Travail pratique 1

Travail présenté à

M. Thierry Duchesne

Dans le cadre du cours

Théorie et applications des méthodes de régression
STT-7125

Réalisé par l'équipe 7 :

Alexandre Lepage
& Amedeo Zito

le 2 novembre 2020

Faculté des sciences et de génie
École d'actuariat
Université Laval

1 Introduction

Les méthodes de régression linéaires sont fort utiles afin d'identifier des variables pouvant expliquer un comportement ou un phénomène et elles peuvent s'avérer efficaces pour faire de la prédiction si les données disponibles sont appropriées.

L'objet de ce travail est d'expérimenter l'utilisation de cette famille de modèles afin de résoudre trois problèmes de nature différente. La premier d'entre eux consiste à réaliser une régression linéaire afin de prédire le taux de mortalité à partir de variables mesurant la pollution environnementale et les caractéristiques socio-démographiques de 60 localités. Le second problème consiste à utiliser un modèle de régression afin de définir les facteurs associés à une hausse du risque d'un diagnostic positif de maladie coronarienne. En ce qui a trait au dernier problème, celui-ci s'inscrit dans un contexte d'assurance automobile et consiste à construire un modèle visant à voir s'il y a une association entre les caractéristiques du véhicule et de l'assuré et le nombre de réclamations.

2 Analyse et traitement de la multicollinéarité

La force des modèles de régression linéaire provient de l'hypothèse que la matrice de schéma $\boldsymbol{X}$ est de plein rang ; c.-à-d. qu'aucune colonne n'est linéairement dépendante des autres colonnes. Ce faisant, on s'assure qu'il n'existe qu'un seul inverse possible à la matrice $\boldsymbol{X}'\boldsymbol{X}$; de ce fait, on s'assure également que le vecteur des paramètres du modèle $\hat{\boldsymbol{\beta}}$ soit unique (un seul minimum à la fonction de perte utilisée pour l'entraînement). Ainsi, s'il existe un problème de multicolinéarité, il en découlerait que la matrice de schéma $\boldsymbol{X}$ ne serait plus de plein rang et il pourrait exister plusieurs minimum locaux à la fonction de perte utilisée pour entraîner le modèle. Il en résulterait alors une instabilité dans la convergence des paramètres et la variance de certaines de ces composantes serait démesurément grande. Pour cette raison, avant même de réaliser la sélection de modèle pour chacune des étapes, il faut d'abord mesurer le degré de multicollinéarité entre les potentielles variables explicatives disponibles.

Un outil bien pratique pour détecter la présence de multicollinéarité est le facteur d'inflation de la variance (VIF). Ce dernier peut être calculé avec la fonction `ols_vif_tol()` du *package* `olsrr` en `R`. Cependant, si les données utilisées comportent une ou plusieurs variables catégorielles, cette mesure n'est plus adéquate (voir [Fox and Monette, 1992]). Dans ce cas, on préférera le facteur d'inflation généralisé (GVIF) et sa version standardisée : $(\text{GVIF}_j)^{1/(2p_j)}$, où p_j correspond au nombre de degrés de libertés (le nombre de paramètres) rattachés à la j^{e} variable explicative du modèle, $j \in \{1, \ldots, p'\}$. À noter que lorsque $p_j = 1$, alors $\text{GVIF}_j = \text{VIF}_j$. Afin de mesurer cette métrique, Fox et Monette ont créé la fonction `vif` du *package* `car`. De façon générale, plusieurs auteurs suggèrent de considérer que $\text{VIF}_\text{j} > 10$ pourrait signaler un problème de multicolinéarité. En suivant cette logique, on peut considérer qu'il y a problème de multicolinéarité si $(\text{GVIF}_j)^{1/(2p_j)} > \sqrt{10} \approx 3.16$, pour $j = 1, \ldots, p'$.

Si, effectivement, un problème de multicolinéarité est détecté, alors la fonction `ols_eigen_cindex()` du *package* `olsrr` permet de faire un diagnostique plus approfondi. En effet, cette fonction permet de calculer les valeurs propres (*eigen values*) associées à la matrice des coefficients de corrélation échantillonnaux $\boldsymbol{X}^{\star\prime}\boldsymbol{X}^{\star}$. À partir de celles-ci, elle calcule des indices de conditionnement définis comme

$$\phi_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}}, \ j = 1, \ldots, p',$$

où λ_j correspond à la j-ème valeur propre de $\boldsymbol{X}^{\star\prime}\boldsymbol{X}^{\star}$ et $\lambda_{max} = \max(\lambda_1, \ldots, \lambda_{p'})$. Ces indices de conditionnement sont des indicateurs de la force de dépendance linéaire unissant certaines variables. Ainsi, on regardera la ligne du tableau correspondant à la valeur de ϕ_j qui est la plus élevée. Une règle du pousse veut que si $\phi_j > 30$, alors on est en présence de multicolinéarité. Dans ce cas, on considérera la proportion de la variance de $\hat{\beta}_l$ qui est expliquée par la j-ème dépendance linéaire, pour $l = 1, \ldots, p'$. C.-à-d.

$$p_{lj} = \frac{v_{lj}^2/\lambda_j}{c_{jj}},$$

où v_{lj} correspond au l-ème élément du vecteur propre $\boldsymbol{v}_j$ associé à la j-ème valeur propre λ_j et $c_{jj} = \sum_{l=1}^{p'} v_{lj}^2/\lambda_j$. Si $p_{lj} > 0.6$, alors on conclu que la l-ème variable explicative contribue à la j^{e} multicolinéarité et cause problème.

