

---

---

# 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



UNIVERSITÉ  
**LAVAL**

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. Le 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  $\mathbf{X}$  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  $\mathbf{X}'\mathbf{X}$  ; de ce fait, on s'assure également que le vecteur des paramètres du modèle  $\hat{\beta}$  soit unique (un seul minimum à la fonction de perte utilisée pour l'entraînement). Ainsi, s'il existe un problème de multicollinéarité, il en découlerait que la matrice de schéma  $\mathbf{X}$  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  $j^{\text{e}}$  variable explicative du modèle,  $j \in \{1, \dots, p'\}$ . 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  $\text{VIF}_j > 10$  pourrait signaler un problème de multicollinéarité. En suivant cette logique, on peut considérer qu'il y a problème de multicollinéarité si  $(\text{GVIF}_j)^{1/(2p_j)} > \sqrt{10} \approx 3.16$ , pour  $j = 1, \dots, p'$ .

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

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

où  $\lambda_j$  correspond à la  $j$ -ème valeur propre de  $\mathbf{X}^{*'}\mathbf{X}^*$  et  $\lambda_{\max} = \max(\lambda_1, \dots, \lambda_{p'})$ . 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  $\phi_j$  qui est la plus élevée. Une règle du pousse veut que si  $\phi_j > 30$ , alors on est en présence de multicollinéarité. Dans ce cas, on considérera la proportion de la variance de  $\hat{\beta}_l$  qui est expliquée par la  $j$ -ème dépendance linéaire, pour  $l = 1, \dots, p'$ . C.-à-d.

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

où  $v_{lj}$  correspond au  $l$ -ème élément du vecteur propre  $v_j$  associé à la  $j$ -ème valeur propre  $\lambda_j$  et  $c_{jj} = \sum_{l=1}^{p'} v_{lj}^2 / \lambda_j$ . Si  $p_{lj} > 0.6$ , alors on conclut que la  $l$ -ème variable explicative contribue à la  $j$ -ème multicollinéarité 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  $j = \underset{j \in \{1, \dots, p'\}}{\operatorname{argmax}} \{\phi_j\}$ , 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 envisageables, 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 :

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

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 :

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

À titre comparatif, nous avons voulu tester s'il était possible d'avoir un modèle plus performant si le traitement de la multicollinéarité 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  $\lambda$  en commençant par celui qui est le plus inclusif (lambda le plus petit). Ainsi, on trouve que la valeur de  $\lambda$  qui est minimal tout en minimisant la statistique de déviance et en éliminant la multicollinéarité est  $\lambda = 5.98761443432345$ . Avec ce dernier, on trouve le modèle suivant :

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

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(1~	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 :

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

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 Calcule 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 à 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 multicollinéarité conformément à la section 2. Comme cette étape ne concerne que la matrice  $\mathbf{X}$ , 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  $(GVIF_j)^{1/(2p_j)} < 3.16$ , pour  $j = 1, \dots, 13$ , alors on déduit qu'il n'y a pas de problème de multicollinéarité 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 de la statistique du  $R^2$  ajusté. Ainsi, on trouve que le meilleur modèle correspond à (5).

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

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 vraisemblances pour un seuil d'inclusion et d'exclusion de 5%. Ce faisant, on trouve le modèle (6).

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

## 4.4 Facteurs explicatifs

Afin de répondre à 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  $\pi$ , 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  $\eta = \mathbf{x}'\boldsymbol{\beta}$  et  $\mathbf{x}'\boldsymbol{\beta}$  est représentée par (6). Par dérivation en chaîne, on a que

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

Comme  $\pi \in [0, 1]$ , alors  $\pi(1 - \pi) \in [0, 1]$ . Ainsi, pour  $\mathbf{x}$  fixé, on déduit avec (7) que la fonction logistique est croissante par rapport aux coefficients. Cela signifie que si on augmente une seule 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) .

$$(1.64857 - 1.06758 \times \text{thal7}) \quad (8)$$

$$(2.00222 - 1.06758 \times \text{ca}) \quad (9)$$

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  $\text{ca} \in \{0, 1, 2, 3\}$ .

## 5 Question 3

Cette question utilise les données d'assurance automobile australienne qui contient les valeurs de 3 caractéristiques du véhicule et 2 caractéristiques de l'assuré. L'objectif est de construire un modèle sur le nombre de réclamation par police. Plus précisément, on veut identifier lesquelles variables explicatives sont associée avec le nombre de réclamation. Puisqu'on cherche de modéliser un dénombrement, on modélise le nombre de réclamation avec une loi Poisson ou Binomial Négative. La fonction de lien pour le GLM est le logarithme. De plus, on applique un offset sur la variables **Exposure**, car elle est proportionnel au nombre de risque

### 5.1 Traitement de la multicollinéarité

On commence par une analyse sur la multicollinéarité. Puisqu'on a principalement des variables catégorielles, il faut utiliser le VIF agrégée. La fonction **vif** du package **car** nous permet de calculer le VIF agrégée. Finalement, on n'observe aucune valeur supérieur a 10, donc la multicollinéarité n'est pas un problème.

