Guide pour une leçon de modélisation stochastique. Modèle linéaire et modèle linéaire gaussien : calculs par moindres carrés et par maximum de vraisemblance. Applications

Sana Louhichi

De l'observation à la théorie..
De la pomme tombée à la loi de la gravitation universelle..

Guide pour une leçon de modélisation stochastique.

Modèle linéaire et modèle linéaire gaussien : calculs par moindres carrés et par maximum de vraisemblance. Applications.

Sana Louhichi,
E.mail : sana.louhichi@univ-grenoble-alpes.fr
Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK Grenoble, France.

Table des matières

Avant-propos **1**

1 Mémento **3**
1.1 Modèle linéaire . . . 3
1.1.1 Mots clés . . . 3
1.1.2 Synthèse . . . 3
1.2 Qualité du modèle . . . 9
1.2.1 Mots clés . . . 9
1.2.2 Synthèse . . . 9
1.3 Modèle linéaire gaussien . . . 10
1.3.1 Mots clés . . . 10
1.3.2 Synthèse . . . 10
1.4 Vérification graphique des hypothèses de modélisation . . . 13
1.4.1 Mots clés . . . 13
1.4.2 Synthèse . . . 13
1.4.3 Illustrations Graphiques . . . 14

2 Problématiques **19**
2.1 Problématique I . . . 19
2.2 Problématique II . . . 21

BIBLIOGRAPHIE **23**

Index **23**

Table des figures **25**

Notations **27**

Avant-propos

Ce manuscrit est sous la forme d'un guide plutôt que d'un cours détaillé classique. Il s'adresse à toute personne ayant une formation en mathématiques et souhaitant approfondir ses connaissances en probabilités et statistique en rapport avec la modélisation.

Dans le même esprit que [5], ce guide se présente sous la forme d'un mémento qui rappelle les résultats théoriques utiles, et parfois indispensables, à maîtriser pour la leçon de modélisation en question. Les résultats sont énoncés sous la forme d'un résumé et donc sans démonstration mais souvent accompagnés par des illustrations graphiques. L'auteur intéressé trouvera sans doute les démonstrations des résultats de son intérêt. Des mots clés sont détaillés pour chaque section du mémento, orientant ainsi vers le cœur du sujet. Des questions et des exercices sont posés au fur et à mesure de la progression du texte. Le mémento sera utile pour la résolution des problèmes de modélisation du second chapitre.

Ce présent guide concerne un problème de régression et plus précisément un problème de régression linéaire. On observe sur n individus les vecteurs $(x_i, y_i)_{1\leq i\leq n}$ avec y_i des mesures quantitatives. Par exemple :

- n étant le nombre de disques, x_i est le rayon du ième disque et y_i son périmètre. L'objectif est d'étudier expérimentalement la relation mathématique,
$$\text{Périmètre d'un disque} = 2\pi\cdot(\text{son rayon}).$$
- n étant le nombre de voitures, x_i la vitesse de la ième voiture, y_i sa distance de freinage. Étudier la relation entre ces deux grandeurs permet de comprendre comment la vitesse influe sur la distance de freinage et aussi de pouvoir prédire une distance de freinage pour une vitesse donnée, ce qui permettra de donner des consignes pratiques aux conducteurs (voir Problématique I). La modélisation linéaire suppose que cette relation est linéaire :
$$\text{La distance de freinage} \sim \text{b}+\ \beta\ \cdot\ \text{la vitesse},$$
($\sim$ pour dire à-peu-prés), il convient donc d'écrire ce modèle et d'estimer entre autres, en se basant sur les observations $(x_i, y_i)_i$, les paramètres b et β.
- n étant le nombre de patients. y_i taux de de glycémie à jeun dans le sang du ième patient, x_i des mesures numériques en rapport avec ce patient, par exemple x_i est le vecteur (âge, poid, nombre d'heures de sports pratiqués par jour, mesure de l'hypertension artérielle) $= (x_i^{(1)}, x_i^{(2)}, x_i^{(3)}, x_i^{(4)})$. Étudier la relation entre y_i et x_i permet de la comprendre, de prédire et de donner des conseils préventifs. Supposer que la relation est linéaire, revient à dire :
$$\text{taux de de glycémie} \sim \text{b} + \beta_1\cdot\text{âge} + \beta_2\cdot\text{poid}+ \beta_3\cdot(\text{nombre d'heures de sports pratiqués par jour})+ \beta_4\cdot\text{hypertension artérielle}.$$
Là aussi, il convient d'écrire ce modèle et d'estimer entre autres, en se basant sur les observations $(x_i, y_i)_i$, les paramètres b et $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$.

Aprés avoir observé et analysé descriptivement les données, l'étape serait de poser un modèle statistique i.e. une formulation mathématique permettant d'expliquer les observations et de décrire le plus possible la réalité. L'étude du modèle nécessite souvent des hypothèses de modélisation i.e. des conditions et des suppositions utiles pour étudier le modèle. Les conclusions qu'on peut tirer du modèle dépendant de ces hypothèses de modélisation. Il convient donc d'étudier la validité de ces hypothèses. Les simulations

ainsi que les outils graphiques permettent d'orienter vers la validité ou non d'une ou de telles hypothèses de modélisation. C'est l'objectif de la section 1.4.

