Devoir 1 STT-7120

Étienne Bessette, Philippe Blouin-Leclerc & William Bourget

2018-11-16

Questions Théoriques

Chapitre 1 - Exercice 1

a) Montrez que $\hat{b}_1 = \frac{c_1\hat{\beta}_1}{c_2}$

Avec l'équation donnée dans l'exercice, on remarque facilement qu'il s'agit d'une régression linéaire de la forme $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ où le $\beta_0^* = \beta_0 + \beta_1 \overline{x}_n$ ce qui résulte à l'équation donnée de l'exercice. Alors, puisque le β_1 demeure inchangé, il est possible de partir de l'équation 1.17 du manuel de cours, où on obtient que la valeur $\hat{\beta}_1$ peut être définie comme suit:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n Y_i(x_i - \overline{x}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}$$

Alors, pour l'estimateur de $\hat{b}_1$, nous obtenons ceci:

$$\begin{aligned}
\hat{b}_1 &= \frac{\sum_{i=1}^n \widetilde{Y}_i(\widetilde{x}_i - \overline{\widetilde{x}}_n)}{\sum_{i=1}^n (\widetilde{x}_i - \overline{\widetilde{x}}_n)^2} \\
&= \frac{\sum_{i=1}^n c_1 Y_i(x_i - \overline{x}_n)c_2}{\sum_{i=1}^n (x_i - \overline{x}_n)^2 c_2^2} \\
&= \frac{c_1}{c_2} \frac{\sum_{i=1}^n Y_i(x_i - \overline{x}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2} \\
&= \frac{c_1}{c_2}\hat{\beta}_1
\end{aligned}$$

b) Montrez que $\widetilde{\sigma}^2 = Var(\widetilde{Y}_i) = c_1^2\sigma^2$ et donc que $\widetilde{s}^2 = c_1^2 s_2$ est un estimateur non biaisé de $\widetilde{\sigma}^2$.

$$\begin{aligned}
Var(\widetilde{Y}_i) &= Var(c_1 Y_i) \\
&= c_1^2 Var(Y_i) \\
&= c_1^2\sigma^2
\end{aligned}$$

Alors, on a que

$$\begin{aligned}
\widetilde{s}^2 &= \frac{\sum_{i=1}^n (\widetilde{Y}_i - \hat{\widetilde{Y}}_i)^2}{n-p} \\
\widetilde{s}^2 &= \frac{\sum_{i=1}^n (c_1 Y_i - (\hat{b}_0 + \hat{b}_1(\widetilde{x}_i - \overline{\widetilde{x}}_n)))^2}{n-p}
\end{aligned}$$

En utilisant le fait que $\hat{b}_1 = \frac{c_1\hat{\beta}_1}{c_2}$ et $\hat{b}_0 = c_1\hat{\beta}_0$, on remplace dans l'équation pour obtenir:

$$
\begin{aligned}
\widetilde{s}^2 &= \frac{\sum_{i=1}^{n}(c_1 Y_i - (c_1\hat{\beta}_0 + \frac{c_1\hat{\beta}_1}{c_2}(\widetilde{x}_i - \overline{\widetilde{x}}_n)))^2}{n-p} \\
&= \frac{\sum_{i=1}^{n}(c_1 Y_i - (c_1\hat{\beta}_0 + \frac{c_1\hat{\beta}_1}{c_2}(x_i - \overline{x}_n)c_2))^2}{n-p} \\
&= c_1^2 \frac{\sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1(x_i - \overline{x}_n)))^2}{n-p} \\
&= c_1^2 \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p} \\
&= c_1^2 s^2
\end{aligned}
$$

c) Montrez que si $(\hat{\beta}_{1L}, \hat{\beta}_{1U})$ est un intervalle de confiance à 95% pour β_1, montrez que $(\frac{c_1}{c_2}\hat{\beta}_{1L}, \frac{c_1}{c_2}\hat{\beta}_{1U})$ est un intervalle de confiance à 95% pour b_1.

$$
\begin{aligned}
&Pr\left(\hat{b}_1 - t_{\frac{\alpha}{2};n-p}\sqrt{\widetilde{s}^2(\widetilde{X}'\widetilde{X})^{-1}} \leq b_1 \leq \hat{b}_1 - t_{\frac{\alpha}{2};n-p}\sqrt{\widetilde{s}^2(\widetilde{X}'\widetilde{X})^{-1}}\right) \\
&= Pr\left(\frac{c_1}{c_2}\hat{\beta}_1 - t_{\frac{\alpha}{2};n-p}\sqrt{\frac{c_1^2}{c_2^2}s^2(X'X)^{-1}} \leq b_1 \leq \frac{c_1}{c_2}\hat{\beta}_1 - t_{\frac{\alpha}{2};n-p}\sqrt{\frac{c_1^2}{c_2^2}s^2(X'X)^{-1}}\right) \\
&= Pr\left(\frac{c_1}{c_2}\left(\hat{\beta}_1 - t_{\frac{\alpha}{2};n-p}\sqrt{s^2(X'X)^{-1}}\right) \leq b_1 \leq \frac{c_1}{c_2}\left(\hat{\beta}_1 + t_{\frac{\alpha}{2};n-p}\sqrt{s^2(X'X)^{-1}}\right)\right) \\
&= Pr\left(\frac{c_1}{c_2}\hat{\beta}_{1L} \leq b_1 \leq \frac{c_1}{c_2}\hat{\beta}_{1U}\right)
\end{aligned}
$$

