2.2 Régression Linéaire Multiple

Les modèles classiques de régession (linéaire, logistique) sont anciens et moins l'occasion de battage médiatique que ceux récents issus de l'apprentissage automatique. Néanmoins, ils présentent un grand intérêt compte tenu de leur robustesse, de leur stabilité face à des fluctuations d'échantillons et de leur capacité à passer à l'échelle pour des données massives. Ils restent ainsi toujours très utilisés en production notamment lorsque la fonction à modéliser est bien linéaire et qu'il serait contre productif de chercher plus compliqué.

2.2.1 Modèle

Une variable quantitative $\mathbf{Y}$ dite à expliquer (ou encore, réponse, exogène, dépendante) est mise en relation avec p variables quantitatives $\mathbf{X^1}, \ldots, \mathbf{X^p}$ dites explicatives (ou encore de contrôle, endogènes, indépendantes, régresseurs, prédicteurs).

Les données sont supposées provenir d'un échantillon statistique de n observations, chacune étant dans $\mathbb{R}^{(p+1)}$ (avec $n > p+1$) :

$$(x_i^1, \ldots, x_i^j, \ldots, x_i^p, y_i)\,, i = 1, \ldots, n$$

L'écriture du modèle linéaire dans cette situation conduit à supposer que l'espérance de $\mathbf{Y}$ appartient au sous-espace de $\mathbb{R}^n$ engendré par $\{\mathbf{1}, \mathbf{X^1}, \ldots, \mathbf{X^p}\}$ où $\mathbf{1}$ désigne le vecteur de $\mathbb{R}^n$ constitué de 1s. C'est-à-dire que les $(p+1)$ variables aléatoires vérifient :

$$Y_i = \beta_0 + \beta_1 X_i^1 + \beta_2 X_i^2 + \ldots + \beta_p X_i^p + \varepsilon_i\,, i = 1, 2, \ldots, n$$

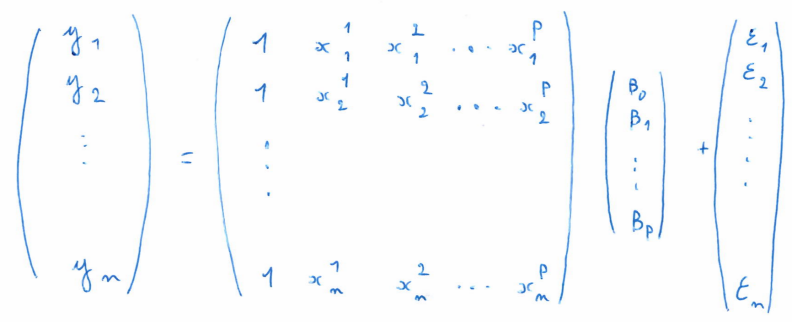
avec les hypothèses suivantes :

- Les ε_i sont des termes d'erreur indépendants et identiquement distribués, *i.e.* $E(\varepsilon_i) = 0$, $Var(\varepsilon) = \sigma^2\mathbf{I}$.
- Les termes de $\mathbf{X^j}$, *i.e.* du vecteur qui contient les observations de la j^{eme} variable, sont supposés déterministes (facteurs contrôlés). Dans certain contextes, on suppose alternativement que l'erreur ε est indépendante de la distribution conjointe de $\mathbf{X^1}, ..., \mathbf{X^p}$. On écrit dans ce cas que $E(\mathbf{Y}|\mathbf{X^1}, \ldots, \mathbf{X^p}) = \beta_0 + \beta_1\mathbf{X^1} + \beta_2\mathbf{X^2} + \ldots + \beta_p\mathbf{X^p}$ et que $Var(Y|\mathbf{X^1}, \ldots, \mathbf{X^p}) = \sigma^2$.
- Les paramètres inconnus $\beta_0, \ldots, \beta_p$ sont supposés constants.
- En option, pour l'étude spécifique des lois des estimateurs, une quatrième hypothèse considère la normalité de la variable d'erreur ε (*i.e.* $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$). Les ε_i sont alors i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

Les données sont rangées dans une matrice $\mathbf{X}$ de taille $(n \times (p+1))$ de terme général X_i^j, dont la première colonne contient le vecteur $\mathbf{1}$ (c'est à dire $X_0^i = 1$), et dans un vecteur $\mathbf{Y}$ de terme général Y_i . En notant les vecteurs $\varepsilon = [\varepsilon_1 \ldots \varepsilon_n]'$ et $\beta = [\beta_0 \beta_1 \ldots \beta_p]'$, le modèle s'écrit matriciellement :

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon\,.$$

Ce modèle est détaillé ci-dessous :



2.2.2 Estimation

Conditionnellement à la connaissance des valeurs des $\mathbf{X}^j$, les paramètres inconnus du modèle, le vecteur β et le paramètre de nuisance σ^2, sont estimés par minimisation des carrés des écarts (M.C.) ou encore par maximisation de la vraisemblance (M.V.) en considèrant l'hypothèse de la normalité de la variable d'erreur. Les estimateurs ont alors les mêmes expressions, l'hypothèse de normalité et l'utilisation de la vraisemblance conférant à ces derniers des propriétés complémentaires.

Etudions l'estimation par moindres carrés. L'expression à minimiser sur $\beta \in \mathbb{R}^{p+1}$ s'écrit :

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i^1 - \ldots - \beta_p X_i^p)^2 = ||\mathbf{Y} - \mathbf{X}\beta\,||^2 \tag{2.1}$$

$$= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta \tag{2.2}$$

Par dérivation matricielle de la dernière équation on obtient les équations normales :

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\beta = 0$$

dont la solution correspond à un minimum car la matrice hessienne $2\mathbf{X}'\mathbf{X}$ est semi définie-positive.

Nous faisons l'hypothèse supplémentaire que la matrice $\mathbf{X}'\mathbf{X}$ est inversible, c'est-à-dire que la matrice $\mathbf{X}$ est de rang $(p+1)$ et donc qu'il n'existe pas de colinéarité entre ses colonnes. Si cette hypothèse n'est pas vérifiée, il suffit en principe de supprimer des colonnes de $\mathbf{X}$ et donc des variables du modèle. Une approche de réduction de dimension (régression ridge, Lasso, PLS...) est en pratique à mettre en oeuvre (voir Section 3.3). Alors, l'estimation des paramètres β_j est donnée par :

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

et les valeurs ajustées (ou estimées, prédites) de $\mathbf{Y}$ ont pour expression :

$$\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

où $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ est connue sous le nom *hat matrix*. Géométriquement, c'est la matrice de projection orthogonale dans $\mathbb{R}^n$ sur le sous-espace $Vect(\mathbf{X})$

engendré par les vecteurs colonnes de $\mathbf{X}$. On note alors :

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\widehat{\beta} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

le vecteur des résidus.

