

Régression linéaire multiple

Frédéric Bertrand et Myriam Maumy¹

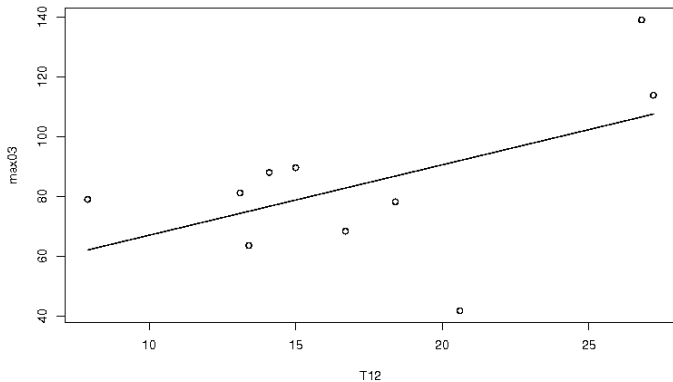
¹IRMA, Université de Strasbourg
Strasbourg, France

Magistère 2^e Année

- **Problème** : Étude de la concentration d'ozone dans l'air.
- **Modèle** : La température (v.a. X) et la concentration d'ozone (v.a. Y) sont liées de manière linéaire :

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

- **Observations** : $n = 10$ mesures de la température et de la concentration d'ozone.
- **But** : Estimer β_0 et β_1 afin de prédire la concentration d'ozone connaissant la température.



Souvent la régression linéaire est trop simpliste. Il faut alors utiliser d'autres modèles plus réalistes mais parfois plus complexes :

- Utiliser d'autres fonctions que les fonctions affines comme les fonctions polynômiales, exponentielles, logarithmiques. . .
- Considérer plusieurs variables explicatives.

Exemple : La température **et** la vitesse du vent

Le principe de la régression linéaire multiple est simple :

- Déterminer la variable expliquée Y .
Exemple : La concentration d'ozone.
- Déterminer $(p - 1)$ variables explicatives X_1, \dots, X_{p-1} .
Exemple : X_1 température, X_2 vitesse du vent. . .
- Il ne reste plus qu'à appliquer un modèle linéaire :

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_{p-1} X_{p-1} + \varepsilon$$

Dans un échantillon de n individus, on mesure $y_i, x_{i,1}, \dots, x_{i,p-1}$ pour $i = 1 \dots n$.

Observations	Y	X_1	\dots	X_{p-1}
1	y_1	$x_{1,1}$	\dots	$x_{1,p-1}$
2	y_2	$x_{2,1}$	\dots	$x_{2,p-1}$
\vdots	\vdots	\vdots	\vdots	\vdots
n	y_n	$x_{n,1}$	\dots	$x_{n,p-1}$

Remarque : Les variables $x_{i,j}$ sont fixes tandis que les variables y_i sont aléatoires.

But :

Estimer les paramètres $\beta_0, \dots, \beta_{p-1}$ du modèle de régression et ce de manière optimale.

Méthode : La méthode des moindres carrés. Cette méthode revient à minimiser la quantité suivante :

$$\sum_{i=1}^n \left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \dots + \hat{\beta}_{p-1} x_{i,p-1} \right) \right)^2 .$$

Le système peut se réécrire :

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Vecteur des résidus : $\boldsymbol{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$.

Remarque : Les variables \mathbf{y} et \mathbf{X} sont mesurées tandis que l'estimateur $\hat{\boldsymbol{\beta}}$ est à déterminer.

