Régression Linéaire et Régression pénalisée

immediate

16 novembre 2018

Note aux lecteurs : Pour une lecture et une compréhension approfondie, nous conseillons aux lecteurs de se référer à Wikistat et Rencher ainsi que les papiers des différentes méthodes présentées dans les sections suivantes, notamment Ridge et Lasso.

1 Introduction

Les modèles classiques de régression (linéaire, logistique) sont anciens et moins l'occasion de battage médiatique que ceux récents issus de l'apprentissage machine. Néanmoins, compte tenu de leur robustesse, de leur stabilité face à des fluctuations des échantillons, de leur capacité à passer à l'échelle des données massives... tout ceci fait qu'ils restent 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 Rappels - Régression multiple

2.1 Modèle

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

Les données sont supposées provenir de l'observation d'un échantillon statistique de taille $n (n > p + 1)$ de $\mathbb{R}^{p+1}$:

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

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 $\mathbf{1}, \mathbf{X}^1, ..., \mathbf{X}^p$ où $\mathbf{1}$ désigne le vecteur de $\mathbb{R}^n$ constitué de 1s. C'est-à-dire que les $(p+1)$ variables aléatoires vérifient :

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

avec les hypothèses suivantes :

- Les ϵ_i sont des termes d'erreur indépendants et identiquement distribués ; $\mathbb{E}(\epsilon_i) = 0$, $Var(\epsilon) = \sigma^2\mathbf{I}$.
- Les termes $\mathbf{X}^j$ sont supposés déterministes (facteurs contrôlés) ou bien l'erreur ϵ est indépendante de la distribution conjointe de $\mathbf{X}^1, ..., \mathbf{X}^p$. On écrit dans ce dernier cas que :
  $\mathbb{E}(\mathbf{Y}|\mathbf{X}^j, ..., \mathbf{X}^j) = \beta_0 + \beta_1\mathbf{X}^1 + ... + \beta_p\mathbf{X}^p$ et $Var(\mathbf{Y}|\mathbf{X}^j, ..., \mathbf{X}^j) = \sigma^2$
- Les paramètres inconnus $\beta_0, ..., \beta_p$ 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 ($\mathcal{N}(0, \sigma^2\mathbf{I})$). Les ϵ_i sont alors i.i.d de loi $\mathcal{N}(0, \sigma^2)$.

Les données sont rangées dans une matrice $\mathbf{X}$ $(n \times (p+1))$ de terme général X_i^j, dont la première colonne contient le vecteur $\mathbf{1}$ ($X_i^0 = 1$), et dans un vecteur $\mathbf{Y}$ de terme général Y_i. En notant les vecteurs $\epsilon = [\epsilon_1...\epsilon_p]'$ et $\beta = [\beta_1...\beta_p]'$, le modèle s'écrit matriciellement :

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

2.2 Estimation

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

Lors d'une estimation par moindres carrés, nous cherchons à résoudre sur $\beta \in \mathbb{R}^{p+1}$:

$$(\mathcal{P}) : \min_{\beta\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i^1 - ... - \beta_p X_i^p)^2 = \min_{\beta\in\mathbb{R}^{p+1}} ||\mathbf{Y} - \mathbf{X}\beta||^2$$

Par dérivation matricielle, on obtient les équations normales $\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\beta = 0$ dont la solution est bien un minimum car la hessienne $2\mathbf{X}'\mathbf{X}$ est bien semi-définie positive.

Sous 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, il est possible d'étudier analytiquement et de résoudre le problème de minimisation. Si cette hypothèse n'est pas vérifiée, il suffit en principe de supprimer des colonnes de X et donc des variables du modèle. Une approche de réduction de dimension (régression ridge, Lasso, PLS...) est alors à mettre en oeuvre. Alors, l'estimation des paramètres β_j est donnée par :

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

Il est alors possible de trouver un estimateur sans biais de σ^2, défini par :

$$\hat{\sigma}^2 = \frac{||\mathbf{Y} - \mathbf{X}\hat{\beta}||^2}{n-p-1}$$

Pour plus de précisions concernant ces estimateurs, le lecteur pourra se référer aux p.141-151 de Linear Models in Statistics (Rencher).

2.3 Prévisions

Connaissant les valeurs des variables $\mathbf{X}^j$ pour une nouvelle observation : $\mathbf{x}_0 = [x_0^1, x_0^2, ..., x_0^p]$ appartenant au domaine dans lequel l'hypothèse de linéarité reste valide, une prévision, notée $\hat{y}_0$ de $\mathbf{Y}$ ou $\mathbb{E}(\mathbf{Y})$ est donnée par :

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_0 x_0^1 + ... + \hat{\beta}_p x_0^p.$$

Les intervalles de confiance des prévisions de $\mathbf{Y}$ et $\mathbb{E}(\mathbf{Y})$, 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 :

$$\hat{y}_0 \pm t_{\frac{\alpha}{2};(n-p-1)}\hat{\sigma}(1 + \mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{\frac{1}{2}}, \hat{y}_0 \pm t_{\frac{\alpha}{2};(n-p-1)}\hat{\sigma}(\mathbf{v}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{v}_0)^{\frac{1}{2}}.$$

où $t_{\frac{\alpha}{2};(n-p-1)}$ est le quantile de la loi de Student au risque α et à $n-p-1$ degrés de liberté. Les variances de ces prévisions, comme celles des estimations des paramètres, dépendent directement du conditionnement de la matrice $\mathbf{X}'\mathbf{X}$.