Alors, puisque nous avons que l'intervalle de confiance à 95% de β_1 est la suivante:

$$
0.95 = Pr\left(\hat{\beta}_1 - t_{\frac{0.05}{2};n-p}\sqrt{s^2(X'X)^{-1}} \leq \beta_1 \leq \hat{\beta}_1 - t_{\frac{0.05}{2};n-p}\sqrt{s^2(X'X)^{-1}}\right)
$$

Nous obtenons que l'intervalle au même niveau de confiance pour b_1 étant $(\frac{c_1}{c_2}\hat{\beta}_{1L}, \frac{c_1}{c_2}\hat{\beta}_{1U})$.

d) Soit t_1, la statistique t qui permet de tester l'hypothèse $H_0 : \beta_1 = \beta_{1,0}$. Montrez que la statistique t pour le test $H_0 : b_1 = b_{1,0}$, où $b_{1,0} = \frac{c_1\beta_{1,0}}{c_2}$, sera égale à t_1.

$$
\begin{aligned}
t &= \frac{\hat{b}_1 - b_{1,0}}{se(\hat{b}_1)} \\
t &= \frac{\frac{c_1\hat{\beta}_{1,0}}{c_2} - \frac{c_1\beta_{1,0}}{c_2}}{\sqrt{\widetilde{s}^2(\widetilde{X}'\widetilde{X})^{-1}}} \\
t &= \frac{\frac{c_1\hat{\beta}_{1,0}}{c_2} - \frac{c_1\beta_{1,0}}{c_2}}{\sqrt{\frac{c_1^2}{c_2^2}s^2(X'X)^{-1}}} \\
t &= \frac{\frac{c_1}{c_2}\hat{\beta}_{1,0} - \frac{c_1}{c_2}\beta_{1,0}}{\frac{c_1}{c_2}\sqrt{s^2(X'X)^{-1}}} \\
t &= \frac{\hat{\beta}_{1,0} - \beta_{1,0}}{\sqrt{s^2(X'X)^{-1}}} \\
t &= t_1
\end{aligned}
$$

e) À la lumière des résultats obtenus en (a)-(d), que pouvez-vous dire quant au choix des unités de mesure pour la variable endogène et la variable exogène ?

Comme il est possible de le constater, un changement d'unité de mesure impacte les valeurs des estimations des coefficients de régression. Alors, lors d'un interprétation des coefficients, il est bien important de considérer l'unité de mesure sélectionnée. Par contre, il est possible de voir au numéro (d) que les unités de mesure n'impacteront pas les tests d'hypothèses et en analysant les résultats, il sera possible de rejeter ou non l'hypothèse nulle avec le même niveau de confiance. Alors, dans tous les cas, lorsqu'il y a un changement d'unité de mesure, la régression linéaire s'ajustera en conséquence et nous obtiendrons les mêmes conclusions.

Chapitre 1 - Exercice 10

a)

Afin de répondre à la question de l'ingénieur, il est possible de faire un test F de l'importance globale de la régression. Ceci revient donc à tester

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_1 : \text{au moins un des coefficients n'est pas 0.}$$

où ceci revient à prendre la statistique F de la table ANOVA et de rejeter H_0 lorsque celle-ci prend de grande valeur.

$$F = \frac{SS_{Reg}/p^{'}}{s^2}$$
$$F = \frac{(SS_{TOT} - SS_{Res})/p^{'}}{s^2}$$
$$F = \frac{(SS_{TOT} - s^2(n-p))/p^{'}}{s^2}$$
$$F = \frac{(\sum_{i=1}^{23} Y_i^2 - n\overline{Y_n}^2 - s^2(n-p))/p^{'}}{s^2}$$
$$F = \frac{(2850 - (23)(\frac{207}{23})^2 - 7.81(23-4))/3}{7.81} = 35.79$$

Alors, on a que

$$Pr[F_{3,19} \geq 35.79] \approx 0$$

Donc on rejete l'hypothèse nulle qu'il n'y a aucun des coefficients qui explique la qualité des stylos. On conclut que les données montrent de l'évidence qu'au moins une des 3 variables exogènes expliquent une partie de la variable endogène.

b)

$$E[Y|x_1^* = x_1 - 1] - E[Y|x_1^* = x_1] = E[Y|x_2^* = x_2 + 1] - E[Y|x_2^* = x_2]$$
$$\beta_0+\beta_1(x_1-1)+\beta_2x_2+\beta_3x_3-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3) = \beta_0+\beta_1x_1+\beta_2(x_2+1)+\beta_3x_3-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3)$$

Ceci revient à tester l'hypothèse suivante

$$-\beta_1 = \beta_2$$

Alors, on peut substituer β_2 par $-\beta 1$ dans l'équation du modèle complet pour obtenir notre hypothèse nulle.

$$Y = \beta_0 + \beta_1 x_1 - \beta_1 x_2 + \beta_3 x_3 + \epsilon$$
$$Y = \beta_0 + \beta_1(x_1 - x_2) + \beta_3 x_3 + \epsilon$$

