

---

---

# 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 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  $\mathbf{X}$  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 utilisé 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  $j = \operatorname{argmax}_{j \in \{1, \dots, p'\}} \{\phi_j\}$ , 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 2 à 6 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 2 à 6, 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 :

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

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. 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                |      |
|--------------------|--------------------|--------------------|----|----|----|----|---------|---------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|------|
| <i>&lt;int&gt;</i> | <i>&lt;int&gt;</i> | <i>&lt;chr&gt;</i> |    |    |    |    |         |         | <i>&lt;dbl&gt;</i> | <i>&lt;dbl&gt;</i> | <i>&lt;dbl&gt;</i> | <i>&lt;dbl&gt;</i> | <i>&lt;dbl&gt;</i> | <i>&lt;dbl&gt;</i> | <i>&lt;dbl&gt;</i> | <i>&lt;dbl&gt;</i> | <i>&lt;dbl&gt;</i> | <i>&lt;dbl&gt;</i> |      |
| 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). Conséquemment, 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 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

### A Illustrations

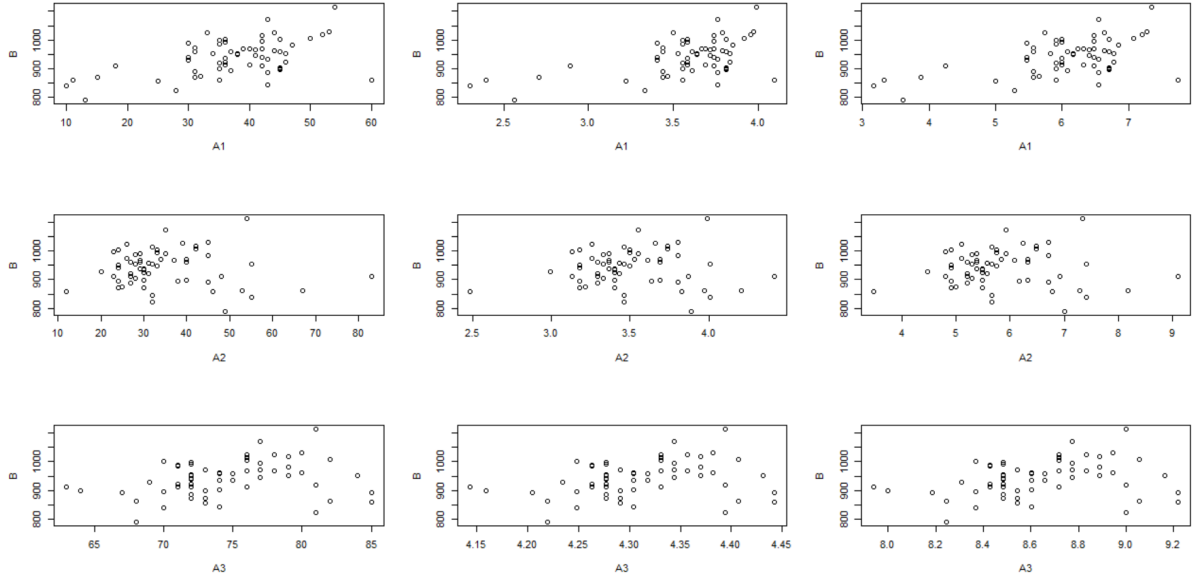


Illustration 2 – À gauche, on compare la variable endogène B avec les variables exogènes A1 à A3. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