### 5.2 Traitement de la surdispersion

On sait que la loi Poisson a une espérance égal a sa variance. Si on calcule la moyenne et la variance empirique du nombre de réclamation, on obtient 0.07276 et 0.07739737 respectivement. Sachant que la nombre de réclamation dominant la base de donnée est égal a 0, on peut conclure en premier vue que la variance et la moyenne ne sont pas égal. Il est donc probable que la loi Binomiale Négative est plus appropriée. La valeur phi du modèle complet avec la loi Poisson est de 0.373606. La valeur devrait être proche de 1 pour la loi Poisson, c est donc un autre indice de sous-dispersion. Afin de choisir définitivement en la loi appropriée, on fait un test du ratio de vraisemblance. L'hypothèse nulle est le modèle complet GLM avec la loi Poisson. La contre-hypothèse est le modèle complet GLM avec la loi Binomiale Négative. La valeur p de la statistique est proche de 0, on rejette l'hypothèse nulle. La vraisemblance est donc significativement meilleur pour la loi Binomiale Négative. Utiliser l'AIC ou le BIC comme critère donne la même conclusion.

### 5.3 Sélection des variables explicatives

Comme pour la question 2, section 4.3, on utilise la fonction **R stepAIC** et les trois approches différents pour trouver le meilleur sous-modèle. Les trois approches donne le meme sous-modèle (10).

$$\text{ClaimNb} \sim \text{DrivAge} + \text{VehAge} + \text{VehBody} + \text{offset}(\log(\text{Exposure})) \quad (10)$$

On a donc trois variables explicatives qui sont associées au nombre de réclamation.

## 5.4 Association avec la valeur relative du véhicule

Le modèle finale n'inclue pas la valeur relative du véhicule. Avec un teste de ratio de vraisemblance, la variable de valeur relative du véhicule n'ajoute pas assez de valeur et est donc écarté du modèle.

## A Illustrations

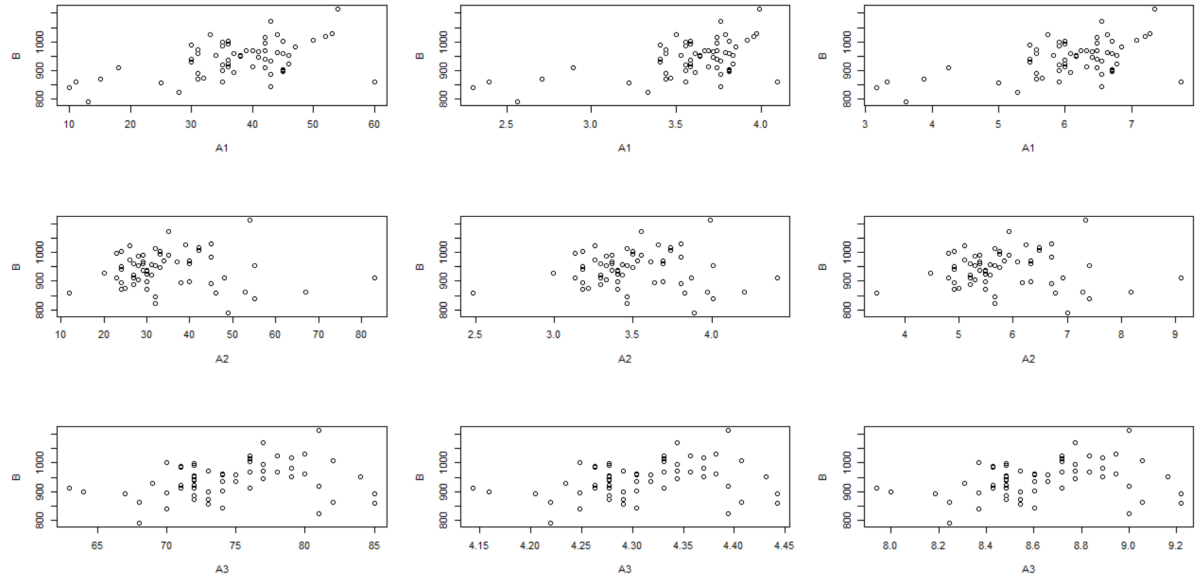


Illustration 4 – À gauche, on compare la variable endogène B avec les variables exogènes A1 à A3. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

