Analyse de Données - Régression

Antoine Lavault[1,2]
Ouneïs Gloton[3]

[1]Apeira Technologies

[2]UMR CNRS 9912 STMS, IRCAM, Sorbonne Université

[3]Institut de Mathématiques de Bourgogne, uB

19 janvier 2024

Contenu

Contenu

Outline

Problème général

- On cherche à prédire une variable Y (variable *expliquée*) à partir d'autres variables $X = (X_1, \ldots, X_p)$ (variables *explicatives*).
- On fait une hypothèse sur la relation entre Y et les X_k. Ici on se limite au cadre des méthodes *paramétriques*, c'est à dire qu'on suppose

$$Y \sim \mathcal{D}(X, \beta)$$

pour un certain β inconnu, avec $\{\mathcal{D}(X, \beta)\}_{X,\beta}$ une famille paramétrique de distributions de probabilité.

Remarque

En d'autres termes, à une valeur de X fixée, la valeur mesurée de Y est une variable aléatoire dont la distribution dépend de X et du paramètre inconnu β. On suppose ici que la variable X est mesurée avec une précision infinie.

- A partir d'un jeu de données fini $(x_i, y_i)_{1 \leq i \leq n}$, où les y_i sont supposés être des échantillons de variables aléatoires $Y_i \sim \mathcal{D}(x_i, \beta)$ indépendantes, le but est d'estimer β, c'est à dire de calculer un *estimateur* $\hat{\beta}$ qui permettra de prédire la moyenne, variance, etc. de Y pour d'autres valeurs de X.

Estimation par maximum de vraisemblance

- Si Y est une variable discrète :
  On suppose que Y est à valeurs dans un ensemble fini $\mathcal{Y}$. On note $p(y|x,\beta)$ la probabilité que Y prenne la valeur $y \in \mathcal{Y}$ lorsque $Y \sim \mathcal{D}(x,\beta)$. En inversant le point de vue, on note $L(\beta|x,y) = p(y|x,\beta)$ que l'on appelle maintenant la *vraisemblance* (likelihood) de β pour l'observation (x,y).
- Si Y est une variable continue :
  On suppose que Y est à valeurs dans R^k, et que la distribution $\mathcal{D}(x,\beta)$ est donnée par une densité $f(y|x,\beta)$. La fonction de vraisemblance est alors $L(\beta|x,y) = f(y|x,\beta)$.
- L'estimation par *maximum de vraisemblance* (Maximum Likelihood Estimation, ou MLE) consiste à définir

$$\hat{\beta} = \text{argmax}_\beta \prod_{i=1}^{n} L(\beta|x_i,y_i) = \text{argmin}_\beta \sum_{i=1}^{n} -\ell(\beta|x_i,y_i)$$

où $\ell(\beta|x,y) = \log L(\beta|x,y)$ est la *log-vraisemblance* (log-likelihood).

Outline

Intuition

- La dépendance affine entre variables est parmi les modèles les plus simples qu'on peut imaginer. Il est pertinent dans de nombreuses situations (même s'il ne doit pas être utilisé automatiquement !).
- Même dans les cas où le modèle est pertinent, voire exact, les mesures que l'on effectue sont toujours entachées d'erreur. La théorie de la régression linéaire indique comment estimer les paramètres du modèle de manière optimale (pour différentes interprétations d'"optimal").

Définition du problème

- Les variables expliquée Y et explicatives $X_1, \ldots, X_p$ sont des variables réelles.
- On suppose que Y suit une loi normale de variance σ^2 et de moyenne $\mu = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$, c'est-à-dire que la moyenne de Y dépend de manière affine des X_k. Une astuce permet de se ramener à un problème linéaire : on écrit $\mu = X\beta$ avec $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$ et $X = \begin{pmatrix} 1 & X_1 & \ldots & X_p \end{pmatrix}$.

