Régression Linéaire

L3 FLEX - Machine Learning
Responsable UE : Romain Thoreau

Université Paul Sabatier - 2024-2025

Ce polycopié est fourni en supplément des notes manuscrites données en cours. Il ne peut se substituer aux notes de cours, et ne reprend pas nécessairement l'intégralité de leur contenu. Par ailleurs, il sera peut-être enrichi et mis à jour au cours de l'UE.

Plan

1 Régression

La régression est une méthode pour étudier la relation entre deux variables aléatoires X et Y. Généralement, on cherche à expliquer Y à partir de X. En d'autres termes, la variable Y est la variable d'intérêt : sachant X, on souhaite prédire Y. On appelle souvent Y la variable à expliquer, et X la variable explicative. L'objectif de la régression est d'estimer l'espérance de Y sachant X, notée $\mathbb{E}(Y|X)$, à partir de données, souvent appelées observations, c'est-à-dire des réalisations des variables

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

indépendantes et identiquement distribuées selon la loi jointe de X et Y.

1.1 Exemple introductif

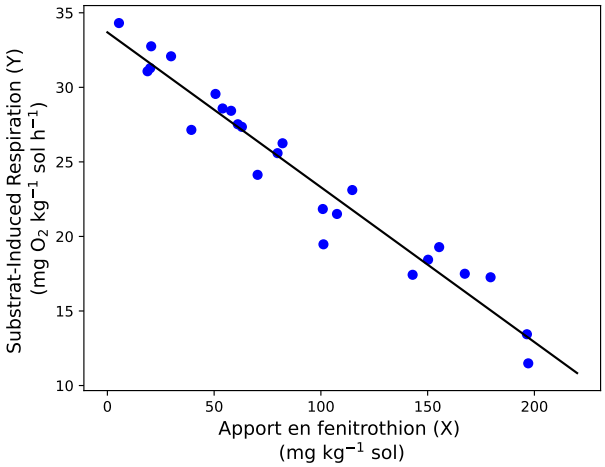

Figure 1: Mesures de *substrat-induced respiration* en fonction de l'apport en *fenitrothion*. La droite représente la fonction de régression estimée à partir des données.

Un laboratoire de recherche s'intéresse au lien entre l'usage de pesticides et la diversité bactérienne des sols agricoles en France. Précisément, elle étudie la relation entre la *substrat-induced respiration* (la variable à expliquer Y), une grandeur physique, exprimée en mg d'O_2 par kg de sol par heure, qui caractérise la diversité bactérienne des sols, et l'apport moyen en pesticide *fenitrothion* (la variable explicative X). Comme il n'existe pas de modèle théorique qui exprime la dépendance entre ces deux variables, l'équipe de recherche procède à une étude statistique. Pour cela, elle réalise une série de mesures sur plusieurs parcelles agricoles en France, illustrées sur la Fig. 1. A partir de ces données, une fonction de régression, également représentée sur la Fig. 1, est estimée.

2 Modèle linéaire simple

2.1 Définition

Le modèle linéaire simple considère deux variables aléatoire réelles X et Y, liées par une relation affine :

$$\mathbb{E}[Y|X] = \beta_0 + \beta_1 X \tag{1}$$

où $\beta_0, \beta_1 \in \mathbb{R}$ sont les paramètres du modèle.

Hypothèses de modélisation On suppose par ailleurs, pour la i^{eme} variable observée, que

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{2}$$

où ϵ_i est un terme d'erreur aléatoire et non observé tel que :

- **[H1]** Les erreurs sont centrées : $\mathbb{E}[\epsilon_i] = 0$,
- **[H2]** La variance des erreurs est constante : $\mathbb{V}[\epsilon_i] = \sigma^2$,
- **[H3]** Les erreurs sont non corrélées : $\forall i \neq j, \mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$.

Remarque : *En toute rigueur, on devrait noter l'espérance et la variance des erreurs ϵ_i conditionnellement à X_i, e.g. $\mathbb{E}[\epsilon_i|X_i] = 0$, à moins de considérer les X_i comme des variables non aléatoires.*

2.2 Optimisation

L'optimisation du modèle consiste à estimer les paramètres du modèle : l'ordonnée à l'origine β_0, le coefficient directeur β_1 et la variance σ^2.

Notations On note x_i et y_i les réalisations des variables aléatoires X_i et Y_i, respectivement.

Hypothèses sur les données Pour l'optimisation, on suppose que $n > 2$ et qu'il existe $i \neq j$ tels que $x_i \neq x_j$.