Donc nous pouvons tester l'hypothèse nulle selon le principe de somme de carrés résiduelle additionnelle

$$H_0 : Y_i = \beta_0 + \beta_1(x_{i1} - x_{i2}) + \beta_3 x_{i3} + \epsilon_i$$
$$H_1 : Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

ce qui revient à comparer l'énoncé (v) de l'exercice contre l'énoncé (i) et nous rejeterons l'hypothèse nulle si la valeur de la statistique F suivante prendra une grande valeur

$$F = \frac{SS_{Res}^{H_0} - SS_{Res}^{H_1}}{\Delta_{dl} s_{H_1}^2}$$
$$F = \frac{s_{H_0}^2(23-3) - s_{H_1}^2(23-4)}{1 * s_{H_1}^2}$$
$$F = \frac{8.12(23-3) - 7.81(23-4)}{1 * 7.81} = 1.7939$$

On obtient ensuite que

$$Pr[F_{1,23-4} \geq 1.7939] = 0.1963$$

Nous obtenons que les données ne montrent pas d'évidence afin de pouvoir rejeter l'hypothèse nulle.

c)

Ceci revient à tester l'hypothèse suivante:

$$\begin{aligned} H_0 : \beta_3 = 0 \quad &\Rightarrow Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i \\ H_1 : \beta_3 \neq 0 \quad &\Rightarrow Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i \end{aligned}$$

Et il est possible de faire le test en utilisant le principe de somme de carrés résiduelle additionnelle. On compare alors l'équation (iv) à l'équation (i) du numéro.

$$\begin{aligned} &= \frac{SS_{Res}^{H_0} - SS_{Res}^{H_1}}{\Delta_{dl} s_{H_1}^2} \\ &= \frac{s_{H_0}^2(23-3) - s_{H_1}^2(23-4)}{1 * s_{H_1}^2} \\ &= \frac{18.12(23-3) - 7.81(23-4)}{1 * 7.81} \\ &= 27.40 \end{aligned}$$

Les données montrent alors de la forte évidence contre H_0 puisque $Pr[F_{1,19} \geq 27.40] \approx 0$. On répond alors à l'administrateur que oui, les données montrent de l'évidence que changer le bonus de productivité affectera la qualité de stylos.

Chapitre 1 - Exercice 12

a)

$$E[Y_i | m_{iA} = 1, m_{iB} = 0] = E[Y_i | m_{iA} = 0, m_{iB} = 1] = E[Y_i | m_{iA} = 0, m_{iB} = 0]$$

$$\beta_0 + \beta_1 x_i + \beta_A + \beta_2 x_i = \beta_0 + \beta_1 x_i + \beta_B + \beta_3 x_i = \beta_0 + \beta_1 x_i$$

$$\beta_A + \beta_2 x_i = \beta_B + \beta_3 x_i = 0$$

Puisque l'énoncé mentionne que ceci doit être valide peu importe le nombre d'heures consacrés au cours (x_i), alors on obtient que le test d'hypothèse revient à tester :

$$\begin{aligned} H_0 &: \beta_A = \beta_B = \beta_2 = \beta_3 = 0 \\ H_1 &: \beta_A \neq \beta_B \neq \beta_2 \neq \beta_3 \neq 0 \end{aligned}$$

b)

Pour le manuel A, on obtient le résultat suivant:

$$E[Y_i | x_i^* = x_i + 1, m_{iA} = 1, m_{iB} = 0] - E[Y_i | x_i^* = x_i, m_{iA} = 1, m_{iB} = 0]$$

$$= \beta_0 + \beta_1(x_i + 1) + \beta_A + \beta_2(x_i + 1) - (\beta_0 + \beta_1 x_i + \beta_A + \beta_2 x_i)$$

$$= \beta_1 + \beta_2$$

Pour le manuel B, on obtient le résultat suivant:

$$E[Y_i|x_i^* = x_i + 1, m_{iA} = 0, m_{iB} = 1] - E[Y_i|x_i^* = x_i, m_{iA} = 0, m_{iB} = 1]$$
$$= \beta_0 + \beta_1(x_i + 1) + \beta_B + \beta_3(x_i + 1) - (\beta_0 + \beta_1 x_i + \beta_B + \beta_3 x_i)$$
$$= \beta_1 + \beta_3$$

Pour le manuel C, on obtient le résultat suivant:

$$E[Y_i|x_i^* = x_i + 1, m_{iA} = 0, m_{iB} = 0] - E[Y_i|x_i^* = x_i, m_{iA} = 0, m_{iB} = 0]$$
$$= \beta_0 + \beta_1(x_i + 1) - (\beta_0 + \beta_1 x_i)$$
$$= \beta_1$$

Afin de tester si les 3 équations ci-dessus sont équivalentes, ceci revient à tester

$$\beta_1 + \beta_2 = \beta_1 + \beta_3 = \beta_1 \quad \Rightarrow \beta_2 = \beta_3 = 0$$

Alors, on doit tester

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_1 : \beta_2 \neq \beta_3 \neq 0$$

c)