Méthode des moindres carrés

- Le jeu de données est de la forme $(x_i, y_i)_{1 \leq i \leq n}$ avec y_i réel et $x_i = \begin{pmatrix} 1 & x_{i,1} & \dots & x_{i,p} \end{pmatrix}$.

  On introduit les notations $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ et

  $$X = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix}.$$

- Cette méthode consiste à estimer β par

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^{n} (y_i - x_i\beta)^2 = \text{argmin}_\beta (Y - X\beta)^T (Y - X\beta)$$

c'est-à-dire qu'on cherche $\hat{\beta}$ qui minimise la somme des carrés des différences entre les valeurs observées y_i et les moyennes prédites $x_i\hat{\beta}$.

Méthode des moindres carrés : calcul de l'estimateur

- Un minimum de $L(\beta) = (Y - X\beta)^T(Y - X\beta)$ est en particulier un point singulier de L. Au premier ordre en $\mathrm{d}\beta$ on a

$$L(\beta + \mathrm{d}\beta) - L(\beta) = 2(Y - X\beta)^T(-X\mathrm{d}\beta)$$

ce qui amène à

$$\begin{aligned}(X\beta - Y)^T X &= 0\\ \beta^T X^T X &= Y^T X\\ X^T X\beta &= X^T Y\\ \beta &= (X^T X)^{-1} X^T Y\end{aligned}$$

sous réserve d'inversibilité de $X^T X$.

Critère d'inversibilité de X^TX

- On note $m = p + 1$, et $p = \operatorname{rg} X$. Comme X est de dimensions $n \times m$, on a $p \leq \min(n, m)$. En utilisant Gram-Schmidt, on peut écrire $X = QP$ avec P matrice $m \times m$ inversible et Q matrice $n \times m$ telle que $Q^TQ = \left(\begin{array}{c|c} I_p & \\ \hline & 0_{m-p} \end{array}\right)$.
  Il devient clair que $X^TX = P^T(Q^TQ)P$, la matrice de Gram de X, est de rang p. Comme X^TX est de dimensions $m \times m$:
- Si $n < m$:
  On a $p < m$ donc X^TX n'est pas inversible.
- Si $n \geq m$: Il faut et il suffit que $\operatorname{rg} X = m$ pour que X^TX soit inversible.
  C'est le cas génériquement.

Comparaison avec résolution d'un système linéaire

- Étant donnés les (x_i, y_i), le problème linéaire correspondant est $y_i = x_i\beta$ pour tout i, en d'autres termes $Y = X\beta$.
- Si $n < m$:
  Le système est *sous-déterminé*. Génériquement, $\operatorname{rg} X = n$ et on peut choisir $Q = \begin{pmatrix} I_n & 0_{n\times(m-n)} \end{pmatrix}$. Le système se réécrit $Y = QP\beta$ dont les solutions sont données par $\beta = P^{-1}\begin{pmatrix} Y \\ Z \end{pmatrix}$ avec Z matrice $(m-n)\times 1$ arbitraire. Il existe une infinité de solutions.

Remarque

La conséquence de la sous-détermination sur la méthode des moindres carrés est qu'elle n'est pas définie. Une manière de le voir est que chaque solution du système linéaire minimise L, et il y en a une infinité. Il faut se tourner vers d'autres méthodes comme la "régression des moindres carrés partiels".

Comparaison avec résolution d'un système linéaire

- Si $n = m$:
  Génériquement, X est alors inversible. Dans ce cas, le système admet une unique solution $\beta = X^{-1}Y$.

Remarque

Cette solution coïncide avec l'estimation des moindres carrés puisque $(X^TX)^{-1}X^TY = X^{-1}(X^T)^{-1}X^TY = X^{-1}Y$. Une autre manière de le voir est que $L(\beta) \geq 0$, et $L(\beta) = 0$ si et seulement si $y_i = x_i\beta$ pour tout i, ce qui est le cas pour $\beta = X^{-1}Y$.

