La REGRESSION RIDGE

La régression Ridge ordinaire ou bornée ordinaire a été proposée par E. Hoerl et Kennard dans " Ridge regression : biaised estimation for nonorthogonal problems" Technometrics, Vol. 1 Février 1970.
On part de la constatation suivante : lorsque l'on se trouve en face d'un problème de forte colinéarité les valeurs propres de ${}^tXX$,les $d_i^2 = \lambda_i$ correspondant à la colinéarité sont très petits. Hoerl et Kennard proposent de les augmenter faiblement en ajoutant à tous les d_i^2 une même constante K>0, ceci dans le but de rendre stables les coeffcients. Regardons ce que donne cette proposition.
Remarque: Ils ont également proposé une autre méthode, la Ridge Généralisée, pour laquelle la constante ajoutée varie suivant les variables; cette méthode est nettement plus complexe.

1 Définition de la Ridge ordinaire (R.O.)

1.1 Sur le modèle canonique

Nous redéfinissons le modèle canonique (voir cours sur la colinéarité) dans un modèle cette fois **centré et réduit**.
Soit le modèle $\overrightarrow{Y} = X\overrightarrow{a} + \overrightarrow{\epsilon}$, V la matrice orthogonale des vecteurs propres ($V^tV = I$) de ${}^tXX$et Λ sa matrice diagonale des valeurs propres λ_i.
On ${}^tXX = V\Lambda{}^tV$ et $\overrightarrow{Y} = X\overrightarrow{a} + \overrightarrow{\epsilon} = XV^tV\overrightarrow{a} + \overrightarrow{\epsilon}$
En notant le vecteur$XV = Z$ et ${}^tV\overrightarrow{a} = \overrightarrow{\gamma}$ le modèle peut s'écrire $\overrightarrow{Y} = Z\overrightarrow{\gamma} + \overrightarrow{\epsilon}$ c'est le modèle sous sa forme canonique

- On constate que ${}^tZZ = {}^t\,V{}^tXXV = {}^t\,VV\Lambda{}^tVV = \Lambda$ matrice diagonale des vecteurs propres ou des carrés des valeurs singulières

$$ {}^tZZ = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_i & 0 \\ 0 & 0 & \lambda_k \end{pmatrix} = \begin{pmatrix} d_1^2 & 0 & 0 \\ 0 & d_i^2 & 0 \\ 0 & 0 & d_k^2 \end{pmatrix} $$

donc $V_{\overrightarrow{\gamma}} = \sigma^2\ ({}^tZZ)^{-1} = \sigma^2\Lambda^{-1}$ matrice diagonale.

- Dans ces conditions la régression Ridge Ordinaire donne le résultat suivant

  Hoerl et Kennard proposent donc d'augmenter "un peu" les d_i^2 en ajoutant à tous (même à ceux qui sont grands) la même constante K>0. Cela conduit à augmenter tous les termes de la valeur K, comme ${}^tZZ$ est une matrice diagonale , on obtient

$$ {}^tzz + KI = \begin{pmatrix} d_1^2 + K & .. & 0 \\ .. & .. & .. \\ 0 & .. & d_k^2 + K \end{pmatrix} $$

L'estimateur canonique de la Ridge s'écrit donc

$$\overrightarrow{\widehat{\gamma r}} = ({}^tZZ + KI)^{-1t} z\,\overrightarrow{y}$$

Nous discuterons par la suite de la valeur à donner à K. Il devra être petit, mais que signifie "petit"? On le cernera mieux en prenant comme on l'a indiqué au début un modèle dans lequel les variables sont centrées et réduites.

