# 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 construire un modèle de régression qui estime la probabilité de diagnostic de maladie coronarienne. L'objectif étant 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 afin de 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 utilisé 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 $j^{\text{e}}$ 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 $\lambda_j$ 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 $\phi_j$ 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 $\lambda_j$ 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 $j^{\text{e}}$ 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 3 à 7 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 3 à 7, 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 profiterait 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 $\lambda$ 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 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. Conséquemment, 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). Conséquemment, 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 d'observer des interactions entre ces variables. Pour identifier celles qui sont d'intérêt, 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 celui-ci. 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 donnée avec 13 variables explicatives qui mesure des métriques médicales du corps humain. L'objectif est de construire un modèle de régression qui estime la probabilité de diagnostic de maladie coronarienne positif. Ainsi, on veut identifier lesquelles des métriques médicales sont associes a une hausse du risque d,un diagnostic positif de la maladie coronarienne.

### 4.1 Analyse préliminaire

Afin d atteindre l'objectif de cette question, on commence par faire une analyse des variables explicatives.

Premièrement, on observe des " ?" dans les donnees indiquant des valeurs manquantes. On décide d'écarter les 6 observations avec des valeurs manquantes, car ils ne sont pas nombreux.

Deuxièmement, on trace des graphiques de boites a moustache pour les variables explicatives continues. On analysant les illustrations 8 et 9, on n'identifie pas une variable explicative dominante et aucune

transformation semble utile. On observes les moyennes différentes entre des cas la maladie coronarienne positive ou négative. Certaines variables sembles avoir du pouvoir discriminant, mais il est difficile de tirée des conclusions claires, surtout avec la volatilité observe. On procède avec une analyse plus rigoureuse.

## 4.2 Traitement de la multicollinéarité

Comme pour la question 1, on commence par identifier de la multicollinéarité dans les données. Puisque on s'intéresse seulement a des dépendances entre les variables explicatives, on peut utiliser un modèle linéaire standard avec toutes les 13 variables et calculer les VIFs pour chaque variable explicative. On observe aucune valeur de VIF supérieur a 10 et donc la multicollinéarité ne pose pas de problème pour cette question.

## 4.3 Sélection des variables explicatives

L'objectif est de construire un model predictive. On cherche a prédire une probabilité de se trouve dans la classe 1, "diagnostic de maladie coronarienne positif". Une telle probabilité peut être modélisé avec la loi binomial. Le modèle choisit est donc un GLM avec la loi binomiale et la fonction de lien `logit`. Afin de trouver les variables explicatives les plus pertinents, on choisit des approches itératives. La fonction R `stepAIC` du package `MASS` permet de itérativement sélectionner une variables, selon le critère AIC. On ajoute donc la variable, si elle réduit le AIC. On essaie trois méthode different : `"backward"`,`"forward"` et `"both"`. La méthode `"backward"` part du modele complet avec 13 variables et retire a chaque itération la variable qui augmente le AIC le plus. La méthode `"forward"` part du modele nulle avec 0 variables et ajoute a chaque itération une variable. Elle choisit la variables qui a le plus petit AIC de toutes les sous-modele possible avec un variables de plus. La méthode `"both"` part du modele complet avec 13 variables et retire ou ajoute a chaque itération une variable qui diminue le AIC du nouveau sous-modèle. On obtient le meilleur (5) AIC avec la méthode `"backward"`.

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

Les coefficients de pouls maximum atteint `thalach` et l'indicatrice indiquant la présence d'angine induite par l'exercice `exang` ne sont pas significative au seuil de 5%. Alors, on continue avec des testes de ratio de vraisemblance au seuil 5% en retirant `thalach` et `exang` individuellement. On décide que les deux variables n'ajoute pas assez de valeur, donc on les écartes du modèle. De plus, on teste des interactions. On identifie une interaction intéressante entre le nombre de vaissaux sanguins majeurs colorés par fluroscopie `ca` et `thal` qui est significative au seuil de 1%. On obtient comme modèle final l'équation (6).

