

Remise à niveau - Statistique

Modèle linéaire

Claire Boyer

1. Modèle linéaire

2. Méthode des moindres carrés

3. Modèle linéaire gaussien

4. Modèles emboîtés

Modèle de régression. L'objectif est d'expliquer une **variable de sortie** Y en fonction de **variables explicatives/d'entrée** $X \in \mathbb{R}^p$, à l'aide d'un échantillon $(X_1, Y_1), \dots, (X_n, Y_n)$ copies de (X, Y) . Dans un modèle de régression, on suppose que le lien entre l'entrée et la sortie est donné par

$$Y_i = f^*(X_i) + \varepsilon_i$$

où

- ▶ $f^* : \mathbb{R}^p \rightarrow \mathbb{R}$ est la fonction de régression
- ▶ $\varepsilon_1, \dots, \varepsilon_n$ sont des termes de bruit, v.a indépendantes et centrées

Exemple 1 :

$$Y_i = ax_i^2 + bx_i + c + \varepsilon_i.$$

Exemple 2 :

$$Y_i = ax_i + b + \varepsilon_i.$$

Modèle linéaire càd avec une fonction de régression linéaire

$$g^*(X) = X^\top \beta^*, \quad \beta^* \in \mathbb{R}^p$$

Le paramètre $\beta^* \in \mathbb{R}^p$ représente les **coefficients de régression**

Forme matricielle. En notant

$$\mathbb{Y} = (Y_1, \dots, Y_n)^\top, \quad \varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top, \quad \mathbb{X} = \begin{pmatrix} X_{1,1} & \cdots & X_{1,p} \\ \vdots & & \vdots \\ X_{n,1} & \cdots & X_{n,p} \end{pmatrix},$$

le modèle s'écrit sous forme matricielle

$$\mathbb{Y} = \mathbb{X}\beta^* + \varepsilon$$

Modèle linéaire simple. Si $X_1, \dots, X_n \in \mathbb{R}$ et pour tout i :

$$Y_i = a + bX_i + \varepsilon_i,$$

alors

$$\mathbb{Y} = \mathbb{X}\beta^* + \varepsilon, \quad \mathbb{X} = \begin{pmatrix} 1 & X_1 \\ \vdots & \vdots \\ 1 & X_n \end{pmatrix}, \quad \beta = \begin{pmatrix} a \\ b \end{pmatrix}$$

La fonction $x \mapsto a + bx$ est dite *droite de régression*

Régression polynomiale. Si $X_1, \dots, X_n \in \mathbb{R}$ et

$$Y_i = a_0 + a_1X_i + \dots + a_pX_i^p + \varepsilon_i,$$

alors

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon, \quad \mathbb{X} = \begin{pmatrix} 1 & X_1 & \dots & X_1^p \\ \vdots & \vdots & & \vdots \\ 1 & X_n & \dots & X_n^p \end{pmatrix}, \quad \beta = \begin{pmatrix} a_0 \\ \vdots \\ a_p \end{pmatrix}.$$

Question : Comment estimer β ?

1. Modèle linéaire

2. Méthode des moindres carrés

3. Modèle linéaire gaussien

4. Modèles emboîtés

Principe. On cherche un estimateur $\hat{\beta}$ de β^* qui rend l'ajustement du modèle "proche" des données observées

On se donne une *fonction de coût* :

$$F(\beta) = \|\mathbb{Y} - \mathbb{X}\beta\|^2 = (\mathbb{Y} - \mathbb{X}\beta)^\top (\mathbb{Y} - \mathbb{X}\beta).$$

L'estimateur des moindres carrés est donné par :

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbb{Y} - \mathbb{X}\beta\|^2.$$

Développement de la fonction de coût.

$$F(\beta) = Y^T Y - 2\beta^T X^T Y + \beta^T X^T X \beta$$

Condition du premier ordre (règle de Fermat). On annule le gradient de F

$$\nabla F(\beta) = -2X^T Y + 2X^T X \beta = 0$$

Équations normales. On obtient

$$X^T X \hat{\beta} = X^T Y$$

Solution. Si $X^T X$ est **inversible** (si X est de **rang** p),

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Théorème (Existence et unicité)

Si $\mathbb{X}$ est de rang p , alors l'estimateur des moindres carrés est unique et

$$\hat{\beta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y}$$

Proposition. La matrice $\mathbb{X}^T \mathbb{X}$ est semi-définie positive. De plus, elle est définie positive si et seulement si $\mathbb{X}$ est de rang p .

Remarque. La matrice $\mathbb{X}$ ne peut être de rang p que si $n \geq p$.