Dans ce cas, les solutions possibles consistent à appliquer une transformation non linéaire sur les variables explicatives (p.ex. transformation logarithmique ou racine carrée), réduire la dimension de la matrice de schéma en retirant la variable la plus problématique. S'il y a plusieurs valeurs de l pour lesquelles $p_{lj} > 0.5$ pour un même j, alors il est possible de les regrouper à l'aide d'une moyenne. Par exemple, pour $j = \underset{j \in \{1,\ldots,p'\}}{\text{argmax}} \{\phi_j\}$, si on a $p_{1j} > 0.5$ et $p_{2j} > 0.5$, alors on peut combiner x_1 et x_2 de la façon suivante : $(x_1 + x_2)/2$.

Après avoir réalisé ces étapes, il faut recommencer itérativement ce processus jusqu'à ce qu'il n'y ait plus de problème soulevé par l'analyse du VIF ou du GVIF.

3 Question 1

Pour la première question de ce travail, on considère un jeu de données présentant des variables mesurant la pollution environnementale et les caractéristiques socio-démographiques de 60 localités. L'objectif de cette question est de valider qu'il est possible de prédire la mortalité d'une région en fonction de ces variables explicatives et de donner un estimé ponctuel ainsi qu'un intervalle de confiance à 95% sur une observation de donnée.

3.1 Traitement de la multicollinéarité

Pour atteindre cet objectif, la première étape consiste à réaliser une analyse de multicollinéarité telle que décrite dans la section 2. Se faisant, on découvre que les VIFs pour les variables A12 et A13 sont supérieures à 10. Afin de traiter ce problème, on peut commencer par regarder s'il est pertinent d'effectuer une transformation non linéaire de certaines variables explicatives. Afin de visualiser les options envisageable, les illustrations 4 à 8 présentent l'effet d'une transformation logarithmique (au centre) et celui d'une transformation racine carrée (à droite) sur la relation existant entre la variable endogène et chacune des variables explicatives. Si on voit que cette transformation améliore la relation de linéarité existant entre les variables en question, alors on procède à la transformation appropriée. Suite à l'analyse des illustrations 4 à 8, on en vient à considérer que les variables A12 et A13 méritent à recevoir une transformation logarithmique, de même que les variables A9 et A14 profiteraient à recevoir une transformation racine carrée. Afin de valider ces observations, on teste différents modèles utilisant plusieurs combinaisons de transformations. Il advient que le modèle complet possédant l'AIC le plus petit est le suivant :

$$\begin{aligned} \text{B} \sim \text{A1} + \text{A2} + \text{A3} + \text{A4} + \text{A5} + \text{A6} + \text{A7} + \text{A8} + \text{I(sqrt(A9))} \\ + \text{A10} + \text{A11} + \text{I(log(A12))} + \text{I(log(A13))} + \text{I(log(A14))} + \text{A15}. \end{aligned} \tag{1}$$

Une fois que ces transformations sont effectuées, on refait l'analyse des VIFs. Puisque celle-ci nous indique que la multicollinéarité n'est toujours pas réglée, on procède aux étapes décrites dans la section 2 après quoi on trouve le modèle suivant :

$$\text{B} \sim \text{A1} + \text{A2} + \text{A8} + \text{A11} + \text{I(log(A14))} + \text{I(log(sqrt(A12} * \text{A13)))}. \tag{2}$$

À titre comparatif, nous avons voulu tester s'il était possible d'avoir un modèle plus performant si le traitement de la multicolinéarité était fait par sélection de variables en utilisant une régression LASSO. Avec la fonction `glm.net` du *package* du même nom et avec le paramètre `alpha=1`, on essaie plusieurs valeurs pour le terme de pénalité λ en commençant par celui qui est le plus inclusif (lambda le plus petit). Ainsi, on trouve que la valeur de λ qui est minimal tout en minimisant la statistique de déviance et en éliminant la multicolinéarité est $\lambda = 5.98761443432345$. Avec ce dernier, on trouve le modèle suivant :

$$\text{B} \sim \text{A1} + \text{A2} + \text{A6} + \text{A7} + \text{A8} + \text{I(log(A13))} + \text{I(log(A14))} + \text{I(sqrt(A9))}. \tag{3}$$

Ainsi, si on compare les modèles (2) et (3), on trouve que le modèle (3) est celui qui minimise l'AIC et qui maximise également la statistique du R^2 de prédiction. Conséquemment, il s'agira du modèle de base utilisé pour la sélection des variables.

3.2 Sélection des variables explicatives

On se rappelle que l'objectif de cette question est de construire un modèle prédictif. Pour se faire, l'idéal est de produire tous les sous-modèles possibles découlant de (3). Cette opération peut être réalisée avec la fonction `ols_step_all_possible` du *package* `olsrr`. Une fois que tous les sous-modèles sont produits, on regarde les 3 modèles qui maximisent la statistique du R^2 de prédiction comme le démontre l'illustration 1.

	mindex	n	predictors	rsquare	adjr	predrsq	cp	aic	sbic	sbc	msep	fpe	apc	hsp
	<int>	<int>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	219	6	A1 A2 A6 A8 I(log(~	0.781	0.756	0.715	8.78	590.	421.	607.	56727.	1054.	0.277	18.2
2	220	6	A1 A2 A7 A8 I(log(~	0.773	0.748	0.703	10.7	592.	422.	609.	58703.	1091.	0.287	18.8
3	247	7	A1 A2 A6 A7 A8 I(l~	0.788	0.760	0.715	8.92	590.	422.	609.	55876.	1053.	0.277	18.2

Illustration 1 – Résultats des trois meilleurs modèles selon la statistique du R^2 de prédiction.