Nous voulons alors tester si

$$E[Y|x_i^* = 1.25x_i, , m_{iA} = 0, m_{iB} = 0] = E[Y|x_i^* = x_i, m_{iA} = 0, m_{iB} = 1]$$
$$\beta_0 + \beta_1 1.25x_i = \beta_0 + \beta_1 x_i + \beta_B + \beta_3 x_i$$

ce qui revient à tester simultanément les 2 conditions suivantes en fonction des termes des coefficients du modèle:

$$H_0 : \beta_B = 0, \quad 0.25\beta_1 = \beta_3$$
$$H_0 : \beta_B \neq 0, \quad 0.25\beta_1 \neq \beta_3$$

d)

On peut écrire ceci sous la forme suivante:

$$E[Y|x_i^* = x_i + 1, m_{iA} = 1, m_{iB} = 0] - E[Y|x_i^* = x_i, m_{iA} = 1, m_{iB} = 0]$$
$$= E[Y|x_i^* = x_i + 1.5, m_{iA} = 0, m_{iB} = 0] - E[Y|x_i^* = x_i, m_{iA} = 0, m_{iB} = 0]$$
$$\beta_0 + \beta_1(x_i + 1) + \beta_A + \beta_2(x_i + 1) - (\beta_0 + \beta_1 x_i + \beta_A + \beta_2 x_i)) = \beta_0 + \beta_1(x_i + 1.5) - (\beta_0 + \beta_1 x_i)$$
$$\beta_1 + \beta_2 = 1.5\beta_1$$

Ce qui revient à tester l'hypothèse suivante:

$$H_0 : 0.5\beta_1 = \beta_2$$

$$H_1 : 0.5\beta_1 \neq \beta_2$$

e)

En gardant la même notation que dans l'exercice et en respectant les 3 contraintes, on peut écrire le modèle de régression linéaire suivant :

$$Y_i = \beta_0 + \beta_1(max(x_i; 5) - 5) + \beta_C(1 - m_{iA} - m_{iB}) + \epsilon_i$$

Chapitre 2 - Exercice 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 calcule les intervalles de confiance de Wald pour nos $\hat{\beta}_j$

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

d)

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

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

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:

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

Chapitre 2 - Exercice 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 $\ell(u; Y) = \sum_{i=1}^{n} ln(\frac{1}{u_i}) - \frac{y_i}{u_i}$

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

Chapitre 2 - Exercice 5

a)

On peut exprimer 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

$$\epsilon = D_0 - D1 = 3.8$$
$$p = 0.5P(\chi_1^2 > \epsilon) \qquad = 0.5P(\chi_1^2 > 3.8) = 0.02562629$$

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

Chapitre 3 - Exercice 1

Commencons par une démonstration qui sera utile à quelques endroits lors de la preuve du théorème miracle. Pour la démonstration il est intéressant de regarder comment chaque valeur de la matrice $(X'X)$ est calculé. Chaque élément sera dénoté $(X'X)_{ij}$

$$(X'X)_{kj} = (X'X)_{jk} = \sum_{n=1}^{N} X_{n,j}X_{n,k}$$

Si une donnée est retirée de la sommation, il restera donc $N-1$ donnée dans la sommation. Par exemple, si la i^e donnée est enlevée de la sommation, on retrouve :

$$(X'X)_{kj} = (X'X)_{jk} = \sum_{n=1}^{N-1} X_{n,j}X_{n,k} + x_{i,j}x_{i,k}$$

$$\sum_{n=1}^{N-1} X_{n,j}X_{n,k} + x_{i,j}x_{i,k} = (X'_{-i,}X_{-i})_{kj} + x_{ij}x'_{ik}$$

On passe de la forme de chaque élément à la forme matricielle.

$$(X'X) = (X'_{-i}X_{-i}) + x_i x'_i \quad , \quad (X'_{-i}X_{-i}) = (X'X) - x_i x'_i \quad eq(3.1)$$

Le résultat 10 de la proposition 0.1 stipule que :

$$(A - vv')^{-1} = A^{-1} + \frac{A^{-1}vv'A^{-1}}{1 - v'A^{-1}v}$$

Il est possible de faire un paralèle avec les matrices de notre modèle linéaire. $A = X'X$ et $v = x_i$ En remplacent ces matrices dans l'équations précédente, on obtient

$$(X'X - x_i x'_i)^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i x'_i (X'X)^{-1}}{1 - x'_i(X'X)^{-1}x_i}$$

Il est possible de remplacer le terme à gauche de l'équation par l'équation 3.1. Et $h_{ii} = x'_i(X'X)^{-1}x_i$

$$(X'_{-i}X_{-i})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i x'_i (X'X)^{-1}}{1 - hii} \quad eq(3.2)$$

