Régression Ridge : sélection de λ et étude de cas

Chargé du cours
Prof. Mustapha Rachdi

Université Grenoble Alpes
UFR SHS, BP. 47
38040 Grenoble cedex 09
France
Bureau provisoire : 07 au BSHM
e-mail : mustapha.rachdi@univ-grenoble-alpes.fr

- Comme nous en avons discuté la dernière fois, le paramètre λ contrôle le compromis entre la pénalité et l'ajustement du modèle, et a donc un très grand impact sur l'estimation qui en résulte :
  - Quand $\lambda \to 0$, Q approche L et $\widehat{\beta}$ approche l'estimateur MCO.
  - D'autre part, quand $\lambda \to \infty$, la pénalité domine la fonction objective et $\widehat{\beta} \approx 0$
- Clairement, la sélection de λ est un aspect pratique très important de l'ajustement des modèles de régression pénalisés.

- En général, une approche raisonnable pour sélectionner λ de manière objective consiste à choisir la valeur de λ permettant le plus grand pouvoir prédictif : si $\lambda = 1$ peut prédire les observations des co-variables mieux que $\lambda = 5$, ceci est une raison claire de préférer $\lambda = 1$

- Supposons que $\mathbb{E}(y_i) = f(\mathbf{x}_i)$, $var(y_i) = \sigma^2$, et que nous ayons ajusté un modèle pour obtenir $\widehat{f}_\lambda$, un estimateur de f, et soit $\{\widehat{y}_i(\lambda)\}_{i=1}^n$ des valeurs ajustées, où $\widehat{y}_i(\lambda) = \widehat{f}_\lambda(\mathbf{x}_i)$

- Il est clairement trompeur d'évaluer la précision de la prédiction en comparant $\widehat{y}_i(\lambda)$ aux y_i. La valeur observée y_i a déjà été utilisée pour calculer $\widehat{y}_i(\lambda)$, et n'est donc pas une véritable prédiction.

- Le simple calcul de la somme résiduelle des carrés (RSS) sous-estimera la véritable exactitude prédictive du modèle.
- Au lieu de cela, nous devons examiner comment $\widehat{y}_i(\lambda)$ prédit une nouvelle observation y_i^{new} générée à partir du modèle sous-jacent :

$$y_i^{new} = f(\mathbf{x}_i) + \varepsilon_i^{new}$$

- Ensuite, l'erreur de prédiction peut être mesurée par :

$$\text{PE}(\lambda) = \sum_{i=1}^{n} \left(y_i^{new} - \widehat{y}_i(\lambda)\right)^2$$

- Pour clarifier ceci, dans ce cadre nous mesurons les nouvelles réponses $\{y_i^{new}\}_{i=1}^n$, mais en les valeurs originales des prédicteurs $\{\mathbf{x}_i\}_{i=1}^n$.

- Le modèle ayant le plus grand pouvoir prédictif est donc celui qui minimise l'erreur de prédiction moyenne

$$\mathbb{E}(\text{PE}(\lambda)) = \mathbb{E}\sum_{i=1}^{n}\{y_i^{new} - \widehat{y}_i(\lambda)\}^2,$$

où l'espérance est prise sur les observations originales $\{y_i\}_{i=1}^n$ ainsi que sur les nouvelles observations $\{y_i^{new}\}_{i=1}^n$.

-

Theorem

$$\mathbb{E}(PE(\lambda)) = \mathbb{E}\sum_{i=1}^{n}\{y_i - \widehat{y}_i(\lambda)\}^2 + 2\sum_{i=1}^{n} cov\left(y_i, \widehat{y}_i(\lambda)\right)$$

- Ainsi, l'erreur de prédiction moyenne consiste en deux termes :
  - le premier terme est l'erreur d'ajustement intra-échantillon
  - le second terme est un facteur de correction de biais qui est issu de la tendance de l'erreur d'ajustement intra-échantillon pour sous-estimer l'erreur de prédiction hors échantillon, également connu sous le nom d'*optimisme de l'ajustement du modèle*
- Le second terme peut également être considéré comme une mesure de la complexité du modèle, ou des degrés de liberté :

$$\text{df} = \sum_{i=1}^{n} \frac{cov(y_i, \widehat{y}_i)}{\sigma^2} = \frac{tr\{cov(\mathbf{y}, \widehat{\mathbf{y}})}{\sigma^2}$$