1.2 Estimateur Ridge dans le modèle de base

$$\overrightarrow{\widehat{ar}} = V\overrightarrow{\widehat{\gamma r}} = V({}^tV^tXXV + K^tVV)^{-1t}V^tX\,\overrightarrow{y} = V[{}^tV({}^tXX + KI)V]^{-1}V^tX\,\overrightarrow{y}$$

$$\overrightarrow{\widehat{ar}} = V^tV({}^tXX + KI)^{-1}V^tVX\,\overrightarrow{y}$$

$$\overrightarrow{\widehat{ar}} = ({}^tXX + KI)^{-1}X\,\overrightarrow{y}$$

Car la matrice des vecteurs propres V est orthogonale. La ridge consiste donc à ajouter la constante K à tous les éléments diagonaux de ${}^tXX$. C'est la seconde définition de la ridge.

2 Propriétés de l'estimateur R.O.

2.1 Espérance de l'estimateur Ridge

$E(\overrightarrow{\widehat{ar}}) = E(({}^tXX + KI)^{-1t}X\,\overrightarrow{y}) = E(({}^tXX + KI)^{-1t}X(X\,\overrightarrow{a} + \overrightarrow{\epsilon})$
Si on fait les hypothèses classiques des erreurs d'espérance nulle et des variables explicatives non aléatoires, on obtient,

$$E(\overrightarrow{\widehat{ar}}) = E(({}^tXX + KI)^{-1t}X(X\,\overrightarrow{a} + \overrightarrow{\epsilon}) = ({}^tXX + KI)^{-1t}XX\,\overrightarrow{a}$$

$$E(\overrightarrow{\widehat{ar}}) = ({}^tXX + KI)^{-1}({}^tXX + KI - KI)\,\overrightarrow{a} = \overrightarrow{a} - ({}^tXX + KI)^{-1}\,\overrightarrow{a}$$

De même

$$E(\overrightarrow{\widehat{\gamma r}}) = \overrightarrow{\gamma} - ({}^tZZ + KI)^{-1}\,\overrightarrow{\gamma}$$

L'estimateur Ridge est donc biaisé.

2.2 Variance de l'estimateur Ridge

$$V_{\widehat{ar}} = ({}^tXX + KI)^{-1t}XV_{\overrightarrow{y}}X({}^tXX + KI)^{-1}$$

Si on fait l'hypothèse de non autocorrélation et homoscédasticité alors $V_{\overrightarrow{y}} = \sigma^2 I$

$$V_{\widehat{ar}} = \sigma^2({}^tXX + KI)^{-1t}XX({}^tXX + KI)^{-1}$$

de même

$$V_{\widehat{\gamma r}} = \sigma^2({}^tZZ + KI)^{-1t}ZZ({}^tZZ + KI)^{-1}$$

2.3 Conclusion

Nous nous trouvons face à un estimateur biaisé. Est-il meilleur que les MCO ? Mystère car pour l'instant nous ne savons comparer que des estimateurs sans biais : on prend celui de variance Minimum. Mais pour comparer des estimateurs sans biais avec des estimateurs biaisés il faut introduire un nouveau critère celui du M.S.E.

3 Le critère du Mean Square Erreur (M.S.E)

3.1 Définition du MSE

La colinéarité entraine comme nous l'avons vu des estimateurs qui bien que sans biais ont des variances trop grandes donc des intervalles de confiance trop grands. Il est parfois plus judicieux de prendre des estimateurs un peu biaisés mais avec des variances beaucoup plus faibles, ce qui conduira à des intervalles de confiance plus petits

———(————————[——$\overrightarrow{a}$—$\overrightarrow{b}$———]————————)———————

Avec $\overrightarrow{b} = E(\overrightarrow{\widehat{a}})$ où $\overrightarrow{\widehat{a}}$ est un estimateur biaisé de $\overrightarrow{a}$
Ainsi sur ce graphique (très laid) l'estimateur biaisé est bien meilleur (intervalle de confiance entre les deux crochets) que l'estimateur sans biais mais qui a un intervalle de confiance (entre les deux parenthèses) beaucoup plus grand. Mais comment comparer ces deux estimateurs? Quand des estimateurs sont sans biais le meilleur est celui qui a la plus petite variance. Mais quand l'un est biaisé et l'autre sans biais comment choisir un critère sachant que le biais ne doit pas être grand.
Il existe un critère, celui du MSE (mean square estimator), erreur quadratique moyenne.
Définition : Si le vecteur $\overrightarrow{a}$ est estimé par $\overrightarrow{\widehat{a}}$ estimateur biaisé de $\overrightarrow{a}$ donc d'espérance $E(\overrightarrow{\widehat{a}}) \neq \overrightarrow{a}$, on va définir le MSE par la matrice

$$MSE(\overrightarrow{\widehat{a}}) = E[(\overrightarrow{\widehat{a}} - \overrightarrow{a})\ {}^t(\overrightarrow{\widehat{a}} - \overrightarrow{a})]$$

