ENSAI - 2A — Année 2024-2025

Chargés de TD/TP :

Julien Jamme — julien.jamme@insee.fr

Théo Leroy — theo.leroy@insee.fr

Yves Ngounou — yves.ngounou@ensai.fr

Clotilde Patarin — clotilde.patarin@square-management.com

Responsable du cours :

Frédéric Lavancier — frederic.lavancier@ensai.fr

Régression linéaire et généralisé - TD 1

Exercice 1. *Moyenne empirique*

Soit $z_1, ..., z_n$ des observations d'une variable Z.

1. Déterminer la valeur $\hat{m}$ qui minimise la distance quadratique aux données $S(m) = \sum_{i=1}^{n}(z_i - m)^2$.

2. La quantité $\hat{m}$ correspond en fait à l'estimation par moindres carrés ordinaires dans un modèle de régression linéaire : $Y = X\beta + \epsilon$. Préciser ce que valent Y, X, β et ϵ.

3. Retrouver le résultat de la première question à partir de la formule générale de l'estimateur des moindres carrés : $\hat{\beta} = (X'X)^{-1}X'Y$.

Exercice 2. *Reconnaître un modèle de régression linéaire*

Les modèles suivants sont-ils des modèles de régression linéaire? Si non, peut-on appliquer une transformation pour s'y ramener? Pour chaque modèle de régression linéaire du type $Y = X\beta + \epsilon$, on précisera ce que valent Y, X, β et ϵ.

1. On observe $(x_i, y_i), i = 1, ..., n$ liés théoriquement par la relation $y_i = a + bx_i + \epsilon_i,\ i = 1, ..., n$, où les variables ϵ_i sont centrées, de variance σ^2 et non-corrélées. On désire estimer a et b.

2. On observe $(x_i, y_i), i = 1, ..., n$ liés théoriquement par la relation $y_i = a_1 x_i + a_2 x_i^2 + \epsilon_i,\ i = 1, ..., n$, où les variables ϵ_i sont centrées, de variance σ^2 et non-corrélées. On désire estimer a_1 et a_2.

3. On relève pour différents pays ($i = 1, ..., n$) leur production P_i, leur capital K_i, leur facteur travail T_i qui sont théoriquement liées par la relation de Cobb-Douglas $P = \alpha_1 K^{\alpha_2} T^{\alpha_3}$. On désire vérifier cette relation et estimer α_1, α_2 et α_3.

4. Le taux de produit actif y dans un médicament est supposé évoluer au cours du temps t selon la relation $y = \beta_1 e^{-\beta_2 t}$. On dispose des mesures de n taux y_i effectués à n instants t_i. On désire vérifier cette relation et estimer β_1 et β_2.

5. Même problème que précedemment mais le modèle théorique entre les observations s'écrit $y_i = \beta_1 e^{-\beta_2 t_i} + u_i$, $i = 1, \ldots, n$, où les variables u_i sont centrées, de variance σ^2 et non-corrélées.

Exercice 3. *Régression simple*
On considère le modèle de régression linéaire simple où l'on observe n réalisations $(x_i, y_i), i = 1, ..., n$ liées par la relation $y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, ..., n$. On suppose que les x_i sont déterministes et que les variables ϵ_i sont centrées, de variance σ^2 et non-corrélées entre elles.

1. Ecrire le modèle sous forme matricielle.

2. De quel problème de minimisation l'estimateur des moindres carrés $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ est-il la solution?

3. On peut trouver $\hat{\beta}$ en annulant le gradient de la fonction à minimiser précédente. Cela a déjà été fait et les solutions sont à connaitre par coeur : que valent-elles ?

4. Retrouver $\hat{\beta}$ en utilisant la formule générale $\hat{\beta} = (X'X)^{-1}X'Y$.

5. Justifier pourquoi la droite de régression passe nécessairement par le point $(\bar{x}_n, \bar{y}_n)$.

6. On souhaite prédire la valeur y_o associée à la valeur x_o d'un nouvel individu, en supposant que ce dernier suit exactement le même modèle que les n individus précédents. Que vaut la prévision $\hat{y}_o$ de y_o ?

7. Montrer que l'espérance de l'erreur de prévision $y_o - \hat{y}_o$ est nulle.

8. Pour un modèle de régression linéaire générale, la variance de l'erreur de prévision associée à un nouveau vecteur de régresseur x, de dimension p, vaut (cf cours) : $\sigma^2 \left(x'(X'X)^{-1}x + 1\right)$. Montrer qu'ici cette variance peut se récrire :
$$\sigma^2 \left(1 + \frac{1}{n} + \frac{(x_o - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2}\right).$$

9. Discuter de la qualité de la prévision selon que x_o est proche ou non de la moyenne empirique $\bar{x}_n$.

10. Qu'en est-il si n est grand ?

Exercice 4. *D'autres petites questions sur la régression linéaire simple*
On considère le modèle de régression linéaire simple où l'on observe n réalisations $(x_i, y_i), i = 1, ..., n$ liées par la relation théorique $y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, ..., n$.

1. Quelle sont les hypothèses standards sur les erreurs de modélisation ϵ_i ?
2. Sous quelle hypothèse le modèle est-il identifiable, au sens où β_0 et β_1 sont définis de manière unique ?
3. Sous quelle hypothèse l'estimateur par MCO de β_0 et β_1 existe-t-il ?
4. Les variables y_i ont-elles même espérance ?
5. La droite de régression estimée à partir des observations (x_i, y_i) passe-t-elle toujours par le point $(\bar{x}_n, \bar{y}_n)$?
6. Les estimateurs par MCO des coefficients β_0 et β_1 sont-ils indépendants ?
7. Est-il possible de trouver des estimateurs des coefficients de régression de plus faible variance que celle des estimateurs par MCO ?