Comme les modèles 219 et 247 sont très comparables, on choisira celui qui est le plus simple étant donné que le jeu de données d'entraînement comporte très peu d'observations (60). Ainsi, le modèle sélectionné est le suivant :

$$\mathrm{B} \sim \mathrm{A1} + \mathrm{A2} + \mathrm{A6} + \mathrm{A8} + \mathrm{I}(\log(\mathrm{A13})) + \mathrm{I}(\mathrm{sqrt}(\mathrm{A9})). \tag{4}$$

À ce stade, il pourrait être intéressant de voir si des interactions entre les variables amélioreraient le pouvoir de prédiction. Pour identifier celles qui sont intéressantes, on applique le test F partiel sur toutes les interactions de 1er ordre possible. Se faisant, on trouve qu'aucune des interactions n'est significative au seuil de 1%. Donc, le modèle (4) correspond à notre modèle final.

Avec la fonction `ols_regress` du *package* `olsrr`, on peut calculer plusieurs statistiques d'intérêt pour décrire les performances du modèle final. D'une part, on a un R^2 de prédiction qui est de 71%, ce qui est très bon considérant que la mortalité est un phénomène complexe auquel il est impossible de décrire à 100% avec des variables explicatives. D'autre part, la statistique de Wald nous indique que chacune des variables explicatives incluses dans le modèle est significative à un seuil de 5%.

3.3 Calcul de la prédiction

Maintenant que l'on a un modèle appréciable, on aimerait calculer un estimé ponctuel ainsi qu'un intervalle de confiance à 95% pour le taux de mortalité à un endroit pour lequel les variables A1 à A15 valent respectivement

$$40\ 30\ 80\ 9\ 3\ 10\ 77\ 4100\ 13\ 46\ 15\ 25\ 26\ 145\ 55.$$

Avec la fonction `predict` du *package* `stats`, on obtient un estimé de $\hat{B} = 999.4799$ ainsi qu'un intervalle de prédiction correspondant à $B \in [936.0751, 1062.885]$.

4 Question 2

Pour la deuxième question de ce travail, on présente une base de données avec 13 variables explicatives, chacune mesurant une métrique médicales du corps humain. L'objectif est de construire un modèle de régression qui estime la probabilité de diagnostic positif pour la maladie coronarienne afin d'identifier les variables associées a une hausse du risque de développer la maladie.

4.1 Analyse préliminaire

D'abord, en jetant un œil sur les données, on se rend compte qu'un prétraitement est nécessaire. Effectivement, les variables `ca` et `thal` contiennent des points d'interrogation `"?"` puisque, pour certaines observations, ces métriques n'ont pas été calculées. Comme ces cas sont peu nombreux (6 observations), ils sont simplement écartés du jeu de données. D'autres modifications ont été apportées aux variables afin que le format des données soit cohérent avec leur nature. Pour plus de détail, voir le code informatique fournit en pièce jointe.

4.2 Traitement de la multicollinéarité

Comme pour la première question, afin de construire un modèle de prédiction, la première étape est de réaliser une analyse de multicolinéarité conformément à la section 2. Comme cette étape ne concerne que la matrice $\boldsymbol{X}$, il n'est pas nécessaire d'entraîner un GLM à ce stade pour réaliser cette analyse. Ainsi, il est possible d'utiliser la fonction `ols_vif_tol()` en entraînant un modèle linéaire sur une variable réponse bidon. Se faisant, on trouve qu'aucune variable explicative ne semble a priori dépendre linéairement des autres. Cependant, on se rappelle que, pour les variables catégorielles, cette méthode est peu fiable. Pour cette raison, on regarde également la métrique du VIF généralisé (GVIF). Comme $(\text{GVIF}_j)^{1/(2p_j)} < 3.16$, pour $j = 1, \ldots, 13$, alors on déduit qu'il n'y a pas de problème de multicolinéarité et on peut procéder à la sélection des variables explicatives.

4.3 Sélection des variables explicatives

Pour cette question, comme la variable endogène est catégorielle, le type de modèle de régression sélectionné pour réaliser la tâche est le modèle linéaire généralisé de la famille binomiale. De plus, comme l'objectif est d'interpréter les paramètres du modèle, la fonction de lien retenue correspond à la fonction canonique, c.-à-d. la fonction logit.

Dans un premier temps, avant de retirer des variables explicatives du modèle complet, il est pertinent de voir si certaines transformations appliquées sur les variables explicatives numériques pourraient améliorer l'explicabilité de la variable endogène. Pour se faire, on peut entraîner un modèle additif généralisé (GAM) à l'aide de la fonction `gam` qui provient du *package* du même nom. Par la suite, il suffit de passer le modèle entraîné dans la fonction `plot` afin de visualiser graphiquement les transformations réalisées par la fonction `gam`. Si la courbe affichée est linéaire, cela signifie qu'aucune transformation n'est nécessaire. Autrement, la forme de la courbe donne une idée de la transformation à effectuer. Dans le cas présent, aucune transformation n'est nécessaire.

Dans un deuxième temps, afin de sélectionner les variables explicatives, la fonction `glmbb` du *package* du même nom permet de tester tous les sous-modèles possibles en utilisant l'AIC comme critère de performance. Après avoir exécuté cette fonction, on sélectionne les cinq modèles minimisant ce critère. Étant donné que l'objectif est d'expliquer un phénomène et non de réaliser une prédiction, les cinq modèles restant sont comparés à l'aide du la statistique du R^2 ajusté. Ainsi, on trouve que le meilleur modèle correspond à (5).

$$\text{Y} \sim \text{thal} + \text{ca} + \text{cp} + \text{oldpeak} + \text{slope} + \text{sex} + \text{trestbps} + \text{exang} + \text{thalach}. \tag{5}$$

Les méthodes algorithmiques d'inclusion (*forward*), d'exclusion (*backward*) et pas-à-pas (*stepwise*) ont également été testées avec la fonction `stepAIC` du *package* `MASS` ; ce faisant, on obtient le même résultat dans tous les cas.