Les références, à ce sujet, sont nombreuses. Une liste non exhaustive est donnée à la fin de ce manuscrit.

Chapitre 1

Mémento

1.1 Modèle linéaire

1.1.1 Mots clés

Fonction de régression, modèle linéaire simple, modèle linéaire multiple, variables explicatives, variable à expliquer, bruit, estimation des paramètres par la méthode des moindres carrées, droite des moindres carrés, estimateurs des moindres carrés, estimateur BLUE (Best Linear Unbiased Estimator), résidus.

1.1.2 Synthèse

Soient $(X_i, Y_i)_{1\leq i\leq n}$ des v.a. i.i.d. de même loi que le couple (X, Y) à valeurs dans $\mathbb{R}^p \times \mathbb{R}$, $p \in \mathbb{N} \setminus \{0\}$.

On cherche à modéliser la relation entre X et Y afin de comprendre et de prédire.

— Y est la variable à expliquer ou la variable réponse

— Les composantes du vecteur X sont les variables explicatives ou prédicteurs.

Un modèle linéaire est un modèle statistique paramétrique qui suppose que la relation entre Y et X est linéaire. Afin de mieux s'approcher de la réalité, cette relation linéaire est supposée imparfaite i.e. on ajoute à cette relation un « bruit »

Définition 1.1. *On appelle modèle linéaire liant X à Y, le modèle*

$$Y = \beta^t X + b + \epsilon,$$

avec $\mathbb{E}(\epsilon^2) < \infty$,

$$\mathbb{E}(\epsilon|X) = 0, \;\; Var(\epsilon|X) = \sigma^2, \;\; ps.^{1}$$

Le modèle linéaire est dit simple si $p = 1$ sinon il est dit multiple.

1. Noter que $\mathbb{E}(\epsilon|X)$ et $Var(\epsilon|X)$ sont deux v.a. $\sigma(X)$-mesurable.

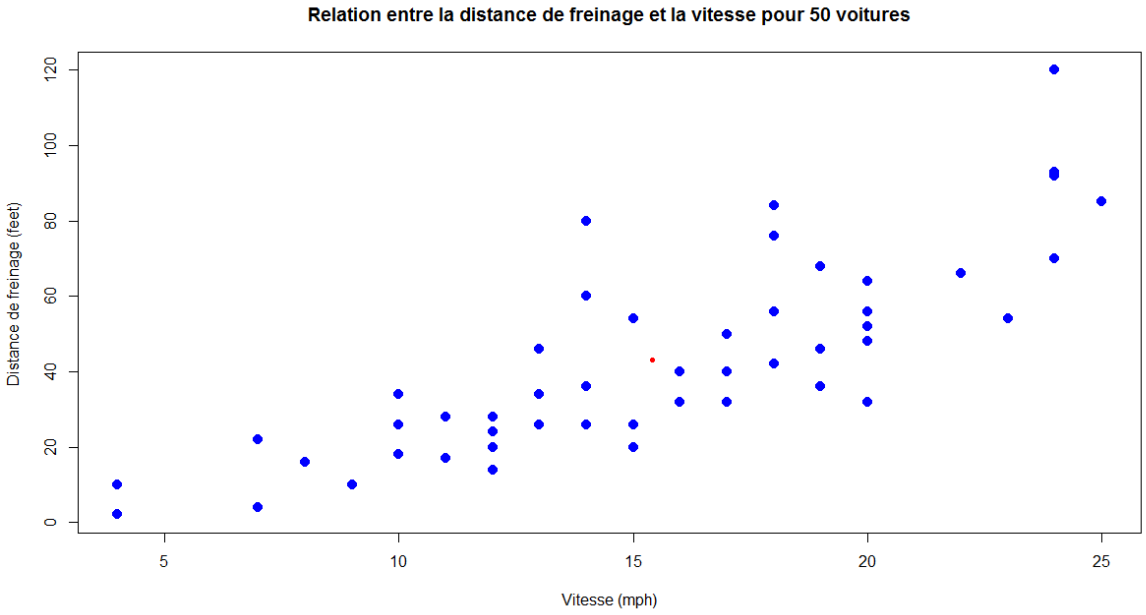

FIGURE 1.1 – Nuage de points et point moyen du nuage. La relation semble être linéaire. On peut la modéliser par un modèle linéaire simple (ici $p = 1$) et étudier ensuite les pertinences de ce modèle.

Remarque. La régression linéaire se caractérise par des variables explicatives quantitatives. L'ANOVA, quant à elle, concerne des variables explicatives qualitatives.

Notons que,

- (X, Y) est de loi inconnue mais on dispose d'un échantillon aléatoire observable $(X_i, Y_i)_{1\leq i\leq n}$ de loi (X, Y).
- $\beta \in \mathbb{R}^p$ et $b \in \mathbb{R}$ sont les paramètres inconnus et non aléatoire du modèle, β est le vecteur poids, b est dit le biais (intercept en anglais).
- ϵ est une v.a. non observable qu'on appelle bruit, elle décrit l'erreur.