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

On utilise aussi la fonction `"glmbb"` qui calcules toutes la sous-modèles possibles (en incluant des interactions).On choisit le modèle avec le AIC minimal et on obtient le modèle donnée par (7).

$$\mathrm{Y} \sim \mathrm{sex} + \mathrm{cp} + \mathrm{trestbps} + \mathrm{thalach} + \mathrm{oldpeak} + \mathrm{slope} + \mathrm{ca} * \mathrm{thal} \tag{7}$$

On remarque que la seule différence est la variable `thalach` qu'on a écartée auparavant au seuil de 5% avec le teste de ratio de vraisemblance. On garde donc le modèle (6) comme modèle finale.

## 4.4 Facteur explicatives du modèle

Afin de répondre a la question on analyse la sortie du modèle final 2.

```
glm(formula = Y ~ sex + cp + trestbps + oldpeak + slope + ca +
    thal + ca:thal, family = binomial("logit"), data = data,
    x = T, y = T)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3802  -0.4414  -0.1305   0.3708   2.9822

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -11.84372    2.27309  -5.210 1.88e-07 ***
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 ***
thal3        -35.50289 2090.59348  -0.017 0.986451
thal4          3.66740    1.35074   2.715 0.006625 **
ca:thal3      17.14674 1045.29619   0.016 0.986912
ca:thal4      -0.88747    0.51038  -1.739 0.082060 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 409.95  on 296  degrees of freedom
Residual deviance: 191.83  on 283  degrees of freedom
AIC: 219.83

Number of Fisher Scoring iterations: 17
```

Illustration 2 – Output R modèle finale Q2

Selon notre modèle, les variables suivantes on tous une association positive avec un diagnostic positif de la maladie coronarienne : le sexe `sex`, la nature des douleurs a la poitrine `cp`, la tension artérielle au repos `trestbps`, la baisse dans ST lors de l exercice maximal `oldpeak` et la pente du segment de ST lors de l'exercice maximal `slope`. Toutes les coefficients sont supérieurs a 0. Pour les variables explicatives `ca` et `thal`, la réponse est moins claire. Il y a des valeurs postivies et negatives pour les coefficient. Pour le le nombre de vaissaux sanguins majeurs colorés par fluroscopie `ca`, association pour une unitee positive

de `ca` est donnée par l équation

$$1.48959 + 17.14674 \times \texttt{thal3} - 0.88747 \times \texttt{thal4} \tag{8}$$

qui est donc toujour positive car $\texttt{thal3} \geq 0$ et $\texttt{thal3} \geq 0$