Notons finalement qu'il est possible d'inférer sur l'estimation des paramètres β_j comme dans la régression linéaire simple mais nous nous intéresserons dans ce cours à d'autres aspects de la régression multiple, notamment la sélection de modèle.

2.2.3 Prévision

Connaissant les valeurs des variables $\mathbf{X}^j$ pour une nouvelle observation : $x_0 = [x_0^1, x_0^2, \ldots, x_0^p]'$ appartenant au domaine dans lequel l'hypothèse de linéarité reste valide, une prévision, notée $\widehat{y_0}$ de $\mathbf{Y}$ ou $E(\mathbf{Y})$ est donnée par :

$$\widehat{y_0} = \widehat{\beta_0} + \widehat{\beta_1} x_0^1 + \cdots + \widehat{\beta_p} x_0^p \,.$$

Les intervalles de confiance des prévisions de $\mathbf{Y}$ et $E(\mathbf{Y})$, pour une valeur $\mathbf{x}_0 \in \mathbb{R}^p$ et en posant $\mathbf{v}_0 = (1|\mathbf{x}_0')' \in \mathbb{R}^{p+1}$, sont respectivement

$$\begin{aligned} \widehat{y_0} &\pm t_{\alpha/2;(n-p-1)} \widehat{\sigma} (1 + \mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}, \\ \widehat{y_0} &\pm t_{\alpha/2;(n-p-1)} \widehat{\sigma} (\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{1/2}. \end{aligned}$$

Il est intéressant de remarquer que ces intervalles de confiance dépendent d'une loi de Student à $n-p-1$ degrés de liberté (voir appendice A). A dimension des observations fixée p, plus n est grand, plus les valeurs de la loi de Student seront faible, et ainsi les marges seront réduites. Cependant, plus p est proche de n, plus les marges sont élevées. En parallèle, les variances de ces prévisions, comme celles des estimations des paramètres, dépendent aussi directement du conditionnement de la matrice $\mathbf{X}'\mathbf{X}$ de par le terme $\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0$.

2.2.4 Qualité d'ajustement

Tout comme dans le modèle linéaire simple (Sous-section 2.1.5), la qualité d'ajustement du modèle peut être mesurée avec $p > 1$ variables par coefficient de détermination R^2. On note SSE la somme des carrés des résidus (sum of squared errors) :

$$SSE = ||\mathbf{Y} - \widehat{\mathbf{Y}}||^2 = ||e||^2 \,.$$

On définit également la somme totale des carrés (total sum of squares) par

$$SST = ||\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1}||^2 = \mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{Y}}^2 \,.$$

et la somme des carrés de la régression (regression sum of squares) par

$$SSR = ||\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\mathbf{1}||^2 = \widehat{\mathbf{Y}}'\widehat{\mathbf{Y}} - n\bar{\mathbf{Y}}^2 = \mathbf{Y}'\mathbf{H}\mathbf{Y} - n\bar{\mathbf{Y}}^2 = \widehat{\beta}'\mathbf{X}'\mathbf{Y} - n\bar{\mathbf{Y}}^2 \,.$$

où $\bar{\mathbf{Y}}\mathbf{1}$ est le vecteur de même taille que $\mathbf{Y}$ dont tous les termes sont égaux à la moyenne des valeurs observées de $\mathbf{Y}$. Le *coefficient de détermination* est alors le rapport

$$R^2 = \frac{SSR}{SST} = \frac{||\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\mathbf{1}||^2}{||\mathbf{Y} - \bar{\mathbf{Y}}\mathbf{1}||^2}$$

qui est donc la part de variation de $\mathbf{Y}$ expliquée par le modèle de régression. La quantité R est appelée coefficient de corrélation multiple entre $\mathbf{Y}$ et les variables explicatives, c'est le coefficient de corrélation usuel entre $\mathbf{Y}$ et sa prévision $\widehat{\mathbf{Y}}$.

Notons que le coefficient de détermination croît avec le nombre p de variables par construction. D'une manière générale, plus un modèle est complexe plus il va pouvoir coller aux données, mais moins il sera explicable et sera généralisable.

Chapitre 3

Sélection de modèle en régression linéaire multiple

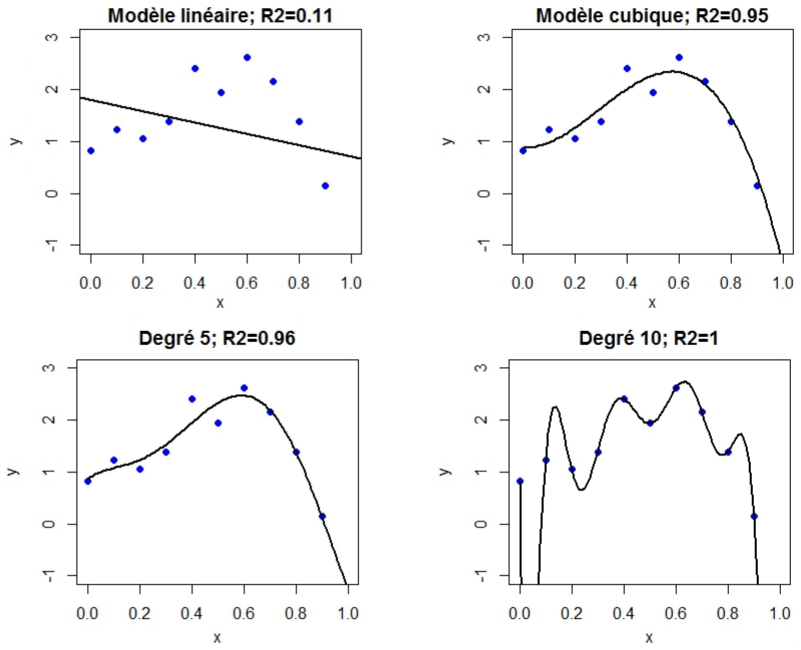
3.1 Introduction

La pratique de la modélisation statistique vise trois objectifs éventuellement complémentaires.

