INSA – MS ESD Année 2019-2020
BRUNEVAL Guillaume
ROUFFIAC Jean-Eudes

STAT2 - Régression pénalisée

Rapport de TP n°5

Titre : *Étude par simulation de l'intérêt de la régression Ridge*

1 Introduction

L'objectif de ce TP est de découvrir une méthode de régression pénalisée qui est la méthode ridge.

Dans un premier temps, nous utiliserons des données simulées afin d'étudier les conditions sous lesquelles la régression Ridge est plus efficace que les moindres carrées ordinaires.

Dans un second temps, nous étudierons le même jeu de données que le TP précédent et nous comparerons les performances obtenues avec le modèle linéaire classique et celles obtenues avec le modèle Ridge.

2 Données simulées

Dans cette partie, nous simulerons des ensembles de données plus ou moins grands, prenant en compte plus ou moins de variables, tels que certaines variables soient liées entre elles et d'autres non, et tels que les valeurs à estimer soient proches de 0 et d'autres non. On estime ensuite les coefficients de la régression par la méthode des moindres carrés ordinaires et par la méthode Ridge. On calcule ensuite les erreurs quadratiques des coefficients pour chaque échantillon simulé ; on compare la distribution de ces erreurs. On peut enfin visualiser l'effet de la pénalisation sur les prévisions. On déterminera par ailleurs la valeur optimale de λ pour chaque échantillon à l'aide du GCV.

Dans le code R, des ensembles de données sont générés chacun avec n observations et p variables explicatives, suivant une distribution normale à p dimensions de moyenne $E(X_j) = 0$ et de covariance $Cov(X_j, X_t) = \rho^{|(j-l)|}$. La vraie valeur de β est k pour les 10 premières variables explicatives et zéro pour toutes les autres. Le but va être de faire varier les facteurs en jeu n, p, k et ρ dans le but de déterminer les conditions pour lesquelles la régression Ridge est préférable aux moindres carrés ordinaires.

Influence de la taille n de l'échantillon

Influence sur les valeurs optimales de lambda

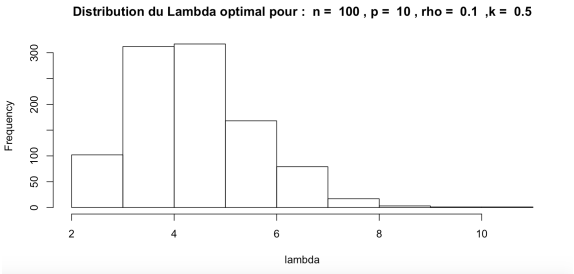


FIGURE 1 – Histogramme des valeurs de lambda optimales pour n = 100

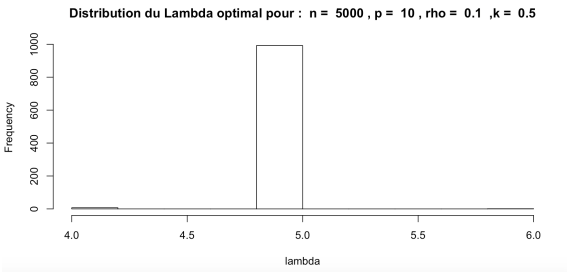


FIGURE 2 – Histogramme des valeurs de lambda optimales pour n = 5000

Lorsque n augmente, on peut observer une nette concentration des valeurs de lambda autour de 5. Pour n petit, la valeur optimale de lambda est distibuée avec une ventilation plus grande. On note cependant également une concentration accrue autour de 5. On détermine donc le lambda optimal de manière plus certaine lorsque l'échantillon est grand.