Trouvons maitenant l'équation de $\hat{\beta}_{-i}$ en fonction de matrice contenant la donnée i

$$\hat{\beta}_{-i} = (X'_{-i}X_{-i})^{-1}X'_{-i}Y_{-1}$$

On peut remplacer $(X'_{-i}X_{-i})^{-1}$ par l'équation 3.2. De plus, en partant du même principe que l'équation 3.1, il est possible de prouver que $(X'_{-i}Y_{-i}) = (X'Y) - x_i y_i$. En remplacent ces deux termes dans l'équation précédente on obtient :

$$\hat{\beta}_{-i} = ((X'X)^{-1} + \frac{(X'X)^{-1}x_i x'_i (X'X)^{-1}}{1 - hii})(X'Y - x_i y_i)$$

$$\hat{\beta}_{-i} = (X'X)^{-1}X'Y - (X'X)^{-1}x_i y_i + \frac{(X'X)^{-1}x_i x'_i (X'X)^{-1}X'Y}{1 - hii} - \frac{(X'X)^{-1}x_i x'_i (X'X)^{-1}x_i y_i}{1 - hii}$$

$$\hat{\beta}_{-i} = \hat{\beta} - (X'X)^{-1}x_i y_i + \frac{(X'X)^{-1}x_i x'_i \hat{\beta}}{1 - hii} - \frac{(X'X)^{-1}x_i y_i h_{ii}}{1 - hii}$$

$$\hat{\beta}_{-i} = \hat{\beta} - (\frac{(X'X)^{-1}x_i}{1 - h_{ii}})(y_i(1 - h_{ii}) - x'_i\hat{\beta} + y_i h_{ii})$$

$$\hat{\beta}_{-i} = \hat{\beta} - (\frac{(X'X)^{-1}x_i}{1-h_{ii}})(y_i - x_i'\hat{\beta})$$

$$\hat{\beta}_{-i} = \hat{\beta} - (\frac{(X'X)^{-1}x_i}{1-h_{ii}})(e_i) \quad eq(3.3)$$

On peut aussi écrire l'équation de $\hat{\beta}_{-i}$ de cette façon :

$$\hat{\beta}_{-i} = \hat{\beta} - (\frac{(X'X)^{-1}x_i y_i}{1-h_{ii}}) + \frac{x_i'(X'X)^{-1}x_i\hat{\beta}}{1-h_{ii}})$$

$$\hat{\beta}_{-i} = \hat{\beta}(1 + \frac{h_{ii}}{1-h_{ii}}) - (\frac{(X'X)^{-1}x_i y_i}{1-h_{ii}})$$

$$\hat{\beta}_{-i} = \frac{(X'X)^{-1}X'Y - (X'X)^{-1}x_i y_i}{1-h_{ii}}$$

$$\hat{\beta}_{-i} = \frac{(X'X)^{-1}(X'Y - x_i y_i)}{1-h_{ii}}$$

On sait que :

$$e_{i,-i} = Y_i - x_i'\hat{\beta}_{-i}$$

En remplacent la dernière forme de l'équation de $\hat{\beta}_{-i}$ dans l'équation ci-haut, on obtient la preuve de la partie c)

$$e_{i,-i} = \frac{Y_i(1-h_{ii}) - x_i'(X'X)^{-1}(X'Y - x_i y_i)}{1-h_{ii}}$$

En remplacent le $\hat{\beta}_{-i}$ dans l'équation par l'équation 3.3, on obtient :

$$e_{i,-i} = Y_i - x_i'(\hat{\beta} - (\frac{(X'X)^{-1}x_i}{1-h_{ii}})(e_i))$$

$$e_{i,-i} = Y_i - x_i'\hat{\beta} + (\frac{(x_i'X'X)^{-1}x_i}{1-h_{ii}})(e_i)$$

$$e_{i,-i} = e_i + \frac{h_{ii}e_i}{1-h_{ii}}$$

$$e_{i,-i} = \frac{e_i}{1-h_{ii}}$$

Questions Pratiques

P1

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 certains indices de conditionnement qui sont calculés à partir des valeurs propres atteignent des valeurs supérieur à 30. Finalement, on constate que les variables WT9 et LG9 sont probablement en multicollinéarité, car leurs dépendances linéaires 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.