On désire maintenant voir si des interactions amélioreraient l'explicabilité de la variable endogène. On procède donc avec la méthode pas-à-pas avec le test du ratio des vraisemblance pour un seuil d'inclusion et d'exclusion de 5%. Ce faisant, on trouve le modèle (6).

$$\text{Y} \sim \text{sex} + \text{cp} + \text{trestbps} + \text{oldpeak} + \text{slope} + \text{ca} + \text{thal} + \text{ca} : \text{thal}. \tag{6}$$

4.4 Facteurs explicatifs

Afin de répondre a la question on analyse la sortie `R` de la fonction `summary` pour le modèle (6) tel que présenté dans l'illustration 2.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -8.86454    1.93915  -4.571 4.85e-06 ***
sex1           1.40717    0.49583   2.838 0.004539 **
cp2            1.61196    0.81927   1.968 0.049120 *
cp3            0.40454    0.71319   0.567 0.570563
cp4            2.95468    0.72985   4.048 5.16e-05 ***
trestbps       0.02424    0.01068   2.270 0.023186 *
oldpeak        0.51880    0.22521   2.304 0.021242 *
slope2         1.51399    0.45766   3.308 0.000939 ***
slope3         0.64047    0.89281   0.717 0.473145
ca             1.48959    0.33258   4.479 7.50e-06 ***
thal6         -1.20941    0.99214  -1.219 0.222848
thal7          1.89246    0.48427   3.908 9.31e-05 ***
ca:thal6      17.14674 1045.29619   0.016 0.986912
ca:thal7      -0.88747    0.51038  -1.739 0.082060 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Illustration 2 – Sortie `R` de la fonction `summary` pour le modèle (6).

La première chose que l'on remarque dans l'illustration 2, c'est que l'écart-type de l'estimateur associé à l'interaction `ca:thal6` est démesurément grand. Pourtant, une analyse de la multicollinéarité ne soulève aucun problème. On en déduit que cette variable n'est pas utile et pourrait probablement être regroupée avec `thal3`. Dans ce cas, l'interprétation de la variable `thal` ainsi modifiée serait que le défaut soit réparable ou non. Si on procède à cette modification, le sommaire du modèle devient tel que présenté dans l'illustration 3.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.97445    1.88804  -4.753 2.00e-06 ***
sex1         1.40297    0.47341   2.964 0.003041 **
cp2          1.66449    0.80518   2.067 0.038713 *
cp3          0.62225    0.69737   0.892 0.372240
cp4          2.98252    0.70695   4.219 2.46e-05 ***
trestbps     0.02370    0.01047   2.264 0.023587 *
oldpeak      0.49459    0.21834   2.265 0.023502 *
slope2       1.59109    0.44473   3.578 0.000347 ***
slope3       0.51161    0.86810   0.589 0.555630
ca           1.64857    0.32852   5.018 5.22e-07 ***
thal7        2.00222    0.46940   4.266 1.99e-05 ***
ca:thal7    -1.06758    0.50348  -2.120 0.033972 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Illustration 3 – Sortie `R` de la fonction `summary` pour le modèle (6) lorsque les classes 3 et 6 sont regroupées pour la variable catégorielle `thal`.

Désormais, toutes les classes de la variable *thal* et de son interaction sont significative au seuil de 5% selon le test de Wald et le problème de l'écart-type démesuré est réglé. On peut donc procéder à l'analyse de l'association des variables explicatives avec la variable endogène.

Soit π, la probabilité qu'un diagnostic de maladie coronarienne soit positif. On a

$$\pi = \text{logistique}(\eta) = \text{logit}^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}},$$

où $\eta = \boldsymbol{x}'\boldsymbol{\beta}$ et $\boldsymbol{x}'\boldsymbol{\beta}$ est représentée par (6). Par dérivation en chaîne, on a

$$\frac{\partial \pi}{\partial \boldsymbol{\beta}} = \frac{\partial \pi}{\partial \eta}\frac{\partial \eta}{\partial \boldsymbol{\beta}} = \frac{e^{\eta}(1+e^{\eta}) - e^{2\eta}}{(1+e^{\eta})^2}\boldsymbol{x} = \frac{e^{\eta}}{(1+e^{\eta})^2}\boldsymbol{x} = \pi(1-\pi)\boldsymbol{x}. \tag{7}$$

Comme $\pi \in [0,1]$, alors $\pi(1-\pi) \in [0,1]$. Ainsi, si $\boldsymbol{x}$ est un vecteur de valeurs positives, on déduit avec (7) que la fonction logistique est croissante par rapport aux coefficients. Cela signifie que si on augmente une variable explicative d'une unité, la prédiction de la probabilité de diagnostic positif augmentera si le coefficient visé est positif et elle diminuera dans le cas contraire. Conséquemment, les variables exogènes ayant un estimateur positifs seront associées à une hausse du risque d'un diagnostic positif de maladie coronarienne. Ces dernières sont :

— `sex` : le sexe,

— `cp` : la nature des douleurs à la poitrine,

— `trestbps` : la tension artérielle au repos,

— `oldpeak` : la baisse du segment ST induite par l'exercice par rapport au repos,

— `slope` : la pente du segment ST lors de l'exercice maximal.