Comparaison avec résolution d'un système linéaire

- Si $n > m$:
  Génériquement, $\operatorname{rg} X = m$ et on peut écrire $Q = UX'$ avec U matrice $n \times n$ orthogonale et $X' = \begin{pmatrix} I_m \\ 0_{(n-m)\times m} \end{pmatrix}$. Le système s'écrit $U^{-1}Y = X'P\beta$ qui n'admet une solution que si $U^{-1}Y$ est de la forme $\begin{pmatrix} Y' \\ 0 \end{pmatrix}$ pour une certaine matrice Y' de dimensions $m \times 1$, ce qui n'est pas le cas génériquement. Si c'est le cas, $\beta = P^{-1}Y'$.

Remarque

La méthode des moindres carrés devient intéressante puisque même si le système linéaire n'as pas de solution, elle donne une meilleure approximation dans un certain sens. Elle consiste à écrire $U^{-1}Y = \begin{pmatrix} Y' \\ Z \end{pmatrix}$ avec Y' de dimensions $m \times 1$, et de définir $\beta = P^{-1}Y'$. En d'autres termes, on projette Y sur $\operatorname{Im}(X)$ avant de résoudre le système linéaire. Pour preuve, il suffit de faire le calcul $(X^TX)^{-1}X^TY = P^{-1}X'^TU^{-1}Y = P^{-1}Y'$.

Estimation par maximum de vraisemblance

- Rappelons que l'on suppose $Y \sim \mathcal{N}(x\beta, \sigma^2)$. On a alors

$$L(\beta|x, y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - x\beta)^2}{2\sigma^2}\right)$$

d'où il est clair que

$$\text{argmin}_\beta \sum_i -\ell(\beta|x_i, y_i) = \text{argmin}_\beta \sum_i (y - x\beta)^2$$

ce qui montre l'équivalence avec la méthode des moindres carrés dans le cas gaussien.

Biais et efficacité de l'estimateur des moindres carrés

- Sous les hypothèses énoncées, $\hat{\beta}$ est un estimateur *non biaisé*, c'est-à-dire que $\mathbb{E}[\hat{\beta} - \beta] = 0$. L'espérance est prise sur les $Y_i \sim \mathcal{N}(x_i\beta, \sigma^2)$ indépendants.
- $\hat{\beta}$ est un estimateur *linéaire*, c'est-à-dire $\hat{\beta}(X, Y) = M(X)Y$ pour une certaine matrice $M(X)$ (qui elle dépend non-linéairement de X).
- Sous les hypothèses énoncées, $\hat{\beta}$ est le meilleur estimateur non biaisé dans le sens où $\mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T]$, la variance de l'erreur d'estimation, est minimale parmi tous les estimateurs non-biaisés (non nécessairement linéaires). C'est-à-dire, étant donné un autre estimateur non biaisé $\hat{\beta}_1$, la différence $\mathbb{E}[(\hat{\beta}_1 - \beta)(\hat{\beta}_1 - \beta)^T] - \mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T]$ est une matrice autoadjointe positive. C'est une conséquence de l'hypothèse de normalité et de l'inégalité de Cramer-Rao. La variance de $\hat{\beta}$ est $\sigma^2(X^TX)^{-1}$.

Théorème de Gauss-Markov

- Sous des hypothèses plus faibles, on a un résultat similaire. On ne suppose plus les Y_i indépendants. On suppose que Y_i a pour moyenne $x_i\beta$, que les Y_i ont la même variance σ^2, et qu'elles sont non-corrélées, c'est-à-dire $\mathbb{E}[(Y_i - x_i\beta)(Y_j - x_j\beta)] = 0$ lorsque $i \neq j$.
- Le théorème de Gauss-Markov énonce alors que $\hat{\beta}$ est le meilleur estimateur parmi les estimateurs linéaires non biaisés.