On note $\text{Im}(\mathbb{X}) \subseteq \mathbb{R}^n$ le sous-espace vectoriel engendré par les colonnes de $\mathbb{X}$. Le problème de minimisation s'écrit

$$\min_{\beta' \in \mathbb{R}^p} \|\mathbb{Y} - \mathbb{X}\beta'\|^2 = \min_{h \in \text{Im}(\mathbb{X})} \|\mathbb{Y} - h\|^2,$$

et chercher le minimum revient donc à chercher la projection orthogonale de $\mathbb{Y}$ sur $\text{Im}(\mathbb{X})$

Remarque. $\dim(\text{Im}(\mathbb{X})) = \text{rang}(\mathbb{X})$.

Proposition. Si $\text{rang}(\mathbb{X}) = p$ (i.e. $\mathbb{X}^\top \mathbb{X}$ est inversible), alors

$$P_{\text{Im}(\mathbb{X})} = \mathbb{X}(\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top$$

est la matrice de projection orthogonale sur $\text{Im}(\mathbb{X})$ et $\text{rang}(P_{\text{Im}(\mathbb{X})}) = p$.

Propriétés de l'EMC. Supposons $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)$ avec $\mathbb{E}[\varepsilon] = 0_{\mathbb{R}^n}$ et $\text{Var}[\varepsilon] = \sigma^2 I_n$, $\sigma^2 > 0$, et $\text{rang}(\mathbb{X}) = p$. Alors

$$\mathbb{E}[\hat{\beta}] = \beta \quad (\text{sans biais}) \quad \text{et} \quad \text{Var}[\hat{\beta}] = \sigma^2 (\mathbb{X}^\top \mathbb{X})^{-1}$$

Définition (résidus).

1. Le vecteur des résidus est $\hat{\varepsilon} = \mathbb{Y} - \mathbb{X}\hat{\beta}$.
2. La somme des carrés des résidus est

$$\text{SCR} = \|\hat{\varepsilon}\|^2 = \|\mathbb{Y} - \mathbb{X}\hat{\beta}\|^2.$$

Estimation de σ^2 . Sous les mêmes hypothèses et si $\text{rang}(\mathbb{X}) = p$,

$$\hat{\sigma}^2 = \frac{\|\mathbb{Y} - \mathbb{X}\hat{\beta}\|^2}{n - p} = \frac{1}{n - p} \sum_{i=1}^n (Y_i - X_i^\top \hat{\beta})^2$$

est un **estimateur sans biais** de σ^2

1. Modèle linéaire

2. Méthode des moindres carrés

3. Modèle linéaire gaussien

4. Modèles emboîtés

Définition. On suppose que les erreurs suivent une loi normale

$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \quad \sigma^2 > 0.$$

Alors

$$Y = X\beta + \varepsilon \sim \mathcal{N}(X\beta, \sigma^2 I_n).$$

Propriétés des estimateurs. Pour un modèle linéaire gaussien, si $\text{rang}(X) = p$, alors

1. $\hat{\beta} = (X^\top X)^{-1} X^\top Y \sim \mathcal{N}(\beta, \sigma^2 (X^\top X)^{-1})$
2. $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}^2$
3. $\hat{\beta}$ et $\hat{\sigma}^2$ sont **indépendants**

Loi de Student. Soit $u \in \mathbb{R}^p$. On définit

$$T = \frac{u^\top \hat{\beta} - u^\top \beta}{\hat{\sigma} \sqrt{u^\top (\mathbb{X}^\top \mathbb{X})^{-1} u}}.$$

Alors

$$T \sim \mathcal{T}_{n-p}$$

Intervalle de confiance. Soit $u \in \mathbb{R}^p$. Un intervalle de confiance bilatère de niveau $1 - \alpha$ pour $u^\top \beta$ est donné par

$$\left[u^\top \hat{\beta} \pm t_{1-\alpha/2, n-p} \hat{\sigma} \sqrt{u^\top (\mathbb{X}^\top \mathbb{X})^{-1} u} \right]$$

avec $t_{1-\alpha/2, n-p}$ le quantile d'ordre $1 - \alpha/2$ d'une loi de Student $\mathcal{T}_{n-p}$