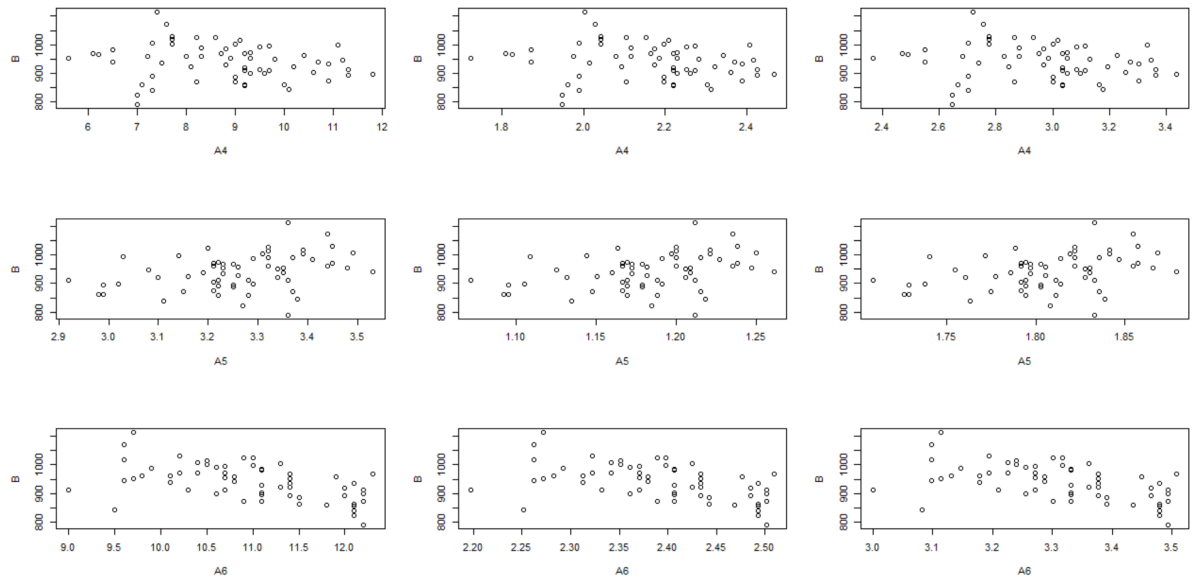


Illustration 5 – À gauche, on compare la variable endogène B avec les variables exogènes A4 à A6. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.



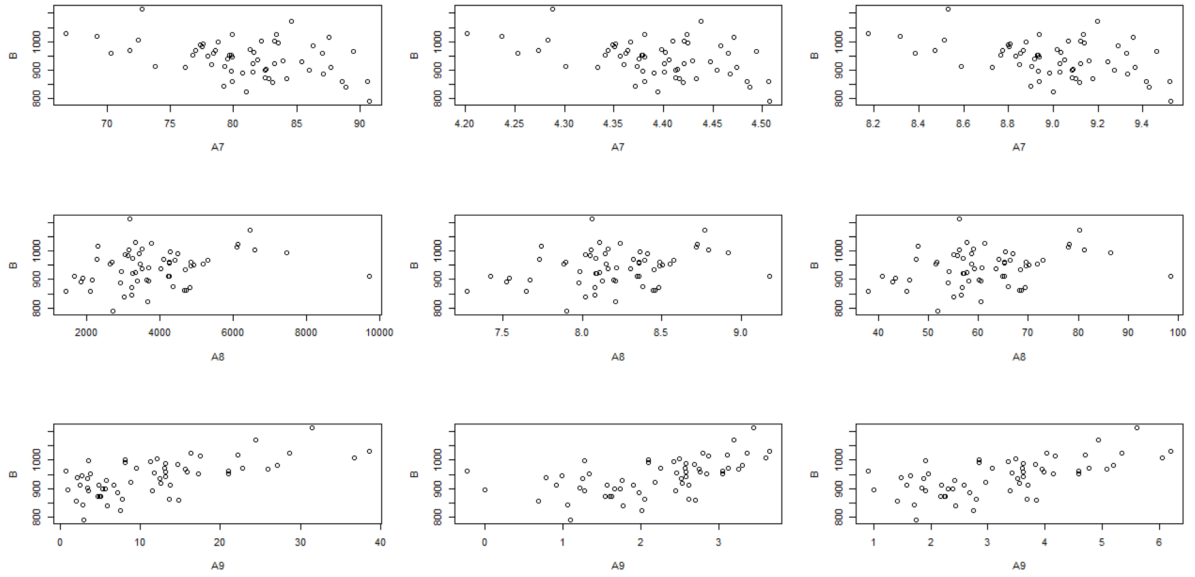


Illustration 6 – À gauche, on compare la variable endogène B avec les variables exogènes A7 à A9. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

