Régression linéaire multiple

Kaoutar IKKANE

FPO
Mathématiques appliquées pour la science de données

December 26, 2024

1. Introduction
2. Modèle
3. Estimation
4. Ecart au modèle : Résidus
5. Décomposition de la variabilité
6. Conslusion

Introduction

Introduction

Le but de la régression simple (resp. multiple) est d'expliquer une variable Y à l'aide d'une variable X (resp. plusieurs variables $X_1, \ldots, X_q$). La variable Y est appelée variable dépendante, ou variable à expliquer et les variables X_j ($j = 1, \ldots, q$) sont appelées variables indépendantes, ou variables explicatives.

Modèle

Modèle

Une variable quantitative Y est mise en relation avec p variables quantitatives $X^1, \ldots, X^p$. Les données sont supposées provenir de l'observation d'un échantillon statistique de taille n $(n > p+1)$ de $\mathbb{R}^{(p+1)}$:

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

Modèle

Le modèle s'écrit :

$$Y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i \quad i = 1, 2, \ldots, n.$$

Hypothèses :

$$L(\varepsilon_i) = \mathcal{N}(0, \sigma^2) \quad \text{et} \quad \text{cov}(\varepsilon_i, \varepsilon_k) = 0 \quad \forall i, k$$

Modèle

FACULTÉ POLYDISCIPLINAIRE DE OUARZAZATE

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_j x_{1j} + \cdots + \beta_p x_{1p} \\
&\vdots \\
Y_i &= \beta_0 + \beta_1 x_{i1} + \cdots + \beta_j x_{ij} + \cdots + \beta_p x_{ip} \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_j x_{nj} + \cdots + \beta_p x_{np}
\end{aligned}
$$

Modèle

Ecriture matricielle du modèle : $Y = X\beta + \epsilon$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_p \end{pmatrix}$$

Modèle

الكلية متعددة التخصصات -ورزازات
ⵜⴰⵖⵉⵎⵎⴰⵜ ⵜⴰⵙⵜⵉⵎⵎⴰⵙⵜ – ⵡⴰⵔⵣⴰⵣⴰⵜ
FACULTÉ POLYDISCIPLINAIRE DE OUARZAZATE

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1j} & \dots & x_{1p} \\ \\ 1 & x_{i1} & \dots & x_{ij} & \dots & x_{ip} \\ \\ 1 & x_{n1} & \dots & x_{nj} & \dots & x_{np} \end{pmatrix}$$

Ecriture matricielle des hypothèses sur les résidus :
$Y = X\beta + \epsilon$

$$\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix} \qquad E(\varepsilon) = \begin{pmatrix} E(\varepsilon_1) \\ \vdots \\ E(\varepsilon_i) \\ \vdots \\ E(\varepsilon_n) \end{pmatrix} = 0$$

Modèle

FACULTÉ POLYDISCIPLINAIRE DE OUARZAZATE

$$V(\varepsilon) = \begin{pmatrix} V(\varepsilon_1) & & \\ & & \\ \operatorname{cov}(\varepsilon_i, \varepsilon_1) & V(\varepsilon_i) & \\ & & \\ & & V(\varepsilon_n) \end{pmatrix} = \sigma^2 I_d$$

Estimation

Méthode des moindres carrés

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_j x_{ij} + \cdots + \widehat{\beta}_p x_{ip} + e_i$$

Ecriture matricielle : $Y = X\widehat{\beta} + E$

$$\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_j \\ \vdots \\ \widehat{\beta}_p \end{pmatrix} \qquad E = \begin{pmatrix} e_1 \\ \vdots \\ e_i \\ \vdots \\ e_n \end{pmatrix}$$

Méthode des moindres carrés

Trouver $\widehat{\beta}$ tel que : $\sum e_i^2$ minimum (MMC)

solution:

Equations normales : $X'X\widehat{\beta} = X'Y$

1. n = p+1 et X inversible: $\widehat{\beta} = X^{-1}Y$ et $E = 0$
2. n > p+1 et X'X inversible: $\widehat{\beta} = (X'X)^{-1}X'Y$
3. n < p+1 : infinité de solution avec E = 0

