Université de Strasbourg
M1 - Magistère
Statistique - études de cas

Ségolen Geffray
geffray@math.unistra.fr
Année 2016/2017

Sujet 12: Etude de l'estimateur des coefficients de régression par moindres carrés ordinaires et ridge lorsque p devient grand devant n

---

• **Quelques rappels sur la régression linéaire**:
De manière générale, le but de tout modèle de régression est de spécifier une relation (de nature stochastique) entre une variable Y appelée variable à expliquer et un certain nombre de variables explicatives $X^{(1)}, ..., X^{(p)}$. Dans toute la suite, on supposera que les variables explicatives sont de loi continue.
Notons $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ pour $i = 1, ..., n$, les observations correspondant à n individus que l'on suppose distribuées comme $(Y, X^{(1)}, ..., X^{(p)})$. Le modèle de régression linéaire gaussien standard s'écrit sous la forme générique suivante:

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_i^{(k)} + \varepsilon_i, \qquad \text{où} \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2). \tag{1}$$

Le modèle de régression linéaire standard repose ainsi sur les hypothèses suivantes:

($\mathbf{H_1}$) **linéarité**: l'espérance conditionnelle de la variable réponse vaut

$$\mathbb{E}[Y_i | X_i^{(1)}, ..., X_i^{(p)}] = \beta_0 + \sum_{k=1}^{p} \beta_k X_i^{(k)}.$$

Attention, bien que l'espérance conditionnelle de la variable réponse soit une combinaison affine des covariables, la linéarité s'apprécie relativement à la linéarité de $\mathbb{E}[Y_i | X_i^{(1)}, ..., X_i^{(p)}]$ en les paramètres.
On supposera toujours que le modèle est identifiable, à savoir

$$\beta := (\beta_0, ..., \beta_p)^t = (\beta'_0, ..., \beta'_p)^t := \beta' \quad \Longrightarrow \quad \mathbb{E}_\beta[Y_i | X_i^{(1)}, ..., X_i^{(p)}] = \mathbb{E}_{\beta'}[Y_i | X_i^{(1)}, ..., X_i^{(p)}].$$

($\mathbf{H_2}$) **centrage des erreurs**: $\mathbb{E}[\varepsilon_i] = 0$ pour $i = 1, ..., n$.

($\mathbf{H_3}$) **exogénéité = non-endogénéité**: $(X_i^{(1)}, ..., X_i^{(p)})$ et ε_i sont indépendants pour $i = 1, ..., n$.

($\mathbf{H_4}$) **non-colinéarité des covariables**: les covariables sont non-corrélées entre elles, ce qui se traduit en pratique par le fait que les vecteurs $(X_1^{(k)}, ..., X_n^{(k)})$ sont non-colinéaires pour $k = 1, ..., p$. On écartera également le cas trivial où l'un de ces vecteurs serait colinéaire au vecteur de longueur n dont toutes les composantes sont égales à 1.

($\mathbf{H_5}$) **non-corrélation**: les vecteurs $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ sont non-corrélés pour $i = 1, ..., n$.

($\mathbf{H_6}$) **homoscédasticité**: les lois conditionnelles de Y_i sachant $X_i^{(1)}, ..., X_i^{(p)}$ ont même variance donc on a $\sigma_i^2 = \sigma^2$ pour $i = 1, ..., n$.

Le modèle de régression linéaire standard est gaussien lorsqu'on effectue l'hypothèse supplémentaire suivante:

($\mathbf{H_7}$) **normalité**: conditionnellement à $X_i^{(1)}, ..., X_i^{(p)}$, la variable de réponse Y_i suit la loi normale.

Notons I_n=matrice identité de taille $n \times n$ et avec

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & X_1^{(1)} & \dots & X_1^{(p)} \\ \vdots & \vdots & & \vdots \\ 1 & X_n^{(1)} & \dots & X_n^{(p)} \end{pmatrix} = \begin{pmatrix} \mathbb{X}_1 \\ \vdots \\ \mathbb{X}_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Le modèle de régression linéaire homoscédastique se réécrit matriciellement sous la forme:

$$\mathbb{Y} = \mathbb{X}.\beta + \boldsymbol{\varepsilon}, \qquad \text{où} \quad \boldsymbol{\varepsilon} \sim (0_n, \sigma^2 I_n) \text{ et } \boldsymbol{\varepsilon} \perp\!\!\!\perp \mathbb{X}.$$

Le modèle de régression linéaire gaussien homoscédastique s'obtient sous forme matricielle lorsque l'on ajoute l'hypothèse $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0_n, \sigma^2 I_n)$.