Influences sur les performances de OLS et Ridge

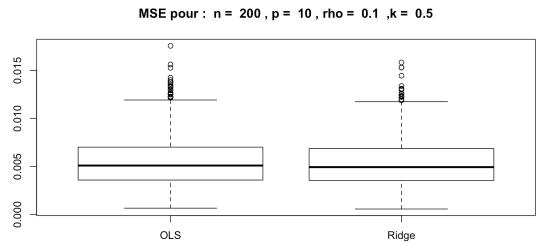


FIGURE 3 – Diagrammes en boîte comparatif des indicateurs MSE de OLS et Ridge pour n = 200

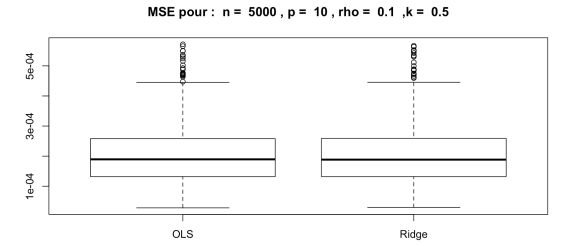


FIGURE 4 – Diagrammes en boîte comparatif des indicateurs MSE de OLS et Ridge pour n = 5000

Lorsque n augmente, l'indicateur MSE diminue fortement, de l'ordre de 40 fois lorsqu'on passe d'une taille d'échantillon de 200 à 5000. La dispersion des MSE elle aussi diminue puisque l'écart interquartile est environ divisé par 40 également. On constate en revanche que les deux modèles adoptent un comportement similaire lorsque n évolue. La valeur de n n'est donc pas un facteur de choix entre OLS et Ridge.

Remarque Nous faisons ici le choix de ne pas commenter les boxplot de ychap car nous ne comprenons pas exactement en quoi consiste la variable ychap dans le code. Il semble qu'elle ne reprenne que la somme des coefficient calculées par le modèle et n'est donc pas vraiment une prédiction. Par ailleurs, les boxplot demeurent semblables pour OLS et Ridge quels que soient les paramètres modifiés, or nous devrions théoriquement constater une variabilité accrue des estimations de Y par OLS lorsque nos variables sont corrélées. Par ailleurs le test de Wilcoxon ne retient jamais l'hypothèse H_0, nous déciderons donc de ne pas l'utiliser. Nous ne commenterons les résultats donnés par l'erreur standard implémentée dans le script que lorsqu'ils seront non nuls, ce qui est rarement le cas.

Explications On vient de constater finalement que plus on a de données, plus on pourra déterminer de manière certaine notre modèle et meilleur il sera en estimation. C'est assez intuitif. Que cela soit le cas pour OLS comme pour Ridge, l'est aussi.

Influence du nombre de variables explicatives p et étude de son rapport avec la taille de l'échantillon.

Influence sur les valeurs optimales de lambda

Lorsque p est petit, on peut observer une nette concentration des valeurs de lambda autour de 5. Lorsque p augmente, la valeur optimale de lambda est distribuée de manière quasi normale autour d'une valeur de 45. On détermine donc le lambda optimal de manière plus certaine lorsque le nombre de variables est petit. Ce phénomène d'incertitude sur la valeur optimale de lambda reste vrai quelque soit la valeur de n.