Cas simple : régression affine

- Dans ce cas, on peut écrire $X_i = (1\ x_i)$, $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$, avec Y_i de moyenne $\beta_0 + \beta_1 x_i$.
- L'estimation par moindres carrés donne

$$\beta_1 = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

où

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\mathrm{Var}(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\mathrm{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Outline

Idée générale

- La méthode des moindres carrés est trop sensible à une caractéristique fréquente des jeux de données : la présence de *points aberrants* (outliers), qui invalident l'hypothèse de normalité des erreurs.
- La *régression robuste* fournit des estimateurs qui sont peu affectés par la présence de points aberrants, jusqu'à une certaine proportion de contamination du jeu de données (le *breakdown point*).

Un exemple : la méthode RANSAC

- Il s'agit d'un algorithme itératif très simple. On suppose qu'on a un jeu de données $(x_i, y_i)_{1 \leq i \leq N}$. On fixe : n, le nombre de points pour une première estimation de β ; d, le nombre de points nécessaire pour valider la première estimation ; ϵ, l'erreur au-delà de laquelle un point est considéré comme un outlier ; et k, le nombre d'itérations de l'algorithme.
- Une itération consiste à prendre n points au hasard dans le jeu de données, de calculer un premier $\hat{\beta}_1$ à partir de ce sous-ensemble, et pour chaque point (x_i, y_i) du jeu données :
  - de le considérer comme un outlier si $(x_i \hat{\beta}_1 - y_i)^2 > \epsilon$
  - de le considérer comme un inlier si $(x_i \hat{\beta}_1 - y_i)^2 \leq \epsilon$

  Puis, si le nombre d'inliers est supérieur ou égal à d, on calcule à partir d'eux un deuxième modèle $\hat{\beta}_2$. On calcule son erreur totale sur le jeu de données complet : $e = \sum_{i=1}^{N} (y_i - x_i \hat{\beta}_2)^2$.
- On définit $\hat{\beta}$ comme le $\hat{\beta}_2$ de ces itérations qui la plus petite erreur totale.

Outline

Régression logistique : définition du problème

- La variable expliquée Y est une variable binaire, qui peut représenter une propriété qualitative. Elle ne peut prendre que les valeurs 0 ou 1.
- Les variables explicatives $X_1, \ldots, X_p$ sont réelles, on les regroupe encore comme $X = (1\ X_1 \ldots X_p)$.
- On suppose $Y \sim \mathrm{Bernoulli}(p(X\beta))$ où $p(z) = \frac{1}{1+e^{-z}}$.
  L'idée étant que $\log \frac{p}{1-p} = X\beta$, c'est-à-dire que le logarithme du rapport des chances (log odds) de Y dépende de manière affine des variables explicatives.

Estimation par MLE

- On estime β par MLE. La vraisemblance de β pour l'observation (x, y) est

$$L(\beta|x,y) = p(y|x,\beta) = \begin{cases} p(x\beta) = 1/(1+e^{-x\beta}) & \text{si } y = 1 \\ 1 - p(x\beta) = e^{-x\beta}/(1+e^{-x\beta}) & \text{si } y = 0 \end{cases}$$

  Ce qui donne

$$-\ell(\beta|x,y) = \begin{cases} \log(1+e^{-x\beta}) & \text{si } y = 1 \\ x\beta + \log(1+e^{-x\beta}) & \text{si } y = 0 \end{cases}$$

- Pour n observations $(x_i, y_i)_{1 \le i \le n}$, on définit $\hat{\beta} = \text{argmin}_\beta\, C(\beta)$ où $C(\beta) = \sum_{i=1}^{n} -\ell(\beta|x_i, y_i)$. On ne peut pas résoudre l'équation $C'(\beta) = 0$ explicitement car elle est non-linaire.
- On peut utiliser une méthode itérative comme la méthode de Newton pour minimiser $C(\beta)$.