École d'actuariat
Université Laval

# Modèles linéaires en actuariat
# ACT-2003
# Exercices Chapitre 4

Marie-Pier Côté
Automne 2018

**Question 1.** Est-ce que les compagnies d'assurance utilisent la race comme un facteur déterminant dans leur décision de rendre de l'assurance disponible ? Fienberg (1985) a rassemblé des données d'un rapport de la *U.S. Commission on Civil Rights* sur le nombre de polices d'assurance habitation émises à Chicago entre Décembre 1977 et Février 1978. Les polices d'assurance étaient placées dans 2 catégories :

- polices émises dans le marché standard, volontaire
- polices émises dans le marché sous-standard, involontaire

Les polices du marché sous-standard sont émises selon un programme gouvernemental d'accès à l'assurance. Les personnes qui contractent ce type d'assurance se sont vues refuser une police d'assurance sur le marché volontaire. On s'intéresse à l'accessibilité à l'assurance selon la race et on utilise le nombre de polices émises (ou renouvellées) sur le marché sous-standard comme mesure de "non-accessibilité".

La ville de Chicago a été divisée en 45 régions (selon le code postal). Pour chaque région, on a les informations suivantes :

| Variable | Description |
|---|---|
| `race` | pourcentage de la population de la région provenant d'une minorité raciale |
| `fire` | Nombre d'incendies par millier de maisons |
| `theft` | Nombre de vols par millier de maisons |
| `age` | Pourcentage des maisons construites avant 1940 |
| `involact` | Nouvelles polices et renouvellements dans le marché sous-standard, par centaine de maisons |
| `income` | Revenu familial moyen |

On s'intéresse majoritairement à l'effet de la variable explicative `race`, mais on veut aussi tenir compte des autres facteurs qui pourraient être en cause, et des intéractions entre ces facteurs. Les modèles considérés sont :

**Modèle A :** `involact∼race`

**Modèle B :** `involact∼race+I(log(income))`

**Modèle C :** `involact∼race+fire+age`

**Modèle D :** `involact∼race+fire+theft+age`

**Modèle E :** `involact∼race+I(log(income))+fire+theft+age`

**Modèle F :** `involact∼race+I(log(income))*age+fire+theft`

**Modèle G :** `involact∼I(log(income))*(age+race)+fire+theft`

**Modèle H :** `involact∼I(log(income))*age+race*(fire+theft+I(log(income)))`

Note : `A*B` représente `A+B+A:B`, c'est-à-dire les effets principaux et les intéractions entre les variables explicatives `A` et `B`.
On a les informations suivantes sur les modèles A à H :

| Modèle | $p'$ | PRESS | $R_p^2$ | $C_p$ de Mallows | AIC | BIC | $R_a^2$ |
|---|---|---|---|---|---|---|---|
| A | 2 | 9.6344 | 0.4735 | 63.24 | -69.86 | -66.25 | 0.5126 |
| B | 3 | 8.8248 | 0.5177 | 49.55 | -75.20 | -69.78 | 0.5761 |
| C | 4 | 5.2083 | 0.7154 | 8.58 | -103.09 | -95.87 | 0.7765 |
| D | 5 | 4.5727 | 0.7501 | 7.97 | -103.75 | -94.71 | 0.7840 |
| E | 6 | 4.8985 | 0.7323 | 9.88 | -101.84 | -91.00 | 0.7790 |
| F | 7 | 4.8999 | 0.7322 | 9.64 | -102.25 | -89.61 | 0.7850 |
| G | 8 | 4.7528 | 0.7403 | 8.46 | -103.92 | -89.47 | 0.7964 |
| H | 10 | 5.4817 | 0.7004 | 10.00 | -102.98 | -84.91 | 0.7989 |

Les facteurs d'inflation de la variance pour ces modèles sont présentés dans le tableau suivant :

| | C | D | E | F | G | H |
|---|---|---|---|---|---|---|
| `race` | 1.73 | 1.81 | 3.81 | 3.83 | 2191 | 5449 |
| `fire` | 2.03 | 2.03 | 2.16 | 2.48 | 2.50 | 19 |
| `age` | 1.25 | 1.39 | 2.08 | 4070 | 5247 | 6316 |
| `theft` | | 1.23 | 1.63 | 1.64 | 1.68 | 4.05 |
| `I(log(income))` | | | 4.66 | 21 | 21 | 22 |
| `I(log(income)):age` | | | | 3793 | 4932 | 5919 |
| `I(log(income)):race` | | | | | 2064 | 5155 |
| `race:theft` | | | | | | 24 |
| `race:fire` | | | | | | 40 |

On sait également que les postulats de la régression linéaire multiple sont vérifiés.

a) Quel est le meilleur modèle selon
   (i) le critère PRESS ?

(ii) le critère du coefficient de détermination de prévision $R^2_p$ ?
(iii) le $C_p$ de Mallows ?
(iv) le critère d'information d'Akaike ?
(v) le critère d'information de Bayes ?
(vi) le coefficient de détermination ajusté $R^2_a$ ?

b) Que peut-on remarquer en regardant les facteurs d'inflation de la variance pour les modèles C à H ?

c) Selon vous, quel serait le meilleur modèle à utiliser pour ces données ? Pourquoi ?

**Question 2. Un cas simplifié de régression ridge et lasso.** Cet exercice est inspiré de James et al. (2013). Considérons le cas simplifié où $n = p$ et la matrixce d'incidence $X$ est diagonale, avec des 1 sur la diagonale et des 0 pour tous les éléments hors-diagonale.

