Rapport

Questions Théoriques

T-2

#2

La fonction réciproque est la suivante: $g(u) = \frac{1}{u}$. Par conséquent, on obtient: $u_i = \frac{1}{\beta_0+\beta_1 x_{i1}+x_{i2}}$

a)

Dans le cas où x_{i1} augmente de 1, on obtient que Y augmente de:

$$E[Y_i; x_{i1} = x_{i1}^* + 1] - E[Y_i; x_{i1} = x_{i1}^*] = \frac{1}{\beta_0 + \beta_1(x_{i1} + 1) + x_{i2}} - \frac{1}{\beta_0 + \beta_1 x_{i1} + x_{i2}}$$

b)

Dans le cas où x_{i2} augmente de 1, on obtient que Y augmente de:

$$E[Y_i; x_{i2} = x_{i2}^* + 1] - E[Y_i; x_{i2} = x_{i2}^*] = \frac{1}{\beta_0 + \beta_1 x_{i1} + x_{i2} + 1} - \frac{1}{\beta_0 + \beta_1 x_{i1} + x_{i2}}$$

c)

On calcul les intervalles de confiance de Wald pour nos $\hat{\beta}_j$

$$\begin{aligned}
IC(\hat{\beta}_j, \alpha) &= \hat{\beta}_j \pm z_{\alpha/2}\sqrt{V\hat{ar}(\hat{\beta}_j)} \\
IC(\hat{\beta}_0, 0.95) &= 0.1 \pm 1.96\sqrt{0.000625} = [0.051, 0.149] \\
IC(\hat{\beta}_1, 0.95) &= -0.01 \pm 1.96\sqrt{0.000016} = [0.01784, -0.00216]
\end{aligned}$$

d)

L'intervalle de confiance pour notre prédicteur linéaire $\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$ est $\boldsymbol{x}_0'\hat{\boldsymbol{\beta}} \pm z_{\alpha/2}\sqrt{v^2(\boldsymbol{x}_0)}$

$$\begin{aligned}
v^2(\boldsymbol{x}_0) &= V\hat{ar}\left(\sum_{j=0}^{p'} x_{0j}\hat{\beta}_j\right) \\
&= V\hat{ar}(\hat{\beta}_0) + x_{01}^2 V\hat{ar}(\hat{\beta}_1) + 2x_{01}C\hat{ov}(\hat{\beta}_0, \hat{\beta}_1) \\
&= 0.000625 + 5^2 * 0000016 + 2 * 5 * -0.0001 \\
&= 0.000025
\end{aligned}$$

On obtient donc:

$IC(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}, 0.95) = 0.1 - 5 * 0.01 + 0.5 \pm 1.96\sqrt{0.000025} = [0.5402, 0.5598]$

On peut maintenant trouver notre intervalle de confiance pour Y_i:

$IC(Y_i, 0.95) = \frac{1}{IC(\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}, 0.95)} = [1.786, 1.851166]$

#4

a)

Selon notre fonction, on obtient:

$$\begin{aligned}
\ell(\beta_0, \beta_1; Y) &= \sum_{i=1}^{n} ln(f(y_i|x_i; \beta_0, \beta_1)) \\
&= \sum_{i=1}^{n} ln(\frac{1}{u_i} e^{-\frac{y_i}{u_i}}) \\
&= \sum_{i=1}^{n} ln(\frac{1}{u_i}) - \frac{y_i}{u_i} \\
&= \sum_{i=1}^{n} ln(\beta_0 + \beta_1 x_i) - y_i(\beta_0 + \beta_1 x_i)
\end{aligned}$$

b)

On cherche la déviance. On sait que $\ell(u; Y) = \sum_{i=1}^{n} ln(\frac{1}{u_i}) - \frac{y_i}{u_i}$

$$\begin{aligned}
D(Y, u) &= 2(\ell(y; Y) - \ell(u; Y)) \\
&= 2(\ell(y; Y) - \ell(g^{-1}(\beta_0 + \beta_1 x_i; Y)) \\
&= 2\left(\sum_{i=1}^{n} ln(\frac{1}{y_i}) - \frac{y_i}{y_i} - \sum_{i=1}^{n} ln(\beta_0 + \beta_1 x_i) - y_i(\beta_0 + \beta_1 x_i)\right) \\
&= 2\left(\sum_{i=1}^{n} ln(\frac{1}{y_i}) - 1 - ln(\beta_0 + \beta_1 x_i) + y_i(\beta_0 + \beta_1 x_i)\right)
\end{aligned}$$