Sur l'échantillon $(X_i, Y_i)_{1\leq i\leq n}$, le modèle linéaire s'écrit,

$$Y_i = \beta^t X_i + b + \epsilon_i,$$

et $\epsilon_1, \cdots \epsilon_n$ sont i.i.d. et vérifient $\mathbb{E}(\epsilon_1^2) < \infty$,

$$\mathbb{E}(\epsilon_i|X_i) = 0, \;\; \mathrm{Var}(\epsilon_i|X_i) = \sigma^2, \; ps.$$

L'hypothèse de modélisation $\mathrm{Var}(\epsilon_i|X_i) = \sigma^2$ est dite d'homoscédasticité .

Remarque. Le modèle linéaire s'écrit, sous une forme matricielle,

$$\mathbb{Y} = \mathbb{X}\tilde{\beta} + \epsilon,$$

avec,

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} \qquad \epsilon = \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_n \end{pmatrix} \qquad \tilde{\beta} = \begin{pmatrix} b \\ \beta \end{pmatrix}$$

$$\mathbb{X} = \begin{pmatrix} 1 & X_1^t \\ 1 & X_2^t \\ \cdots & \cdots \\ 1 & X_n^t \end{pmatrix}$$

$\mathbb{X}$ est une matrice à n lignes et $(p+1)$ colonnes.

Question 1.1. *Pourquoi l'hypothèse de modélisation $\mathbb{E}(\epsilon|X) = 0$ ps est-elle raisonnable pour décrire un modèle linéaire ?*

Exercice 1.1.
On suppose que $\mathbb{E}(|Y|) < \infty$. On appelle la fonction de régression de Y sur X, la fonction réelle r définie sur $\mathbb{R}^p$, pour $x \in \mathbb{R}^p$, par

$$r(x) = \mathbb{E}(Y|X = x).$$

On considère le modèle introduit dans la définition (1.1). Montrer que

$$r(x) = \beta^t x + b,$$

et que p.s.,

$$Y = r(X) + \epsilon.$$

La question qui se pose maintenant est : quel est le critère à utiliser afin d'estimer (à l'aide des observations $(X_i, Y_i)_{1\leq i\leq n}$), les paramètres du modèle linéaire introduit dans la définition 1.1 ?.

Définition 1.2. *Le critère des moindres carrés ordinaires (MCO) est défini par,*

$$\mathcal{C}(\beta, b, (X_i, Y_i)_{1\leq i\leq n}) = \sum_{i=1}^{n} (Y_i - \beta^t X_i - b)^2.$$

Les estimateurs des moindres carrés ordinaires de β et b sont les minimiseurs (lorsqu'ils existent) de la fonction :

$$(\beta, b) \mapsto \mathcal{C}(\beta, b, (X_i, Y_i)_{1\leq i\leq n}),$$

en d'autres termes,

$$(\hat{\beta}_n, \hat{b}_n) \in argmin_{\beta\in\mathbb{R}^p, b\in\mathbb{R}} \mathcal{C}(\beta, b, (X_i, Y_i)_{1\leq i\leq n}). \tag{1.1}$$

Question 1.2. *Que représente intuitivement le critère des moindres carrés (on peut considérer le cas $p = 1$ et représenter le nuage de points $(x_i, y_i)_{1\leq i\leq n}$).*

Question 1.3. *Discuter l'existence et l'unicité de la solution de (1.1).*

Exercice 1.2.
On suppose, dans cet exercice, que $p = 1$ (i.e. cas du modèle linéaire simple).

1. Montrer que
$$\hat{\beta}_n = \frac{S_{x,y}}{S_x^2},\ \hat{b}_n = \overline{Y}_n - \hat{\beta}_n\, \overline{X}_n,$$
avec
$$S_{x,y} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \overline{X}_n \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)$$
$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$
On rappelle que $\overline{X}_n$ (de même pour $\overline{Y}_n$) est la moyenne empirique de l'échantillon $(X_1, \cdots, X_n)$, S_x^2 est sa variance empirique, le point $(\overline{X}_n, \overline{Y}_n)$ est dit le point moyen du nuage des points et $S_{x,y}$ est la covariance empirique des deux échantillons.
2. La droite d'équation $y = \hat{\beta}_n x + \hat{b}_n$ est appelée droite des moindres carrés . Vérifier que cette droite passe par le point moyen $(\overline{X}_n, \overline{Y}_n)$ du nuage des points $(X_i, Y_i)_{1\le i\le n}$.
3. Montrer qu'un estimateur sans biais de σ^2 est donnée par
$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{\epsilon}_i^2,$$
avec $\hat{\epsilon}_i = Y_i - \hat{Y}_i$ et $\hat{Y}_i = \hat{\beta}_n X_i + \hat{b}_n$.

Théorème 1.1. *On suppose maintenant que $p \in \mathbb{N} \setminus \{0\}$. Alors $(\hat{b}_n, \hat{\beta}_n)^t$ existe et est unique si $\mathbb{X}^t\mathbb{X}$ est inversible et dans ce cas,*
$$(\hat{b}_n, \hat{\beta}_n)^t = \left(\mathbb{X}^t\ \mathbb{X}\right)^{-1} \mathbb{X}^t\mathbb{Y}.$$

La démonstration du théorème 1.1 est en exercice.