1. Descriptif : Cherche de façon exploratoire les liaisons entre $\mathbf{Y}$ et d'autres variables, potentiellement explicatives, $\mathbf{X}^j$ qui peuvent être nombreuses afin, par exemple d'en sélectionner un sous-ensemble. L'Analyses en Composantes Principales peut contribuer à cette recherche (voir cour de Statistique). Si p est grand, des algorithmes de recherche moins performants mais économiques en temps de calcul sont aussi à considérer.
2. Explicatif : Le deuxième objectif est sous-tendu par une connaissance a priori du domaine concerné et dont des résultats théoriques peuvent vouloir être confirmés, infirmés ou précisés par l'estimation des paramètres. Dans ce cas, les résultats inférentiels permettent de construire le bon test conduisant à la prise de décision recherchée (voir cour de Statistique).
3. Prédictif : Dans le troisième cas, l'accent est mis sur la qualité des prévisions. C'est la situation rencontrée en apprentissage. Ceci conduit à rechercher des modèles parcimonieux (sparse) c'est-à-dire avec un nombre volontairement restreint de variables explicatives. Dans ce contexte, un bon modèle n'est pas celui qui explique le mieux les données au sens d'un Coefficient de détermination R^2 maximum, mais celui conduisant aux prévisions les plus fiables.

3.1.1 Intérêt de modèles parcimonieux

Ceci est illustré ceci par un exemple simple en régression polynomiale sur un jeu de données simulées. Notons qu'on sort ici du cadre linéaire pour illustrer sur un graphique 2D l'intérêt d'un modèle parcimonieux. On approche les (x_i, y_i) à l'aide de polynomes de degrés K : $y_i = \beta_0 + \sum_{k=1}^{K}(\beta_k x_i^k) + \varepsilon$. Les résultats après estimation des β_k et le coefficient de détermination R^2 sont donnés ci-dessous pour $K = 1, 3, 5, 10$.



L'ajustement du modèle mesuré par le R^2 croît naturellement avec le nombre de paramètres K et atteint la valeur 1 lorque le polynôme interpole les observations (quand $K = n$). Dans un but de prédire de nouvelles valeurs de y, le meilleur modèle n'est cependant clairement pas celui ayant le polynôme le plus élevé. Il vaudra mieux utiliser un modèle plus contrains qui ne capturent pas le bruit inhérent aux données. Ils sont ainsi plus génériques et moins spécifique aux données observées.

Ce phénomène est connu sous le nom de sur-apprentissage en apprentissage automatique. Selectionner les variables/dimensions les plus pertinentes peut aussi être particulièrement souhaitable pour expliquer au mieux les données. Ce principe est connu sous le nom de parcimonie (sparseness) en apprentissage automatique et conduit à avoir peu de $\beta_k > 0$.

Nous nous focaliserons ici sur des critères de qualité de prévision. Le C_p de Mallows, le critère d'information d'Akaïke (AIC), celui bayésien de Sawa (BIC) sont les plus classiques. Ils sont équivalents avec le R^2 lorsque le nombre de variables à sélectionner (ou complexité du modèle) est fixé. Le choix du critère est déterminant lorsqu'il s'agit de comparer des modèles de complexitée différentes. Certains critères se ramènent, dans le cas gaussien, à l'utilisation d'une expression pénalisée de la fonction de vraisemblance afin de favoriser des modèles parcimonieux.

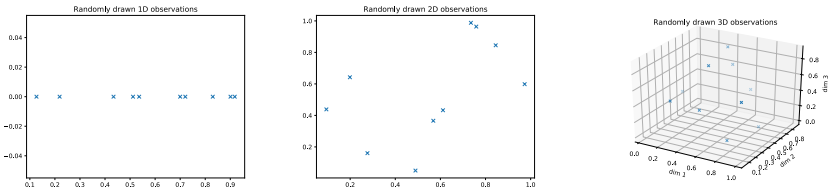
3.1.2 Fléau de la dimension

Un autre angle d'attaque pour motiver le besoin de parcimonie en apprentissage automatique est celui du *fléau de la dimension*. L'idée générale de cette

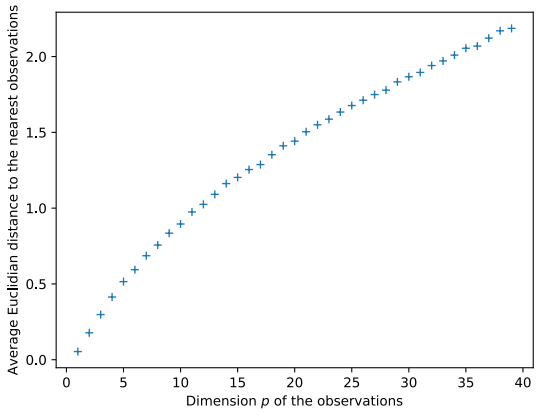
notion est que lorsque la dimension p des données augmente, le volume de l'espace dans lequel les données vivent croît rapidement, si bien que les données se retrouvent isolées et deviennent éparses. Imaginons par exemple que l'on souhaite mettre en lien les profils $x_i = (x_i^1, \ldots, x_i^p)$ de n étudiants $i = 1, \ldots, n$ à leur note à un concours $y_i \in \mathbb{R}$. On peut représenter dans x_i la moyenne des étudiants au cours de l'année passée en Français et en Mathématiques, ce qui donne un profil général avec seulement $p = 2$ variables. On peut aussi utiliser les moyennes dans toutes les matières pour être plus fin, ce qui permet d'avoir environ $p = 10$ variables. On peut aussi utiliser des variables variées, comme le revenu moyen dans le quartier des parents, la taille, des indicateurs issus des sites internet visités, pour arriver à $p \sim 50$ variables.

Chaque fois qu'une dimension sera ajoutée, on aura plus d'information sur les élèves mais chaque élève sera aussi de plus en plus unique. Si on veut deviner la note y_{test} d'un nouvel élève, il sera alors de plus en plus difficile de trouver plusieurs élèves qui lui ressemblent dans la base d'apprentissage afin d'interpoler leurs notes avec le modèle de prédiction. Il semble alors totalement intuitif de sélectionner quelles variables sont les plus pertinentes dans le jeu d'apprentissage pour prédire les y_{test}.