La méthode des moindres carrés consiste à trouver le vecteur $\hat{\boldsymbol{\beta}}$ qui minimise $\|\boldsymbol{\varepsilon}\|^2 = \boldsymbol{\varepsilon}^t \boldsymbol{\varepsilon}$.

$$\begin{aligned}\|\varepsilon\|^2 &= {}^t(\mathbf{y} - \mathbf{X}\beta)(\mathbf{y} - \mathbf{X}\beta) \\ &= {}^t\mathbf{y}\mathbf{y} - {}^t\hat{\beta}{}^t\mathbf{X}\mathbf{y} - {}^t\mathbf{y}\mathbf{X}\hat{\beta} + {}^t\hat{\beta}{}^t\mathbf{X}\mathbf{X}\hat{\beta} \\ &= {}^t\mathbf{y}\mathbf{y} - 2{}^t\hat{\beta}{}^t\mathbf{X}\mathbf{y} + {}^t\hat{\beta}{}^t\mathbf{X}\mathbf{X}\hat{\beta}\end{aligned}$$

car ${}^t\hat{\beta}{}^t\mathbf{X}\mathbf{y}$ est un scalaire. Donc il est égal à sa transposée.

La dérivée par rapport à $\hat{\beta}$ est alors égale à :

$$-2{}^t\mathbf{X}\mathbf{y} + 2{}^t\mathbf{X}\mathbf{X}\hat{\beta}.$$

- **Problème** : On cherche $\hat{\beta}$ qui annule cette dérivée. Donc on doit résoudre l'équation suivante :

$${}^t\mathbf{XX}\hat{\beta} = {}^t\mathbf{Xy}.$$

- **Solution** : On trouve après avoir inversé la matrice ${}^t\mathbf{XX}$ (il faut naturellement vérifier que ${}^t\mathbf{XX}$ est carrée et inversible c'est-à-dire qu'aucune des colonnes qui compose cette matrice ne soit proportionnelle aux autres colonnes)

$$\hat{\beta} = ({}^t\mathbf{XX})^{-1}{}^t\mathbf{Xy}.$$

Retrouvons les résultats de la régression linéaire simple ($p = 2$)

$${}^t\mathbf{X}\mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}; \quad {}^t\mathbf{X}\mathbf{y} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}.$$

Donc :

$$\begin{aligned} ({}^t\mathbf{X}\mathbf{X})^{-1} &= \frac{1}{n \sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \\ &= \frac{1}{\sum (x_i - \bar{x})^2} \begin{pmatrix} \sum x_i^2 / n & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}. \end{aligned}$$

Finalement on retrouve bien :

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \frac{\bar{y} \sum x_i^2 - \bar{x} \sum x_i y_i}{\sum (x_i - \bar{x})^2} \\ \frac{\sum x_i y_i - n \bar{x} \bar{y}}{\sum (x_i - \bar{x})^2} \end{pmatrix}$$

ce qui correspond aux estimateurs de la régression linéaire simple que nous avons déjà rencontrés dans le cours 1.

```
> a <- lm(max03 ~ T12 + VX)
> summary(a)
```

Call:

```
lm(formula = max03 ~ T12 + VX)
```

Residuals:

```
Min 1Q Median 3Q Max
-47.860 -10.561  5.119 10.645 26.506
```

Coefficients:

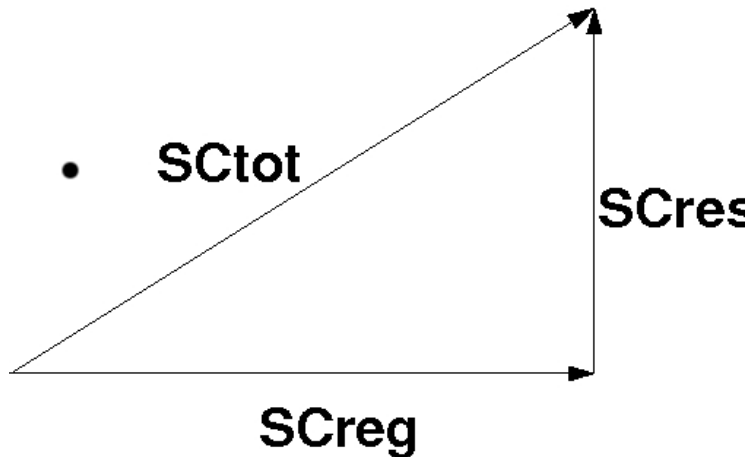
```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.6520 26.5324  1.381 0.210
T12  2.6623  1.4202  1.875 0.103
VX  0.5431  0.7775  0.699 0.507
Residual standard error: 24.78 on 7 degrees of freedom
Multiple R-Squared:  0.3351, Adjusted R-squared:  0.1452
F-statistic: 1.764 on 2 and 7 DF, p-value: 0.2396
```