Question 1.4.
1. *Montrer que $\mathbb{X}^t\ \mathbb{X}$ est inversible si et seulement le rang de $\mathbb{X}$ est $p+1$.*
2. *Vérifier que si $p+1 > n$ alors la matrice $\mathbb{X}^t\mathbb{X}$ n'est pas inversible. Interpréter concrètement la condition $p+1 > n$.*

Proposition 1.1. *On suppose que la matrice $\mathbb{X}$ est déterministe. Alors,*
$$\mathbb{E}\left[\begin{pmatrix}\hat{b}_n\\ \hat{\beta}_n\end{pmatrix}\right] = \begin{pmatrix} b\\ \beta\end{pmatrix}, \qquad \mathbb{V}ar\left[\begin{pmatrix}\hat{b}_n\\ \hat{\beta}_n\end{pmatrix}\right] = \sigma^2\left(\mathbb{X}^t\ \mathbb{X}\right)^{-1}.$$

La démonstration de la proposition 1.1 est en exercice (pour rappel : l'espérance d'un vecteur aléatoire est le vecteur des espérances. La variance d'un vecteur aléatoire est la matrice de covariances (donc dont les termes diagonaux sont les variances des marginales du vecteur)).

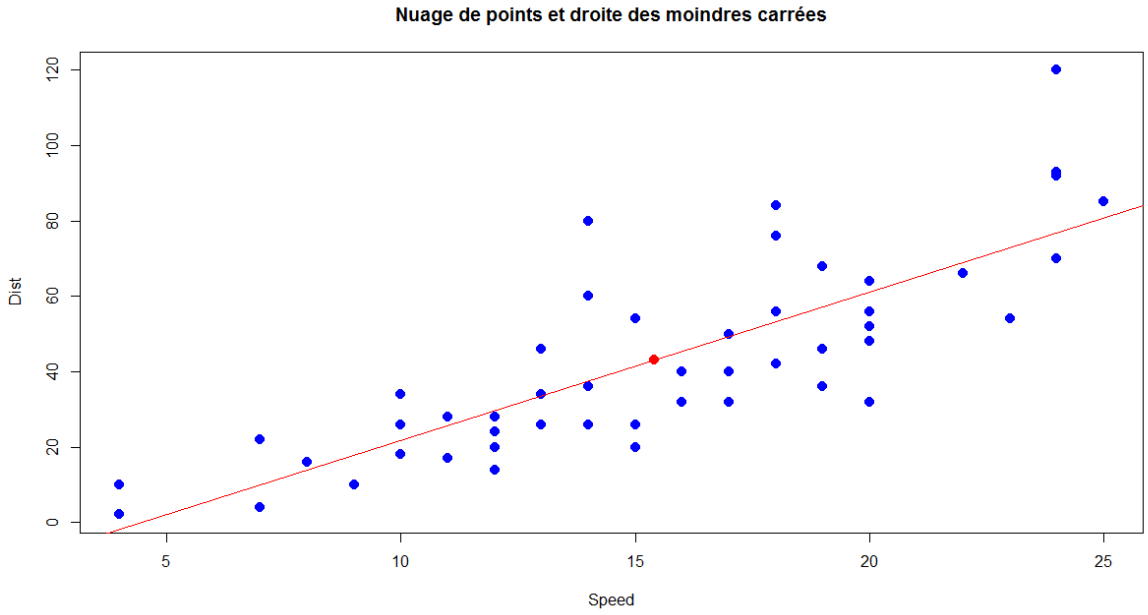

FIGURE 1.2 – Nuage de points et droite MCO

Exercice 1.3.
Montrer que pour tout $x \in \mathbb{R}^p$,

$$\hat{b}_n + \hat{\beta}_n^t x = \sum_{i=1}^{n} w_{n,i}(x) Y_i,$$

$w_{n,i}(x)$ sont des poids à déterminer, ne dépendant que de $(X_i)_{1\leq i\leq n}, x, n$. On dit que l'estimateur $(\hat{b}_n, \hat{\beta}_n)^t$ est linéaire (à ne pas confondre avec la définition de modèle linéaire).

Remarque. Une prédiction de Y_{new} non observé, selon ce modèle, pour un X_{new} donné est donc,

$$\hat{Y}_{new} = \hat{b}_n + \hat{\beta}_n^t X_{new}$$

Question 1.5. *On pose,*

$$\hat{Y}_i = \hat{b}_n + \hat{\beta}_n^t X_i$$

Quelle est la différence entre Y_i et $\hat{Y}_i$?. Illustrer graphiquement la réponse. Que représente $Y_i - \hat{Y}_i$?