#5

a)

On peut peut expremer u_i de la façon suivante: $u_i = exp\,(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ln(v_i)) = e^{\beta_0} e^{\beta_1 x_{i1}} e^{\beta_2 x_{i2}} v_i$. Par conséquent, si on remplace v_i par $v_i^* = 1.05 v_i$, on obtient:

$$\begin{aligned}
u_i^* &= e^{\beta_0} e^{\beta_1 x_{i1}} e^{\beta_2 x_{i2}} v_i^* \\
u_i^* &= e^{\beta_0} e^{\beta_1 x_{i1}} e^{\beta_2 x_{i2}} v_i 1.05 \\
u_i^* &= u_i * 1.05
\end{aligned}$$

b)

$$\begin{aligned}
u_i^* &= e^{\beta_0} e^{0.24(x_{i1}+1)} e^{\beta_2 x_{i2}} v_i \\
u_i^* &= e^{\beta_0} e^{0.24 x_{i1}} e^{0.24} e^{\beta_2 x_{i2}} v_i \\
u_i^* &= u_i e^{0.24} \\
u_i^* &= u_i 1.27
\end{aligned}$$

Lorsque x_{i1} augmente de une unité, alors la moyenne augmente d'environ 27% ($e^{0.24}$).

c)

On procède au test d'hypothèse suivant:

H_0: Le modèle poisson est suffisant

H_1: Le modèle binomiale négatif est nécessaire

$$
\begin{aligned}
\epsilon &= D_0 - D1 = 3.8 \\
p &= 0.5P(\chi_1^2 > \epsilon) && = 0.5P(\chi_1^2 > 3.8) = 0.02562629
\end{aligned}
$$

Comme la p-value est inférieure à 0.05, on rejète l'hypothèse H_0. On doit utilise le modèle négative binomial.

P-1

Analyse de multicolinéarité

Avant de faire notre modèle on peut procéder à une analyse de multicolinéarité en calculant nos facteur d'inflation de la variance (VIF):

```
## # A tibble: 6 x 3
##   Variables Tolerance   VIF
##   <chr>         <dbl> <dbl>
## 1 WT2          0.518   1.93
## 2 HT2          0.304   3.29
## 3 WT9          0.0709 14.1
## 4 HT9          0.277   3.61
## 5 LG9          0.0883 11.3
## 6 ST9          0.675   1.48
```

```
## Tolerance and Variance Inflation Factor
## ---------------------------------------
## # A tibble: 6 x 3
##   Variables Tolerance   VIF
##   <chr>         <dbl> <dbl>
## 1 WT2          0.518   1.93
## 2 HT2          0.304   3.29
## 3 WT9          0.0709 14.1
## 4 HT9          0.277   3.61
## 5 LG9          0.0883 11.3
## 6 ST9          0.675   1.48
##
##
## Eigenvalue and Condition Index
## ------------------------------
##   Eigenvalue Condition Index    intercept        WT2          HT2
## 1 6.96547210           1.000 0.0000048538 0.00014278 0.0000070895
## 2 0.01558744          21.139 0.0024597226 0.00056037 0.0021884104
## 3 0.01068856          25.528 0.0017927623 0.17933042 0.0011040227
## 4 0.00639981          32.991 0.0000885548 0.62817767 0.0000089499
## 5 0.00147774          68.656 0.0162934970 0.00042912 0.0269772209
## 6 0.00025142         166.446 0.0000564458 0.07143476 0.7028342227
## 7 0.00012293         238.041 0.9793041636 0.11992489 0.2668800838
##           WT9          HT9         LG9        ST9
## 1 0.000026084 0.0000082937 0.000008255 0.00029103
```

```
## 2 0.009866383 0.0022007833 0.000019912 0.48670716
## 3 0.033968630 0.0011253071 0.000496706 0.37674138
## 4 0.061207422 0.0009060785 0.003410444 0.09996265
## 5 0.051248011 0.0541522718 0.142343281 0.00027082
## 6 0.036956039 0.7594853714 0.006226504 0.00330584
## 7 0.806727430 0.1821218940 0.847494898 0.03272112
```

Avec cette analyse, on réalise qu'il a certains facteurs d'inflation de la variance supérieur à 10 (WT9 et LG9). On peut aussi observer les indices de conditionnement qui sont calculés à partir des valeurs propres qui atteignent des valeurs supérieur à 30. Finalement, on constate que les variables WT9 et LG9 sont probablement en multicollinéarité, car leur dépendance linéaire p_{lj} sont supérieur à 60% pour l'indice de conditionnement le plus élevé (238.041>30).