Voila une petite expérience pour illustrer l'influence de la dimension sur le niveau d'isolement d'observations. On tire suivant une loi uniforme $n = 10$ observations dans un domaine $[0, 1]^p$, avec différentes valeurs de p. Les figures ci-dessous représentent des tirages dans $[0, 1]$, dans $[0, 1]^2$ et dans $[0, 1]^3$.



On voit bien que le $[0, 1]^p$ est de moins en moins densément rempli quand p augmente. Mesurons alors empiriquement la distance euclidienne moyenne d'une nouvelle observation dans $[0, 1]^p$ au point le plus proche, parmi dix points tirés suivant cette loi uniforme. Nous obtenons les distances moyennes à un point en fonction de p ci-dessous :



La distance sera 3.8 fois plus grande si $p = 2$ que $p = 1$, elle sera plus de 10 fois plus grande si $p = 5$ que $p = 1$, plus de 40 fois plus grande si $p = 30$ que $p = 1$, ... bref il sera rapidement difficile de trouver un modèle pouvant interpoler efficacement des observations à n fixé.

Ceci est encore pire en posant le problème à l'envers. Pour avoir la même distance moyenne à un des 10 points dans $[0, 1]$, il faudra environ 140 observations dans $[0, 1]^2$, environ 2800 observations dans $[0, 1]^3$, ... En particulier, quand le nombre d'observations ne peut pas être trop grand et que les observations sont en grande dimension, il faudra définir un modèle efficace pour interpoler les observations (voir les réseaux de neurones convolutionnels), soit réduire leur dimension préalablement (voir l'ACP, la PLS ou encore les espaces latents de réseaux de neurones), soit sélectionner les variables les plus pertinentes. Dans le cadre du cours de modèle linéaire, nous allons voir cette troisième option qui est extrêmement classique et la plus interprétable.

3.1.3 Compromis biais-variance

Avant de rentrer dans les méthodes de sélection de modèle, discutons de la formalisation du problème de compromis biais-variance. Considérons les données d'apprentissage (x_i, y_i), $i = 1, \dots, n$. De manière générale, on suppose que les x_i peuvent expliquer partiellement les y_i, et que d'autres paramètres indépendants des x_i entrent aussi en jeu. Généralement, on modélisera alors le problème sous cette forme :

$$y_i = f(x_i) + \epsilon_i \,,$$

où f est une fonction inconnue et ϵ_i suit une loi Normale de moyenne nulle et d'écart type σ. Le but de la regression est alors de trouver une fonction $\hat{f}$ qui approxime au mieux f. Ceci se fait en fixant d'abord un modèle (linéaire, polynôme, arbre de décision, réseau de neurones, ...) puis en apprenant ses q paramètres à partir de ce que l'on connait, c'est à dire les (x_i, y_i). Le problème qui émerge naturellement est le suivant : Comment simultanément estimer f au mieux et tenir le moins possible compte du bruit ϵ sachant que les deux sont inconnus ? C'est la question clé du compromis biais-variance.

Plus formellement, on minimise l'esperance empirique de $(y - \hat{f}(x))^2$ sur les (x_i, y_i), c'est à dire l'erreur au carré moyenne (Mean Squared Error – MSE). Elle peut être décomposée sous cette forme :

$$\mathbb{E}\left[(y - \hat{f}(x))^2\right] = \underbrace{\mathbb{E}\left[\hat{f}(x) - f(x)\right]^2}_{biais[\hat{f}(x)]} + \underbrace{\mathbb{E}\left[\hat{f}(x)^2\right] - \mathbb{E}\left[\hat{f}(x)\right]^2}_{variance[\hat{f}(x)]} + \sigma^2 \tag{3.1}$$

Cette représentation de la MSE peut être démontré en utilisant les relations suivantes :

- $\mathbb{E}[f(x)] = f(x)$ car $f(x)$ est déterministe
- $\mathbb{E}[y] = \mathbb{E}[f(x) + \epsilon] = \mathbb{E}[f(x)] + \mathbb{E}[\epsilon] = \mathbb{E}[f(x)] = f(x)$
- $Var[\epsilon] = \mathbb{E}[\epsilon^2] + (\mathbb{E}[\epsilon])^2 = \mathbb{E}[\epsilon^2] = \sigma^2$
- $Var[y] = \mathbb{E}[(y - \mathbb{E}[y])^2] = \mathbb{E}[(y - f(x))^2] = \mathbb{E}[(f(x) + \epsilon - f(x))^2] = \sigma^2$

Plus intéressant ici, les différents termes d'Eq. (3.1) peuvent être interprétés comme suit :

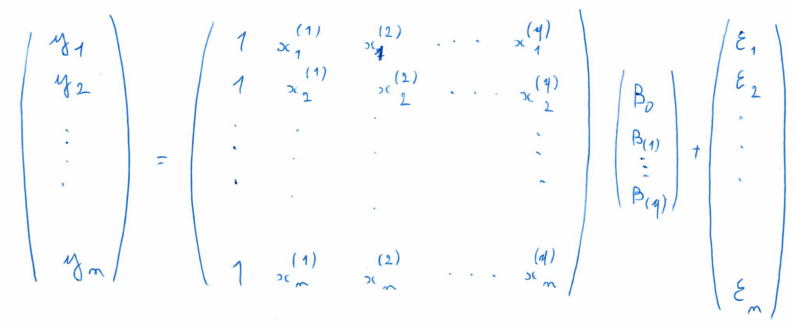
- Le terme de biais $\mathbb{E}[\hat{f}(x) - f(x)]^2$ représente à quel point le modèle $\hat{f}$ approxime la fonction inconnue f.
- Le terme de variance $\mathbb{E}[\hat{f}(x)^2] - \mathbb{E}[\hat{f}(x)]^2 = Var[\hat{f}(x)]$ représent le niveau de variabilité de $\hat{f}$, sans tenir compte de f.
- Le terme σ^2 représente enfin le niveau de bruit dans les données (x_i, y_i), qui tout comme f est inconnu.

Pour une MSE (i.e. $\mathbb{E}[(y - \hat{f}(x))^2]$) donnée, un $\hat{f}$ représentera alors un compromis entre qualité d'approximation de f au niveau des observations $\{x_i\}_{i=1,\dots,n}$ et sa stabilité. Une trop grande qualité d'approximation au niveau des observations impliquera alors des fonctions $\hat{f}$ instables et ainsi moins généralisables en dehors des $\{x_i\}_{i=1,\dots,n}$ (sur-apprentissage). A contrario, des fonctions $\hat{f}$ trop stables captureront mal les relations entre les x_i et les y_i et auront de même un faible pouvoir prédictif.