2.4 Diagnostics

La validité d'un modèle de régression multiple et donc la fiabilité des prévisions, dépendent de la bonne vérification des hypothèses :

- homoscédasticité : variance σ^2 des résidus constante,
- linéarité du modèle : paramètres β_j constant,
- absence de points influents par la distance de Cook :

$$D_i = \frac{1}{s^2(p+1)}(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)}),$$

  où $\hat{\mathbf{y}}_{(i)}$ est l'estimation de $\mathbf{y}$ en n'ayant pas pris en compte la i-ème observation dans le calcul des estimateurs ;
- éventuellement la normalité des résidus,
- le conditionnement de la matrice $\mathbf{X}'\mathbf{X}$.

Tracer le graphe des résidus standardisés en fonction des valeurs ajustés montre leur plus ou moins bonne répartition autour de l'axe $y = 0$. La forme de ce nuage est susceptible de dénoncer une absence de linéarité ou une hétéroscédasticité.

Le conditionnement de la matrice $\mathbf{X}'\mathbf{X}$ est indiqué par le rapport $\kappa = \frac{\lambda_1}{\lambda_p}$, où $\lambda_1, ..., \lambda_p$ sont les valeurs propres de la matrice des corrélations rangées par ordre

décroissant. Ainsi, des problèmes de variances excessives voire même de précision numérique apparaissent dès que les dernières valeurs propres sont relativement trop petites.

3 Régression régularisée ou pénalisée

3.1 Régression Ridge

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

$$\tilde{\mathbf{X}} = \begin{pmatrix} 1 & X_1^1 & X_1^2 & . & X_1^p \\ 1 & X_2^1 & X_2^2 & . & X_2^p \\ . & . & . & . & . \\ 1 & X_n^1 & X_n^2 & . & X_n^p \end{pmatrix}, \tilde{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_p \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_p \end{pmatrix}$$

où $\mathbf{X}^0 = (1, 1, ..., 1)'$, 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 $\mathbb{L}^2$. En d'autres termes, on cherche à résoudre le problème d'optimisation suivant :

$$(\mathcal{P}_{ridge}) : \min_{\beta \in \mathbb{R}^{p+1}} (\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i^1 - ... - \beta_p X_i^p)^2 + \lambda \sum_{j=1}^{p} \beta_j^2)$$

On notera que β_0 n'est pas pénalisé. La régression ridge revient encore à estimer le modèle par les moindres carrés sous la contrainte que la norme du vecteur β des paramètres ne soit pas trop grande. 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. Supposons maintenant que $\mathbf{X}$ et $\mathbf{Y}$ sont 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 :

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

La solution est donc explicite et linéaire en $\mathbf{Y}$.

Remarques :

- $\mathbf{X}'\mathbf{X}$ est une matrice symétrique positive. Il en résulte que pour tout $\lambda > 0$, $\mathbf{X}'\mathbf{X} + \lambda\mathbf{I}_p$ est nécessairement 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 : $\hat{\beta}_0 = \bar{\mathbf{Y}}$; ajouter une constante à $\mathbf{Y}$ ne modifie pas les $\hat{\beta}_j$ 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) les vecteurs avant de minimiser le critère.

3.2 Régression Lasso

La régression ridge permet donc de contourner les problèmes de colinéarité mme en présence d'un nombre important de variables explicatives ou prédicteurs ($p > n$). La principale faiblesse de cette méthode est liée aux difficultés d'interprétation car, sans sélection, toutes les variables sont concernées dans le mod !le. D'autres approches par pénalisation permettent également une sélection, c'est le cas de la régression Lasso.

La méthode Lasso Tibshirani correspond, encore une fois, à la minimisation d'un critère des moindres carrés avec une pénalité de type l_1 (et non plus l_2 comme dans la régression ridge). Soit le modèle linéaire :

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

L'estimateur Lasso de β est défini en résolvant le problème d'optimisation suivant :

$$(\mathcal{P}_{Lasso}) : \min_{\beta \in \mathbb{R}^{p+1}} (\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i^1 - ... - \beta_p X_i^p)^2 + \lambda \sum_{j=1}^{p} |\beta_j|),$$

où λ est un paramètre positif à choisir. Comme pour la régression ridge, on peut montrer que ceci équivaut à un problème d'optimisation sous contrainte tel que la norme l_1 du vecteur β des paramètres ne soit pas trop grande. Le paramètre λ joue encore le rle de paramètre de régularisation : si $\lambda = 0$, on retrouve l'estimateur des moindres carrés, et s'il tend vers l'infini, on annule tous les coefficients $\beta_j,\ j = 1, ..., p$. La solution obtenue est dite parcimonieuse (sparse en anglais), car elle comporte des coefficients nuls.

3.3 Elastic Net

La méthode Elastic Net permet de combiner la régression ridge et la regression Lasso, en introduisant les deux types de pénalités simultanément. Le critère à minimiser devient alors :

$$(\mathcal{P}_{EN}) : \min_{\beta \in \mathbb{R}^{p+1}} (\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i^1 - ... - \beta_p X_i^p)^2 + \lambda(\alpha \sum_{j=1}^{p} |\beta_j| + (1-\alpha) \sum_{j=1}^{p} \beta_j^2),$$

Pour $\alpha = 1$, on retrouve la méthode Lasso, et pour $\alpha = 0$ on retrouve la régression ridge.

3.4 Choix des paramètres de régularisation

Comme dans tout problème de régularisation, le choix de la valeur des paramètres de régularisation (λ pour ridge et lasso, λ et α pour Elastic Net) est crucial et déterminera le choix de modèle. La validation croisée est généralement utilisée pour optimiser le choix. Le principe de la validation croisée est expliquée dans cette vignette.

3.5 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 X : c'est la régression sur les composantes principales associées aux plus grandes valeurs propres. L'autre approche ou régression PLS (partial least square 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. Ce deux méthodes sont développées dans une vignette spécifique.