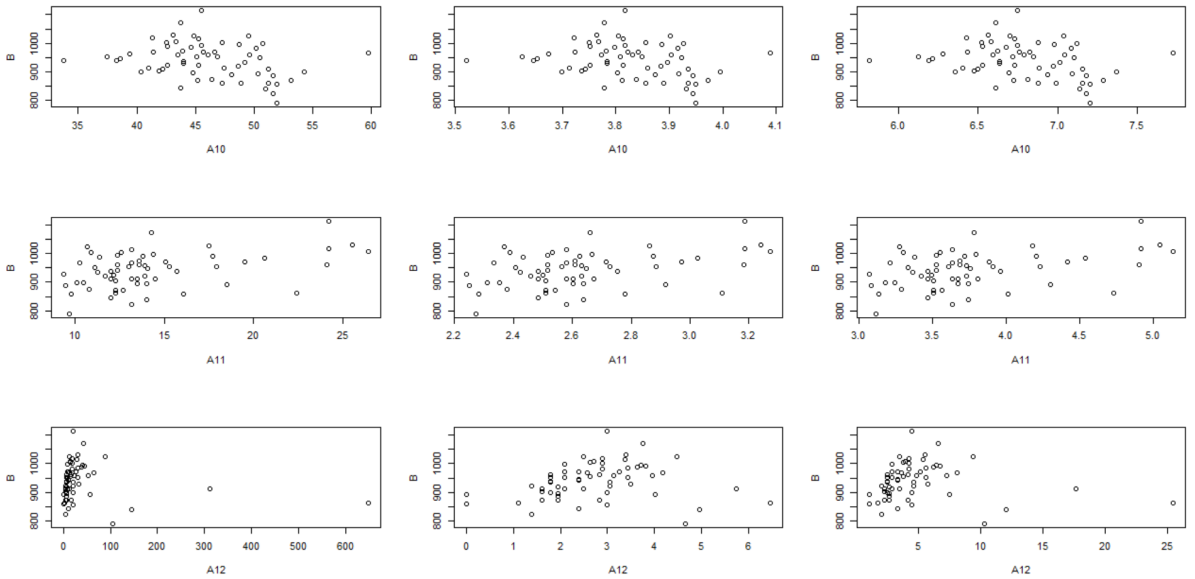


Illustration 7 – À gauche, on compare la variable endogène B avec les variables exogènes A10 à A12. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

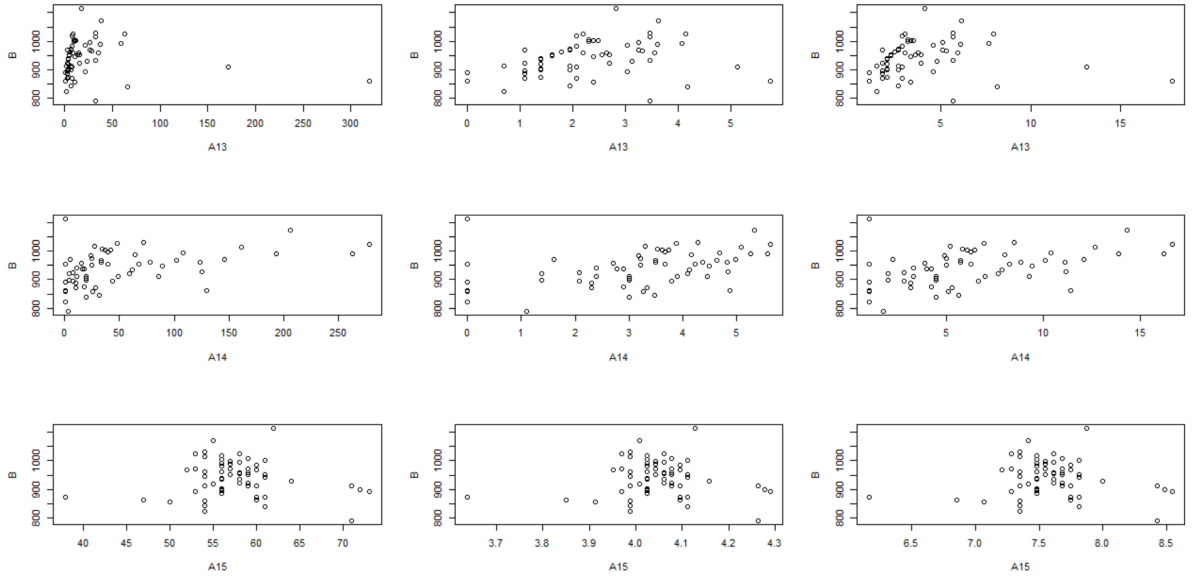


Illustration 8 – À gauche, on compare la variable endogène B avec les variables exogènes A13 à A15. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

## Références

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