Influences sur les performances de OLS et Ridge

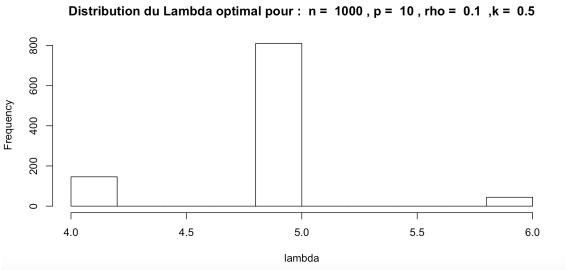


FIGURE 5 – Histogramme des valeurs de lambda optimales pour p = 10, n= 1000

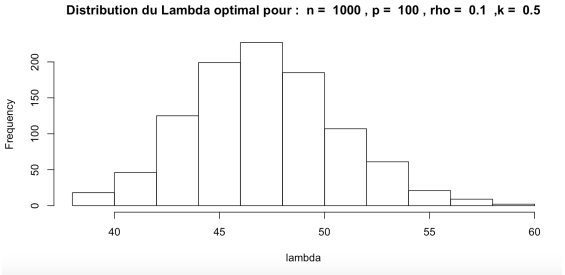


FIGURE 6 – Histogramme des valeurs de lambda optimales pour p = 100, n = 1000

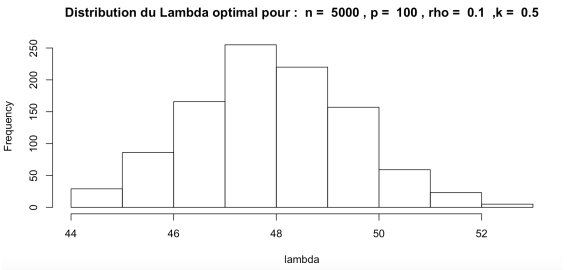


FIGURE 7 – Histogramme des valeurs de lambda optimales pour p = 100, n = 5000

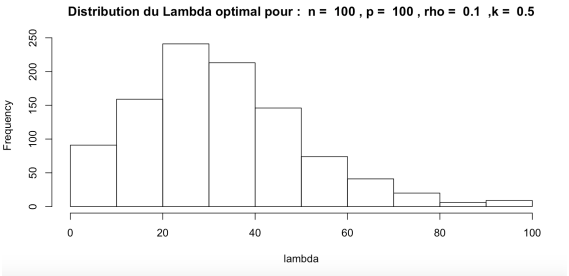


FIGURE 8 – Histogramme des valeurs de lambda optimales pour p = 100, n = 100

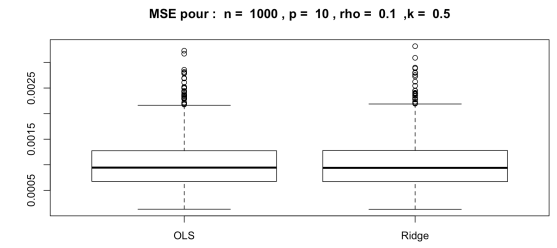


FIGURE 9 – Diagrammes en boîte comparatif des indicateurs MSE de OLS et Ridge pour p = 10 et n = 1000

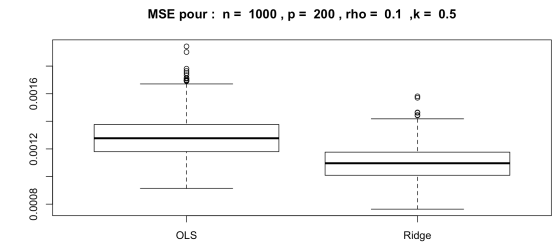


FIGURE 10 – Diagrammes en boîte comparatif des indicateurs MSE de OLS et Ridge pour p = 200 et n = 1000

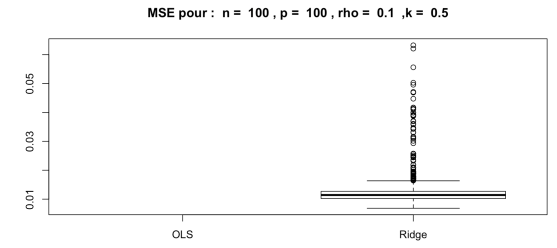


FIGURE 11 – Diagrammes en boîte comparatif des indicateurs MSE de OLS et Ridge pour p = 100 et n = 100

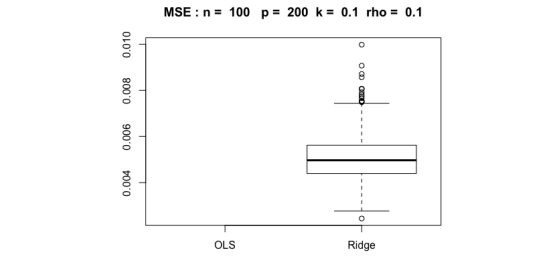


FIGURE 12 – Diagrammes en boîte comparatif des indicateurs MSE de OLS et Ridge pour p = 100 et n = 200

Lorsque p est petit, on observe des performances similaires pour OLS et Ridge, avec tous deux des MSE moyens à 0.009 et des distributions semblables. Lorsque p augmente, Ridge continue de fonctionner aussi bien alors que OLS va voir son indicateur MSE croître en moyenne de 30 à 40%. Lorsque la valeur de p se rapproche de la valeur de n, OLS se met à disfonctionner de manière plus flagrante encore pour finir par ne plus fonctionner du tout lorsque $p >= n$, comme l'indiquent les figures 11 et 12. Le modèle ridge quant à lui ne semble pas significativement impacté par l'accroissemment du nombre de variable. On en conclut dont qu'il est préférable de choisir Ridge pour les jeux de données comportant beaucoup de variables, à fortiori lorsque ce nombre se rapproche ou depasse la taille de l'échantillon.