Résultats préliminaires :

- $\sum \hat{y}_i^2 = \sum \hat{y}_i y_i$ ou (forme matricielle) ${}^t \hat{\mathbf{y}} \hat{\mathbf{y}} = {}^t \mathbf{y} \hat{\mathbf{y}}$
- $\sum \hat{y}_i = \sum y_i$

Propriété des moindres carrés :

$$\begin{aligned} \sum (y_i - \bar{y})^2 &= \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2 \\ \text{SC}_{tot} &= \text{SC}_{reg} + \text{SC}_{res} \end{aligned}$$



Le **coefficient de détermination** est défini par :

$$R^2 = \frac{SC_{reg}}{SC_{tot}}$$

Intuitivement ce coefficient de détermination quantifie la capacité du modèle à expliquer les variations de Y .

- Si R^2 est proche de 1 alors le modèle est proche de la réalité.
- Si R^2 est proche de 0 alors le modèle explique très mal la réalité. Il faut alors trouver un meilleur modèle.

On fait les hypothèses suivantes :

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

où le vecteur aléatoire ε suit une loi *multinormale* qui vérifie les hypothèses suivantes :

- $\mathbb{E}[\varepsilon] = 0$
- $\text{Var}[\varepsilon] = \sigma^2 \mathbf{I}_n$,

où σ^2 est la variance de la population et \mathbf{I}_n est la matrice identité de taille n .

Ceci implique que :

- $\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta$
- $\text{Var}[\mathbf{y}] = \sigma^2\mathbf{I}_n$.

On peut alors démontrer, **sous ces hypothèses** :

- $\mathbb{E}[\hat{\beta}] = \beta$. Ce qui signifie que $\hat{\beta}$ est un estimateur sans biais
- $\text{Var}[\hat{\beta}] = \sigma^2(\mathbf{t}\mathbf{X}\mathbf{X})^{-1}$.

Il reste un **problème** : Estimer la variance σ^2 qui est a priori une quantité inconnue.

Un estimateur sans biais de la variance σ^2 est défini par :

$$s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - p} = \frac{SC_{res}}{n - p} = \frac{SC_{tot} - SC_{reg}}{n - p},$$

où

- n est le nombre d'individus/d'observations,
- p est le nombre de variables explicatives.

On appelle la quantité $(n - p)$ **le nombre de degrés de liberté**.

But : Tester l'hypothèse nulle

$$\mathcal{H}_0 : \beta_j = b_j \quad \text{pour } j = 0, \dots, p - 1$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \beta_j \neq b_j \quad \text{pour } j = 0, \dots, p - 1.$$

Méthode :

- Calculer la statistique

$$t_{obs} = \frac{\hat{\beta}_j - b_j}{s(\hat{\beta}_j)}$$

où $s^2(\hat{\beta}_j)$ est l'élément diagonal d'indice j de $s^2(t\mathbf{XX})^{-1}$.

- Si l'hypothèse nulle \mathcal{H}_0 est vraie, alors t_{obs} suit une loi de Student avec $(n - p)$ degrés de liberté.

- Valeur critique : $t_{(\alpha/2, n-p)}$ le $(1 - \alpha/2)$ -quantile d'une loi de Student avec $(n - p)$ degrés de liberté (cf table de la loi de Student).
- On rejette l'hypothèse nulle \mathcal{H}_0 si $|t_{obs}| \geq t_{(\alpha/2, n-p)}$.

Cas particulier : Tester si « $\beta_j = 0$ » pour un certain j .
Si l'hypothèse nulle \mathcal{H}_0 : « $\beta_j = 0$ » est acceptable alors la variable X_j n'est pas significative au sein du modèle. On peut simplifier le modèle, ...et recommencer !