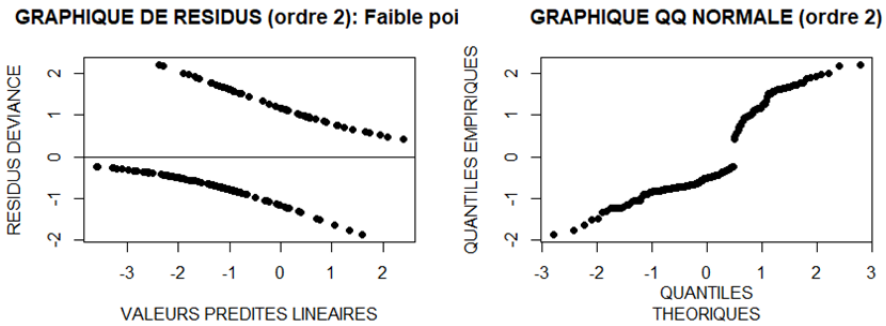
Les courbes ROC correspondantes :



On choisit AIC, car c'est elle qui a la plus grande aire sous la courbe. Finalement, en faisant le test d'adéquation, on obtient une p-valeur de 0.23.

```
glm(formula = faible.poids ~ acpp + iuter + htension + fumeur +
    race + I(acpp^2) + I(dpoids^2), family = binomial(logit),
    data = poidsNaissance.dat)

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 193.69  on 180  degrees of freedom
AIC: 211.69

> 1-pchisq(193.69, 180)
[1] 0.2299736
```

Le graphique des résidus suivant présentent des patrons typiques de données non-groupée. C'est une bonne indication que nous avons un bon modèle. De plus, on voit qu'il n'y a pas de valeurs aberrantes.



Par la suite, nous faisons aussi une **analyse bayésienne**. Nous venons de sélectionner un modèle de régression logistique basé sur les méthodes vues en cours, on a obtenu le modele AIC.glm:

```
AIC.glm = glm(faible.poids~acpp + iuter + htension + fumeur + race + I(acpp^2) +
I(dpoids^2), family = binomial(logit))
```

Pour approfondir l'analyse de ce modèle, nous proposons de réaliser une courte analyse bayésienne. Cette approche nous permettra d'incorporer des connaissances préalables sur les paramètres du modèle choisi et d'obtenir des distributions a posteriori pour les paramètres d'intérêt. Ce faisant, nous pouvons mieux comprendre les coefficients estimés et prendre des décisions plus éclairées en fonction des résultats.Nous commençons par eliciter une loi à priori en base aux parametres presentes dans notre modèle de regression logistique. Nous posons les variables beta.prior et Sigma.prior qui sont respectivement les estimations fait pour chaque variable explicative et sa variance multiplié par 5.

$$\beta.prior = esimateur \tag{1}$$

$$Sigma.prior = Variance * 5 \tag{2}$$

Ceci est particulièrement important car notre loi a priori est basée sur peu d'informations sur les données. En faisant ce procès, nous pouvons obtenir des valeurs plausibles pour les paramètres qui couvrent un large éventail de possibilités, tout en restant cohérents avec nos connaissances a priori limitées. On realise notre adequation bayesienne du modele avec le prior indiqué avant.

```
model.bayesian.prior <- stan_glm(faible.poids ~ acpp + iuter + htension + fumeur + race +
                                   I(acpp^2) + I(dpoids^2),
                                 data = poidsNaissance.dat,
                                 family = binomial(link = "logit"),
                                 prior = normal(beta.prior, sqrt(diag(Sigma.prior))),
                                 prior_intercept = normal(0,100))
```

Et on obtient les resultats suivants avec l'analyse bayesienne:

```
Estimates:
              mean   sd   10%   50%   90%
(Intercept) -1.2    0.5 -1.8  -1.2  -0.6
acpp         3.0    0.8  1.9   3.0   4.1
iuter1       0.9    0.4  0.4   0.9   1.5
htension1    2.0    0.7  1.2   2.0   2.9
fumeur1      0.9    0.4  0.5   0.9   1.4
race2        1.3    0.5  0.7   1.3   1.9
race3        0.9    0.4  0.4   0.9   1.4
I(acpp^2)   -1.5    0.5 -2.2  -1.5  -0.9
I(dpoids^2)  0.0    0.0  0.0   0.0   0.0

Fit Diagnostics:
           mean   sd   10%   50%   90%
mean_PPD 0.3    0.0  0.3   0.3   0.4
```