2.2.1 Méthode des moindres carrés

La méthode des moindres carrés consiste à estimer β_0 et β_1 en minimisant la somme des erreurs au carré :

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0,\beta_1}{\arg\min} \sum_{i=1}^{n} \big(y_i - (\beta_0 + \beta_1 x_i)\big)^2$$

$\hat{\beta}_0$ et $\hat{\beta}_1$ désignent les estimateurs des paramètres du modèle β_0 et β_1, respectivement. Les prédictions sont désignées par $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, et on définit les résidus par $\hat{\epsilon}_i = y_i - \hat{y}_i$.

Théorème 1 (Estimateurs des moindres carrés)**.** *Les estimateurs des moindres carrés sont*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{3}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{4}$$

où $\bar{x}$ et $\bar{y}$ désignent la moyenne des x_i et des y_i, respectivement. Un estimateur non biaisé de σ^2 est

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

Sous les hypothèses H1, H2 et H3, les estimateurs des moindres carrés sont les meilleurs estimateurs linéaires non biaisés (théorème de Gauss-Markov). Les estimateurs sont non biaisés dans le sens où $\mathbb{E}[\hat{\beta}_0] = \beta_0$ et $\mathbb{E}[\hat{\beta}_1] = \beta_1$. De plus, ce sont les meilleurs dans le sens où leur variance est minimale.

2.2.2 Méthode du maximum de vraisemblance

On fait l'hypothèse supplémentaire **[H4]** que les erreurs sont distribuées selon une loi normale : $\epsilon_i | X_i \sim \mathcal{N}(0, \sigma^2)$; ce qui implique que :

$$Y_i | X_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$$

Définition 1 (Vraisemblance)**.** *La vraisemblance du modèle linéaire, notée $L(\beta_0, \beta_1, \sigma^2)$, est la densité jointe des (X_i, Y_i) évaluée en (x_i, y_i) :*

$$L(\beta_0, \beta_1, \sigma^2) = f_{(X_1,Y_1),\dots,(X_n,Y_n)}\big((x_1, y_1), \dots, (x_n, y_n)\big)$$

La vraisemblance du modèle linéaire traduit à quel point il est *vraisemblable* d'observer les données $\{(x_1, y_1), \dots, (x_n, y_n)\}$ étant donné que le modèle décrit la dépendance de Y à X. La méthode du maximum de vraisemblance consiste donc à estimer les paramètres du modèle en maximisant la vraisemblance :

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2 = \underset{\beta_0,\beta_1,\sigma}{\arg\max}\, L(\beta_0, \beta_1, \sigma^2)$$

Puisque les (X_i, Y_i) sont indépendants et identiquement distribués, la vraisemblance peut s'écrire comme suit :

$$\begin{aligned} L(\beta_0, \beta_1, \sigma^2) &= \prod_{i=1}^{n} f_{X,Y}(x_i, y_i) \\ &= \prod_{i=1}^{n} f_X(x_i) f_{Y|X}(y_i|x_i) \\ &= \underbrace{\prod_{i=1}^{n} f_X(x_i)}_{=L_1} \underbrace{\prod_{i=1}^{n} f_{Y|X}(y_i|x_i)}_{=L_2} \end{aligned}$$

L_1 est la vraisemblance marginale des X_i, et ne dépend pas des paramètres du modèle. Ainsi, maximiser L est équivalent à maximiser L_2, la vraisemblance conditionnelle, qui est le produit des densités conditionnelles (par hypothèse, des lois normales $\mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2)$) :

$$\begin{aligned}
\underset{\beta_0,\beta_1,\sigma}{\arg\max}\, L(\beta_0,\beta_1,\sigma^2) &= \underset{\beta_0,\beta_1,\sigma}{\arg\max}\, L_2(\beta_0,\beta_1,\sigma^2) \\
&= \underset{\beta_0,\beta_1,\sigma}{\arg\max} \prod_{i=1}^{n} f_{Y|X}(y_i|x_i) \\
&= \underset{\beta_0,\beta_1,\sigma}{\arg\max} \log\Big[\prod_{i=1}^{n} f_{Y|X}(y_i|x_i)\Big] \\
&= \underset{\beta_0,\beta_1,\sigma}{\arg\max} \sum_{i=1}^{n} \log\, f_{Y|X}(y_i|x_i) \\
&= \underset{\beta_0,\beta_1,\sigma}{\arg\max} \sum_{i=1}^{n} \log\Big[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2}\frac{\big(y_i - (\beta_0+\beta_1 x_i)\big)^2}{\sigma^2}\Big)\Big] \\
&= \underset{\beta_0,\beta_1,\sigma}{\arg\max} -\frac{n}{2}\log\, 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(y_i - (\beta_0+\beta_1 x_i)\big)^2
\end{aligned}$$