- Par exemple, considérons la régression MCO, avec

$$\widehat{\mathbf{y}} = \mathbf{X}({}^{\mathbf{t}}\mathbf{X}\mathbf{X})^{-1}{}^{\mathbf{t}}\mathbf{X}\mathbf{y}$$

- *Résultat* :

$$\mathrm{df} = \mathit{rang}(\mathbf{X})$$

- Ainsi, notre définition basée sur la covariance est en accord avec la notion habituelle de degrés de liberté comme étant le nombre de paramètres (dans un modèle non pénalisé)

- Une méthode d'ajustement du modèle est dite *linéaire* si l'on peut écrire $\widehat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ pour une matrice $\mathbf{S}$.
- *Résultat* : pour toute méthode linéaire,

$$\mathrm{df} = tr(\mathbf{S})$$

- La régression ridge (de crête) est une méthode d'ajustement linéaire, avec

$$\mathbf{S} = n^{-1}\mathbf{X}(n^{-1}\,{}^t\mathbf{X}\mathbf{X} + \lambda\mathbf{I})^{-1}\,{}^t\mathbf{X}$$

ainsi,

$$\mathrm{df}(\lambda) = tr\left(n^{-1}\mathbf{X}(n^{-1}\,{}^t\mathbf{X}\mathbf{X} + \lambda\mathbf{I})^{-1}\,{}^t\mathbf{X}\right) = \sum_{j=1}^{p} \frac{d_j}{d_j + \lambda}$$

où $d_1, \ldots, d_p$ sont les valeurs propres de $n^{-1}\,{}^t\mathbf{X}\mathbf{X}$.

- Ce résultat illustre le fait qu'en régression pénalisée, la sélection du modèle est continue
- Lorsque nous changeons λ, nous augmentons progressivement la complexité du modèle, et de petits changements dans λ entraînent de petits changements dans l'estimation.
- Ceci contraste fortement avec la sélection de "meilleurs sous-ensembles, où la complexité est ajoutée par des sauts discrets lorsque nous introduisons des paramètres, et l'ajout d'un seul paramètre peut introduire de grands changements dans les estimations du modèle.

- Maintenant que nous avons généralisé le concept de degrés de liberté, nous allons décrire divers critères de sélection de modèles qui peuvent être utilisés pour sélectionner λ.
- Ceci sera bref, puisque vous avez probablement rencontré ces critères dans d'autres cours [1].
- Pour commencer, revenons à $\mathbb{E}(\text{PE})$. Rappelons qu'il s'agissait de deux termes : un terme d'erreur dans l'échantillon et un terme de complexité du modèle.
- En utilisant RSS/σ^2 pour le premier terme et $\text{df}(\lambda)$ pour le second, on obtient un critère connu sous le nom de statistique C_p (de Mallows) :

$$C_p = \frac{\text{RSS}(\lambda)}{\sigma^2} + 2\,\text{df}(\lambda)$$

1. d'ailleurs le TP2 est consacré à la mise en oeuvre de ces techniques sur les données du cancer

- Une approche alternative est de laisser de côté (hors du processus d'ajustement) quelques observations et de les utiliser pour évaluer la précision de la prédiction. En général, ceci est connu sous le nom de *validation-croisée*, dont nous parlerons plus tard.
- Cependant, pour les méthodes d'ajustement linéaire, il existe une solution élégante et très proche de lerreur leave-one-out (validation-croisée) qui ne nécessite pas de réajuster le modèle.
- Soit $\widehat{f}_{(-i)}$ le modèle ajusté pour lequel lobservation i est laissée de côté

$$\sum_i \left\{ y_i - \widehat{f}_{(-i)}(x_i) \right\}^2 = \sum_i \left(\frac{y_i - \widehat{f}(x_i)}{1 - S_{ii}} \right)^2$$

où S_{ii} est le $i$ème élément de la diagonale de $\mathbf{S}$.