Trouver un bon compromis entre biais et variance pourra se faire en réduisant explicitement la dimension d'un modèle (Section 3.2) ou en régularisant l'estimation des paramètres d'un modèle (Section 3.3). Dans tous les cas, il sera plus que recommandé d'estimer à quel point le modèle appris est généralisable à l'aide d'une technique de validation croisée (Section 3.4).

3.2 Sélection de modèle par sélection de variables et minimisation de critères pénalisés

Considérons un modèle linéaire $\mathcal{M}$ à q variables $\mathbf{X}^{(j)}$, $j = 1, \dots, q$. Dans ce modèle $q < p$ et chaque $\mathbf{X}^{(j)}$ correspond à une des p variables observées $\mathbf{X}^k$, $k = 1, \dots, p$. Ce modèle s'écrit :



La selection de modèle consiste à la fois à choisir les meilleur variables explicatives des y_i et à estimer les paramètres β_i optimaux. Nous développons dans cette section plusieurs critères de sélection de modèle.

Critère C_p de Mallows

On rappelle que la somme des carrés des résidus $SSE = ||\mathbf{Y} - \widehat{\mathbf{Y}}||^2 = ||e||^2$. On dénote alors la *mean square error* :

$$MSE = \frac{SSE}{n-p-1},$$

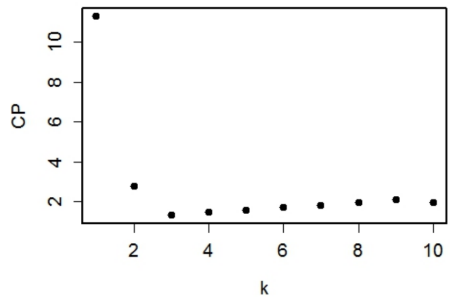
où $n-p-1$ est le nombre de degrés de liberté du modèle compet à p variables et n observations.

L'indicateur proposé par Mallows en 1973 pour evaluer la qualité d'un modèle donné $\mathcal{M}$ à q variables est alors

$$C_p = (n-(q+1))\frac{MSE_{\mathcal{M}}}{MSE} - (n-2(q+1))$$

où $MSE_{\mathcal{M}}$ est la MSE calculée pour le modèle $\mathcal{M}$.

Il est alors d'usage de rechercher un modèle qui minimise le C_p. Ceci revient à considérer que le "vrai" modèle complet est moins fiable qu'un modèle réduit donc biaisé mais d'estimation plus précise. A qualité de modèle constant $MSE_{\mathcal{M}}/MSE$, plus q est faible, plus C_p est faible. Par contre si l'erreur du modèle $\mathcal{M}$ augmente à q fixé, C_p augmente. Voila ci-dessous l'évolution de C_p en fonction de K dans l'exemple introductif du chapitre. Ici, le meilleur modèle contient $q = 3$ variables.



Critères AIC, BIC et PRESS

Dans le cas du modèle linéaire, et si la variance des observations est supposée connue, le critère AIC (Akaïke's Information criterium) est équivalent au critère C_p de Mallows. Le critère BIC (Bayesian Information Criterium) est une extension d'AIC dans lequel le terme de pénalité est plus important. Le PRESS (somme des erreurs quadratiques) de Allen (1974) est l'introduction historique de la validation croisée ou leave-one-out (loo). Ces critères peuvent être résumés par :

- AIC : $AIC(\mathcal{M}) = n \log MSE_{\mathcal{M}} + 2(q+1)$
- BIC : $AIC(\mathcal{M}) = n \log (MSE_{\mathcal{M}}) + \log (n)(q+1)$
- PRESS : On désigne par $\widehat{y_{(-i)j}}$ la prévision de y_j calculée sans tenir compte de la $i$ème observation lors de l'estimation des paramètres alors : $PRESS = \sum_{i=1}^{n} (y_i - \widehat{y_{(-i)i}})^2$

et permettent de comparer les capacités prédictives de différents modèles.

Algorithmes de sélection de variables

Dans le cas général les variables ne sont pas pré-ordonnées par importance. C'est d'ailleurs le cas le plus courant en pratique ! Lorsque p est grand, il n'est pas raisonnable d'explorer les 2^p modèles possibles afin de sélectionner le meilleur au sens de l'un des critères ci-dessus. Différentes stratégies existent pour explorer efficacement les modèles possibles. Elles doivent être choisies en fonction de l'objectif recherché, de la valeur de p et des moyens de calcul disponibles. Deux types d'algorithmes sont résumés ci-dessous par ordre croissant de temps de calcul nécessaire, c'est-à-dire par nombre croissant de modèles considérés explorés parmi les 2^p et ainsi par capacité croissante d'optimalité.

Sélection (forward) A l'état initial $q = 1$ et toutes les p variables sont testées. La variable qui permet de réduire au mieux le critère du modèle obtenu est selectionnée, on la dénote (1). On teste alors si une des $p - 1$ variables restantes améliore la qualité du modèle avec $q = 2$ et (1) déjà sélectionné... et ainsi de suite. La procédure s'arrête lorsque toutes les variables sont introduites ou lorsque le critère ne décroît plus.

Elimination (backward) L'algorithme démarre cette fois du modèle complet. À chaque étape, la variable dont l'élimination conduit la valeur du critère la plus faible est supprimée. La procédure s'arrête lorsque la valeur du critère ne décroît plus.

Mixte (stepwise) Cet algorithme introduit une étape d'élimination de variable après chaque étape de sélection afin de retirer du modèle d'éventuels variables qui seraient devenues moins indispensables du fait de la présence de celles nouvellement introduites.

3.3 Sélection de modèle par régularisation

Les méthodes de régression régularisée sont à utiliser quand le probleme est mal conditioné, et typiquement quand le nombre d'observations n est plus petit

que la dimension des observations p. Ce cas est très courant en pratique, par exemple quand chaque observation coute cher a obtenir mais est en très grande dimension, comme c'est le cas en génomique ou dans de nombreuses applications industrielles.