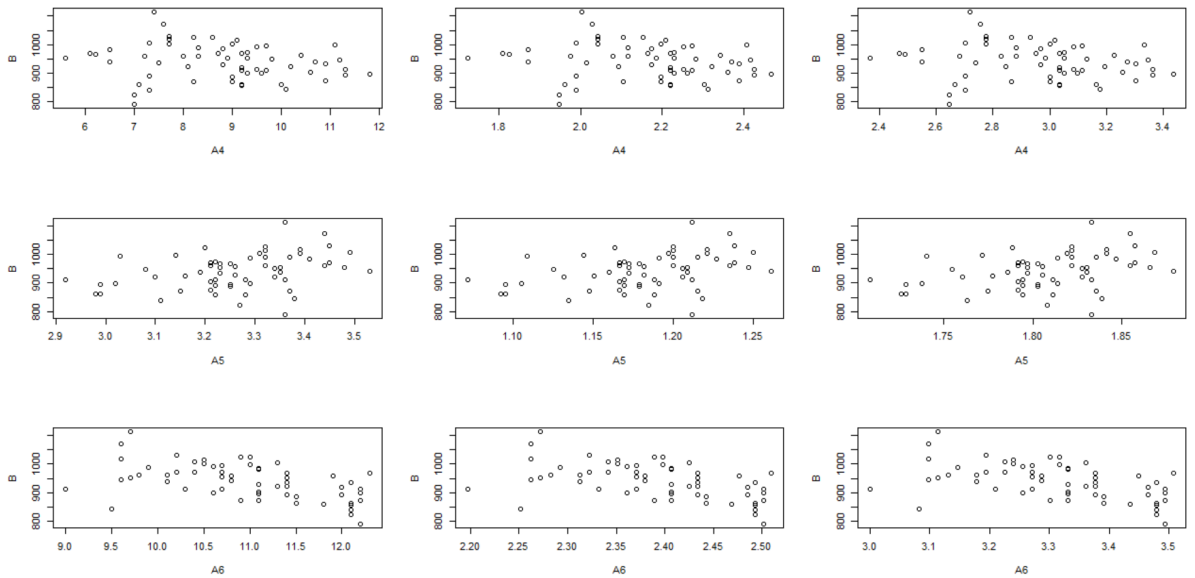


Illustration 3 – À gauche, on compare la variable endogène B avec les variables exogènes A4 à A6. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

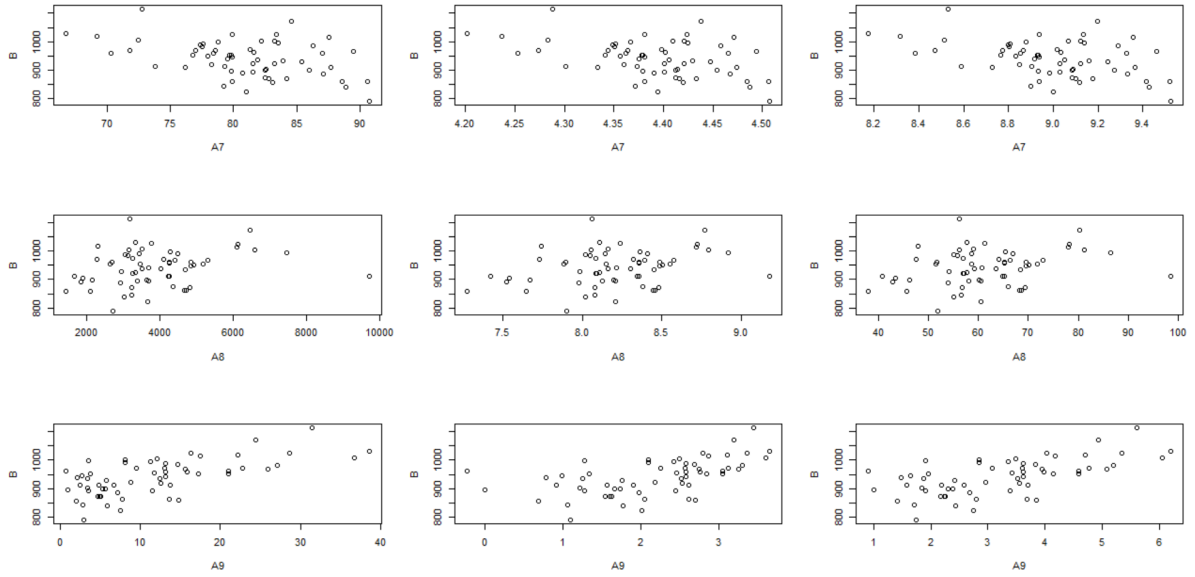


Illustration 4 – À gauche, on compare la variable endogène B avec les variables exogènes A7 à A9. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