# 5 Question 3

# A Illustrations

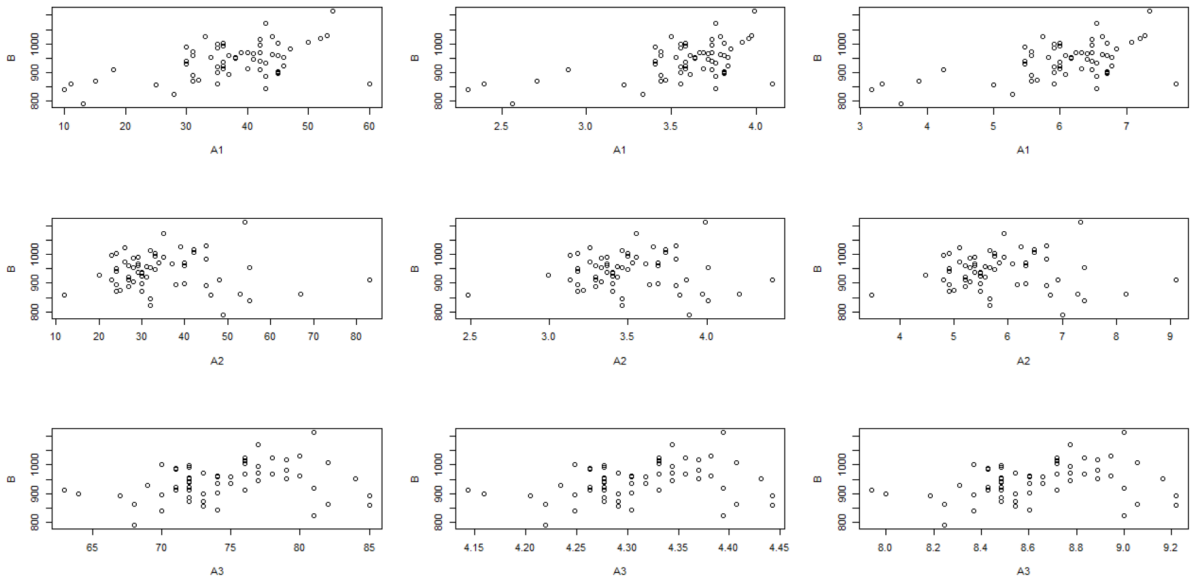


Illustration 3 – À gauche, on compare la variable endogène B avec les variables exogènes A1 à A3. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

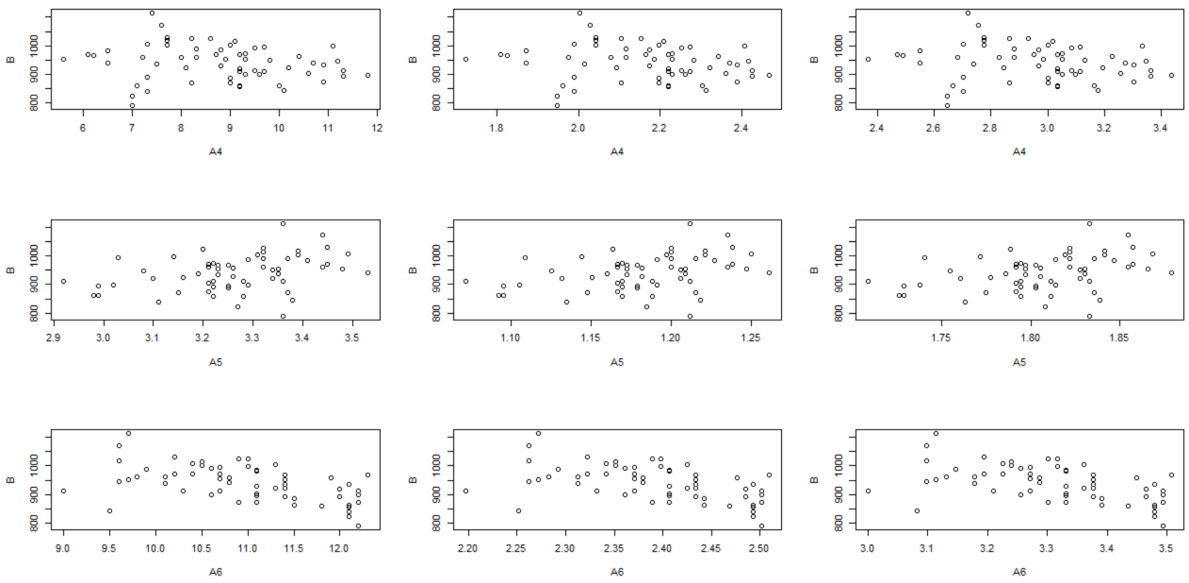


Illustration 4 – À gauche, on compare la variable endogène B avec les variables exogènes A4 à A6. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