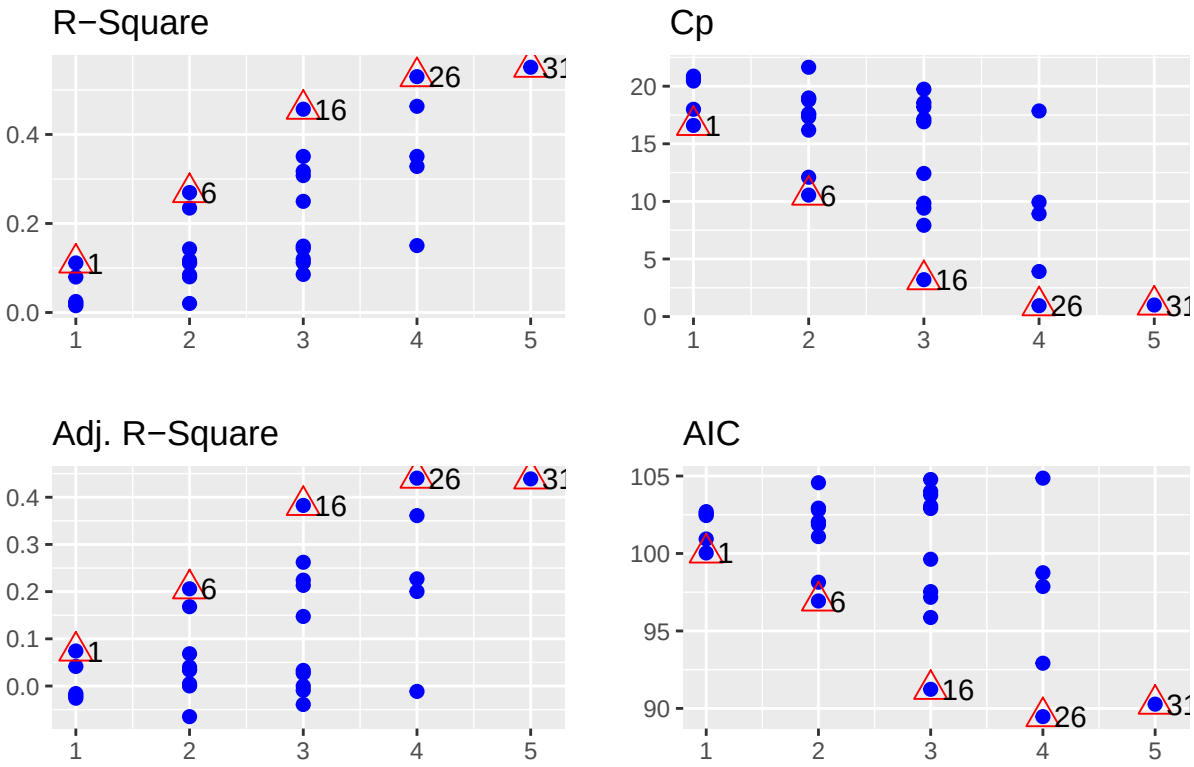
Modèle

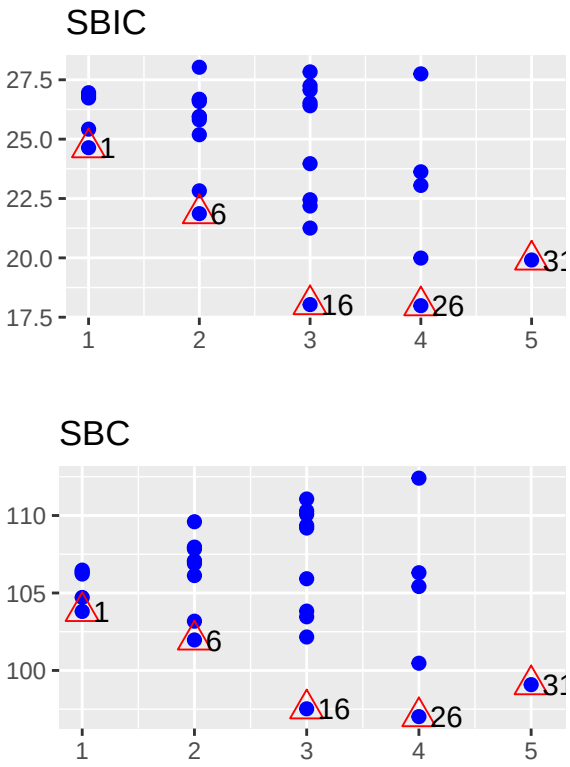
Puisque qu'on a peu de variables dans nos données ($p' = 5$), on peut se permettre de trouver tous les sous-modèles possibles pour ensuite choisir le meilleur:

```
##    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
```


R−Square
0.0
0.2
0.4
1
2
3
4
5
1
6
16
26
31
Cp
0
5
10
15
20
1
6
16
26
31
Adj. R−Square
0.0
0.1
0.2
0.3
0.4
1
6
16
26
31
AIC
90
95
100
105
1
6
16
26
31



À partir de ces informations, on peut voir que, basé sur la valeur du R^2_{adj}, le meilleur modèle est celui qui inclut 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 adéquat.

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]

P2

Il est de mise de débuter en vérifiant s'il y a de la multicolinéarité dans la matrice schéma X.

```
## # A tibble: 18 x 3
##    Variables           Tolerance   VIF
##    <chr>                   <dbl> <dbl>
##  1 age                     0.665  1.50
##  2 sex                     0.745  1.34
##  3 as.factor(cp)2          0.347  2.88
##  4 as.factor(cp)3          0.282  3.54
##  5 as.factor(cp)4          0.230  4.34
##  6 trestbps                0.815  1.23
##  7 chol                    0.869  1.15
##  8 fbs                     0.887  1.13
##  9 as.factor(restecg)1     0.919  1.09
## 10 as.factor(restecg)2     0.892  1.12
## 11 thalach                 0.576  1.74
## 12 exang                   0.697  1.43
## 13 oldspeak                0.536  1.87
## 14 as.factor(slope)2       0.605  1.65
## 15 as.factor(slope)3       0.641  1.56
## 16 as.numeric(ca)          0.726  1.38
## 17 as.factor(thal)6.0      0.791  1.26
## 18 as.factor(thal)7.0      0.632  1.58
```

Alors, puisqu'aucun VIF est supérieur à 10, il n'y a pas présence de multicolinéarité entre les variables de la matrice X.