Explications Soit un jeu de données représenté ici par une matrice X. X est de dimension $n \times p$. Que se passe-t-il lorsque $p >= n$? X est au plus de rang n et notre système est nécessairement lié (que le jeu soit corrélé ou pas), on est au moins dans une situation de multi-colinéarité et $det(X^T X) = 0$. La matrice $X^T X$ ne sera pas inversible et on ne pourra tout simplement pas calculer nos coefficient à l'aide de l'équation normale. OLS ne peut pas donner de résultat. En revanche, la regression Ridge, par essence, applique une pénalité pour rendre la matrice inversible. En effet, on comprend aisément que même si $det(X^T X) = 0$, $det(X^Y X + \lambda I_p)$ sera différent de 0. Cela explique également que l'on trouve des valeurs de λ optimales plus grandes lorsque p grandit (voir Figures 5 et 6).

Influence de la vraie valeur k de β

Remarque : Nous faisons le choix ici de nous placer dans le cas où p est égale à 10, et où il n'y a donc pas de valeurs à estimer égales à 0 afin de bien observer l'effet de la seule variation de k.

Influence sur les valeurs optimales de lambda

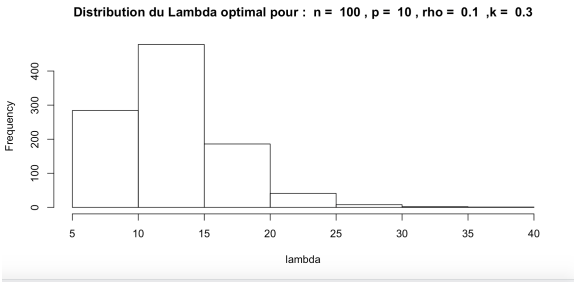


FIGURE 13 – Histogramme des valeurs de lambda optimales pour k = 0.3

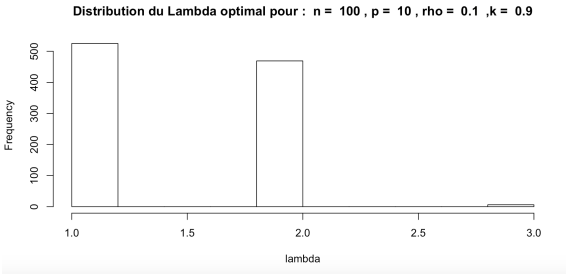


FIGURE 14 – Histogramme des valeurs de lambda optimales pour k = 0.9

Pour les petites valeurs de k, c'est à dire pour des jeux générées avec des petites valeurs de β, les valeurs de lambda sont plus incertaines et se regroupent autour d'environ 12, pour des valeur de k assez élevées, on observe un phénomène assez étrange avec une concentration de valeurs proches de 1 et de 2. Il semble en tout cas établi que des échantillons simulés avec des valeur de k petite nécessitent une pénalité plus lourde.

Influences sur les performances de OLS et Ridge

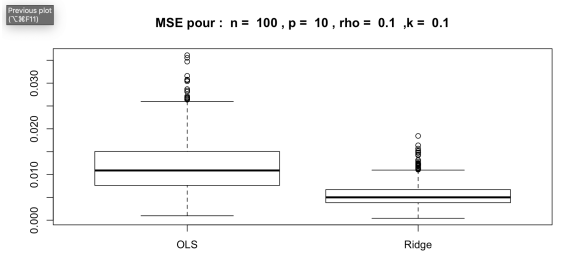


FIGURE 15 – Diagrammes en boîte comparatif des indicateurs MSE de OLS et Ridge pour k = 0.1

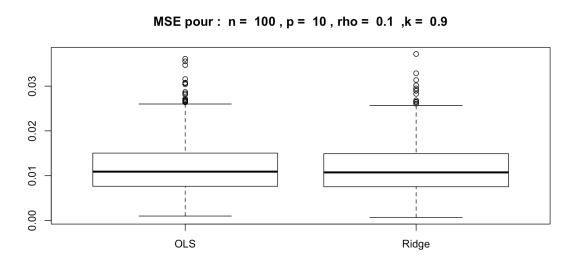


FIGURE 16 – Diagrammes en boîte comparatif des indicateurs MSE de OLS et Ridge pour k = 0.9