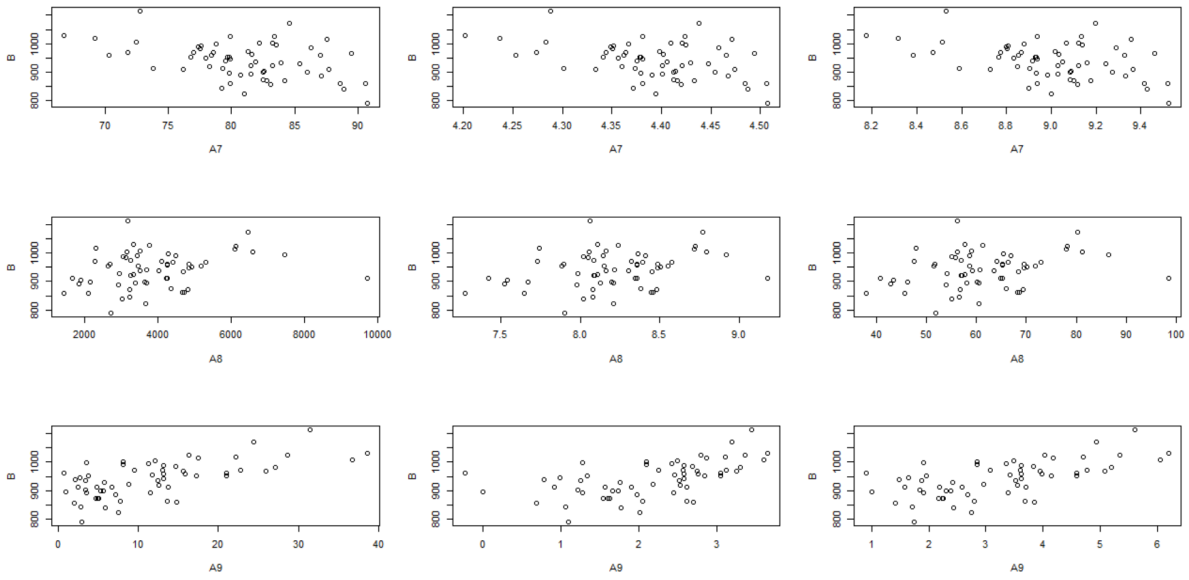


Illustration 5 – À gauche, on compare la variable endogène B avec les variables exogènes A7 à A9. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

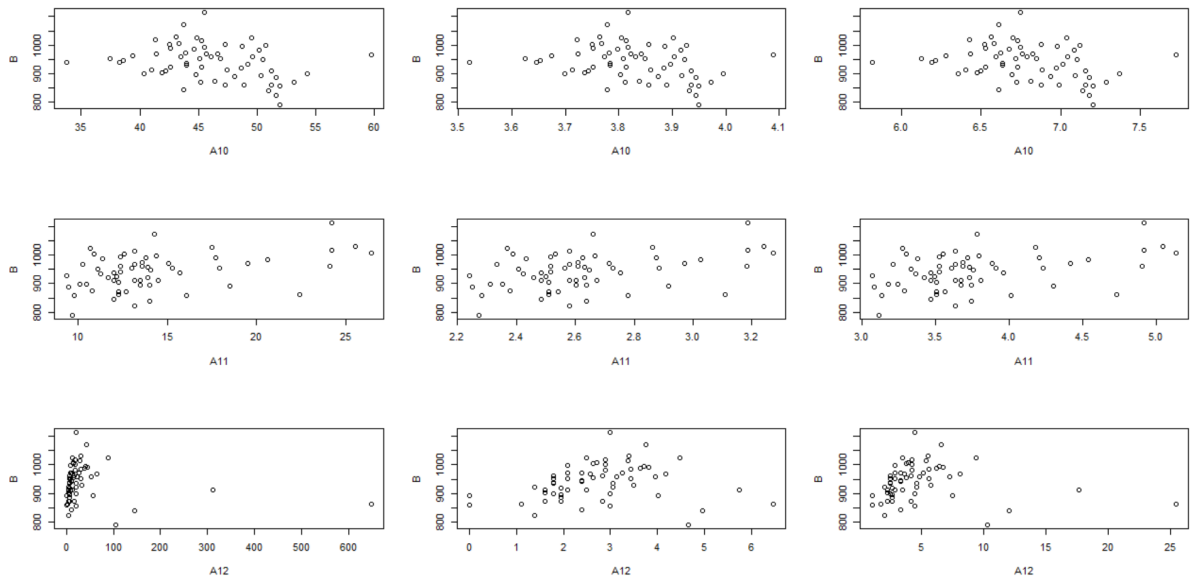


Illustration 6 – À gauche, on compare la variable endogène B avec les variables exogènes A10 à A12. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