Exercice 5. *Convergence des estimateurs*
On se place comme dans l'exercice précédent dans le cadre d'un modèle de régression simple. On rappelle que la matrice de design X et la matrice $(X'X)^{-1}$ valent dans ce cas :

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad (X'X)^{-1} = \frac{1}{\sum_{i=1}^n (x_i - \overline{x}_n)^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\overline{x}_n \\ -\overline{x}_n & 1 \end{pmatrix}.$$

On va examiner quelques exemples de design, c'est à dire de répartition des valeurs de $x_1, \ldots, x_n$, et vérifier la convergence (ou non) des estimateurs par MCO des paramètres β_0 et β_1 dans chaque cas.

1. Rappeler ce que vaut l'erreur quadratique moyenne de $\hat{\beta}$, l'estimateur par MCO de $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$.
2. Dans ce premier exemple, on se place dans le cas où les observations ont lieu de façon régulièrement espacées, et deviennent de plus en plus nombreuses avec n. Quitte à renormaliser, on suppose ainsi que $x_i = i$ pour tout $i = 1, \ldots, n$.

a) Donner la limite de la matrice $\mathbb{V}(\hat{\beta})$ lorsque $n \to \infty$.

b) En déduire le comportement asymptotique en moyenne quadratique de $\hat{\beta}_0$ et $\hat{\beta}_1$.

3. Même question lorsque les observations deviennent de plus en plus dense dans un intervalle (pour simplifier : l'intervalle $[0, 1]$). On suppose ainsi que $x_i = i/n$ pour tout $i = 1, \ldots, n$.

4. On se place ici dans un cas où les observations sont mal dispersées : on suppose que $x_i = 1/i$ pour tout $i = 1, \ldots, n$. Ainsi les observations se concentrent en 0. Qu'en est-il du comportement asymptotique en moyenne quadratique de $\hat{\beta}_0$ et $\hat{\beta}_1$?

5. Dans les exemples précédents, les x_i étaient déterministes. On suppose ici que les x_i sont aléatoires, i.i.d, de carré intégrable et de variance non nulle. On suppose également que les x_i et les erreurs de modélisation ϵ_i sont indépendants. Cette situation peut être vue comme l'équivalent aléatoire des situations déterministes traitées dans les questions 2 et 3 (selon que la loi des x_i est discrète ou continue).

a) Exprimer $\hat{\beta} - \beta$ en fonction de la matrice X et du vecteur ϵ.

b) En déduire que $\hat{\beta}$ converge presque sûrement vers β lorsque $n \to \infty$.

Exercice 6. *Consommation de confiseries*
Des données, publiées par le Chicago Tribune en 1993, montrent la consommation de confiseries en millions de livres et la population en millions d'habitants dans 17 pays, en 1991. On note y_i la consommation et x_i la population du i-ème pays, $i = 1, \ldots, 17$. On donne les valeurs suivantes :

$$\sum_{i=1}^{17} x_i = 751.8, \quad \sum_{i=1}^{17} y_i = 13683.8,$$
$$\sum_{i=1}^{17} x_i^2 = 97913.92, \quad \sum_{i=1}^{17} y_i^2 = 36404096.44,$$
$$\sum_{i=1}^{17} x_i y_i = 1798166.66.$$

On désire lier à l'aide d'un modèle de régression linéaire (avec constante) la consommation de confiseries en fonction de la population de chaque pays.

1. Ecrire l'équation du modèle envisagé, pour chaque pays, en précisant les hypothèses effectuées.

$\alpha \backslash q$	14	15	16	17	18
0.01	2.62	2.60	2.58	2.57	2.55
0.025	2.14	2.13	2.12	2.11	2.10
0.05	1.76	1.75	1.75	1.74	1.73
0.10	1.35	1.34	1.34	1.33	1.33

Table 1: Quantiles d'ordre $1-\alpha$ d'une loi de Student à q degrés de liberté, pour différentes valeurs de α et de q.

2. Donner les expressions des estimateurs par MCO de la pente et de l'ordonnée à l'origine du modèle, en fonction des sommes données ci-dessus. En déduire leurs valeurs.

3. Donner l'expression d'un estimateur sans biais de la variance de l'erreur de modélisation, en fonction des sommes données ci-dessus. En déduire sa valeur.

4. Que vaut la variance théorique des estimateurs par MCO ? Comment l'estimer ? En déduire une estimation de l'écart-type de chaque estimateur (on pourra s'appuyer sur l'expression de $(X'X)^{-1}$ rappelée dans l'exercice 5).

5. Tester si la pente de la régression est significativement différente de 0, en rappelant les hypothèses sous jacentes. Pour l'application numérique, on fera un test au niveau 5% en s'appuyant sur les quantiles donnés dans la table 1.

6. Donner l'expression de la p-valeur du test précédent. On ne demande pas d'effectuer l'application numérique, mais au moins d'en donner une valeur approximative.

7. Tester de même si l'ordonnée à l'origine est significativement différente de 0, d'une part en fixant le niveau à 5% et d'autre part en évaluant grossièrement la p-valeur associée.