Que k soit petit ou grand, la régression OLS fonctionne de manière similaire avec une distribution similaire des MSE. Lorsque k est grand, il semble qu'on puisse utiliser indifféremment OLS et Ridge avec une MSE moyenne autour 0.01 pour k = 0.9. Ridge devient en revanche plus intéressant pour les jeux de données simulées avec des valeurs de k petites, avec une MSE divisée par deux pour k = 0.1 et une variabilité bien plus faible.

Prenons maintenant une valeur plus grande de p : par exemple p = 100.

On observe alors que le phénomène observé précédemment, à savoir un meilleur fonctionnement de Ridge pour les valeurs de k, persiste mais que son effet est amoindri.

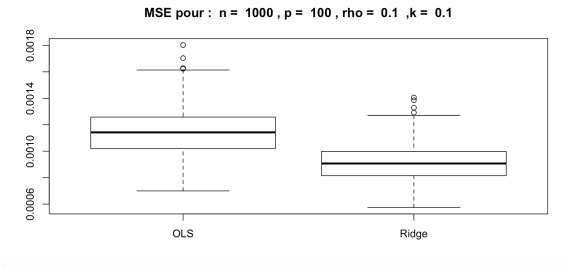


FIGURE 17 – Diagrammes en boîte comparatif des indicateurs MSE de OLS et Ridge pour k = 0.1 et p = 100

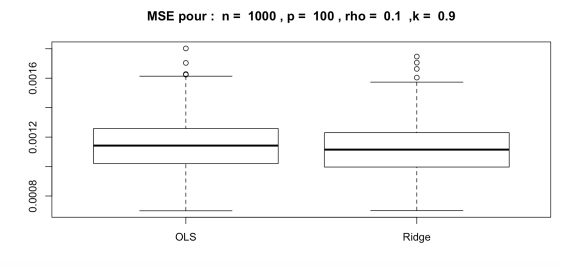


FIGURE 18 – Diagrammes en boîte comparatif des indicateurs MSE de OLS et Ridge pour k = 0.9 et p = 100

Explications (en tout cas tentatives d'explications) L'introduction de la pénalité Ridge est un compromis gagnant. On introduit un biais dans l'estimation de nos paramètres β mais on en fait baisser la variance. Pour la regression Ridge on a les formules suivantes pour nos estimateurs de nos regresseurs β :
$E(\beta_\lambda) = (X^T X + \lambda I_{p+1})^{-1} X^T X \beta$
$Var(\beta_\lambda) = \sigma^2 (X^T X + \lambda I_p)^{-1} X^T X (X^T X + \lambda I_p)^{-1}$

Plus notre paramètre k sera petit et plus nos régresseurs β_{vrai} seront petits, et plus de ce fait nos β estimés seront petits et plus notre biais se rapprochera de celui d'une régression classique, c'est à dire qu'il diminuera. De plus des valeurs de β petites autoriseront des valeur de lambda optimales plus hautes comme le montrent les figures 13 et 14. En augmentant λ, on fait également baisser la variance (dont on voit bien qu'elle ne dépend par ailleurs pas de β) et on améliore globalement notre modèle. C'est du moins ce qu'il me semble observer sur les figures 15 et 16.