On constate que maximiser la vraisemblance par rapport à β_0 et β_1 est équivalent à minimiser la somme des erreurs au carré :

$$\underset{\beta_0,\beta_1}{\arg\max}\, L(\beta_0,\beta_1,\sigma^2) = \underset{\beta_0,\beta_1}{\arg\max} -\sum_{i=1}^{n}\big(y_i - (\beta_0+\beta_1 x_i)\big)^2 = \underset{\beta_0,\beta_1}{\arg\min} \sum_{i=1}^{n}\big(y_i - (\beta_0+\beta_1 x_i)\big)^2$$

Théorème 2 (Estimateurs du maximum de vraisemblance)**.** *Sous l'hypothèse **[H4]** de normalité, les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ du maximum de vraisemblance sont aussi les estimateurs des moindres carrés. De plus, l'estimateur $\hat{\sigma}^2$ qui maximise la vraisemblance $L(\beta_0,\beta_1,\sigma^2)$ par rapport à σ^2 est :*

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

C'est un estimateur biaisé de σ^2.

3 Modèle linéaire multiple

3.1 Définition

Le modèle linéaire multiple généralise le modèle linéaire simple en considérant une variable explicative X de dimension $d > 1$:

$$X_i = (X_{i1}, \ldots, X_{id})$$

On préfère alors considérer la forme matricielle du modèle de régression :

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

où :

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix},\ \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1d} \\ 1 & X_{21} & \ldots & X_{2d} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \ldots & X_{nd} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_d \end{bmatrix} \text{ et } \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Hypothèses de modélisation On fait, un peu près, les mêmes hypothèses que pour le modèle linéaire simple :

- **[H1′]** Les erreurs sont centrées : $\mathbb{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}_n$,
- **[H2′]** La variance des erreurs est constante et les erreurs sont non corrélées : $\mathrm{Cov}[\boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{I}_n$.

3.2 Optimisation

L'optimisation du modèle linéaire multiple est équivalent à l'optimisation du modèle linéaire simple, par la méthode des moindres carrés ou par la méthode du maximum de vraisemblance.

Hypothèses sur les données Pour l'optimisation, on suppose que 1) $n > d+1$ et 2) que les colonnes de $\boldsymbol{X}$ sont linéairement indépendantes, c'est-à-dire que le rang de $\boldsymbol{X}$ est égal à $d+1$. L'hypothèse 2) garantie que la matrice $\boldsymbol{X}^T\boldsymbol{X}$ est inversible. Ces hypothèses sont une généralisation des hypothèses faites pour la régression simple.

Théorème 3 (Estimateurs des moindres carrés du modèle linéaire multiple)**.** *En supposant que la matrice $\boldsymbol{X}^T\boldsymbol{X}$ est inversible, l'estimateur des moindres carrés du modèle linéaire multiple est*

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{5}$$

Un estimateur non biaisé de σ^2 est

$$\hat{\sigma}^2 = \frac{\|\hat{\boldsymbol{\epsilon}}\|_2^2}{n-(d+1)}.$$

3.3 Inférence

Il faut garder à l'esprit que les estimateurs des paramètres du modèle sont des variables aléatoires. Ce sont des estimateurs non biaisés (pour la variance, seulement pour l'estimateur des moindres carrés) :

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} \quad \text{et} \quad \mathbb{E}[\hat{\sigma}^2] = \sigma^2.$$

et la covariance de $\boldsymbol{\beta}$ est

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Sous l'hypothèse supplémentaire **[H4′]** de normalité, on a :

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

3.3.1 Inférence sur les coefficients

On peut dériver des statistiques intéressantes pour valider la significativité statistique des coefficients β_i du modèle. En particulier, il résulte de l'hypothèse de normalité que la statistique

$$T_n = \frac{\hat{\beta}_i - \beta_i}{\sigma_{\hat{\beta}_i}} \sim St(n-d-1). \tag{6}$$

où $St(k)$ désigne la loi de Student à k degrés de liberté, et $\sigma_{\hat{\beta}_i}$ désigne le $(i+1)$-eme élément de la diagonale de $\hat{\sigma}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Test d'hypothèse sur les coefficients β_i On peut utiliser la statistique T_n de l'équation 6 pour réaliser un test d'hypothèse sur les coefficients β_i. On formule l'hypothèse nulle, c'est-à-dire l'hypothèse par défaut, notée H_0 :

$$H_0 : \beta_i = b$$

Par exemple, dans l'exemple de la section 1.1, l'hypothèse nulle serait $H_0 : \beta_1 = 0$, c'est-à-dire, "l'apport en pesticide n'a pas d'effet sur la diversité bactérienne des sols". On parle d'hypothèse par défaut, ou d'hypothèse conservatrice, car elle pourrait avoir des conséquences si elle était réfutée. L'hypothèse alternative, notée H_1, est :