On ajuste une régression linéaire multiple passant par l'origine avec de telles données, c'est-à-dire que $\beta_0 = 0$ est connu et on ne l'estime pas.

Sous ces hypothèses,

a) Trouvez les estimateurs des moindres carrés $\hat{\beta}_1, \ldots, \hat{\beta}_p$.

b) Écrivez l'expression à minimiser pour trouver les estimateurs sous la régression ridge.

c) Trouvez l'expression de l'estimateur ridge.

d) Écrivez l'expression à minimiser pour trouver les estimateurs sous la régression lasso.

e) Montrez que l'estimateur lasso a la forme

$$\hat{\beta}_j^{\text{lasso}} = \begin{cases} y_j - \lambda/2, & \text{si } y_j > \lambda/2 \\ y_j + \lambda/2, & \text{si } y_j < -\lambda/2 \\ 0, & \text{si } |y_j| < \lambda/2. \end{cases}$$

f) Interprétez les effets des pénalités ridge et lasso à la lumière de vos réponses aux sous-questions précédentes.

# SOLUTIONS

**Question 1.**

a) (i) modèle D
(ii) modèle D
(iii) modèle G
(iv) modèle G
(v) modèle C
(vi) modèle H

b) Il y a un très gros problème de multicolinéarité pour les modèle F, G et H, car certains VIFs sont beaucoup plus grands que 10. Ce problème augmente inutilement la variance des paramètres estimés.

c) On évite les modèles F G et H pour ne pas avoir de problème de multicolinéarité. Le modèle D est préférable selon les critères PRESS et $R^2_p$. De plus, ses critères AIC et BIC sont les deuxièmes plus petits. Le $C_p$ est 8, donc 8-5=3. Ce n'est pas parfait, mais ce n'est pas si mal, etc.

**Question 2.**

a) Puisque $n = p$, $\beta_0 = 0$ et que la matrice d'incidence est diagonale, on a $\hat{y}_i = \hat{\beta}_i$ pour $i = 1, \ldots, n$. On minimise $S(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \beta_i)^2$ et on trouve pour $i \in \{1, \ldots, n\}$,

$$\frac{\partial}{\partial \beta_i} S(\boldsymbol{\beta})\Big|_{\hat{\beta}_i} = -2(y_i - \hat{\beta}_i) = 0 \quad \Rightarrow \quad \hat{\beta}_i = y_i.$$

b) On minimise, pour une valeur $\lambda > 0$,

$$S^{\text{ridge}}(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \beta_i)^2 + \lambda \sum_{i=1}^n \beta_i^2.$$

c) On a

$$\frac{\partial}{\partial \beta_i} S^{\text{ridge}}(\boldsymbol{\beta}) = -2(y_i - \beta_i) + 2\lambda\beta_i.$$

On pose égal à 0 et on trouve

$$y_i - \hat{\beta}_i^{\text{ridge}} = \lambda \hat{\beta}_i^{\text{ridge}} \quad \Rightarrow \quad \hat{\beta}_i^{\text{ridge}} = \frac{y_i}{1+\lambda}.$$

d) On minimise, pour une valeur $\lambda > 0$,

$$S^{\text{lasso}}(\boldsymbol{\beta}) = \sum_{i=1}^n (y_i - \beta_i)^2 + \lambda \sum_{i=1}^n |\beta_i|.$$

e) On a

$$\frac{\partial}{\partial \beta_i} S^{\text{lasso}}(\boldsymbol{\beta}) = -2(y_i - \beta_i) + \lambda \,\text{signe}(\beta_i).$$

On utilise les EMV trouvés en a) pour définir le signe. Supposons d'abord que $\hat{\beta}_i = y_i > 0$. Alors, on a aussi $\hat{\beta}_i^{\text{lasso}} > 0$ (sinon, changer le signe donnera une valeur plus petite de l'équation à minimiser). On pose la dérivée égale à 0 et on trouve

$$2(y_i - \hat{\beta}_i^{\text{lasso}}) = \lambda \quad \Rightarrow \quad \hat{\beta}_i^{\text{ridge}} = y_i - \lambda/2,$$

ce qui tient seulement si $\hat{\beta}_i^{\text{lasso}} > 0$, alors on a $\hat{\beta}_i^{\text{ridge}} = \max(0, y_i - \lambda/2)$. Supposons ensuite que $\hat{\beta}_i = y_i < 0$. Alors, on a aussi $\hat{\beta}_i^{\text{lasso}} < 0$. On pose la dérivée égale à 0 et on trouve

$$2(y_i - \hat{\beta}_i^{\text{lasso}}) = -\lambda \quad \Rightarrow \quad \hat{\beta}_i^{\text{ridge}} = y_i + \lambda/2,$$

sous la contrainte que ce soit négatif, donc dans ce cas, $\hat{\beta}_i^{\text{ridge}} = \min(0, y_i + \lambda/2)$. On combine les deux cas et on obtient l'équation donnée.

f) On peut voir que la façon de rapetisser les paramètre est bien différente pour les deux méthodes. Avec ridge, chaque coefficient des moindres carrés est réduit par la même proportion. Avec lasso, chaque coefficient des moindres carrés est réduit vers 0 d'un montant constant $\lambda/2$ ; ceux qui sont plus petits que $\lambda/2$ en valeur absolue sont mis exactement égaux à 0. C'est de cette façon que le lasso permet de faire la sélection des variables explicatives.