3.3.1 Régression ridge

Modèle et estimation

Ayant diagnostiqué un problème mal conditionné mais désirant conserver toutes les variables explicatives pour des raisons d'interprétation, il est possible d'améliorer les propriétés numériques et la variance des estimations en considérant un estimateur biaisé des paramètres par une procédure de régularisation. Soit le modèle linéaire :

$$\mathbf{Y} = \tilde{\mathbf{X}}\tilde{\beta} + \epsilon$$

où :

$$\widetilde{\mathbf{X}} = \begin{pmatrix} 1 & X_1^1 & X_1^2 & . & X_1^p \\ 1 & X_2^1 & X_2^2 & . & X_2^p \\ . & . & . & . & . \\ 1 & X_n^1 & X_n^2 & . & X_n^p \end{pmatrix},$$

$$\widetilde{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{pmatrix}$$

où $\mathbf{X}^0 = (1, 1, \ldots, 1)'$ et $\mathbf{X}$ désigne la matrice $\tilde{\mathbf{X}}$ privée de sa première colonne. L'estimateur ridge est défini par un critère des moindres carrés, avec une pénalité de type $\mathbb{L}^2$ par :

$$\widehat{\beta}_{ridge} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left(\sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} X_i^{(j)} \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right)$$

où λ est un paramètre positif. Notez que le paramètre β_0 n'est pas pénalisé.

En supposant $\mathbf{X}$ et $\mathbf{Y}$ centrés, l'estimateur ridge est obtenu en résolvant les équations normales qui s'expriment sous la forme :

$$\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)\beta$$

Conduisant à :

$$\widehat{\beta}_{ridge} = (\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{Y}$$

La solution est donc explicite et linéaire en $\mathbf{Y}$. Remarquons alors que :

- $\mathbf{X}'\mathbf{X}$ est une matrice symétrique positive, *i.e.* pour tout vecteur $\mathbf{u}$ de $\mathbb{R}^p$: $\mathbf{u}'(\mathbf{X}'\mathbf{X})\mathbf{u} \geq 0$. Il en résulte que pour tout $\lambda > 0$, $\mathbf{X}'\mathbf{X} + \lambda \mathbf{I}_p$ est inversible.
- La constante β_0 n'intervient pas dans la pénalité, sinon, le choix de l'origine pour $\mathbf{Y}$ aurait une influence sur l'estimation de l'ensemble des paramètres. Alors : $\widehat{\beta_0} = \bar{\mathbf{Y}}$; ajouter une constante à $\mathbf{Y}$ ne modifie pas les $\widehat{\beta}_j$ pour $j \geq 1$.

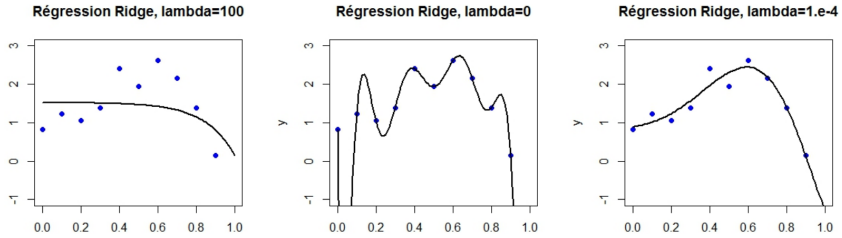
— L'estimateur ridge n'est pas invariant par renormalisation des vecteurs $X^{(j)}$, il est préférable de normaliser (réduire les variables) des vecteurs avant de minimiser le critère.
— La régression ridge est aussi équivalent à estimer le modèle par les moindres carrés sous la contrainte que la norme du vecteur β des paramètres ne soit pas trop grande :

$$\widehat{\beta}_{ridge} = \underset{\beta \in \mathbb{R}^p \ , \ ||\beta||^2 < c}{\arg\min} ||\mathbf{Y} - \mathbf{X}\beta||^2$$

La régression ridge conserve toutes les variables mais, contraignant la norme des paramètres β_j, elle les empêche de prendre de trop grandes valeurs et limite ainsi la variance des prévisions.

Optimisation de la pénalisation

La figure ci-dessous montre quelques résultats obtenus par la méthode ridge en fonction de la valeur de la pénalité λ sur l'exemple de la régression polynomiale (toujours pour pouvoir représenter les résultats dans un graphique 2D mais le principe est le même dans le cas linéaire multiple).

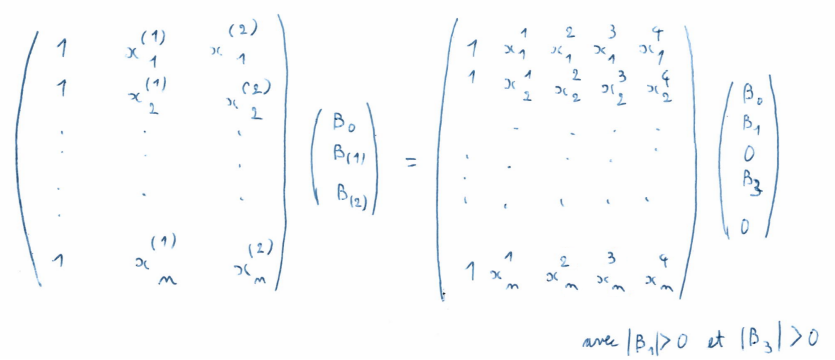


On peut remarquer que plus la pénalité λ augmente et plus la solution obtenue est régulière ou encore, plus le biais augmente (on s'éloigne des données) et la variance diminue (les estimation varient moins) :
— Il y a sur-ajustement avec une pénalité nulle : le modèle passe par tous les points mais oscille dangereusement.
— Il y a par contre sous-ajustement avec une pénalité trop grande.
Comme dans tout problème de régularisation, le choix de la valeur du paramètre λ est alors crucial est déterminera le choix de modèle. La validation croisée est généralement utilisée pour optimiser le choix (voir Section 3.4). La lecture du graphique montrant l'évolution des paramètres en fonction du coefficient ou chemins de régularisation ridge est suffisant pour définir un bon choix mais n'est pas suffisante pour déterminer une valeur optimale et est de plus laborieuse.

3.3.2 Régression LASSO

La régression ridge permet de contourner les problèmes de colinéarité même en présence d'un nombre important de variables explicatives ou prédicteurs ($p > n$). La principale faiblesse de cette méthode est cependant liée à la difficulté d'interprétation. Sans sélection, toutes les variables sont concernées dans le modèle : elles ont une valeur non-nulle et on ne peut pas se ramener au problème posé au début de Section 3.2.