Définition 1.3. *On appelle les résidus, les v.a. $\hat{\epsilon}_i = Y_i - \hat{Y}_i$. Le vecteur des résidus est le*

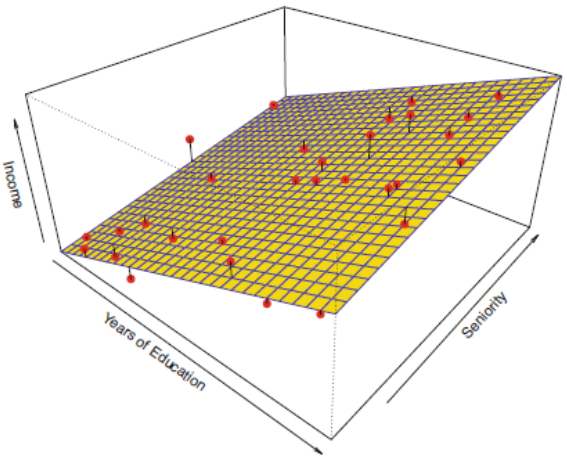

FIGURE 1.3 – Ici $p = 2$. On voudrait expliquer le revenu Y (Income) à l'aide du nombre d'années d'études $X^{(1)}$ (Years of Education) et de l'ancienneté $X^{(2)}$ (seniority). Les points rouges sont le nuage des points. L'hyperplan est celui obtenu, en estimant les paramètres du modèle linéaire multiple par MCO. Il permet de déduire des prédictions d'un revenu pour une ancienneté et un nombre d'années d'études donnés $\hat{Y} = \hat{b} + \hat{\beta}_1 X^{(1)} + \hat{\beta}_2 X^{(2)}$. Les traits verticaux noirs sont les résidus i.e. $Y_i - \hat{Y}_i$. Cet exemple est tiré de [4].

vecteur

$$\hat{\epsilon} = \begin{pmatrix} \hat{\epsilon}_1 \\ \vdots \\ \hat{\epsilon}_i \\ \vdots \\ \hat{\epsilon}_n \end{pmatrix} = \mathbb{Y} - \hat{\mathbb{Y}} = (\mathbb{I}_n - \mathbb{H})\mathbb{Y},$$

$\mathbb{I}_n$ est la matrice identité d'ordre n et $\mathbb{H} = \mathbb{X}\left(\mathbb{X}^t\,\mathbb{X}\right)^{-1}\mathbb{X}^t$. La matrice $\mathbb{H}$ est dite matrice chapeau, elle transforme $\mathbb{Y}$ en $\hat{\mathbb{Y}}$,

$$\hat{\mathbb{Y}} = \mathbb{H}\mathbb{Y}.$$

Proposition 1.2. *Un estimateur sans biais de σ^2 est donné par,*

$$\hat{\sigma}_n^2 = \frac{1}{n-p-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Théorème 1.2. *Théorème de Gauss-Markov. Parmi tous les estimateurs linéaires non biaisés, l'estimateur des moindres carrés présente une variance minimale. (On dit que l'estimateur des moindres carrés est BLUE (Best Linear Unbiaised Estimator)).*

La démonstration est en exercice.

1.2 Qualité du modèle

1.2.1 Mots clés

Décomposition de la variance, coefficient de détermination R^2

1.2.2 Synthèse

La Somme des Carrés Totale SCT est définie par :

$$SCT = \sum_{i=1}^{n} (Y_i - \overline{Y_n})^2$$

c'est la variation totale des $(Y_i)_{1 \leq i \leq n}$ autour de leur moyenne $\overline{Y_n}$,
La Somme des Carrés des Résidus est :

$$SCR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2,$$

représente la variance résiduelle ou non expliquée.
La Somme des Carrés expliquées est :

$$SCE = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y_n})^2$$

représente la variance expliquée par le modèle (donne la variation des valeurs ajustées autour de la moyenne).

Exercice 1.4.

1. Montrer que
$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y_n}) = 0, \ \ ps.$$
2. Vérifier que le vecteur $\hat{Y} = (\hat{Y}_1, \cdots, \hat{Y}_n)^t$ est la projection orthogonale du vecteur $\mathbb{Y}$ sur l'espace vectoriel engendré par $(\mathbf{1}, X^{(1)}, \cdots, X^{(p)})$ avec $\mathbf{1} = (1, 1, \cdots, 1)^t$.

Théorème 1.3. *Montrer, sous les notations précédentes, que SCT= SCE+ SCR*

Le coefficient de détermination R^2 est défini par,

$$R^2 = \frac{SCE}{SCT}$$

1. $0 \leq R^2 \leq 1$
2. Lorsque $R^2 = 0$, le modèle n'explique rien, les variables X et Y ne semblent pas être linéairement corrélées.

3. Lorsque $R^2 = 1$, les points sont alignés sur la droite.
4. Une valeur de R^2 proche de 1 indique une forte corrélation linéaire.

Exercice 1.5.

1. Montrer que pour un modèle linéaire simple, le coefficient de détermination R^2 n'est autre que $\hat{r}^2$ le carré du coefficient de corrélation empirique, $\hat{r}$, définit par :

$$\hat{r} = \frac{\sum_{i=1}^n (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \overline{X}_n)^2 \sum_{i=1}^n (Y_i - \overline{Y}_n)^2}}$$

2. Montrer que $|\hat{r}| \leq 1$. Dans quels cas l'égalité est atteinte ? En déduire une interprétation du $\hat{r}^2$ en rapport avec le modèle linéaire.

1.3 Modèle linéaire gaussien

Afin de pouvoir construire des intervalles de confiance et de faire des tests d'hypothèse sur les paramètres du modèle linéaire, on devrait avoir plus d'informations sur la loi du bruit ϵ_1. C'est l'objectif de cette section.