Pour remédier au problème, on peut retirer la variable LG9:

```
## # A tibble: 5 x 3
##   Variables Tolerance   VIF
##   <chr>         <dbl> <dbl>
## 1 WT2           0.587  1.70
## 2 HT2           0.346  2.89
## 3 WT9           0.409  2.44
## 4 HT9           0.322  3.10
## 5 ST9           0.704  1.42
```

On constate qu'aucun VIF n'est maintenant supérieur à 10. On peut commencer à faire notre modèle avec ces variables.

Modèle

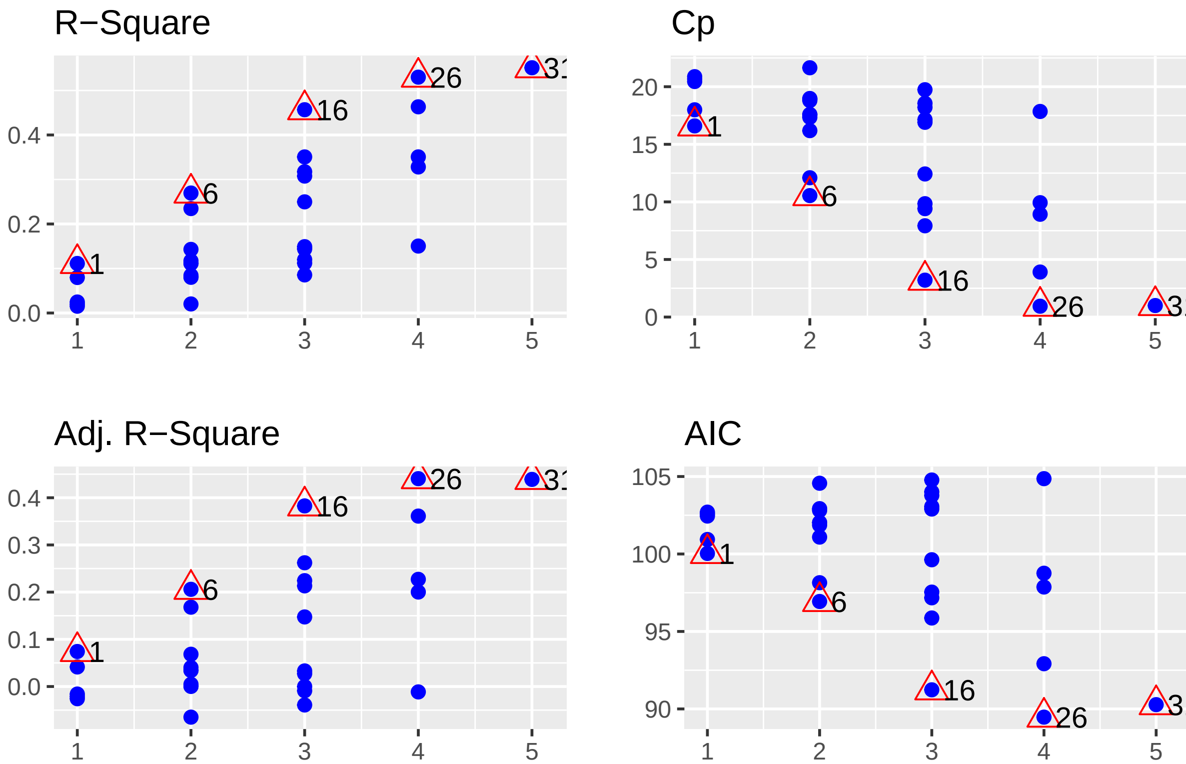
Puisque qu'on a peu de variables dans nos données ($p' = 5$), on peut se pemettre de trouver tous les sous-modèles possibles pour ensuite choisir le meilleur:

```
##    mindex          predictors  rsquare      adjr  predrsq     aic
## 17     25         WT2 HT2 HT9 0.085569 -0.039126 -0.32010 104.774
## 29     26     WT2 WT9 HT9 ST9 0.529970  0.440441  0.25032  89.471
## 27     27     WT2 HT2 WT9 ST9 0.463331  0.361109  0.15655  92.918
## 30     28     HT2 WT9 HT9 ST9 0.350771  0.227109 -0.10651  97.868
## 26     29     WT2 HT2 WT9 HT9 0.328191  0.200227 -0.15439  98.757
## 28     30     WT2 HT2 HT9 ST9 0.150409 -0.011417 -0.33771 104.862
## 31     31 WT2 HT2 WT9 HT9 ST9 0.551089  0.438861  0.17547  90.275
```