- En remplaçant S_{ii} par sa moyenne, $tr(\mathbf{S})/n = \text{df}(\lambda)/n$, on arrive au *critère de validation croisée généralisée* :

$$GCV = \frac{\text{RSS}(\lambda)}{(1 - \text{df}(\lambda)/n)^2}$$

- Comme pour C_p, le critère GCV combine $\text{RSS}(\lambda)$ avec un terme de complexité de modèle, bien que dans GCV il prenne la forme d'un facteur multiplicatif, facteur d'inflation, $(1 - \text{df}(\lambda)/n)^2$ plutôt qu'un terme additif.
- Un aspect intéressant de GCV par opposition à la statistique C_p est qu'il ne nécessite pas une estimation de σ^2.

- Les deux critère C_p et GCV sont développés avec les idées de moindres carrés. Le critère d'information dAkaike (AIC) est une généralisation de C_p aux *modèles généraux du maximum vraisemblance*.
- Plutôt que de considérer la valeur moyenne de $\{y^{new} - \widehat{y}_i(\widehat{\theta})\}^2$, Akaike a proposé d'estimer la valeur moyenne de $\log \mathbb{P}(y_i^{new}|\widehat{\theta})$, où $\widehat{\theta}$ représente les estimations des paramètres du modèle basées sur les données d'origine $\{y_i\}_{i=1}^n$.
- Asymptotiquement, on peut montrer que pour l'estimation du maximum de vraisemblance,

$$AIC = 2L(\widehat{\theta}(\lambda)|\mathbf{X}, \mathbf{y}) + 2\,\mathrm{df}(\lambda)$$

- Pour la distribution normale,

$$AIC = n \log \sigma^2 + \frac{\mathrm{RSS}(\lambda)}{\sigma^2} + 2\,\mathrm{df}(\lambda) + \textit{constante}$$

- Ainsi, dans le cas d'erreurs normalement distribuées et de variance connue σ^2, AIC et C_p sont équivalents à une constante près.

- Une approche plutôt différente consiste à considérer la sélection de modèles à partir d'une perspective bayésienne.
- En supposant que M_λ dénote le modèle avec le paramètre de régularisation λ. Nous sommes intéressés par le calcul de la probabilité a posteriori de M_λ sachant les données : $\mathbb{P}(M_\lambda|\mathbf{X}, \mathbf{y})$.
- Si nous supposons une loi a priori uniforme pour tous les modèles, alors

$$\mathbb{P}(M_\lambda|\mathbf{X}, \mathbf{y}) \propto \mathbb{P}(\mathbf{y}|\mathbf{X}, M_\lambda)$$

- En général, le calcul de cette quantité implique une intégration numérique, mais cette intégrale peut être approximée. Ce qui donne

$$\log \mathbb{P}(\mathbf{y}|\mathbf{X}, M_\lambda) \approx -L(\widehat{\theta}(\lambda)|\mathbf{X}, \mathbf{y}) - \frac{1}{2}\,\mathrm{df}(\lambda)\log(n)$$

- Le *critère d'information bayésien* (BIC) est défini comme -2-fois cette quantité :

$$BIC = 2L(\widehat{\theta}(\lambda)|\mathbf{X}, \mathbf{y}) + \mathrm{df}(\lambda)\log(n)$$

- Ainsi, choisir le modèle avec le plus petit BIC est équivaut (approximativement) au choix du modèle avec la plus forte probabilité a posteriori.

- Notons que, les équations pour AIC et BIC sont étonnamment similaires. la seule différence est $\log(n)$ au lieu de 2 comme facteur multiplicatif pour $\text{df}(\lambda)$.
- En pratique, cela signifie que BIC applique une pénalité plus lourde que l'AIC à la complexité du modèle (à condition que $n \geq 8$) et favorise donc des modèles plus parcimonieux.

Régression ridge utilisant `hdrm` : Etude de cas (données de pollution)

Gestion des données par le package `hdrm`

- Après avoir installé le package `hdrm`, on appelle downloadData pour installer les ensembles de données/datasets (cela ne doit être fait qu'une seule fois) :

code R

```
downloadData() # Download all data sets
downloadData(bcTCGA) # Download a specific data set
```

- Une fois les jeux de données téléchargés, on a deux options pour les charger :

code R

```
Data <- readData(bcTCGA) # Call Data$X, Data$y, etc.
attachData(bcTCGA) # Call X, y, etc.
```

Régression ridge utilisant `hdrm` : Etude de cas (données de pollution)

- La fonction principale de Ridge, qui peut être utilisée de deux manières :

code R

```
ridge(y ~ x1 + x2:x3, data) # comme en lm()
ridge(X, y) # comme en glmnet()
```