On dit que le modèle est bien spécifié lorsqu'il correspond effectivement au mécanisme ayant servi à généré les données. Dans le cas contraire, on dit qu'il est mal spécifié.

• **Estimation de l'effet des variables explicatives**:
Dans le cadre du modèle linéaire gaussien standard présenté ci-dessus, on se demande si les variables explicatives introduites dans le modèle ont un effet ou non sur la variable à expliquer. Cet effet est quantifié par le coefficient de régression correspondant.
L'estimateur des moindres carrés ordinaires de β est $\widehat{\beta} = (\mathbb{X}^t.\mathbb{X})^{-1}.\mathbb{X}^t.\mathbb{Y}$. Sous les hypothèses $(H_1) - (H_4)$, cet estimateur est sans biais

$$\mathbb{E}[\widehat{\beta}|\mathbb{X}] = \beta\,.$$

Sous les hypothèses $(H_1) - (H_6)$, la variance de $\widehat{\beta}$ est

$$\mathrm{Var}(\widehat{\beta}|\mathbb{X}) = \sigma^2(\mathbb{X}^t.\mathbb{X})^{-1}\,.$$

Notons $\widehat{\sigma}^2$ l'estimateur de la variance de σ^2 défini par:

$$\widehat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = \frac{\widehat{\boldsymbol{\varepsilon}}^t.\widehat{\boldsymbol{\varepsilon}}}{n-(p+1)}$$

en notant

$$\widehat{\varepsilon}_i = Y_i - \mathbb{X}_i.\widehat{\beta}$$

et

$$\widehat{\boldsymbol{\varepsilon}} = \mathbb{Y} - \mathbb{X}.\widehat{\beta}\,.$$

Sous les hypothèses (H_1)-(H_6), l'estimateur $\widehat{\sigma}^2$ est sans biais pour σ^2.

Introduisons une hypothèse supplémentaire:
($\mathbf{H_8}$): $\dfrac{\mathbb{X}^t.\mathbb{X}}{n} \xrightarrow{\mathbb{P}} Q$ lorsque $n \to \infty$ où Q est une matrice symétrique définie positive.

Sous les hypothèses (H_1)-(H_6) et (H_8), la convergence $\widehat{\beta} \xrightarrow{\mathbb{P}} \beta$ a lieu lorsque $n \to \infty$.

Sous les hypothèses $(H_1) - (H_7)$, la loi de l'estimateur $\widehat{\beta}$ de β est

$$\widehat{\beta}_{EMV} \sim \mathcal{N}_{p+1}(\beta, \sigma^2(\mathbb{X}^t.\mathbb{X})^{-1}),$$

et, pour $j = 0, 1, ..., p$, on a

$$\frac{\left(\widehat{\beta}^{(j)} - \beta^{(j)}\right)}{\sqrt{\widehat{\sigma}^2[(\mathbb{X}^t.\mathbb{X})^{-1}]_{j,j}}} \sim T(n - (p+1)).$$

• **La régression pénalisée**:
L'idée des méthodes de régression biaisée est la suivante. Une mesure usuelle de la qualité d'un estimateur consiste à calculer le risque quadratique de l'estimateur des coefficients de régression. Ce risque quadratique est défini en dimension $(p+1)$ par une matrice de taille $(p+1) \times (p+1)$:

$$\begin{aligned}
R_\beta\left(\widehat{\beta}\right) &= \mathbb{E}\left[\underbrace{\left(\widehat{\beta} - \beta\right)}_{(p+1)\times 1} . \underbrace{\left(\widehat{\beta} - \beta\right)^t}_{1\times(p+1)}\right] \\
&= \mathbb{E}\left[\left(\widehat{\beta} - \mathbb{E}(\widehat{\beta}) + \mathbb{E}(\widehat{\beta}) - \beta\right) . \left(\widehat{\beta} - \mathbb{E}(\widehat{\beta}) + \mathbb{E}(\widehat{\beta}) - \beta\right)^t\right] \\
&= \underbrace{\mathbb{E}\left[\left(\widehat{\beta} - \mathbb{E}(\widehat{\beta})\right) . \left(\widehat{\beta} - \mathbb{E}(\widehat{\beta})\right)^t\right]}_{\mathrm{Var}(\widehat{\beta})} + \underbrace{\left(\mathbb{E}(\widehat{\beta}) - \beta\right) . \left(\mathbb{E}(\widehat{\beta}) - \beta\right)^t}_{\simeq \text{ carré du biais}} \\
&\quad + \underbrace{\mathbb{E}\left[\left(\widehat{\beta} - \mathbb{E}(\widehat{\beta})\right) . \left(\mathbb{E}(\widehat{\beta}) - \beta\right)^t\right]}_{=0_{(p+1)\times(p+1)}} + \underbrace{\mathbb{E}\left[\left(\mathbb{E}(\widehat{\beta}) - \beta\right) . \left(\widehat{\beta} - \mathbb{E}(\widehat{\beta})\right)^t\right]}_{=0_{(p+1)\times(p+1)}}
\end{aligned}$$

Lorsque $\mathbb{X}$ est de rang plein, ie $\mathrm{rang}(\mathbb{X}) = p+1$, ce qui équivaut à $\mathbb{X}^t.\mathbb{X}$ inversible, alors l'EMCO de β est BLUE (Best Linear Unbiased Estimator) pourvu que $\mathbb{Y} = \mathbb{X}\beta + \varepsilon$ avec $\varepsilon \perp\!\!\!\perp \mathbb{X}$, $\mathbb{E}[\varepsilon] = 0_n$ et $\mathrm{Var}(\varepsilon) = \sigma^2 I_n$. Dans ce cas,

$$R_\beta(\widehat{\beta}) = 0_{(p+1)\times(p+1)} + \mathrm{Var}\left(\widehat{\beta}\right) = \sigma^2(\mathbb{X}^t.\mathbb{X})^{-1}.$$

Que vaut-il mieux?

1. solution A: pas de biais mais une variance qui est susceptible d'exploser, typiquement lorsque $\mathbb{X}$ n'est pas de rang plein à cause de la multicolinéarité des colonnes de $\mathbb{X}$ ou de la dimension du problème.

2. solution B: un petit peu de biais mais une variance contrôlée.

Idée: ne pourrait-on pas rajouter artificiellement un petit peu de biais pour faire désenfler la variance?

• **La régression ridge**:

L'estimateur ridge de β est défini comme étant la solution en β des équations normales modifiées comme suit:

$$(\mathbb{X}^t.\mathbb{X} + \phi I_p).\beta = \mathbb{X}^t.\mathbb{Y}\,.$$

Intuitivement, on cherche à faire en sorte que $\mathbb{X}^t.\mathbb{X}$ se rapproche de I_p. Pour cela, on a rajouté $\phi \geq 0$ à chaque élément diagonal de $\mathbb{X}^t.\mathbb{X}$. On obtient:

$$\widehat{\beta}_{ridge} = (\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1}.\mathbb{X}^t.\mathbb{Y}\,.$$

Le paramètre ϕ ou de manière équivalente s contrôle la complexité du modèle et s'appelle le *shrinkage parameter* = paramètre de réduction ou paramètre de rétrécissement puisque son rôle consiste à rétrécir (en variance) les coefficients estimés de β.
Le paramètre ϕ, inhérent à la méthode ridge, est à choisir dans $\mathbb{R}^+$ et non pas à estimer. C'est pourquoi on l'appelle parfois hyperparamètre pour mieux le distinguer des paramètres de la régression qui eux sont à estimer.
Lorsque $\phi = 0$, l'estimateur ridge coïncide avec l'EMCO.
Lorsque ϕ tend vers $+\infty$, tous les coefficients sont annulés!