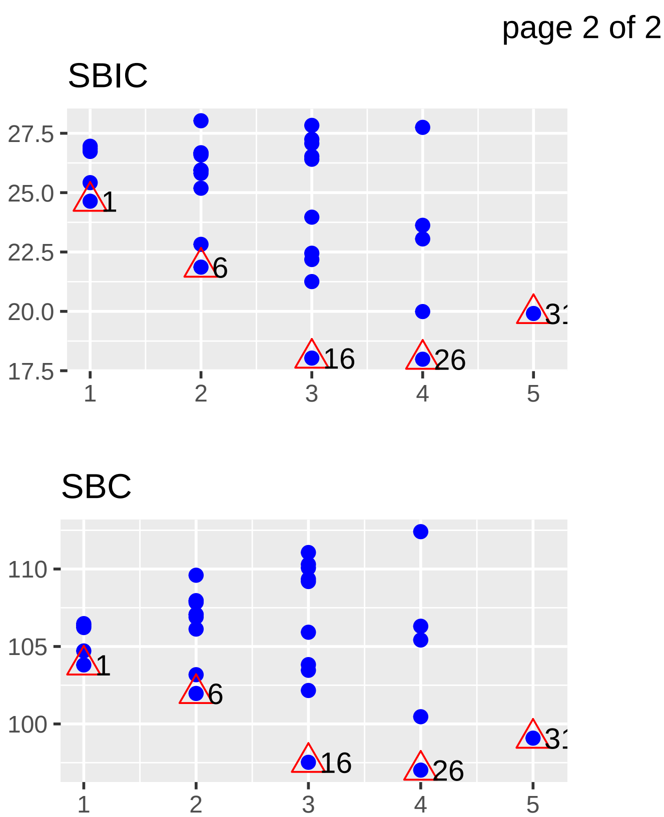
```
## [1] "Modèle avec meilleur R2 ajusté"
```

```
##    mindex n      predictors rsquare    adjr predrsq     cp    aic   sbic
## 29     26 4 WT2 WT9 HT9 ST9 0.52997 0.44044 0.25032 4.9409 89.471 17.985
##       sbc   msep   fpe     apc      hsp
## 29 97.019 1.7778 1.701 0.69385 0.071333
```

page 1 of 2



R−Square
0.4
0.2
0.0
1
2
3
4
5
1
6
16
26
Cp
20
15
10
5
0
Adj. R−Square
0.3
0.1
AIC
105
100
95
90



À partir de ces informations, on peut voir que, basé sur la valeur du R^2_{adj}, le meilleur modèle est le modèle qui inclus les variables WT2, WT9, HT9 et ST9 (on laisse tomber HT2). On peut toutefois voir que le modèle complet a un meilleur R^2 que notre modèle réduit. On peut confirmer que notre modèle réduit est toutefois le meilleur, il a également le plus faible AIC (89.471) et la plus grande valeur de R^2_{prev} (basée sur PRESS) (0.25032). Les graphiques comfirment les mêmes résultats, notre modèle réduit (#26) est le plus addéquat.

On peut maintenant faire notre prédiction:

```
## [1] "Prédiction pour Y"
```

```
##      fit    lwr    upr
## 1 6.5168 3.5355 9.4981
```

Avec notre modèle, on peut prédire qu'un enfant de 9 ans ayant ces caractéristiques aurait un somatotype de 6.5168 à l'âge de 18 ans. L'intervalle de confiance 95% de cette estimé ponctuel est de [3.5355, 9.4981]