Le travail se fera plutôt avec la trace de cette matrice en utilisant les propriétés de la trace qui est la somme des éléments de la diagonale du MSE

$$trMSE(\overrightarrow{\widehat{a}}) = E[{}^t(\overrightarrow{\widehat{a}} - \overrightarrow{a})\ (\overrightarrow{\widehat{a}} - \overrightarrow{a})]$$

$$\begin{aligned} trMSE(\overrightarrow{\widehat{a}}) &= E[{}^t(\overrightarrow{\widehat{a}} - E(\overrightarrow{\widehat{a}}) + E(\overrightarrow{\widehat{a}}) - \overrightarrow{a})\ (\overrightarrow{\widehat{a}} - E(\overrightarrow{\widehat{a}}) + E(\overrightarrow{\widehat{a}}) - \overrightarrow{a})] \\ &= E[{}^t(\overrightarrow{\widehat{a}} - E(\overrightarrow{\widehat{a}}))\ (\overrightarrow{\widehat{a}} - E(\overrightarrow{\widehat{a}})] + E[{}^t(E(\overrightarrow{\widehat{a}}) - \overrightarrow{a})\ (\overrightarrow{\widehat{a}} - E(\overrightarrow{\widehat{a}}))] \\ &\quad + E[{}^t(\overrightarrow{\widehat{a}} - E(\overrightarrow{\widehat{a}}))\ (E(\overrightarrow{\widehat{a}}) - \overrightarrow{a})] + E[{}^t(E(\overrightarrow{\widehat{a}}) - \overrightarrow{a})\ (E(\overrightarrow{\widehat{a}}) - \overrightarrow{a})] \end{aligned}$$

or $E[{}^t(\overrightarrow{\widehat{a}} - E(\overrightarrow{\widehat{a}}))\ (E(\overrightarrow{\widehat{a}}) - \overrightarrow{a})] = [{}^t(E(\overrightarrow{\widehat{a}}) - E(\overrightarrow{\widehat{a}}))\ (E(\overrightarrow{\widehat{a}}) - \overrightarrow{a})] = \overrightarrow{0} = E[{}^t(E(\overrightarrow{\widehat{a}}) - \overrightarrow{a})\ (\overrightarrow{\widehat{a}} - E(\overrightarrow{\widehat{a}}))]$
donc

$$trMSE(\overrightarrow{\widehat{a}}) = E[{}^t(\overrightarrow{\widehat{a}} - E(\overrightarrow{\widehat{a}}))\ (\overrightarrow{\widehat{a}} - E(\overrightarrow{\widehat{a}})] + E[{}^t(\overrightarrow{\widehat{a}} - E(\overrightarrow{\widehat{a}}))\ (E(\overrightarrow{\widehat{a}}) - \overrightarrow{a})]$$

$$trMSE(\overrightarrow{\widehat{a}}) = \sum_1^k var(\widehat{a}_i) + \sum_1^k (biais)^2$$

Dans le cas d'un estimateur sans biais la trace du MSE= $\sum_1^k var(\widehat{a}_i)$

3.2 Propriétés du MSE

Le MSE permet de comparer deux estimateurs qu'ils soient sans biais ou biaisés. On prendra l'estimateur qui a la plus petite trMSE.
Théorème : la trMSE d'un modèle est identique à celui de son modèle canonique qui est beaucoup plus facile à calculer.