Influence de la corrélation ρ entre les variables explicatives

Influence sur les valeurs optimales de lambda

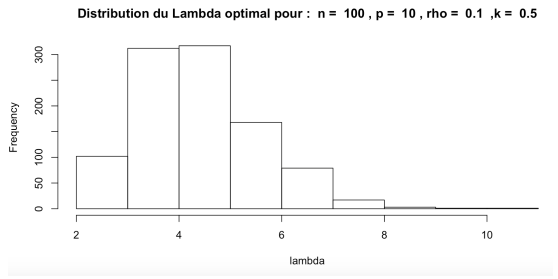


FIGURE 19 – Histogramme des valeurs de lambda optimales pour $\rho = 0.1$

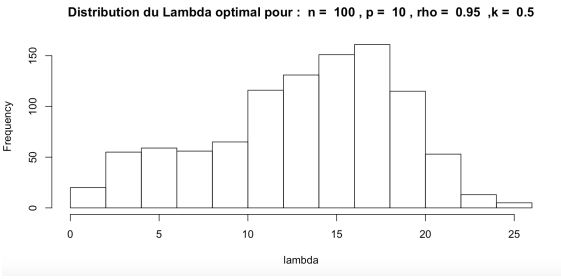


FIGURE 20 – Histogramme des valeurs de lambda optimales pour $\rho = 0.95$

Pour les petites valeurs de ρ, c'est à dire pour des jeux faiblement corrélés, les valeurs de lambda se concentrent autour de 5 et sont distribuées de manière quasi normale avec une variabilité faible, pour des valeur de ρ plus élevées, ici 0.95, on obesrve des valeurs beaucoup plus volatiles, avec un nombre important de valeurs de lambda entre 15 et 20.

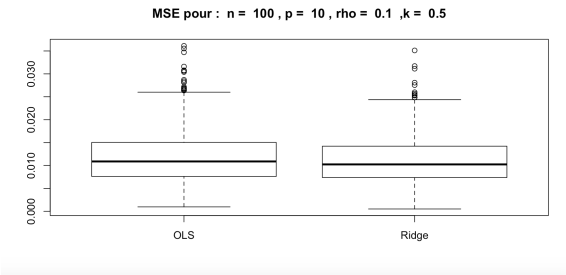


FIGURE 21 – Diagrammes en boîte comparatif des indicateurs MSE de OLS et Ridge pour $\rho = 0.1$

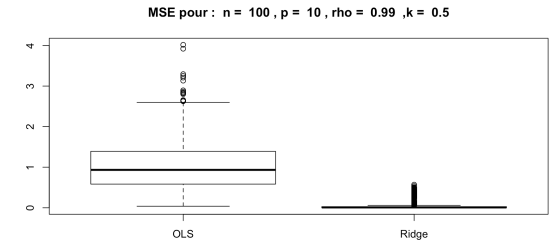


FIGURE 22 – Diagrammes en boîte comparatif des indicateurs MSE de OLS et Ridge pour $\rho = 0.95$

Influences sur les performances de OLS et Ridge

Pour ρ petit, c'est à dire pour des jeux de données faiblement corrélées, OLS et Ridge adoptent un comportement similaire avec une MSE proche de 0.01. Pour les jeu de données fortement corrélés, dans le cas de la figure 20 pour $\rho = 0.99$, ridge conserve un niveau de performance comparable alors qu'OLS disfonctionne complètement et voit sa MSE approcher 1, toutes choses égales par ailleurs. Il est intéressant de mentionner ici que lorsque $\rho = 0.99$, l'erreur standard (telle qu'elle est calculée dans le script avec lequel nous travaillons) pour OLS est de 0.019 alors qu'elle est de 0.003 pour Ridge, cette erreur standard est par ailleurs toujours nulle pour les autres test. Ces chiffres illustrent bien l'impact de la corrélation sur le fonctionneent de OLS.

Explications Soit un jeu de données , que l'on désignera par X, la régression linéaire classique doit inverser la matrice X^TX. Or si le jeu de données est corrélé, $det(X^TX)$ est proche de 0 et il sera plus difficile d'inverser la matrice X^TX. L'essence même de la regression Ridge est d'assurer que cette matrice soit facilement inversible. En effet, on comprend aisément que même si $det(X^TX)$ est proche de 0, $det(X^YX + \lambda I_p)$ ne le sera pas pour λ suffisamment grand. Cela explique également que l'on trouve des valeurs de λ optimales plus grandes pour des jeux de données corrélés (voir Figure 18).

Conclusion

L'intérêt premier d'une régression Ridge est qu'elle est efficace et opérante sur les jeux de données corrélés et lorsque p est égale ou se rapproche de n. Nous avons également observé qu'elle fonctionnait encore mieux lorsque les valeurs à estimer sont petites. La réserve que l'on peut formuler est que pour fonctionner de manière satisfaisante, il faut pouvoir déterminer une valeur de λ optimale ou tout du moins une région de "bonnes valeurs" de λ, ce qui est moins le cas pour de petits jeux de données trop petits par exemple.