On peut observer l'allure d'un modèle complet avant de commencer avec les méthodes de sélections de variables.

```
##                       Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)         -8.6524873  3.0061581 -2.878254 0.0039988267
## age                 -0.0137630  0.0247451 -0.556189 0.5780816490
## sex                  1.5460138  0.5299949  2.917035 0.0035337613
## as.factor(cp)2       1.2395662  0.7708745  1.608000 0.1078351582
## as.factor(cp)3       0.2459588  0.6633123  0.370804 0.7107836447
## as.factor(cp)4       2.0864796  0.6665471  3.130281 0.0017463922
## trestbps             0.0243642  0.0112694  2.161985 0.0306193273
## chol                 0.0044481  0.0039927  1.114042 0.2652609974
## fbs                 -0.5962455  0.6078480 -0.980912 0.3266360155
## as.factor(restecg)1  0.8102016  2.4351018  0.332718 0.7393473442
## as.factor(restecg)2  0.4738946  0.3835184  1.235650 0.2165886486
## thalach             -0.0177229  0.0111095 -1.595295 0.1106463598
## exang                0.7094557  0.4400182  1.612333 0.1068895957
## oldspeak             0.3578748  0.2300701  1.555503 0.1198262843
## as.factor(slope)2    1.1552865  0.4737939  2.438373 0.0147535291
## as.factor(slope)3    0.5251469  0.9196610  0.571022 0.5679845384
## as.numeric(ca)       1.3115099  0.2792756  4.696113 0.0000026516
## as.factor(thal)6.0  -0.0109736  0.7902099 -0.013887 0.9889201490
## as.factor(thal)7.0   1.3927153  0.4251935  3.275486 0.0010548044
```

L'AIC du modèle complet est de 229.64.

On peut voir que quelques variables ne semblent pas statistiquement significatives. Alors, effectuons des méthodes algorithmiques pour sélectionner les variables et afin de s'assurer que nous obtenons le modèle

avec les meilleures statistiques. Pour les méthodes pas-à-pas, d'inclusion et d'exclusion, on obtient le même modèle, soit le modèle contenant les variables:

- *thal*
- *ca*
- *cp*
- *oldspeak*
- *slope*
- *sex*
- *trestbps*
- *exang*
- *thalach*

Et ce modèle nous donne un AIC de 223.98 ce qui est mieux que le modèle complet avec un AIC de 229.64.

Par la suite, on peut poursuivre avec la méthode LASSO. Ce modèle avec le meilleur λ revient à enlever la variable *age* et *fbs* et le modèle en prenant le meilleur λ plus sont écart type revient à enlever les variables *age*, *trestbps*, *chol*, *fbs* et *restecg*. Pour ces 2 modèles, on obtient un AIC de 226.96 et de 227.11 respectivement, ce qui est supérieur (donc moins bon) que les modèles trouvés plus haut.

Par ailleurs, puisque le nombre de variables dans le jeu de données nous le permet, il est possible d'essayer tous les modèles possibles.

On voit que, parmis tous les modèles possibles, on obtient le même modèle qu'obtenu avec les méthodes pas-à-pas, d'inclusion et d'exclusion lorsque notre critère de sélection est l'AIC.

```
##                     Coefficients
## (Intercept)           -8.146468
## sex                    1.430366
## as.factor(cp)2         1.291514
## as.factor(cp)3         0.203036
## as.factor(cp)4         2.173721
## trestbps               0.022862
## thalach               -0.014969
## exang                  0.669450
## oldspeak               0.392538
## as.factor(slope)2      1.200462
## as.factor(slope)3      0.451356
## as.numeric(ca)         1.237046
## as.factor(thal)6.0    -0.238138
## as.factor(thal)7.0     1.363032
```

Alors, les facteurs qui semblent être associés à une hausse d'un diagnostic positif de la maladie coronarienne sont:

- Être un homme
- La nature des douleurs à la poitrine
  - Le fait d'avoir soit une angine atypique, une douleur non anginienne ou une douleur asymptomatique augmente la probabilité
- Plus la tension artérielle au repos à l'admission à l'hôpital est élevée
- Plus le pouls maximum atteint a une valeur réduite
- La présence d'angine induite par l'exercice
- Plus il y a une baisse dans ST induite par l'exercise par rapport au repos
- Une pente plate ou descendante du segment de ST lors de l'exercice maximal
- Plus il y a un nombre élevé de vaissaux sanguins majeurs colorés par fluroscopie
- Avoir un défaut réparable (lorsque *thal* vaut 7)

P3

Il est préférable de regarder les facteurs d'inflations de la variance avant de commencer la modélisation.

```
## # A tibble: 22 x 3
##    Variables                Tolerance    VIF
##    <chr>                        <dbl>  <dbl>
##  1 VehValue                   0.412     2.43
##  2 VehAgeoldest cars          0.622     1.61
##  3 VehAgeyoung cars           0.654     1.53
##  4 VehAgeyoungest cars        0.588     1.70
##  5 VehBodyConvertible         0.364     2.75
##  6 VehBodyCoupe               0.0586   17.1
##  7 VehBodyHardtop             0.0302   33.1
##  8 VehBodyHatchback           0.00350 285.
##  9 VehBodyMinibus             0.0634   15.8
## 10 VehBodyMotorized caravan   0.274     3.65
## # ... with 12 more rows
```

Il y a des variables avec un VIF plus grand que 10. Normalement il serait utile de retirer certaines variables. Par contre, ce sont seulement des catégories d'une variable qui sont affectées. Lorsqu'il y a une variable catégorielle avec *n* catégories dans le modèle, il y a *n-1* variables qui sont créées, qui sont des indicateurs 0,1 selon la catégorie représentée par la variable. Donc il est normal que ces variable souffrent de multicolinéarité. Ce n'est pas un problème, donc on garde tous les variables pour la modélisation.