Déterminons la valeur moyenne de l'estimateur ridge:

$$\begin{aligned}
\mathbb{E}\left[\widehat{\beta}_{ridge}|\mathbb{X}\right] &= (\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1}.\mathbb{X}^t.\mathbb{E}[\mathbb{Y}|\mathbb{X}] \\
&= (\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1}.\mathbb{X}^t.\mathbb{X}.\beta \\
&= (\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1}.\Big((\mathbb{X}^t.\mathbb{X} + \phi I_p) - \phi I_p\Big).\beta \\
&= \beta - \phi(\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1}.\beta\,.
\end{aligned}$$

On en déduit le biais de l'estimateur ridge:

$$b_\beta\left[\widehat{\beta}_{ridge}|\mathbb{X}\right] = -\phi(\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1}.\beta\,.$$

Déterminons la variance de l'estimateur ridge:

$$\begin{aligned}
\mathrm{Var}\left(\widehat{\beta}_{ridge}|\mathbb{X}\right) &= (\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1}.\mathbb{X}^t.\underbrace{\mathrm{Var}(\mathbb{Y}|\mathbb{X})}_{\sigma^2 I_p}\mathbb{X}.(\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1} \\
&= \sigma^2(\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1}.\mathbb{X}^t.\mathbb{X}.(\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1}\,.
\end{aligned}$$

On a déjà montré que l'EMCO est sans biais et sa variance est $\mathrm{Var}\left(\widehat{\beta}\right) = \sigma^2(\mathbb{X}^t.\mathbb{X})^{-1}$.
Il n'est pas aisé de comparer deux matrices symétriques, aussi nous prendrons une mesure de la qualité globale via la trace. La trace de la variance de l'EMCO est:

$$\mathrm{Tr}\left(\mathrm{Var}(\widehat{\beta})\right) = \mathrm{Tr}\left(\sigma^2(\mathbb{X}^t.\mathbb{X})^{-1}\right) = \sigma^2\sum_{j=1}^{p}\frac{1}{\lambda_j}$$

où les λ_j sont les valeurs propres (≥ 0) de la matrice symétrique semi-définie positive $\mathbb{X}^t.\mathbb{X}$.
La trace de la variance de l'estimateur ridge est:

$$\sum_{j=1}^{p}\mathrm{Var}\left(\widehat{\beta}_{j,ridge}\right) = \sigma^2\sum_{j=1}^{p}\frac{\lambda_j}{(\lambda_j+\phi)^2}$$

où les λ_j sont les valeurs propres (≥ 0) de la matrice symétrique semi-définie positive $\mathbb{X}^t.\mathbb{X}$ et où P est la matrice de passage telle que $\mathbb{X}^t.\mathbb{X} = P.\text{diag}(\lambda_j)_{j=1,\ldots,p}.P^t$.
Ainsi, choisir un $\phi > 0$ entraine une réduction de la variance (sommée) des composantes de l'estimateur ridge puisque, pour tout $\phi > 0$, on a

$$\sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + \phi)^2} < \sum_{j=1}^{p} \frac{1}{\lambda_j}\,.$$

Ainsi, plus ϕ est grand, plus la variance sommée est faible. Mais, plus ϕ est grand, plus le biais est grand. Comment calibrer au mieux la valeur du paramètre ϕ? Poursuivons par la comparaison des traces des risques quadratiques.

Déterminons le risque quadratique de l'estimateur ridge:

$$\begin{aligned} R_\beta(\widehat{\beta}_{ridge}) &= b_\beta\left[\widehat{\beta}_{ridge}|\mathbb{X}\right].b_\beta\left[\widehat{\beta}_{ridge}|\mathbb{X}\right]^t + \text{Var}\left(\widehat{\beta}_{ridge}|\mathbb{X}\right) \\ &= (\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1}.(\phi^2\beta.\beta^t + \sigma^2(\mathbb{X}^t.\mathbb{X}))(\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1}. \end{aligned}$$

On peut en déduire que:

$$\text{Tr}\left(R_\beta(\widehat{\beta}_{ridge})\right) = \sum_{j=1}^{p} \frac{\sigma^2\lambda_j + \phi^2([P^t.\beta]_j)^2}{(\lambda_j + \phi)^2}\,.$$