Pour comprendre l'équivalence entre sélectionner explictement des variables, comme dans Section 3.2, et sélectionner des variables en ne considérant que les $|\beta_i| > 0$, imaginons que l'on ai 4 variables $\{1, 2, 3, 4\}$ et que les deux variables sélectionnées soient $(1) = 1$ et $(2) = 3$. Alors on a :



D'autres approches par pénalisation permettent une sélection, c'est le cas de la régression Lasso.

Modèle et estimation

La méthode Lasso (Tibshirani, 1996) correspond à la minimisation d'un critère des moindres carrés avec une pénalité de type L_1 (et non L_2 comme dans la régression ridge). Soit $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$.

L'estimateur Lasso de β dans le modèle $\mathbf{Y} = \tilde{\mathbf{X}}\tilde{\beta} + \epsilon$ est alors défini par :

$$\widehat{\beta}_{LASSO} = \underset{\beta \in \mathbb{R}^{p+1}}{\arg\min} \left(\sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} X_i^{(j)} \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

où λ est un paramètre positif. On peut montrer que ceci équivaut au problème de minimisation suivant

$$\widehat{\beta}_{LASSO} = \underset{\beta \in \mathbb{R}^{p}\ ,\ ||\beta||_1 < t}{\arg\min} ||\mathbf{Y} - \mathbf{X}\beta||^2$$

pour un t convenablement choisi. Comme dans le cas de la régression ridge, le paramètre λ est un paramètre de régularisation :

— Si $\lambda = 0$, on retrouve l'estimateur des moindres carrés.

— Si λ tend vers l'infini, on annule tous les $\hat{\beta}_j$, $j = 1, \ldots, p$.

La solution obtenue est dite parcimonieuse (sparse en anglais), car elle comporte des coefficients nuls.

Pourquoi la pénalisation L_1 sélectionne-elle les variables ?

On se place dans un cadre général dans lequel on minimse une fonction d'erreur sur l'attache aux données $f(\beta_1, \ldots, \beta_p)$. Cette fonction est continue et

deux fois dérivable et les β_i ont une pénalité soit L_1 ou soit L_2 :

$$\widehat{\beta}_{L_1} = \arg\min_{\beta \in \mathbb{R}^p} f(\beta_1, \ldots, \beta_p) + \lambda \sum_{j=1}^{p} |\beta_j|$$
$$\widehat{\beta}_{L_2} = \arg\min_{\beta \in \mathbb{R}^p} f(\beta_1, \ldots, \beta_p) + \lambda \sum_{j=1}^{p} (\beta_j)^2$$

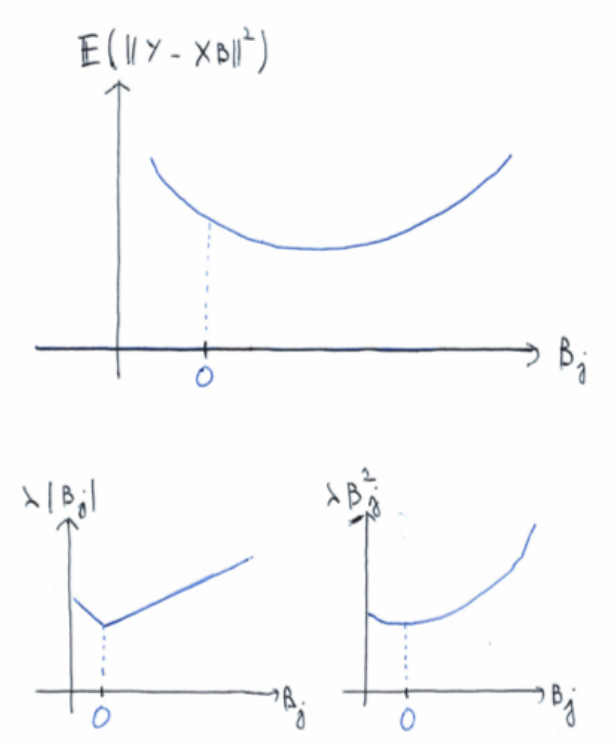
A l'état optimal, *i.e.* pour $\beta = \widehat{\beta}_{L_1}$ ou $\beta = \widehat{\beta}_{L_2}$, les gradients des fonctions optimisées sont nulles. Il existe alors un équilibre entre les gradients de f qui tendent minimiser l'erreur sur l'attache aux données, et les gradients du terme de régularisation qui tendent à ramener β vers $\mathbf{0}$. Pour un β_j donné, on a alors :

$$\text{Cas } L_1 : \frac{\partial f(\beta_1, \ldots, \beta_p)}{\partial \beta_j} = \lambda sign(\beta_j)$$
$$\text{Cas } L_2 : \frac{\partial f(\beta_1, \ldots, \beta_p)}{\partial \beta_j} = 2\lambda \beta_j$$

où $sign(\beta_j)$ vaut 1 si $\beta_j > 0$ et -1 si $\beta_j < 0$. Le cas $\beta_j = 0$ n'est pas bien défini puisque $|\beta_j|$ n'est pas dérivable en 0. En pratique sa dérivée peut être approchée par la fonction définie partout $\beta_j/(|\beta_j| + \epsilon)$ avec $\epsilon > 0$, où l'on considère que les valeurs de $\beta_j < \epsilon$ sont négligeables.

Dans le cas L_1, β_j est nul si $|\partial f(\ldots)/\partial \beta_j| < \lambda$ ce qui permet de ne sélectionner que les β_j ayant réellement une influence sur f. Dans le cas L_2, $2\lambda\beta_j$ est à l'équilibre avec $f(\ldots)/\partial \beta_j$, c'est à dire que plus β_j est faible, moins il pénalise l'attache de f au données. Il a ainsi très peu de chances d'être nul.

Pour avoir une meilleur intuition de ces principes, la figure ci-dessous permet d'illustrer sur une dimension β_j la différence d'impact entre les pénalités L_1 et L_2 au regard d'un terme quadratique d'attache aux données :



Dans cette figure et ne s'intéressant qu'à la dimension j, le minimum de $\mathbb{E}(Y - X\beta) + \lambda(\beta_j)^2$ ne sera jamais en $\beta_j = 0$ pour un λ fini. Le minimum de $\mathbb{E}(Y - X\beta) + \lambda|\beta_j|$ sera par contre $\beta_j = 0$ si λ est suffisamment grand par rapport à la dérivée de $\mathbb{E}(Y - X\beta)$ en ce point.