Pour chaque modèle qui sera testé il y aura un terme d'offset. Celui-ci est l'exposure, c'est à dire la proportion de l'année que l'assuré est couvert. L'espérence de son nombre d'accident sera ainsi proportionel à la proportion de l'année couvert par l'assuré. La méthodologie sera la suivante:

1. Faire un glm avec la loi de poisson et un lien log. Ajuster le meilleur modèle en utilisant des techniques algorithmiques tel que le foward, backward et la combinaison des deux.
2. Faire un glm avec la loi de binomial négative et un lien log pour tenir compte de la surdispersion. Ajuster le meilleur modèle en utilisant des techniques algorithmiques tel que le foward, backward et la combinaison des deux.
3. Faire un glm avec la loi poisson et binomial négative pour les variables trouvées retenues en 1 et 2. Pour chaque combinaison, faire un test de rapport de vraissemblance pour voir si le modèle poisson est correct ou si on doit prendre le modèle avec la loi binomial négative.

Peu importe le modèle utilisé et la technique algorithmique utilisée, la sélection de variable est identique. Les variables qui sont utilisées sont *VehBody, DrivAge & VehAge*

À ce stade ci, on a un glm avec la loi de poisson et un avec la loi binomial négative. La loi binomial négative tient compte de la surdispersion. Un test des rapports de vraissemblance peut être fait pour déterminer si l'amélioration du modèle par la loi binomial négative est significative. La déviance du modèle de poisson est de *14 839* et la déviance du modèle avec la loi binomial négative est de *13 312*. La statistique de rapport de vraissemblance est de *1 527*. La pvalue associé est de 0. Donc on rejette l'hypothèse H0, on peut supposer qu'il y a de la variabilité extra poissonienne dans nos données.

Les variables du modèles finales sont *VehBody, DrivAge & VehAge*. Le modèle est un glm avec la loi binomial négative avec le lien log. Le paramètre du surdispersion est de 1.6951. Il y a un terme d'offset qui est l'exposure. Voici le modèle finale :

```
##
## Call:
## glm.nb(formula = ClaimNb ~ DrivAge + VehAge + VehBody, data = data_tp3,
##     weights = offset(Exposure), init.theta = 1.695107772, link = log)
##
```

```
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -0.717  -0.351  -0.269  -0.172   4.038
##
## Coefficients:
##                            Estimate Std. Error z value   Pr(>|z|)
## (Intercept)                 -1.4401     0.4049   -3.56    0.00038 ***
## DrivAgeolder work. people    0.1936     0.0633    3.06    0.00223 **
## DrivAgeoldest people         0.0332     0.0824    0.40    0.68667
## DrivAgeworking people        0.2164     0.0635    3.41    0.00065 ***
## DrivAgeyoung people          0.2914     0.0655    4.45 0.00000868 ***
## DrivAgeyoungest people       0.3872     0.0783    4.95 0.00000076 ***
## VehAgeoldest cars           -0.1194     0.0509   -2.35    0.01897 *
## VehAgeyoung cars             0.1369     0.0495    2.77    0.00569 **
## VehAgeyoungest cars          0.0885     0.0563    1.57    0.11569
## VehBodyConvertible          -1.6946     0.8482   -2.00    0.04573 *
## VehBodyCoupe                -0.7780     0.4332   -1.80    0.07252 .
## VehBodyHardtop              -0.9454     0.4167   -2.27    0.02327 *
## VehBodyHatchback            -1.2011     0.4037   -2.97    0.00293 **
## VehBodyMinibus              -1.0864     0.4450   -2.44    0.01464 *
## VehBodyMotorized caravan    -0.5649     0.5324   -1.06    0.28868
## VehBodyPanel van            -0.8852     0.4284   -2.07    0.03882 *
## VehBodyRoadster             -0.5355     0.8166   -0.66    0.51194
## VehBodySedan                -1.1138     0.4034   -2.76    0.00576 **
## VehBodyStation wagon        -1.0604     0.4037   -2.63    0.00863 **
## VehBodyTruck                -1.0423     0.4167   -2.50    0.01236 *
## VehBodyUtility              -1.2655     0.4093   -3.09    0.00199 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.6951) family taken to be 1)
##
##     Null deviance: 13403  on 67855  degrees of freedom
## Residual deviance: 13312  on 67835  degrees of freedom
## AIC: 20626
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.695
##           Std. Err.:  0.309
##
##  2 x log-likelihood:  -20582.183
```

La valeur relative du véhicule n'est pas associée au nombre de réclamations esperées. La relation entre la fréquence de sinistre et la valeur du véhicule ne fait pas beaucoup de sens. La valeur du véhicule serait plus associée à la sévérité des sinistres qu'à la fréquence.