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^{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 $\lambda_{max} = \max(\lambda_1, \ldots, \lambda_{p'})$. Pour chaque valeur de ϕ_j plus grande que 30, 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 $p_{lj} > 0.6$, $\{j : \phi_j > 30\}$, alors on conclu que la l-ème variable explicative contribue

à la multicolinéarité et cause problème.

Dans ce cas, les solutions possibles consistent à réduire la dimension de la matrice de schéma en les regroupant. Par exemple, on peut les sommer ou calculer leur moyenne. Autrement, il est aussi possible de réaliser une transformation non linéaire de ces variables (p.ex. une transformation logarithmique ou racine carrée).

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.

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 A13 et A12 sont largement supérieures à 10. Conséquemment, il faut aller plus loin dans le diagnostique en utilisant la fonction `ols_eigen_cindex()` du *package* `olsrr`. Le tableau retourné par la fonction nous indique que les indices de conditionnement ϕ_9 à ϕ_{16} sont supérieurs à 30. Le tableau 1 présente les proportions de variance d'intérêt pour $\lambda_j, j = 9, \ldots, 16$.

j	ϕ_j	Variable	p_{lj}
11	53.01	A12	0.79
11	53.01	A13	0.73
14	112.87	A3	0.60
15	131.38	A6	0.69
16	400.62	A5	0.81

Tableau 1 – Variables responsables de l'inflation de la variance.

Références

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