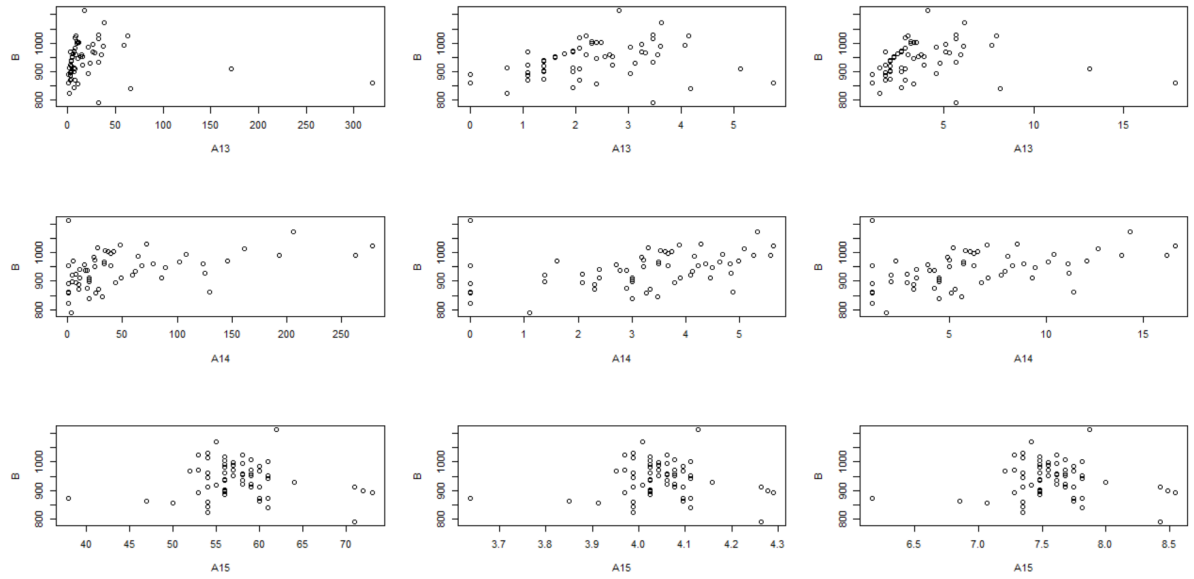


Illustration 7 – À gauche, on compare la variable endogène B avec les variables exogènes A13 à A15. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