Utilisation de la régression Lasso

La pénalisation est optimisée comme en régression ridge par validation croisée (voir Section 3.4).

Grâce à ses solutions parcimonieuses, cette méthode est surtout utilisée pour sélectionner des variables dans des modèles de grande dimension ; on peut l'utiliser si $p > n$ c'est-à-dire s'il y a plus de variables que d'observations. Bien entendu, dans ce cas, les colonnes de la matrice X ne sont pas linéairement indépendantes. Il n'y a donc pas de solution explicite, on utilise des procédures d'optimisation pour trouver la solution. Il faut néanmoins utiliser la méthode avec précaution lorsque les variables explicatives sont corrélées. Pour que la méthode fonctionne, il faut que le nombre de variables influentes (correspondant à des β_j différents de 0) ne dépasse pas n et que les variables non influentes ne soient pas trop corrélées avec celles qui le sont.

3.3.3 Régression Elastic Net

La méthode Elastic Net permet de combiner la régression ridge et la régression Lasso, en introduisant les deux types de pénalités simultanément.

Le critère à minimiser est :

$$\widehat{\beta}_{E.N.} = \underset{\beta\in\mathbb{R}^{p+1}}{\arg\min} \left(\sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} X_i^{(j)} \beta_j)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2 \right) \right)$$

— Pour $\alpha = 1$, on retrouve la méthode LASSO.
— Pour $\alpha = 0$, on retrouve la régression ridge

Il y a dans ce dernier cas deux paramètres à optimiser par validation croisée.

3.3.4 Sélection par réduction de dimension

Le principe de ces approches consiste à calculer la régression sur un ensemble de variables orthogonales deux à deux. Celles-ci peuvent être obtenues à la suite d'une analyse en composantes principales ou par décomposition en valeur singulière de la matrice $\mathbf{X}$: c'est la régression sur les composantes principales associées aux plus grandes valeurs propres.

L'autre approche ou régression PLS (Partial Least Squares, Section 6.2) consiste à rechercher itérativement une composante linéaire des variables de plus forte covariance avec la variable à expliquer sous une contrainte d'orthogonalité avec les composantes précédentes.

3.4 Validation croisée

Considérons la formule générique optimisée dans la section précédente :

$$\widehat{\beta} = \underset{\beta\in\mathbb{R}^{p+1}}{\arg\min} \left(\sum_{i=1}^{n} (Y_i - \sum_{j=0}^{p} X_i^{(j)} \beta_j)^2 + \lambda R(\beta_1, \ldots, \beta_p) \right)$$

où R est la fonction de pénalisation de β qui regularise le problème d'optimisation.

Trois méthodes de validation croisée (cross-validation) pour valider le choix du paramètre λ et éventuellement de α sont largement utilisées en apprentissage automatique (pas seulement en regression linéaire).

3.4.1 Subdivision des observations en deux ensembles de données

La méthode élémentaire est de subdiviser les n observations en deux sous ensembles d'observations :

— Les données d'apprentissage.
— les données de validation.

Les données seront idéalement séparées de maniere aléatoire, par exemple $i = 1, \ldots, n_1$ pour les données d'apprentissage et $i = n_1 + 1, \ldots, n$ pour les données de validation.

Après avoir estimé $\hat{\beta}$ sur les données d'apprentissage, l'erreur d'approximation moyenne peut être estimée sur les données de validation

$$e_{split} = \frac{1}{n - n_1} \sum_{i=n_1+1}^{n} |Y_i - \hat{Y}_i|$$

Si le problème est trop régularisé, les tendances des données seront mal capturées par le modèle et les $\hat{Y}_i$ auront de grandes chances de mal estimer les Y_i. L'erreur e_{split} sera alors élevé. A contrario, si le problème n'est pas assez régularisé, le modèle va trop coller aux données d'apprentissage (sur-apprentissage, overfitting) sans fort pouvoir de prédiction pour d'autres données. L'erreur e_{split} sera alors de même élevé.

Les paramètre optimaux λ et éventuellement α sont alors ceux qui minimisent e_{split}. L'optimisation des paramètres peut typiquement se faire par une *grid search*, une descente de gradient ou un algorithme stochastique (ex : recuit simulé).

3.4.2 K-folds

Afin de quantifier la stabilité de l'estimation des β_j en fonction des données il est intéressant de reproduire plusieurs fois le test de séparation de données en jeu d'apprentissage et jeu d'estimation.

La méthode la plus simple est celle dite des K-folds. Elle consiste à subdiviser les n observations $(Y_i, \mathbf{X_i})$ en K jeux de données de taille similaires δ_k, *i.e.* avec δ_k proche de n/K. Pour simplifier les notations, on suppose ici que $\delta_k = n/K$ est entier.

La méthode d'apprentissage-validation decrite dans la sous-section précedente est alors effectuée K fois, avec pour l'itération k :

— Les données d'apprentissage $(Y_i, \mathbf{X_i})$, $i = 1, \ldots, (k-1)\delta_k, k\delta_k + 1, \ldots, n$ sont utilisées pour estimer les β_j^k.

— Les données de validation $(Y_i, \mathbf{X_i})$, $i = (k-1)\delta_k + 1, \ldots, k\delta_k$. sont utilisées pour calculer e_{split}^k.

$K > 1$ estimation de l'erreur e_{split}^k et des paramètres β_j^k sont alors effectués. Ceci permet d'en mesurer l'erreur de manière plus robuste qu'avec $K = 1$. De plus cela permet de quantifier la variabilité sur l'estimation des β_j : On peut simplement en calculer leur moyenne et écart type. Si une stratégie de sélection de modèle a été effectuée, on peut aussi étudier quels sont les β_j systématiquement sélectionnés et quels sont ceux qui le sont moins.

3.4.3 Leave-one-out

La méthode de validation croisée dite *Leave-one-out* est extremement populaire en apprentissage automatique et est équivalent aux K-folds avec $K = n$. A chaque itération, l'apprentissage est effectué en enlevant une observation du jeu de données et la validation est faite sur cette observation. Cette méthode est plus lente que les K-folds en particulier quand n est grand, mais est la plus robuste et recommandée quand n est petit.