- Une fois que cela est fait, le package offre une variété de fonctions pour intéragir avec l'objet :

code R

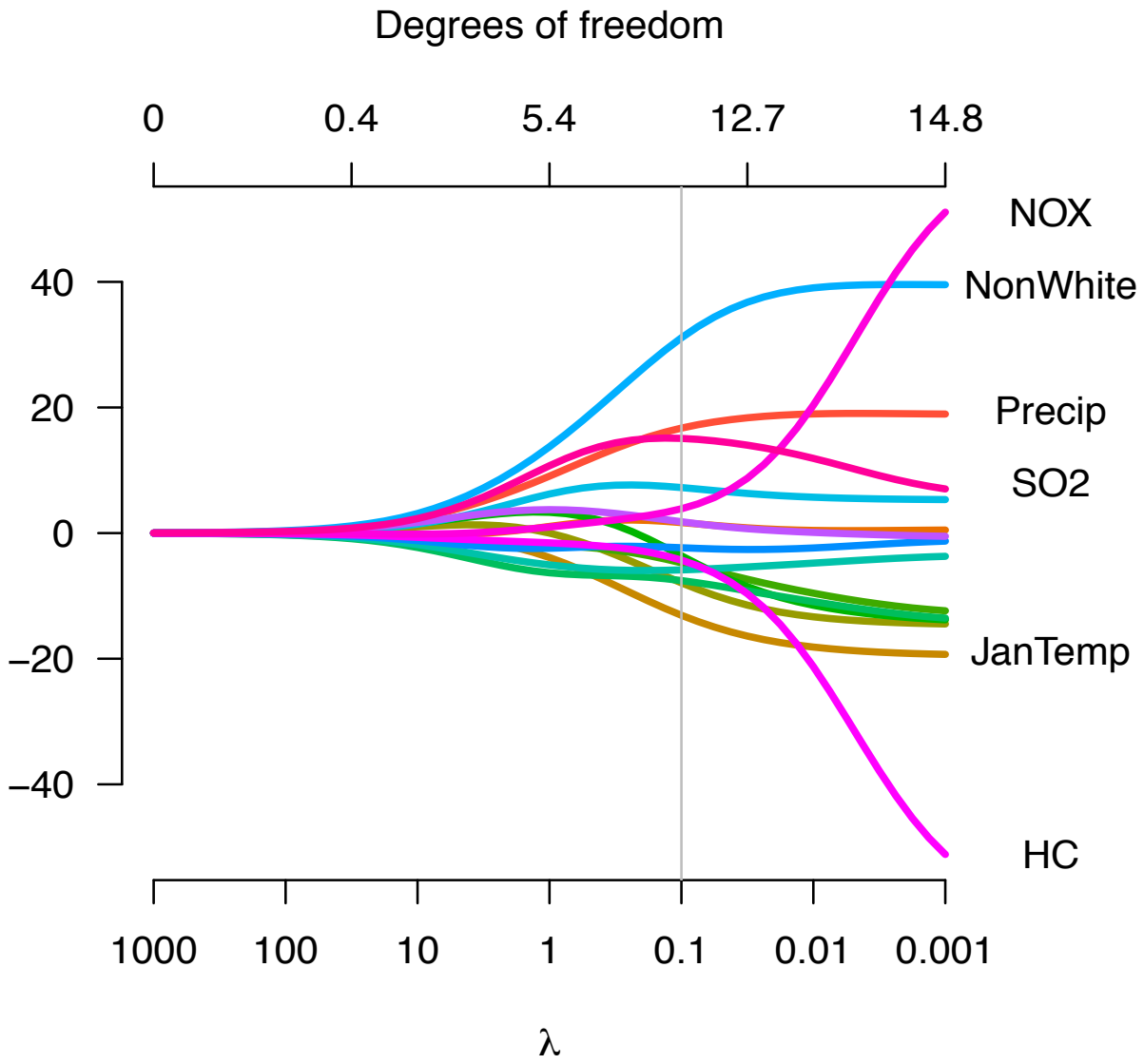
```
plot(fit)
coef(fit)
predict(fit)
summary(fit)
confint(fit)
```