2.1 Données réelles

Dans cette partie, nous allons utiliser les mêmes données que pour le TP4, à savoir les données sur la pollution par les PM10 au Havre.

Nous allons faire une régression Ridge avec ces données. L'idée va être d'ajuster un modèle en gardant l'ensemble des variables explicatives. On travaille sur les données brutes (non centrées et réduites). Dans un premier temps, nous devons déterminer la pénalité (λ) Ridge optimal. En effet, une grande valeur de λ va diminuer la variance, mais augmenter le biais. Pour trouver la valeur du paramètre Ridge λ optimal, on utilise une approche par validation croisée : on se donne une suite de valeurs de λ, et on choisit la valeur qui minimise le critère GCV défini par :

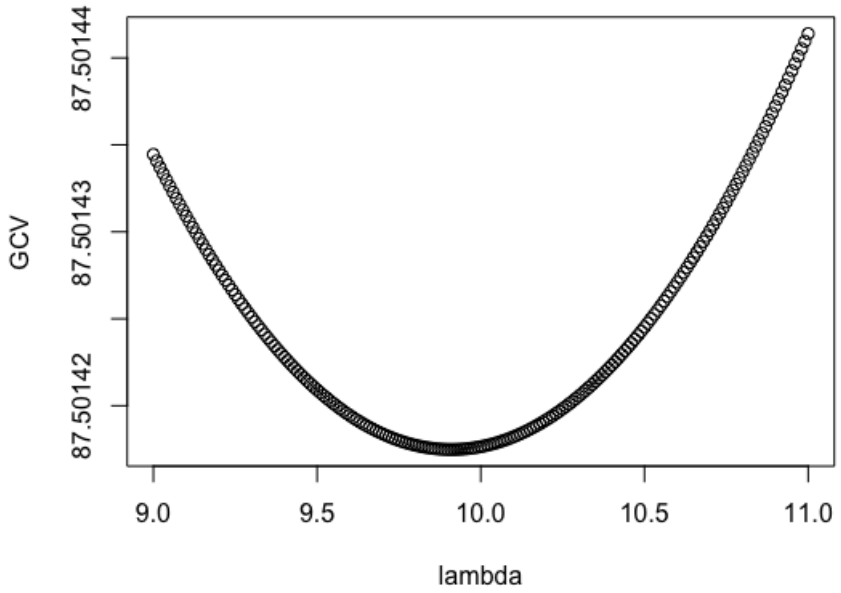
$$GCV(\lambda) = \frac{\left\|Y - X\hat{\beta}_\lambda\right\|^2}{(tr(I_n - A(\lambda)))^2/n} = \frac{\left\|Y - XB_\lambda^{-1}X^TY\right\|^2}{(tr(I_n - A(\lambda)))^2/n} = \frac{\|(I_n - A(\lambda))Y\|^2}{(tr(I_n - A(\lambda)))^2/n}$$

avec $A(\lambda) = XV_\lambda^{-1}X^T$ et $B_(\lambda) = \begin{pmatrix} n & 1_n^T X \\ X^T 1_n & X^T X + \lambda I_p \end{pmatrix}$

On choisit alors la valeur de la pénalité optimale de la manière suivante :

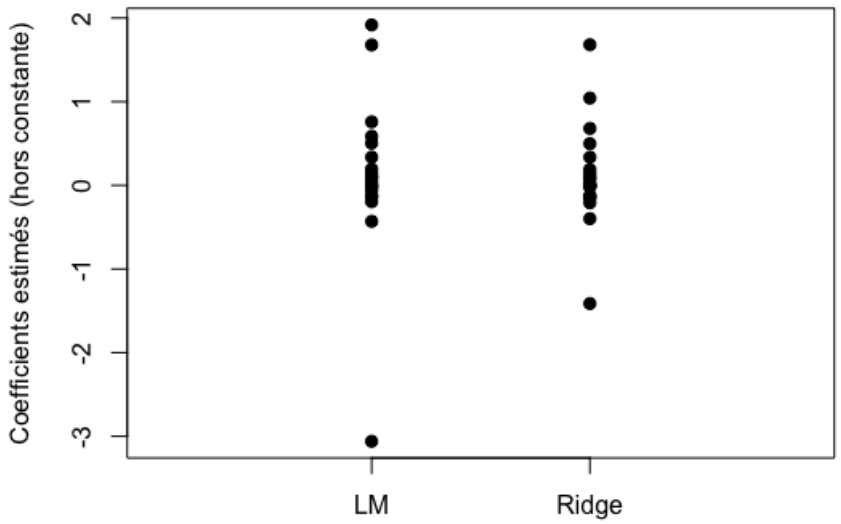
$$\lambda_{opt} = arg\min_\lambda GCV(\lambda)$$