En ce qui attrait aux variables explicatives `ca` et `thal`, l'association positive est moins claire étant donnée l'interaction qui existe entre elles. En effet, les coefficients associés à ces variables sont donnés respectivement par (8) et (9) .

$$(1.64857 - 1.06758 \times \texttt{thal7}) \tag{8}$$

$$(2.00222 - 1.06758 \times \texttt{ca}) \tag{9}$$

On voit donc avec (8) que le coefficient de `ca` est toujours positif, peu importe la valeur de `thal7` puisque $1.64857 - 1.06758 > 0$. Cependant, pour `thal7`, la polarité du coefficient dépendra de la valeur de `ca` puisque $\texttt{ca} \in \{0, 1, 2, 3\}$.

5 Question 3

Pour cette question, on utilise le jeu de données `ausprivauto0405` disponible sur le *package* `R CASdatasets`. Celui-ci contient des données sur les polices d'assurance automobile d'une durée d'un an contractées en Australie entre 2004 et 2005. On y trouve trois caractéristiques du véhicule assuré et deux autres relatives au détenteur de police. On y voit aussi l'exposition (proportion de l'année où la police est active) et le nombre d'accidents survenus lors de la couverture de la police. L'objectif est de construire un modèle de régression qui permettrait d'identifier les variables ayant une association significatives avec le nombre de réclamations. Puisque la variable endogène correspond à un dénombrement, un GLM Poisson est choisi pour réaliser la tâche. Afin de simplifier le modèle et l'interprétation, la fonction de lien log est retenue. Par ailleurs, puisque le dénombrement des accidents est proportionnel à l'exposition, on utilisera cette dernière variable comme terme d'*offset*.

5.1 Traitement de la multicollinéarité

Comme dans la section 4.2, on est en présence de plusieurs variables catégorielles. Conséquemment, la métrique appropriée pour vérifier la multicolinéarité est le VIF généralisé qui est calculé avec la fonction `vif` du *package* `car`. Comme $(\text{GVIF}_j)^{1/(2p_j)} < 3, j = 1, 2, \ldots, 5$, alors on déduit qu'il n'y a pas de problème de multicolinéarité.

5.2 Traitement de la surdispersion

On sait que la loi de Poisson a une espérance égale à sa variance. En termes des lois de la famille exponentielle, cela implique que le paramètre de dispersion ϕ doit être très près de 1. Or, l'estimateur de la méthode des moments de ce paramètre, qui est calculé à l'aide de (10), donne une valeur de 0.373606. On déduit donc que le modèle est sous-dispersé.

$$\hat{\phi}_D = D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})/(n - p), \tag{10}$$

où $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ correspond à la statistique de déviance du modèle, n représente le nombre d'observations, soit $n = 67\,856$, et p correspond au nombre de paramètres du modèle, soit $p = 23$. La façon la plus commune de traiter les données de dénombrement avec variabilité extra-poissonienne est de supposer un modèle de type binomiale négative. Étant donnée que le modèle de Poisson est une simplification du modèle binomial négative (paramètre $\alpha = 0$ dans le cas poissonien), on peut procéder au test du ratio des vraisemblances pour déterminer si le modèle binomial négative améliore significativement la vraisemblance par rapport au GLM Poisson. Cependant comme le test d'hypothèse est hors norme (le test d'hypothèse est unilatéral), il faut adapter le calcul du seuil observé avec (11).

$$p_{\text{value}} = 0.5 \times \mathbb{P}[\chi^2_1 > 2(l_1 - l_0)], \tag{11}$$

où l_1 et l_0 correspondent respectivement à la log-vraisemblance pour le modèle complet (binomial négative) et pour le modèle simplifié (Poisson). Ainsi, en calculant (11) sur le jeu de données, on trouve un seuil observé de 1.904829×10^{-10}. On rejette donc l'hypothèse nulle que le modèle de Poisson est suffisant pour conclure que le modèle binomial négative est mieux adapté aux données.

5.3 Sélection des variables explicatives

Dans un premier temps, afin de vérifier si une transformation de la variable `VehValue` améliorerait le modèle, on peut utiliser la fonction `gam` comme dans la section 4.3. Se faisant, on réalise que non, aucune transformation n'est requise.

Par la suite, pour faire la sélection des variables explicatives, il aurait été idéal de tester tous les sous-modèles possibles pour sélectionner celui qui maximise l'explicabilité de la variable endogène. Cependant, avec `R`, la fonction `glmbb` ne peut être appliqué avec la loi binomiale négative. La fonction `glmulti` provenant du *package* du même nom nécessite l'installation de `JavaScript`. Pour les fins de ce travail, afin d'éviter des difficultés de correction, nous avons décidé de ne pas tester la fonction. Pour ces raisons, les méthodes algorithmiques ont été privilégiées. Ainsi, la fonction `stepAIC` du *package* `MASS` permet de trouver le modèle (12), peu importe qu'on soit sous l'approche d'inclusion, d'exclusion ou pas-à-pas.

$$\text{ClaimNb} \sim \text{DrivAge} + \text{VehAge} + \text{VehBody} + \text{offset}(\log(\text{Exposure})). \tag{12}$$

Il ne reste plus qu'à regarder si l'inclusion d'une interaction améliorerait les performances du modèle. Avec un test du ratio de vraisemblance, on trouve qu'aucune interaction n'est significative au seuil de 5%.