3.2.1 Rappel du modèle canonique

Soit le modèle $\overrightarrow{Y} = X\overrightarrow{a} + \overrightarrow{\epsilon}$, V la matrice orthogonale des vecteurs propres ($V^tV = I$) de ${}^tXX$et Λ sa matrice diagonale des valeurs propres λ_i.
On ${}^tXX = V\Lambda{}^tV$ et $\overrightarrow{Y} = X\overrightarrow{a} + \overrightarrow{\epsilon} = XV{}^tV\overrightarrow{a} + \overrightarrow{\epsilon}$
En notant le vecteur$XV = Z$ et ${}^tV\overrightarrow{a} = \overrightarrow{\gamma}$ le modèle peut s'écrire $\overrightarrow{Y} = Z\overrightarrow{\gamma} + \overrightarrow{\epsilon}$ c'est le modèle sous sa forme canonique
avec $V_{\overrightarrow{\gamma}} = \sigma^2\ ({}^tZZ)^{-1} = \sigma^2\Lambda^{-1}$ matrice diagonale.

- le MSE du modèle canonique s'écrit dans un modèle sans biais:

  trMSE($\overrightarrow{\gamma}$) $= \sigma^2 \sum_1^k 1/\lambda_i = \sigma^2 \sum_1^k 1/d_i^2$ ou les d_i sont les valeurs singulières de la matrice X

3.2.2 Théorème

Le modèle de base et son modèle canonique ont même trMSE. tr$MSE(\overrightarrow{\widehat{a}}) = trMSE(\overrightarrow{\widehat{\gamma}})$
Démonstration:
tr$MSE(\overrightarrow{\widehat{a}}) = E[{}^t(\overrightarrow{\widehat{a}} - \overrightarrow{a})(\overrightarrow{\widehat{a}} - \overrightarrow{a})] = E[{}^t(V\overrightarrow{\widehat{\gamma}} - V\overrightarrow{\gamma})(V\overrightarrow{\widehat{\gamma}} - V\overrightarrow{\gamma})]$
tr$MSE(\overrightarrow{\widehat{a}}) = E[{}^t(\overrightarrow{\widehat{\gamma}} - \overrightarrow{\gamma}){}^tVV(\overrightarrow{\widehat{\gamma}} - \overrightarrow{\gamma})] = E[{}^t(\overrightarrow{\widehat{\gamma}} - \overrightarrow{\gamma})(\overrightarrow{\widehat{\gamma}} - \overrightarrow{\gamma})] = MSE(\overrightarrow{\widehat{\gamma}})$

4 Comparaison des MCO et de la R.O.

On va effectuer les calculs sur les modèles canoniques et toujours sur des données centrées réduites.

4.1 MSE de la R.O.

tr$MSE(\overrightarrow{\widehat{ar}}) = trMSE(\overrightarrow{\widehat{\gamma r}}) = \sum var(\widehat{\gamma r}_i) + \sum biais^2$
$= \sum var(\widehat{\gamma r}_i) + \sum(\gamma r_i - E(\widehat{\gamma r}_i))^2$

or

$$E(\overrightarrow{\widehat{\gamma r}}) = \overrightarrow{\gamma} - \begin{pmatrix} K/(d_1^2+K) & 0 & ... & 0 \\ 0 & K/(d_2^2+K) & ... & 0 \\ 0 & ... & ... & 0 \\ 0 & 0 & ... & K/(d_k^2+K) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ .. \\ \gamma_k \end{pmatrix}$$

$$\sum(\gamma_i - E(\widehat{\gamma r}_i))^2 = \sum_{i=1}^{k} K^2\gamma_i^2/(d_i^2+K)^2$$