- Pour illustrer la régression ridge en pratique, nous allons maintenant considérer une étude qui est conçue pour estimer la relation entre la pollution et la mortalité tout en ajustant pour les effets potentiellement confondants/d'interaction du climat et des conditions socio-économiques.
- Pour quantifier la pollution, le "potentiel de pollution relative a été mesuré pour trois polluants :
  les hydrocarbures (HC),
  les oxydes d'azote (NOX)
  et le dioxyde de soufre (SO_2)
  dans 60 Régions métropolitaines standard aux États-Unis entre $1959 - 1961$.
- Le résultat d'intérêt est la mortalité totale, toutes causes confondues, ajustée sur l'âge, pour 100000 habitants

- Au total, il y a $p = 15$ variables explicatives : les trois variables de pollution, 8 variables démographiques / socioéconomiques et 4 variables climatiques
- Bien que peu considèrent que $p = 15$ constitue une “ grande dimension", le modèle du maximum de vraisemblance se heurte néanmoins à une taille d’échantillon de seulement 60 et à une forte corrélation entre plusieurs variables
- Comme nous le verrons, cela ne permet pas d’apporter une réponse fiable à la question primordiale de la relation entre la pollution et la mortalité.

Régression ridge : Etude des données sur la pollution

Tracé du Ridge / Coefficient paths/ Chemins des coefficients/Trajectoires des coefficients