On considère le test suivant pour un certain $k \in \{1, \dots, p\}$

$$H_0 : \beta_k = 0, \quad H_1 : \beta_k \neq 0,$$

Proposition. Dans le cadre du **modèle linéaire gaussien**, un test de significativité du k -ème coefficient rejettera H_0 au niveau α si

$$0 \notin \left[\hat{\beta}_k \pm \hat{\sigma} \sqrt{v_k} t_{n-p, 1-\alpha/2} \right]$$

où

- ▶ $v_k = (\mathbb{X}^\top \mathbb{X})_{k,k}^{-1}$ est le k -ème coefficient diagonal de $(\mathbb{X}^\top \mathbb{X})^{-1}$
- ▶ $t_{n-p, 1-\alpha/2}$ est le quantile d'ordre $1 - \alpha/2$ d'une loi de Student à $n - p$ degrés de liberté

Test de Fisher : significativité de plusieurs coefficients ^{16 / 19}

On considère la partition suivante des coefficients / matrice de régression

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{et} \quad \mathbb{X} = (\mathbb{X}_0 \mathbb{X}_1)$$

avec $\mathbb{X}_0 \in \mathbb{R}^{n \times p_0}$, $\mathbb{X}_1 \in \mathbb{R}^{n \times p - p_0}$

On veut tester l'hypothèse nulle (seuls les p_0 premiers coefficients sont) ils significatifs ?)

$$H_0 : \beta_1 = 0 \quad \text{contre} \quad H_1 : \beta_1 \neq 0$$

càd

$$H_0 : "Y = \mathbb{X}_0 \beta_0 + \varepsilon" \quad \text{contre} \quad H_1 : "Y = \mathbb{X} \beta + \varepsilon"$$

Estimateurs sous H_0 et sous H_1 .

- ▶ Sous H_1 , l'estimateur est $\hat{\beta} = (\mathbb{X}^\top \mathbb{X})^{-1} \mathbb{X}^\top Y$, avec p paramètres
- ▶ Sous H_0 , l'estimateur est $\hat{\beta}_0 = (\mathbb{X}_0^\top \mathbb{X}_0)^{-1} \mathbb{X}_0^\top Y$, avec p_0 paramètres

Test de Fisher : significativité de plusieurs coefficients^{17 / 19}

Statistique de test.

$$F = \frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_0\|}{(p - p_0)\hat{\sigma}^2} =: \frac{\hat{\sigma}_0}{\hat{\sigma}} \underset{\text{sous } H_0}{\sim} \mathcal{F}_{p-p_0, n-p}$$

où

- ▶ $\hat{\mathbf{Y}} = \mathbb{X}_0 \hat{\beta}_0$
- ▶ $\hat{\beta}_0$ est l'EMC dans le modèle H_0

Tester la significativité des derniers coefficients revient à rejeter H_0 si

$$F > f_{1-\alpha, p-p_0, n-p}$$

avec $f_{1-\alpha, p-p_0, n-p}$ $1 - \alpha$ -quantile d'une loi de Fisher

Interprétation. Si F est trop grand, on rejette H_0 , ce qui signifie que l'ajout des dernières variables améliore significativement le modèle

1. Modèle linéaire
2. Méthode des moindres carrés
3. Modèle linéaire gaussien
4. Modèles emboîtés

Le test de Fisher précédent est un cas particulier de **modèles emboîtés**

Définition. On considère

- ▶ $\mathbb{Y} \sim \mathcal{N}(\mu, \sigma^2 \mathbf{I}_n)$
- ▶ $\mathcal{M}_0 \subset \mathcal{M}_1$ deux sous-espaces vectoriels de $\mathbb{R}^n$

On souhaite tester

$$H_0 : \mathbb{E}[\mathbb{Y}] = \mu \in \mathcal{M}_0 \quad \text{contre} \quad H_1 : \mathbb{E}[\mathbb{Y}] = \mu \in \mathcal{M}_1$$

Statistique. On considère $P_{\mathcal{M}_0}$ et $P_{\mathcal{M}_1}$ les projecteurs orthogonaux sur $\mathcal{M}_0$ et $\mathcal{M}_1$ et la statistique

$$F = \frac{\dim(\mathcal{M}_1^\perp) \|P_{\mathcal{M}_0} \mathbb{Y} - P_{\mathcal{M}_1} \mathbb{Y}\|^2}{(\dim(\mathcal{M}_1) - \dim(\mathcal{M}_0)) \|P_{\mathcal{M}_1^\perp} \mathbb{Y}\|^2}$$
$$F \underset{\text{sous } H_0}{\sim} \mathcal{F}(\dim(\mathcal{M}_1) - \dim(\mathcal{M}_0), \dim(\mathcal{M}_1^\perp))$$

Test. On rejette H_0 si $F > f_{1-\alpha, \dim(\mathcal{M}_1) - \dim(\mathcal{M}_0), \dim(\mathcal{M}_1^\perp)}$