La matrice de Variance-covariance de $\overrightarrow{\widehat{\gamma}}$ s'écrit:

$$\begin{aligned} Var_{\overrightarrow{\widehat{\gamma r}}} &= \sigma^2 \begin{pmatrix} 1/(d_1^2+K) & ... & 0 \\ 0 & ... & 0 \\ 0 & ... & 1/(d_k^2+K) \end{pmatrix} \begin{pmatrix} d_1^2 & & 0 \\ 0 & .. & 0 \\ 0 & & d_k^2 \end{pmatrix} \begin{pmatrix} 1/(d_1^2+K) & & 0 \\ 0 & ... & 0 \\ 0 & & 1/(d_k^2+K) \end{pmatrix} \\ &= \sigma^2 \begin{pmatrix} d_1^2/(d_1^2+K)^2 & & 0 \\ 0 & ... & 0 \\ 0 & & d_k^2/(d_k^2+K)^2 \end{pmatrix} \end{aligned}$$

donc $\sum var(\widehat{\gamma r}_i) = \sigma^2 \sum d_i^2/(d_i^2+K)^2$

$$trMSE(\overrightarrow{\widehat{\gamma r}}) = \sigma^2 \sum_{i=1}^{k} d_i^2/(d_i^2+K)^2 + \sum_{i=1}^{k} K^2\gamma_i^2/(d_i^2+K)^2$$

La trMSE de l'estimateur Ridge s'écrit

$$trMSE(\overrightarrow{\widehat{\gamma r}}) = \sum \frac{\sigma^2 d_i^2 + K^2\gamma_i^2}{(d_i^2+K)^2}$$

4.2 trMSE des MCO

La trMSE de l'estimateur des MCO $\overrightarrow{\widehat{\gamma}}$ est (pour K=0)

$$trMSE(\overrightarrow{\widehat{\gamma}}) = \sum \frac{\sigma^2 d_i^2}{(d_i^2)^2} = \sum \frac{\sigma^2}{d_i^2}$$

4.3 Comparaison

Nous avons vu que la trMSE de l'estimateur Ridge s'écrit

$$trMSE(\overrightarrow{\widehat{\gamma r}}) = \sum \frac{\sigma^2 d_i^2 + K^2\gamma_i^2}{(d_i^2+K)^2}$$

La trMSE de l'estimateur des MCO $\overrightarrow{\widehat{\gamma}}$ est (pour K=0)

$$trMSE(\overrightarrow{\widehat{\gamma}}) = \sum \frac{\sigma^2 d_i^2}{(d_i^2)^2} = \sum \frac{\sigma^2}{d_i^2}$$

Pour que la Ridge soit meilleur estimateur que les MCO il faut suivant le critère du MSE que

$$trMSE(\overrightarrow{\widehat{\gamma r}}) < trMSE(\overrightarrow{\widehat{\gamma}})$$

$$\sum \frac{\sigma^2 d_i^2 + K^2\gamma_i^2}{(d_i^2+K)^2} < \sum \frac{\sigma^2}{d_i^2}$$

Pour cela il suffit de vérifier l'inégalité pour chaque i . Nous allons donc essayer de trouver une condition suffisante pour que la Ridge soit meilleure (au sens du MSE) que les MCO. Il suffit donc d'avoir pour chaque i

$$\frac{\sigma^2 d_i^2 + K^2\gamma_i^2}{(d_i^2+K)^2} < \frac{\sigma^2}{d_i^2}$$

$$\sigma^2 d_i^4 + K^2\gamma_i^2 d_i^2 < \sigma^2(d_i^4 + K^2 + 2Kd_i^2)$$
$$K^2(\gamma_i^2 d_i^2 - \sigma^2) - 2K\sigma^2 d_i^2 < 0$$

C'est une inégalité du second degré en K

- Si $\gamma_i^2 d_i^2 - \sigma^2 > 0$ l'expression est négative entre les racines de l'équation du second degré soit entre 0 et une $K_i^* = \frac{2\sigma^2 d_i^2}{\gamma_i^2 d_i^2 - \sigma^2} > 0$
- Si $\gamma_i^2 d_i^2 - \sigma^2 < 0$ les deux expressions sont négatives et la condition est toujours vérifiée

  On en déduit donc que le MSE de la Ridge est inférieur au MSE des MCO si 0<K<min K_i^*. Ce minimum ne peut être calculé en pratique car il dépend des valeurs théorique des coefficients γ_i^2.