Par ailleurs, puisque l'EMCO est sans biais, le risque quadratique de l'EMCO est:

$$R_\beta(\widehat{\beta}) = \sigma^2(\mathbb{X}^t.\mathbb{X})^{-1}.$$

La trace du risque quadratique de l'EMCO est donc:

$$\text{Tr}\left(R_\beta(\widehat{\beta})\right) = \sigma^2\sum_{j=1}^{p} \frac{1}{\lambda_j}\,.$$

On peut montrer que pour

$$\phi \leq \frac{2\sigma^2}{\beta^t.\beta}$$

la trace du risque quadratique de l'estimateur ridge est inférieure à celle du risque quadratique de l'EMCO. Notons que cette condition est indépendante des variables explicatives mais dépend de β qui est inconnu!

Différentes stratégies sont alors possibles.
1ère stratégie: *HKB ridge estimate* (Hoerl, Kennard et Baldwin, 1975)

- déterminer une première estimation respective de β et σ^2 par MCO notées $\widehat{\sigma^2}$ et $\widehat{\beta}$, pourvu que cela soit possible du point de vue algorithmique,
- prendre $\phi = \dfrac{p\widehat{\sigma^2}}{\widehat{\beta}^t.\widehat{\beta}}$.

2ème stratégie: *LW ridge estimate* (Lawless et Wang, 1976)

- déterminer une première estimation respective de β et σ^2 par MCO notées $\widehat{\sigma^2}$ et $\widehat{\beta}$, pourvu que cela soit possible du point de vue algorithmique,

- prendre $\phi = \dfrac{p\widehat{\sigma^2}}{\widehat{\beta}^t.\mathbb{X}^t.\mathbb{X}.\widehat{\beta}}$.

3ème stratégie: utiliser le critère de validation croisée généralisée

$$GCV(\phi) = \frac{\|\mathbb{Y} - H_\phi.\mathbb{Y}\|^2}{n - \mathrm{Tr}(H_\phi)}$$

avec $H_\phi = (\mathbb{X}^t.\mathbb{X} + \phi I_p)^{-1}.\mathbb{X}^t$ la matrice telle que $\widehat{\mathbb{Y}} = H_\phi.\mathbb{Y}$:

- calculer le critère sur une grille de ϕ,
- choisir la valeur qui minimise le critère GCV.

• Implémentation au moyen du logiciel `R`:
Le logiciel `R` permet d'ajuster simplement un modèle de régression à des données. Pour ajuster un modèle linéaire sur un échantillon $(Y_i, X_i^{(1)}, ..., X_i^{(p)})_{i=1,...,n}$, on stockera les n valeurs des Y_i dans un vecteur `y`, puis, pour $k = 1, \ldots, p$, on stockera les n valeurs des $X_i^{(k)}$ dans un vecteur `xk`. Le modèle linéaire est alors ajusté au moyen de l'instruction suivante (où $p = 3$)

```
mylm<- lm(y~x1+x2+x3)
```

et le résultat est stocké dans l'objet `mylm`. L'instruction suivante permet d'accéder à l'ensemble des quantités calculées:

```
summary(mylm)
```

Sont notamment calculées, les estimations des β_j pour $j = 0, \ldots, p$ et les estimations de leurs écarts-types. L'objet `summary(mylm)` est une liste dont on peut récupérer chacune des composantes qui nous intéresse. Pour cela, l'instruction `str(summary(mylm))` permet de voir le nom et la structure des différentes composantes de cette liste.
La fonction `lm.ridge` du package `MASS` permet d'ajuster une régression ridge. Pour charger le package `MASS`, exécuter l'instruction

```
library(MASS)
```

au début de chaque session.

Exercice 1.
Dans le cadre du modèle de régression linéaire gaussien standard, lorsque p devient grand devant n, illustrer de manière empirique à partir de données simulées le comportement de l'estimateur des moindres carrés ordinaires et de l'estimateur ridge des coefficients de régression, en termes de biais, variance, écart quadratique moyen, consistance et distribution de l'estimateur $\widehat{\beta}$. Vous veillerez à faire varier également:

- le critère de choix de l'hyperparamètre ϕ,
- l'ampleur de p devant n,
- la taille de l'échantillon n simulé,
- la variance de l'erreur résiduelle simulée.