Le code R donné dans l'énoncé du TP permet de calculer le λ optimal. Après quelques essais, on décide de faire varier le paramètre de pénalité entre 9 et 11 par pas de 0.01. Le graphique ci-dessous présente les valeurs du GCV en fonction des valeurs du paramètre de pénalité. On constate que le critère GCV est d'abord décroissant puis croissant. La valeur du paramètre λ qui rend minimum le critère GCV est égale à 9.91.



Une fois le λ optimal trouvé, nous effectuons une régression Ridge avec R en mettant le λ optimal trouvé en paramètre de la fonction. Nous effectuons aussi une régression

linéaire avec la commande *lm*.
Le graphique ci-dessous permet de comparer les coefficients estimés par chacune des méthodes.



On constate que les coefficients estimés par Ridge sont globalement plus petits que ceux obtenus par la méthode des moindres carrés ordinaires.

Nous allons maintenant comparer les performances en estimation et en prévision des modèles "LM", "LM réduit" et "Ridge".

Pour ce faire, nous divisons nos données en deux échantillons : un échantillon d'apprentissage et un échantillon test. Nous allons construire nos modèles en utilisant l'échantillon d'apprentissage, puis nous testerons le modèle sur l'échantillon test. Nous utiliserons les modèles construits dans le TP précédent pour les modèles "LM" et "LM réduit". Ainsi le modèle "LM" est composé de toutes les variables hormis les variables qualitatives et la variable "T.moy" supprimée lors du test VIF. Le modèle "LM réduit" est composé des variables du modèle "LM" hormis les variables "HR.min", "HR.moy", "NO.max" et "SO2.max" supprimées à l'aide d'une méthode de sélection de variables pas à pas descendante. Nous regroupons les critères de performance en estimation et en prévision pour chaque modèle dans le tableau suivant :

	LM		LM réduit		Ridge	
Perf	**Estimation**	**Prévision**	**Estimation**	**Prévision**	**Estimation**	**Prévision**
R	0.73	0.68	0.72	0.68	0.73	0.68
EV	0.53	0.46	0.52	0.46	0.53	0.47
MAE	6.58	6.89	6.58	6.88	6.56	6.86
RMSE	9.09	9.96	9.09	9.95	9.06	9.93
POD	0.26	0.31	0.23	0.28	0.26	0.31
FAR	0.30	0.41	0.27	0.36	0.28	0.41

On remarque que des trois méthodes, c'est la régression Ridge qui est légèrement meilleure. Le RMSE et le MAE sont plus petits en estimation et en prévision, le pourcentage de bonne détection des concentrations de PM10 au dessus de 50 $\mu g/m^3$ est le même que celui du modèle "LM" en prévision et en estimation, mais meilleur que le modèle réduit. Enfin, le taux de fausse alarme (FAR) est plus élevé pour Ridge que pour le modèle réduit.

En conclusion, nous pouvons dire que la régression Ridge permet d'obtenir des résultats très légèrement meilleurs que les résultats obtenus avec les modèles dans le TP précédent. L'avantage de la régression Ridge est qu'elle n'est pas sensible aux colinéarités entre les variables explicatives, donc elle permet de ne pas avoir à faire de VIF ou d'autres tests de colinéarité afin de supprimer des variables. Elle permet donc de garder un maximum de variables, ce qui permet d'avoir de meilleurs résultats. Mais on peut aussi se demander si effectuer la recherche du lambda optimal n'est pas trop coûteux lorsqu'on utilise un jeu de données avec un très grand nombre de données.