Propriétés

Propriétés de l'estimateur $\widehat{\beta} = (X'X)^{-1}X'Y$:

$$E(\widehat{\beta}) = \beta \qquad E(\epsilon) = 0$$

$$V(\widehat{\beta}) = (X'X)^{-1}\sigma^2 \qquad V(\epsilon) = \sigma^2 I_d$$

Propriétés

$$V(\widehat{\beta}) = \begin{pmatrix} V(\widehat{\beta}_1) & & \\ \operatorname{cov}(\widehat{\beta}_j, \widehat{\beta}_1) & V(\widehat{\beta}_j) & \\ & & V(\widehat{\beta}_p) \end{pmatrix}$$

$$\overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}_1 + \cdots + \widehat{\beta}_j \overline{x}_j + \cdots + \widehat{\beta}_p \overline{x}_p$$

Propriétés

Cas des données centrées :

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_j x_{ij} + \cdots + \widehat{\beta}_p x_{ip} + e_i$$

$$y_i - \overline{y} = \widehat{\beta}_1 (x_{i1} - \overline{x}_1) + \cdots + \widehat{\beta}_j (x_{ij} - \overline{x}_j) + \ldots$$
$$+ \widehat{\beta}_p (x_{ip} - \overline{x}_p) + e_i$$

Ecart au modèle : Résidus

Ecart au modèle : Résidus

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_j x_{ij} + \cdots + \widehat{\beta}_p x_{ip}$$

$$e_i = y_i - \widehat{y}_i \qquad \widehat{Y} = X\widehat{\beta} \qquad E = Y - \widehat{Y}$$

Propriétés:

$X'E = 0$ (p+1 équations) $\quad \sum_i e_i = 0 \quad \sum_i e_i x_{ij} = 0 \quad \forall j$

$$\Longrightarrow \; cov[E, x_j] = 0$$

Ecart au modèle : Résidus

$$\hat{\sigma}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n - p - 1} = \frac{\sum_i (e_i)^2}{n - p - 1}$$

$$E[\hat{\sigma}^2] = \sigma^2$$

Ecart au modèle : Résidus

FACULTÉ POLYDISCIPLINAIRE DE OUARZAZATE

Exemple numérique $\hat{y}_i = 0 + 1x_{i1} + 2x_{i2}$

	x_1	x_2	y	$\hat{y}$	e
1	-1	-1	-4	-3	-1
2	-1	1	2	1	1
3	1	-1	0	-1	1
4	1	1	2	3	-1

$$\sum e_i^2 = 4$$

$$\hat{\sigma}^2 = \frac{4}{1}$$

$$\hat{\sigma} = 2$$

Décomposition de la variabilité

Décomposition de la variabilité

FACULTÉ POLYDISCIPLINAIRE DE OUARZAZATE

$$y_i - \overline{y} = (\widehat{y}_i - \overline{y}) + (y_i - \widehat{y}_i)$$

$$\sum_i (y_i - \overline{y})^2 = \sum_i (\widehat{y}_i - \overline{y})^2 + \sum_i (y_i - \widehat{y}_i)^2$$

Variabilité	Totale	Due au modèle	Résiduelle
Notation	SCE_T	SCE_M	SCE_R
Degré de lib	$n-1$	p	$n-p-1$

Décomposition de la variabilité

FACULTÉ POLYDISCIPLINAIRE DE OUARZAZATE

avec:

$$\widehat{y}_i - \overline{y} = \sum_{j=1}^{j=p} \widehat{\beta}_j (x_{ij} - \overline{x}_j)$$

Décomposition de la variabilité

Exemple numérique

variabilité	SCE	ddl
totale	$SCE_T = 24$	$n - 1 = 3$
due au modèle	$SCE_M = 20$	$p = 2$
résiduelle	$SCE_R = 4$	$n - p - 1 = 1$

Décomposition de la variabilité

Coéfficient de détermination :

$$\frac{\mathrm{SCE}_M}{\mathrm{SCE}_T} = R^2$$

$$R^2 = [r(y, \hat{y})]^2$$

Conslusion

conslusion

En conclusion, la régression linéaire multiple est un modèle efficace pour capturer les relations entre une variable cible et plusieurs variables explicatives. À travers l'estimation des coefficients, l'analyse des résidus et la décomposition de la variabilité, on peut évaluer et affiner la pertinence du modèle. Ce modèle reste donc un outil incontournable pour l'analyse de données et la prise de décisions informées.

Merci Pour votre attention