1.3.1 Mots clés

Loi normale, estimation des paramètres par maximum de vraisemblance, intervalles de confiance, intervalles de prédiction, tests d'hypothèse (tests de significativité, tests sur les paramètres)

1.3.2 Synthèse

On considère le modèle linéaire ci-haut définit auquel on ajoute l'hypothèse suivante :

$$\epsilon_1 \sim \mathcal{N}(0, \sigma^2)$$

Le modèle linéaire ainsi posé est dit **Modèle linéaire gaussien.**

Question 1.6. *Lorsque le modèle linéaire est gaussien : quelle est la loi conditionnelle de $Y_i/X_i = x$, pour un x fixé. ?*

Exercice 1.6.

1. Calculer la fonction de vraisemblance, $(\beta, b) \mapsto \prod_{i=1}^n f_{Y_1|X_1=x_i}(y_i)$, $f_{Y_1|X_1=x_i}$ étant la densité conditionnelle de la loi de Y_1 sachant $X_1 = x_i$.
2. Chercher son maximum et conclure.

Proposition 1.3. *On suppose que le modèle linéaire et gaussien et que $\mathbb{X}$ est déterministe. Alors,*

1. *Le vecteur* $\begin{pmatrix}\hat{b}_n\\ \hat{\beta}_n\end{pmatrix}$ *est un vecteur gaussien d'espérance* $\begin{pmatrix}b\\ \beta\end{pmatrix}$ *et de variance* $\sigma^2\left(\mathbb{X}^t\ \mathbb{X}\right)^{-1}$
2. *Le vecteur résidu* $\hat{\epsilon}$ *est un vecteur gaussien centré et de variance* $\sigma^2(\mathbb{I}_n - \mathbb{H})$.
3. $\hat{\mathbb{Y}}$ *est un vecteur gaussien, d'espérance* $\mathbb{X}\begin{pmatrix}b\\ \beta\end{pmatrix}$ *et de variance* $\sigma^2\mathbb{H}$.
4. $\hat{\mathbb{Y}}$ *et* $\hat{\epsilon}$ *sont deux vecteurs aléatoires indépendants.*
5. $(n-p-1)\hat{\sigma}^2/\sigma^2$ *suit la loi de Khi-deux à* $(n-p-1)$ *degrés de libertés.*

La démonstration de la proposition 1.3 est en exercice (elle repose beaucoup sur la définition et les propriétés d'un vecteur gaussien).
Les résultats ci-dessous sont une conséquence immédiate de la proposition 1.3 (démonstration en exercice). On notera par $[(\mathbb{X}^t\ \mathbb{X})^{-1}]_{i,i}$ le ième terme diagonal de la matrice $\left(\mathbb{X}^t\ \mathbb{X}\right)^{-1}$.

Question 1.7. *À quoi correspond le terme* $\sigma^2[(\mathbb{X}^t\ \mathbb{X})^{-1}]_{i,i}$ *pour* $1 \leq i \leq p+1$ *?*

Proposition 1.4. *On a, sous les hypothèses de la proposition 1.3,*

1. $(\hat{b}_n - b)/\sigma\sqrt{[(\mathbb{X}^t\ \mathbb{X})^{-1}]_{1,1}}$ *suit la loi normale centrée réduite sur* $\mathbb{R}$.
2. $(\hat{b}_n - b)/\hat{\sigma}\sqrt{[(\mathbb{X}^t\ \mathbb{X})^{-1}]_{1,1}}$ *suit la loi de Student à* $(n-p-1)$ *degrés de liberté.*
3. *On note par* $(\hat{\beta}_n)_j$ *la j-ième composante du vecteur* $\hat{\beta}_n$ *et par* β_j *celle du vecteur* β. *On a de même :*
   $((\hat{\beta}_n)_j - \beta_j)/\sigma\sqrt{[(\mathbb{X}^t\ \mathbb{X})^{-1}]_{j+1,j+1}}$ *suit la loi normale centrée réduite sur* $\mathbb{R}$.
   $((\hat{\beta}_n)_j - \beta_j)/\hat{\sigma}\sqrt{[(\mathbb{X}^t\ \mathbb{X})^{-1}]_{j+1,j+1}}$ *suit la loi de Student à* $(n-p-1)$ *degrés de liberté.*

Corollaire 1.1.

1. *Un intervalle de confiance pour* β_j *(lorsque* σ *est inconnu) au niveau* $1-\alpha$, *est*
$$\left[(\hat{\beta}_n)_j \pm t_{n-p-1,1-\alpha/2}\ \hat{\sigma}\sqrt{[(\mathbb{X}^t\ \mathbb{X})^{-1}]_{j+1,j+1}}\right],$$
$t_{n-p-1,1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-p-1$ degrés de liberté. Un intervalle est analogue pour la constante b du modèle.

2. *Un intervalle de confiance pour* σ^2 *au niveau* $1-\alpha$, *est*
$$\left[\frac{(n-p-1)\hat{\sigma}^2}{z_{1-\alpha/2}}, \frac{(n-p-1)\hat{\sigma}^2}{z_{\alpha/2}}\right],$$
ici $z_{\alpha/2}$ (resp. $z_{1-\alpha/2}$) désigne le quantile d'ordre $\alpha/2$ (resp. d'ordre $1-\alpha/2$) de la loi de Khi-deux à $n-p-1$ degrés de liberté.