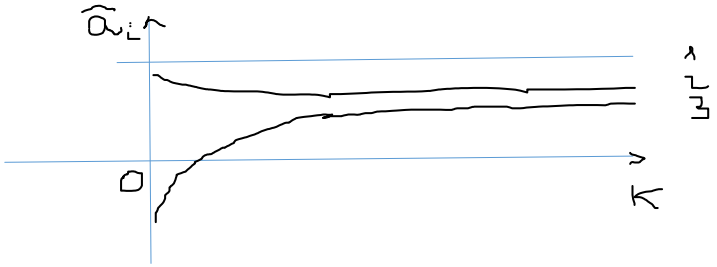
  On prendra des valeurs de K très proches de 0 pour être dans le bon cas.

  On vient donc de démontrer qu'il y a des valeurs de K pour lesquelles la régression Ridge est meilleure que les MCO.

5 Application de la ridge

Sur un modèle à 3 variables. On utilise des valeurs de K très petites et on constate que certains coefficients restent fixes quand K augmente, d'autres bougent au début pour se stabiliser.
On remarque en théorie que pour K=0 on a les MCO et que si K$\rightarrow \infty$ alors les coefficients de la Ridge tendent vers 0, ce qui n'est pas le but recherché car K est choisi très petit d'après la théorie vue ci-dessus.
Interprétation du graphe (toujours aussi laid):
prenons l'exemple d'un modèle à 3 variables 1 2 et 3

- Variable 1 : elle ne joue pas de rôle dans la colinéarité donc si on ajoute K très petit, son coefficient ne change pas



- Variable 2 : elle joue un rôle dans la colinéarité avec la variable 3: son coefficient positif en K=0 (MCO) baisse un peu puis se stabilise.
- Variable 3 : Son coefficient qui était négatif devient positif puis se stabilise.

  Dans la plupart des modèles, on obtient ce genre de résultat. Si les variables ne sont pas colinéaires alors leur coefficient ne bouge pas en appliquant la ridge. Dans le cas des variables colinéaires certains coefficients bougent un peu et d'autres beaucoup avant de se stabiliser, toujours avec K très petit.

6 Choix de K

Nous rappelons que pour des raisons d'élimination des unités, nous travaillons sur des variables CENTREES REDUITES comme le suggèrent Hoerl et Kennard.
Le résultat de la Ridge $\overrightarrow{\widehat{ar}} = ({}^tXX + KI)^{-1}\ {}^tX\overrightarrow{y}$
La matrice à inverser ${}^tXX + KI$ correspond à la matrice

$$X_r = \begin{pmatrix} X \\ \sqrt{K}I \end{pmatrix}$$

telle que ${}^tXX + KI = {}^tX_rX_r$. Les valeurs propres ${}^tXX + KI$ donc sont les valeurs propres de ${}^tXX$ plus K, donc les $d_i^2 + K$
Le conditionnement de la matrice X est Cond X=$d_{\max}/d_{\min}$
Le conditionnement de la matrice X_r est Cond $X_r = \sqrt{d_{\max}^2 + K}/\sqrt{d_{\min}^2 + K}$
S'il y a colinéarité le conditionnement de la matrice X est très grand. Pour faire baisser ce conditionnement, on va passer à X_r en lui imposant un conditionnement beaucoup plus faible (par exemple 50 ou 30) . Cela entraine une valeur de K
Si par exemple on impose 30

$$CondX_r = \sqrt{d_{\max}^2 + K} / \sqrt{d_{\min}^2 + K} = 30$$

$$(d_{\max}^2 + K) = 30^2 (d_{\min}^2 + K)$$

$$K = \frac{d_{\max}^2 - 30^2 d_{\min}^2}{30^2 - 1}$$

On prend cette valeur de K pour l'estimation de la régression Ridge.
On peut aussi (vivement recommandé) faire le graphe de l'évolution des coefficients et regarder quand les cofficients se stabilisent.