Dans l'analyse bayésienne que nous avons effectuée, nous voyons que les estimations des coefficients étaient cohérentes avec l'approche fréquentiste. Cela renforce notre confiance dans le modèle de régression logistique et suggère que le modèle est bien calibré et robuste. De plus, l'analyse bayésienne nous a permis de quantifier l'incertitude des estimations à travers les distributions a posteriori des coefficients. En examinant les moyennes et les écarts-types de ces distributions, nous pouvons voir le degré de variabilité des estimations et le degré de confiance que nous pouvons avoir dans leur exactitude. Globalement, l'analyse bayésienne apporte un complément utile à l'approche fréquentiste.

L'étape suivante de l'analyse consiste à utiliser des distributions a priori moins informatives pour les paramètres. Plus précisément, nous avons utilisé des distributions a priori normales pour les paramètres, avec une moyenne de 0 et un écart type de 100. Ce choix d'a priori permet une gamme plus large de valeurs de paramètres plausibles, reflétant le fait que nous n'avons aucune information a priori sur les paramètres. On utilisera par contre les mêmes variables explicatives utilisés.

```
# Fit the model using the prior distribution
model.bayesian.simple <- stan_glm(faible.poids ~ acpp + iuter + htension + fumeur + race +
                                    I(acpp^2) + I(dpoids^2),
                                  data = poidsNaissance.dat,
                                  family = binomial(link = "logit"),
                                  prior = normal(0, 100),
                                  prior_intercept = normal(0, 100))
```

On obtient les estimations suivantes.

```
Estimates:
              mean   sd   10%   50%   90%
(Intercept) -1.2    0.6 -2.0  -1.2  -0.4
acpp         3.2    1.1  1.8   3.1   4.6
iuter1       0.9    0.5  0.3   0.9   1.6
htension1    2.1    0.8  1.1   2.0   3.1
fumeur1      1.0    0.4  0.4   0.9   1.5
race2        1.3    0.5  0.6   1.3   2.0
race3        0.9    0.5  0.3   0.9   1.5
I(acpp^2)   -1.6    0.6 -2.4  -1.6  -0.8
I(dpoids^2)  0.0    0.0  0.0   0.0   0.0

Fit Diagnostics:
           mean   sd   10%   50%   90%
mean_PPD 0.3    0.0  0.3   0.3   0.4
```

Ces nouvelles estimations obtenues à l'aide de distributions a priori moins informatives fournissent une image similaire à l'analyse précédente. Les estimations moyennes des coefficients sont assez similaires à celles obtenues dans le modèle de régression logistique bayésienne avec les priors informatifs, mais les écarts-types sont plus grands, ce qui indique une plus grande incertitude dans les estimations. Les diagnostics d'ajustement montrent que la distribution prédictive a posteriori (PPD) moyenne est d'environ 0,3, ce qui suggère que le modèle a un ajustement raisonnable aux données. Dans l'ensemble, ces résultats confirment les conclusions de l'analyse précédente, mais avec une plus grande incertitude en raison de l'utilisation d'a priori moins informatifs.

Enfin, sur la base des différentes analyses et tests de diagnostic effectués sur le données, le modèle AIC.glm semble être un modèle adapté et robuste pour nos données. Les estimations des paramètres sont cohérentes et les densités postérieures prédictives ont une moyenne de 0,3 et un écart type de 0,0, indiquant des diagnostics de bonne qualité. De plus, les distributions a priori moins informatives pour les paramètres, telles que les distributions normales, ont produit des résultats similaires, renforçant encore la fiabilité de notre modèle. Par conséquent, après un examen attentif et une évaluation de plusieurs modèles, nous concluons que le modèle AIC.glm est le modèle final et préféré pour analyser notre ensemble de données.