$$H_1 : \beta_i \neq b$$

Le test d'hypothèse consiste à rejeter l'hypothèse nulle avec un risque α, $0 < \alpha < 1$, de se tromper si la réalisation de la statistique T_n, notée t, est *invraisemblable* sous l'hypothèse H_0. Précisément, si H_0 était vraie, alors la probabilité que la valeur absolue de la variable aléatoire T_n soit supérieure à $Q^{St}_{n-d-1}(1-\frac{\alpha}{2})$ est égale à α, où $Q^{St}_{n-d-1}(1-\frac{\alpha}{2})$ désigne le quantile de niveau $1-\frac{\alpha}{2}$ de la loi de Student à $(n-d-1)$ degrés de liberté. Une illustration graphique est donnée dans la Fig. 2. Ainsi, on rejette l'hypothèse nulle avec un risque α si

$$|t| > Q^{St}_{n-d-1}(1-\frac{\alpha}{2})$$

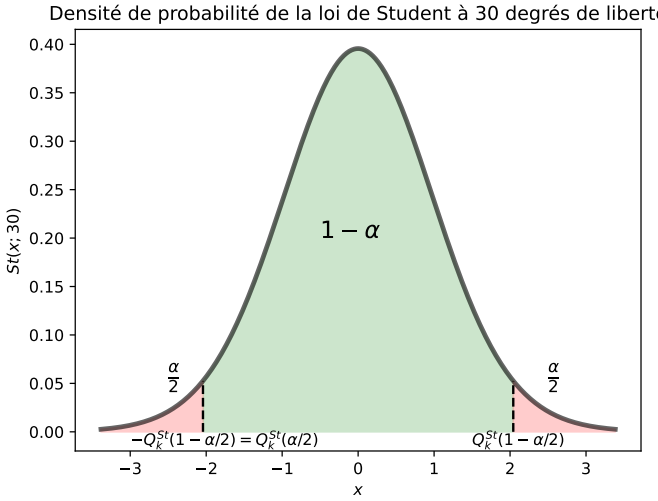

Figure 2: Densité de probabilité de la loi de Student à 30 degrés de liberté. L'aire sous la courbe illustrée en vert est égale à $1-\alpha$, tandis que l'aire sous la courbe illustrée en rouge est égale à α ($\frac{\alpha}{2}$ de chaque côté). Ici, α est égale à 0.05. Les traits pointillés verticaux sont placés aux quantiles de niveau $\alpha/2$ et $1-\alpha/2$, notés $Q_k^{St}(\alpha/2)$ et $Q_k^{St}(1-\alpha/2)$, respectivement. Du fait de la symétrie de la densité, on a : $Q_k^{St}(\alpha/2) = -Q_k^{St}(1-\alpha/2)$.

Dans le cas contraire, on conserve l'hypothèse nulle.

Intervalle de confiance sur les coefficients β_i On peut également utiliser la statistique T_n pour construire des intervalles de confiance. La probabilité que la valeur absolue de la statistique T_n soit inférieure à $Q_k^{St}(1-\alpha/2)$ est égale à $1-\alpha$:

$$\mathbb{P}\big(-Q_k^{St}(1-\alpha/2) \leq T_n \leq Q_k^{St}(1-\alpha/2)\big) = 1-\alpha$$

ce qui est équivalent à

$$\mathbb{P}\big(\hat{\beta}_i - \sigma_{\hat{\beta}_i} Q_k^{St}(1-\alpha/2) \leq \beta_i \leq \hat{\beta}_i + \sigma_{\hat{\beta}_i} Q_k^{St}(1-\alpha/2)\big) = 1-\alpha$$

Ainsi, un intervalle de confiance de β_i de niveau $(1-\alpha)$ est :

$$\mathrm{IC}_{(1-\alpha)} = \big[\hat{\beta}_i - \sigma_{\hat{\beta}_i} Q_k^{St}(1-\alpha/2), \hat{\beta}_i + \sigma_{\hat{\beta}_i} Q_k^{St}(1-\alpha/2)\big] \tag{7}$$

Si on n'avait pas un échantillon statistique, mais une infinité d'échantillons statistiques pour l'optimisation du modèle linéaire, on obtiendrait une infinité d'estimateurs $\hat{\beta}_i$, qui donneraient une infinité d'intervalles de confiance. Alors, il y aurait une fraction $(1-\alpha)$ de ces intervalles de confiance qui contiendraient le paramètre du modèle β_i.