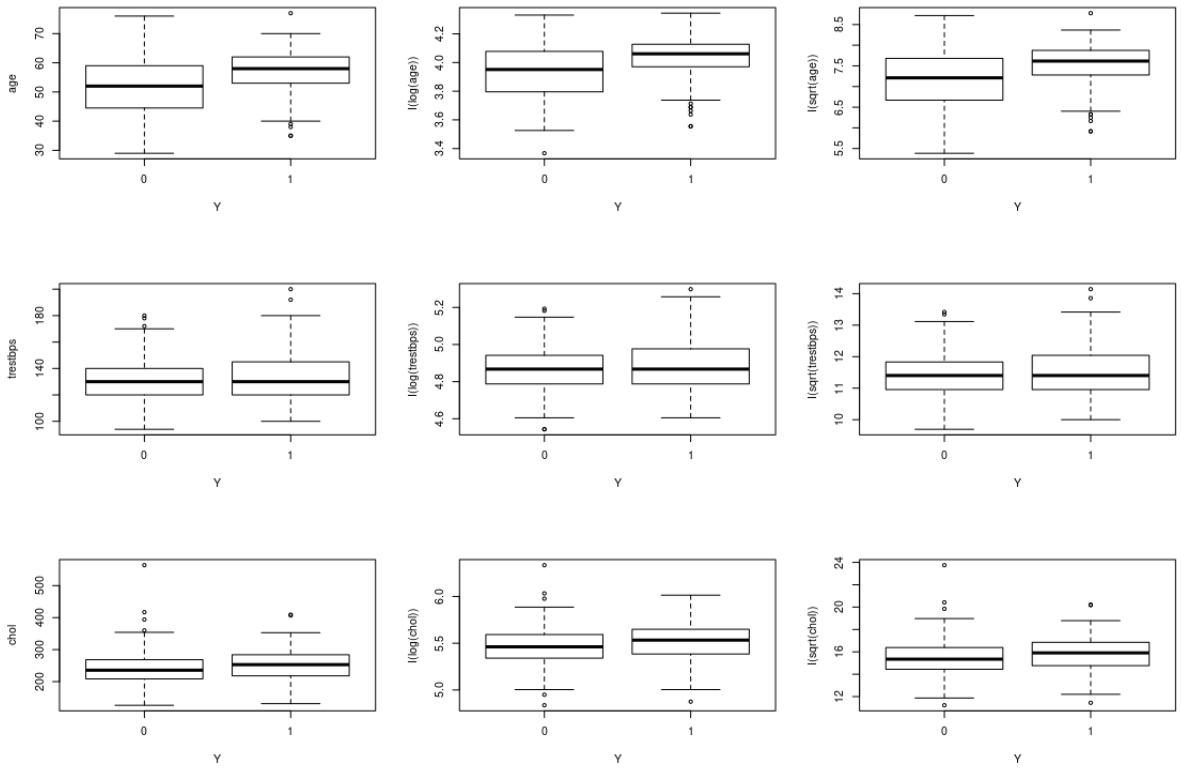


Illustration 8 – Boites a moustache entre la variables explicative et la variable dépendante Y. À gauche, on compare la variable endogène Y avec les variables exogènes. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

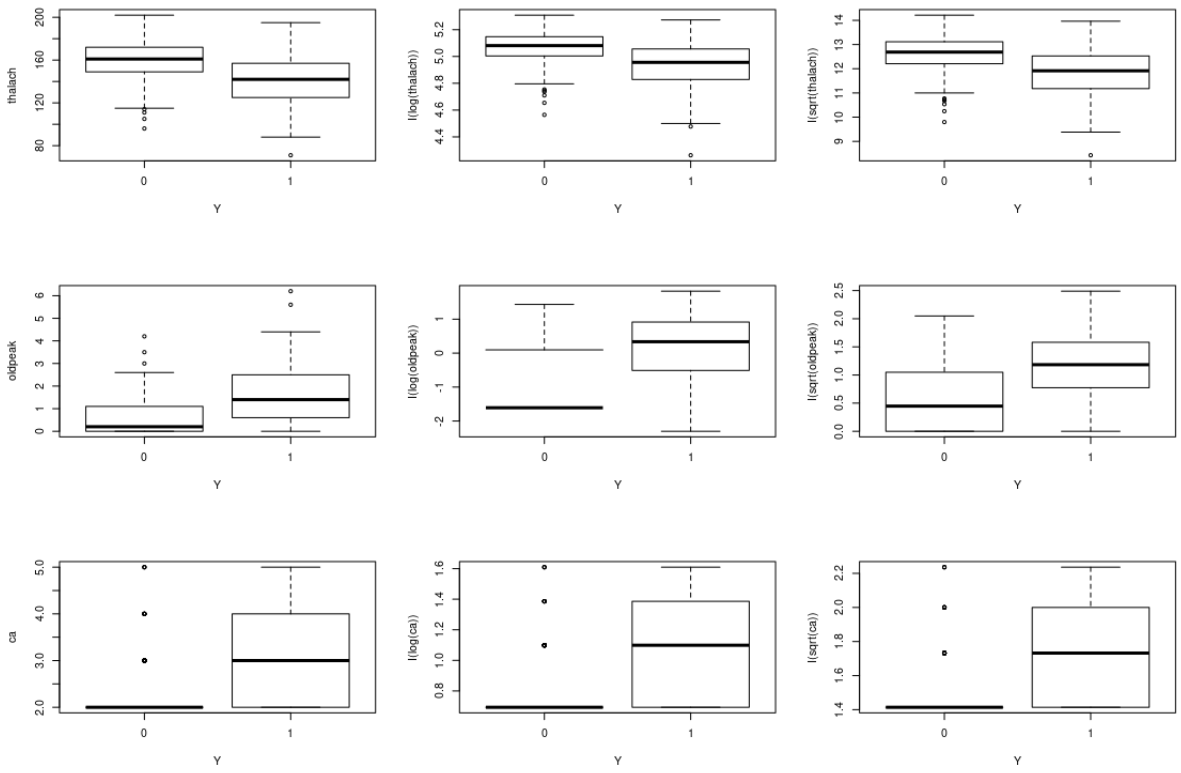


Illustration 9 – Boites a moustache entre la variables explicative et la variable dépendante Y. À gauche, on compare la variable endogène Y avec les variables exogènes. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est 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.