Remarque. On rappelle que le modèle linéaire,
$$Y_i = b + \beta^t X_i + \epsilon_i = b + \sum_{j=1}^{p}\beta_j X_i^{(j)} + \epsilon_i,\ \ 1 \leq i \leq n,$$

Pour voir si la variable explicative $X^{(j)}$ explique bien Y on peut faire le test d'hypothèse suivant (dit test de Student) :

$$H_0 : \ \beta_j = 0$$

contre

$$H_1 : \ \beta_j \neq 0.$$

La statistique du test est

$$T = \frac{((\hat{\beta}_n)_j)}{\hat{\sigma}\sqrt{[(\mathbb{X}^t \ \mathbb{X})^{-1}]_{j+1,j+1}}},$$

dont la loi sous H_0 est la loi de Student à $n-p-1$ degrés de liberté. On note par T_{cal} La valeur calculée de T sur les observations.

1. Si $|T_{cal}| \geq t_{n-p-1,1-\alpha/2}$ alors on rejettera H_0 au seuil α sinon H_0 est conservée.
2. La p-paleur de ce test est $2(1 - F_T(|T_{cal}|))$, F_T étant la fonction de répartition de la loi de Student à $n-p-1$ degrés de liberté.

Test global de significativité de la régression (test de Fisher) : l'objectif est de tester si tous les coefficients sont nuls, excepté la constante b du modèle.

$$H_0 : \ \beta_1 = \cdots = \beta_p = 0.$$

La statistique du test est donnée par :

$$F = \frac{SCE/p}{SCR/(n-p-1)} = \frac{R^2}{1-R^2}\frac{n-p-1}{p},$$

qui suit une loi de Fisher à p et $n-p-1$ degrés de libertés. On rejettra H_0 au seuil α (i.e. on dira que les coefficients du modèle sont conjointement significatifs) si F_{cal} est supérieur au quantile d'ordre $1-\alpha$ d'une loi de Fisher à p et $n-p-1$ degrés de libertés. La p-valeur de ce test est :

$$1 - \Phi_{p,n-p-1}(F_{calc}),$$

$\Phi_{p,n-p-1}$ est la fonction de répartition de la loi de Fisher à p et $n-p-1$ degrés de libertés.

Exercice 1.7.
Soit X_{new} donné et observé, Y_{new} la variable réponse associée qui est non observée mais $\hat{Y}_{new}$ est une prédiction de Y_{new} et est donc calculable. On a,

$$\hat{Y}_{new} = \hat{b}_n + \hat{\beta}_n^t X_{new},$$

on note que $\hat{b}_n$ et $\hat{\beta}_n$ sont construits à partir des observations $(X_i, Y_i)_{1\leq i\leq n}$ supposées indépendantes de (X_{new}, Y_{new}) et que,

$$Y_{new} = b + \beta^t X_{new} + \epsilon_{new}.$$

On suppose que $X_{new}, \mathbb{X}$ sont déterministes.

1. Montrer que la v.a. $\hat{Y}_{new} - Y_{new}$ suit une loi normale centrée et de variance

$$\sigma^2 \left[1 + (1, X_{new})(\mathbb{X}^t\ \mathbb{X})^{-1} \begin{pmatrix} 1 \\ X_{new} \end{pmatrix} \right]$$

2. En en déduire que l'intervalle

$$\left[\hat{Y}_{new} \pm t_{n-p-1,1-\alpha/2}\ \hat{\sigma} \sqrt{\left[1 + (1, X_{new})(\mathbb{X}^t\ \mathbb{X})^{-1} \begin{pmatrix} 1 \\ X_{new} \end{pmatrix} \right]} \right]$$

contient Y_{new} avec la probabilité $1-\alpha$, $t_{n-p-1,1-\alpha/2}$ étant le quantile d'ordre $1-\alpha/2$ de la loi de student à $n-p-1$ degrés de libertés.

Il s'agit d'un intervalle de prédiction pour Y_{new} (à ne pas confondre avec un intervalle de confiance).

1.4 Vérification graphique des hypothèses de modélisation

1.4.1 Mots clés

Hypothèses de modélisation, analyse des résidus, graphes des résidus, QQplot.

1.4.2 Synthèse

On a supposé, les hypothèses de modélisation, ci-dessous :
— la linéarité du modèle,
— l'homoscédasticité i.e. la variance du bruit est une constante qu'on a noté σ^2,
— la normalité du bruit (il suit la loi normale centrée et de variance σ^2),
— l'indépendance des v.a. $(X_i, Y_i)_{1 \leq i \leq n}$

Sans ces hypothèses, les résultats mathématiques développés dans la section 1.3 ne seront pas corrects. Il est très utile de s'assurer donc que ces hypothèses, sont réalistes sur les observations $(x_i, y_i)_{1 \leq i \leq n}$. Le graphique des résidus est un outil important afin d'avoir une idée sur la validité des hypothèses de modélisation. En principe, le graphe des résidus est la première chose à faire une fois l'estimation du modèle linéaire est faite.