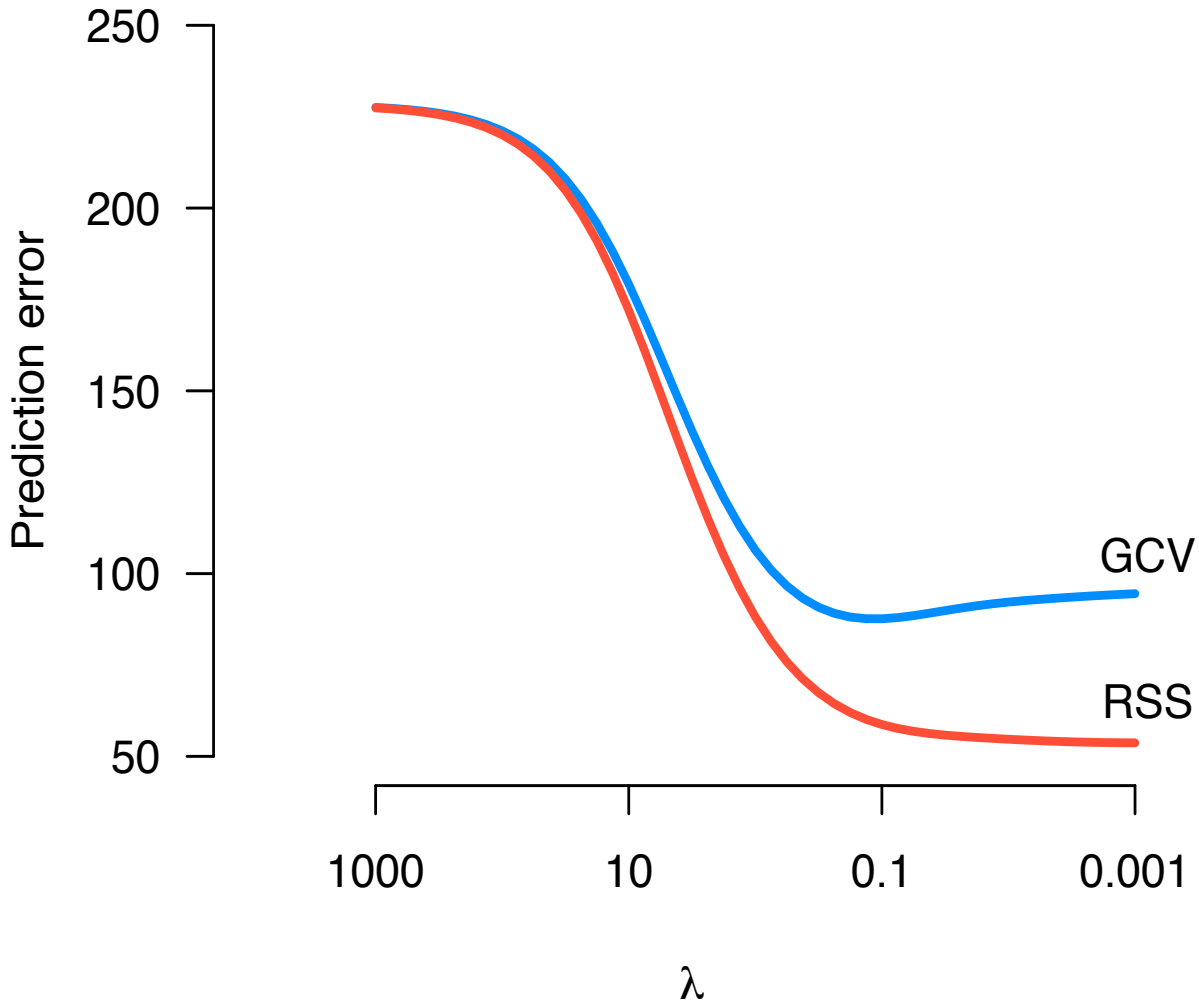


- Il est particulièrement instructif d'examiner les trajectoires des coefficients des trois paramètres de pollution, qui sont tous fortement corrélés les uns avec les autres.
- À de faibles valeurs de λ, les estimations indiquent que la pollution par les NOX a un très fort effet nocif, tandis que la pollution par les HC a un très fort effet protecteur.
- Ce résultat est surprenant, et en effet assez difficile à croire - *l'augmentation de la quantité de HC pollution devrait sauver* 60 *vies par* 100000 ?
- Cependant, à mesure que nous augmentons la pénalité ridge, nous constatons que les effets estimés pour ces deux types de pollution tombent assez rapidement près de zéro.

- Une histoire parallèle est racontée en examinant le chemin du coefficient SO_2.
- SO_2 est corrélé avec HC et NOX (bien que pas aussi fortement corrélés que HC et NOX sont les uns avec les autres), donc sa solution est affectée par les effets estimés pour les deux autres polluants.
- En particulier, alors que la plupart des autres estimations de coefficients augmentent de façon monotone lorsque λ diminue de l'∞ à 0, l'effet estimé de SO_2 augmente, puis diminue.
- Par conséquent, en fonction de la valeur de λ choisie, la pollution par le SO_2 est soit beaucoup plus importante, soit beaucoup moins importante que la pollution par les HC et les NOX.

Régression ridge : Etude des données sur la pollution

Erreur d'ajustement et erreur de prédiction



Régression ridge : Etude des données sur la pollution

t-statistics pour MLE et ridge

Pollution terms:

	Ridge	OLS
SO2	2.78	0.59
NOX	0.37	1.35
HC	-0.41	-1.39

	Ridge	OLS
NonWhite	3.94	3.40
Precip	2.51	2.09
Density	1.45	0.92
Humidity	0.34	0.09
Poor	0.23	-0.05
WhiteCol	-0.39	-0.12
House	-0.46	-1.54
Over65	-0.60	-1.08
Sound	-0.89	-0.38
Educ	-1.04	-1.46
JulyTemp	-1.09	-1.65
JanTemp	-1.77	-1.77

- Pour les t-statistiques de la page précédente, nous avons pris l'erreur standard comme étant la racine carrée des éléments diagonaux de

$$\nabla^2_\beta Q^{-1} = \frac{\sigma^2}{n} \left(n^{-1}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}\right)^{-1},$$

  en utilisant $\widehat{\sigma}^2 = \text{RSS}/(n - \text{df})$ pour estimer σ^2. Ce qui n'est pas la seule possibilité.

- Notons que certains termes deviennent plus significatifs avec une pénalité ridge ajoutée, alors que d'autres deviennent moins significatifs. Bien que les estimateurs soient réduits/rétrécis vers zéro, le fait que la variance soit réduite peut entraîner une augmentation de la signification (c'est-à-dire, contre $\beta = 0$).

- La principale limitation de la régression ridge est le fait que tous ses coefficients sont différents de zéro.
- Cela pose deux problèmes considérables pour la régression en grande dimension :
  - Les solutions deviennent très difficiles à interpréter.
  - Le fardeau/la lourdeur de calcul devient important.
- Il est donc souhaitable d'avoir des modèles qui permettent à la fois le *rétrécissement/shrinkage* et *la sélection*. En d'autres termes, pour conserver les avantages de la régression ridge tout en sélectionnant un sous-ensemble de co-variables importantes.