5.4 Description du modèle final

Dans un premier temps, on s'intéresse à savoir quelles variables contribuent à augmenter l'espérance du nombre de sinistres automobiles qui surviennent dans une année. Pour se faire, on peut calculer la dérivée première de cette espérance par rapport aux coefficients de régression. Soit μ, l'espérance du nombre d'accidents automobiles survenus dans une année. Avec le lien log, on a

$$\mu = e^{\eta}, \tag{13}$$

où $\eta = \boldsymbol{x'\beta}$ et $\boldsymbol{x'\beta}$ est représenté par (12). Par dérivation en chaîne, on a

$$\frac{\partial \mu}{\partial \boldsymbol{\beta}} = \frac{\partial \mu}{\partial \eta}\frac{\partial \eta}{\partial \boldsymbol{\beta}} = e^{\eta}\boldsymbol{x} = \mu\boldsymbol{x}. \tag{14}$$

Comme $\mu \geq 0$, alors, si $\boldsymbol{x}$ est un vecteur de valeurs positives, on déduit avec (14) que l'espérance du nombre de sinistres survenus dans une année est une fonction croissante par rapport aux coefficients de la régression.

L'un des grands avantages de la fonction de lien log pour la régression de Poisson et binomiale négative, c'est sa facilité d'interprétation. En effet, si une variable x_j augmente d'une unité, alors l'espérance du nombre d'accident dans une année sera multipliée par e^{β_j}. Ces valeurs multiplicatives sont calculées pour chacune des variables explicatives dans les tableaux 1 à 3.

Dans un premier temps, interprétons ce que le coefficient $\hat{\beta}_0$ veut dire dans le modèle. Dans cette situation, on considère un assuré relativement âgé (`old people`), qui conduit un vieux bus. Dans ce cas, l'espérance du nombre d'accident pour ce genre d'assuré, pour une année, est de $e^{-1.17389} \approx 0.309$. Par la suite, les facteurs présentés dans les tableaux 1 à 3 viennent augmenter ou diminuer cette espérance selon les caractéristiques de l'assuré et de son véhicule.

Variables	Facteurs
DrivAgeoldest people	1.01
DrivAgeolder work. people	1.25
DrivAgeworking people	1.28
DrivAgeyoung people	1.36
DrivAgeyoungest people	1.62

Tableau 1 – Facteurs multiplicatifs pour la variable `DriveAge`.

Lorsque l'on regarde la variable `DriveAge`, on voit que la catégorie d'âge des assurés qui est la moins risquée correspond aux gens âgés (`old people`) et qui est associée au coefficient β_0. La catégorie d'âge la plus risquée, quant à elle, correspond aux plus jeunes conducteurs (`youngest people`) puisque le facteur multiplicatif correspondant est le plus élevé. On remarque que le risque diminue tandis que l'âge augmente jusqu'à ce que l'on arrive à un âge très respectable (`oldest people`). Dans ce cas, le risque réaugmente étant donnée que les réflexes du conducteur diminuent, de même que sa vision.

Variables	Facteurs
VehBodyConvertible	0.22
VehBodyUtility	0.33
VehBodyHatchback	0.37
VehBodyMinibus	0.38
VehBodyTruck	0.38
VehBodySedan	0.40
VehBodyStation wagon	0.41
VehBodyPanel van	0.42
VehBodyHardtop	0.44
VehBodyRoadster	0.59
VehBodyCoupe	0.61
VehBodyMotorized caravan	0.71

Tableau 2 – Facteurs multiplicatifs pour la variable `VehBody`.

Du point de vue du type de véhicule, on voit que la classe qui maximise l'espérance du nombre d'accident est le bus (associé à l'estimateur β_0) puisque tous les facteurs multiplicatifs sont inférieurs à 1. Le type de véhicule qui minimise cette espérance est le convertible. Si on analyse ces résultats plus en profondeur, on remarque que les véhicules les plus risqués, mis à part le bus, sont des véhicules sports (`roadster`, `coupe`) et ceux minimisant le risque sont des véhicules généralement conduits par des retraités (`convertibles`). Ainsi, bien que nous n'ayons pas observé d'interaction significative avec le test du rapport des vraisemblances, on remarque que certaines catégories de véhicule pourraient être associées à des groupes d'âge.

Variables	Facteurs
VehAgeoldest cars	0.93
VehAgeyoungest cars	1.09
VehAgeyoung cars	1.14

Tableau 3 – Facteurs multiplicatifs pour la variable `VehAge`.

Finalement, en ce qui attrait à l'âge du véhicule, on a que la classe associée au coefficient β_0 est `old cars`. La classe la plus risquée est attribuée à `young cars` et celle qui est la moins sujette à un nombre élevé d'accident est `oldest cars`.

On remarquera que le modèle final n'inclut pas la valeur relative du véhicule (`VehValue`). Si on applique un test du ratio des vraisemblances, on remarque que le seuil observé est de 0.1782. Comme ce seuil est très élevé, on déduit que cette variable n'est pas utile pour prédire le nombre d'accidents automobiles. En revanche, elle aurait pu être significative si on avait étudié le montant des sinistres. Dans ce cas, la valeur du véhicule aurait pu être utilisée comme variable d'exposition.

Références

[Fox and Monette, 1992] Fox, J. and Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, 87(417) :178–183.