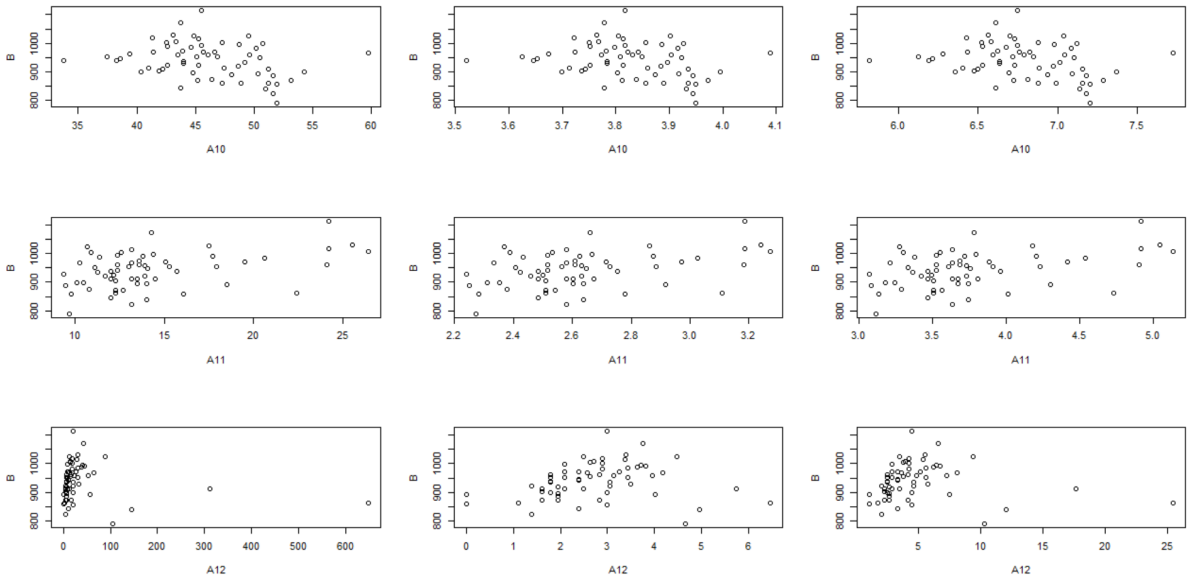


Illustration 5 – À gauche, on compare la variable endogène B avec les variables exogènes A10 à A12. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

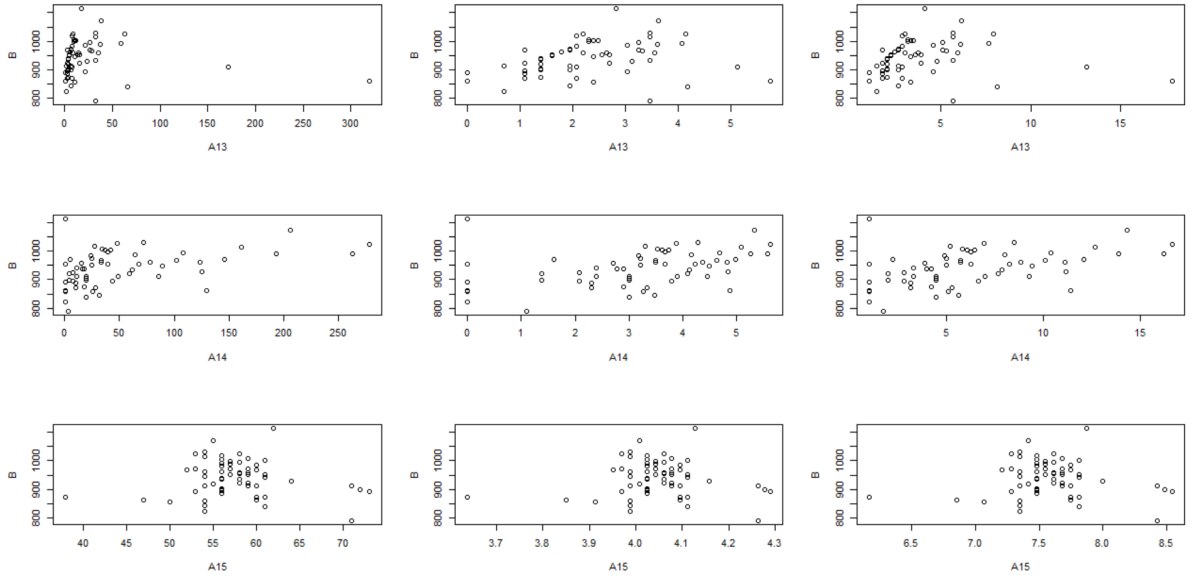


Illustration 6 – À gauche, on compare la variable endogène B avec les variables exogènes A13 à A15. Au centre, on compare la même relation, mais avec une transformation logarithmique effectuée sur les variables exogènes. À droite, c'est la transformation racine carrée qui est appliquée.

## Références

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