A Illustrations

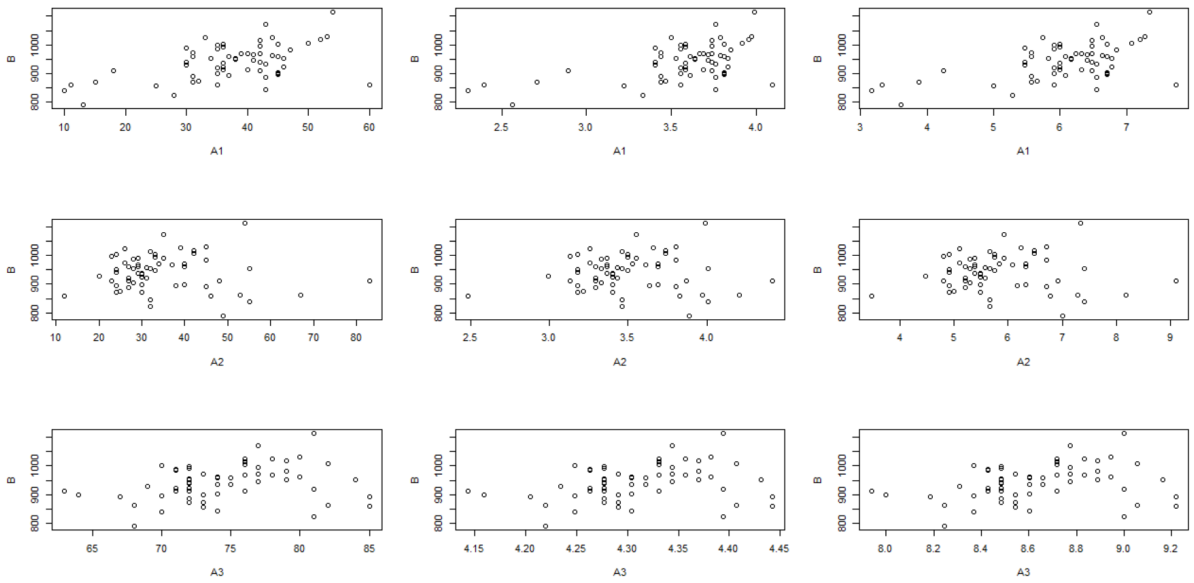

Illustration 4 – À gauche, on compare la variable endogène B avec les variables exogènes A1 à A3. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

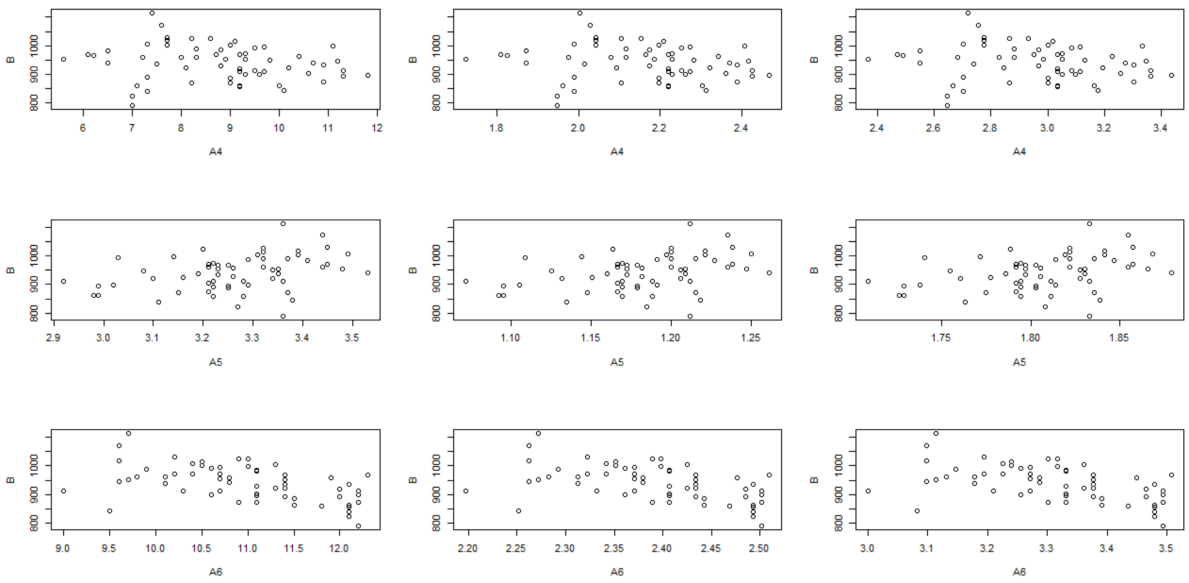

Illustration 5 – À gauche, on compare la variable endogène B avec les variables exogènes A4 à A6. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

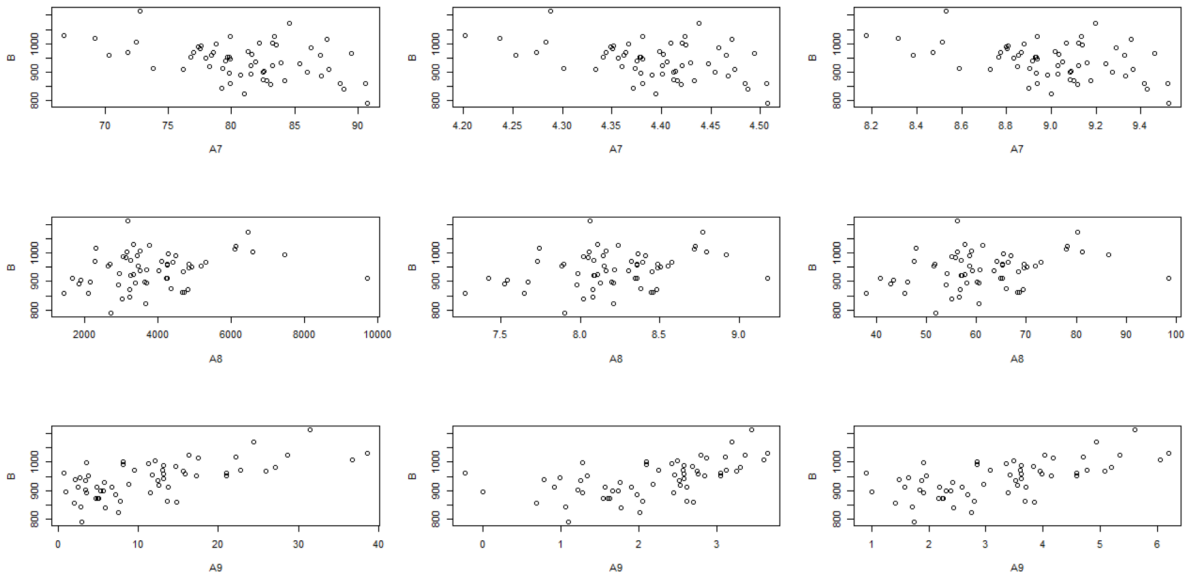

Illustration 6 – À gauche, on compare la variable endogène B avec les variables exogènes A7 à A9. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

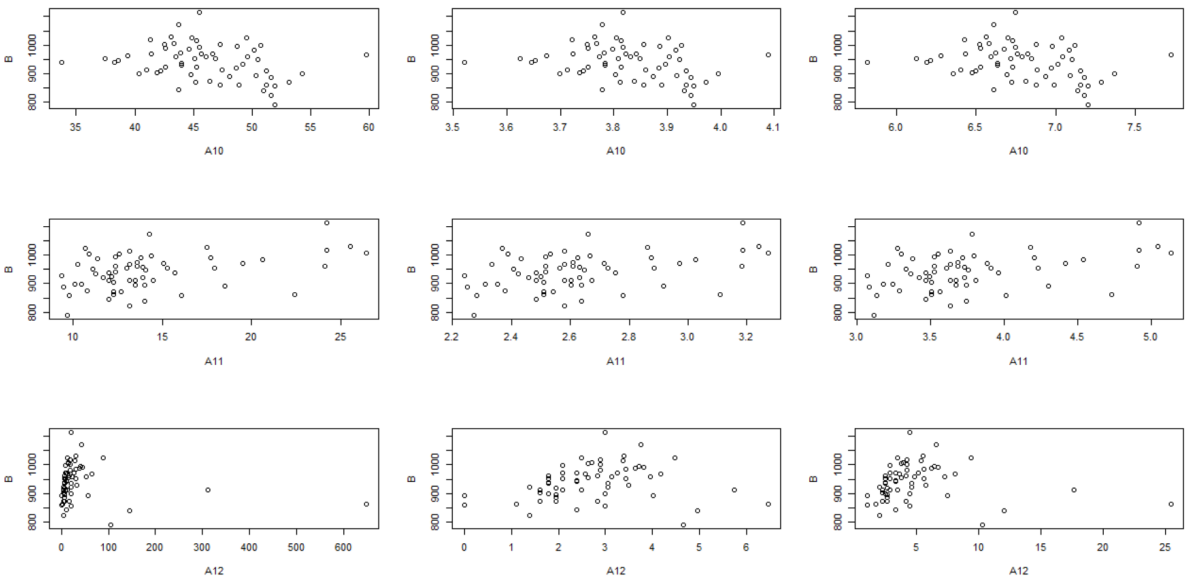

Illustration 7 – À gauche, on compare la variable endogène B avec les variables exogènes A10 à A12. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

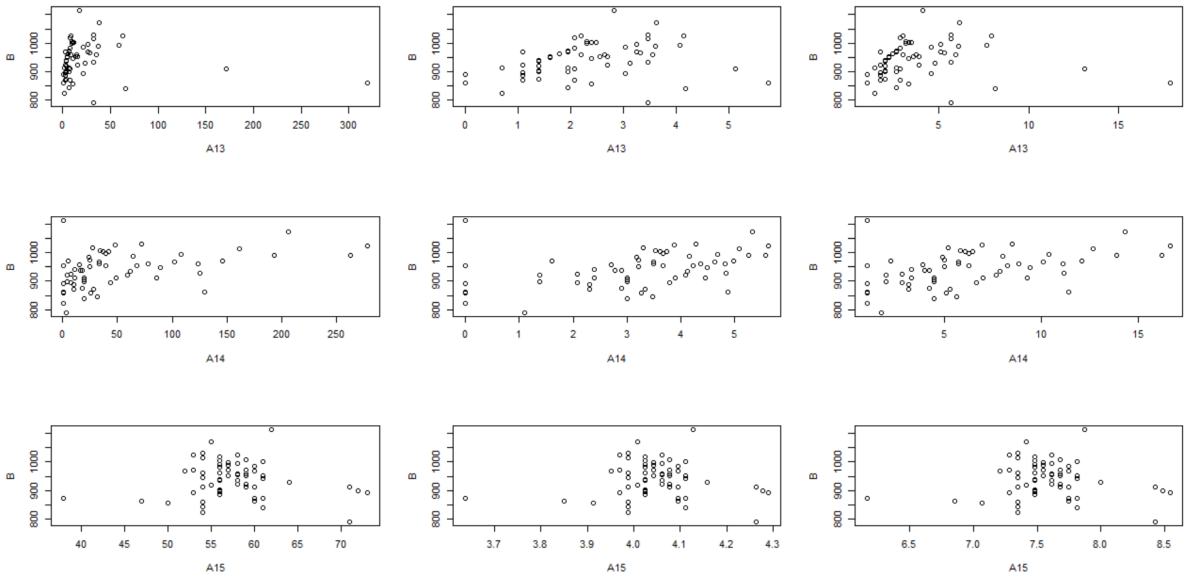

Illustration 8 – À gauche, on compare la variable endogène B avec les variables exogènes A13 à A15. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.