Non linéarité

Si le graphique des résidus (les résidus sont sur l'axe vertical des ordonnées et les valeurs prédites sont sur l'axe horizontal des abscisses) montre un motif curviligne, cela peut indiquer une relation non linéaire entre les variables, c'est-à-dire un modèle de régression non linéaire pourrait être plus approprié. Le graphique doit être approximativement horizontal s'il y a bien une relation linéaire.

Homoscédasticité

L'hypothèse d'homoscédasticité semble être vérifiée si le nuage de points n'a pas de forme particulière, c'est-à-dire des résidus sont homogènes autour de zéro. Le nuage de points aura une dispersion uniforme autour de la ligne horizontale des résidus nuls, indiquant l'homoscédasticité. Sinon cela peut indiquer une hétéroscédasticité, violant l'hypothèse d'homoscédasticité.

Normalité

La distribution des résidus peut être décrite graphiquement par un histogramme ou par un diagramme quantile-quantile. Ce dernier représente les quantiles de la distribution empirique des observations $(y_i - \hat{y}_i)_{1 \leq i \leq n}$ en fonction des quantiles de la distribution normale adéquate : les points doivent être presque alignés.

L'analyse des graphiques donne des idées sur la validation des hypothèses de modélisation qu'on devrait les approfondir avec des outils théoriques qui sortent de l'objectif de ce guide. Par exemple, le test de Breuch et Pagan permet de tester si on peut conserver l'hypothèse d'homoscédasticité ou non, le test de Rainbow permet de tester l'hypothèse de la linéarité du modèle, le test de Shapiro-Wilk pour tester la normalité du bruit.

1.4.3 Illustrations Graphiques

Résidus VS Valeurs prédites

1. Si la relation est linéaire et le bruit est centré, les résidus seront dispersés de façon aléatoire autour de la ligne de 0.
2. Dans le cas de l'homoscédasticité, les résidus forment une bande horizontale approximative autour de la ligne de 0 i.e. la variance des résidus est homogène .
3. Lorsque les résidus sont organisés en forme d'entonnoir, c'est qu'ils ne sont, probablement, pas homoscédastiques.

QQ plots

Le graphique *QQ plots* compare les quantiles (donc la distribution de probabilité) des résidus du modèle à ceux (donc à une distribution de probabilité) d'une loi normale. Il permet donc de vérifier graphiquement l'hypothèse de la normalité.

Échelle localisée

permet de vérifier si la dispersion des résidus augmente pour une valeur prédite donnée (i.e. si la dispersion des résidus est causée par la variable explicative). Si la dispersion augmente, la condition de base d'homoscédasticité n'est pas respectée.

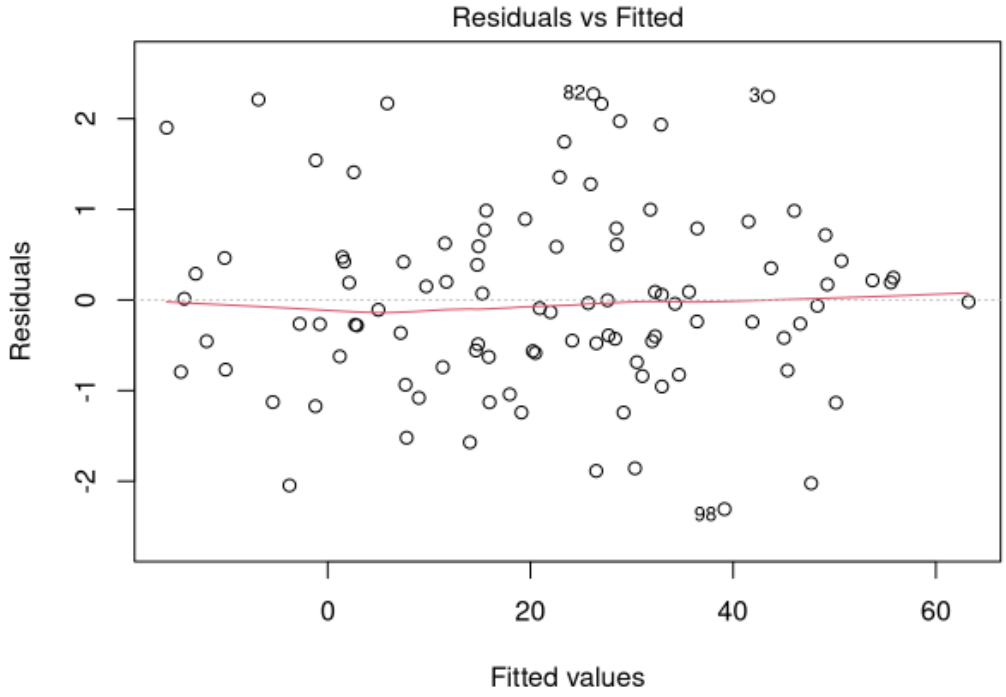

FIGURE 1.4 – Graphique des résidus : en abscisses les valeurs $(\hat{y}_i)_i$ et en ordonnées $(\hat{\epsilon}_i)_i$

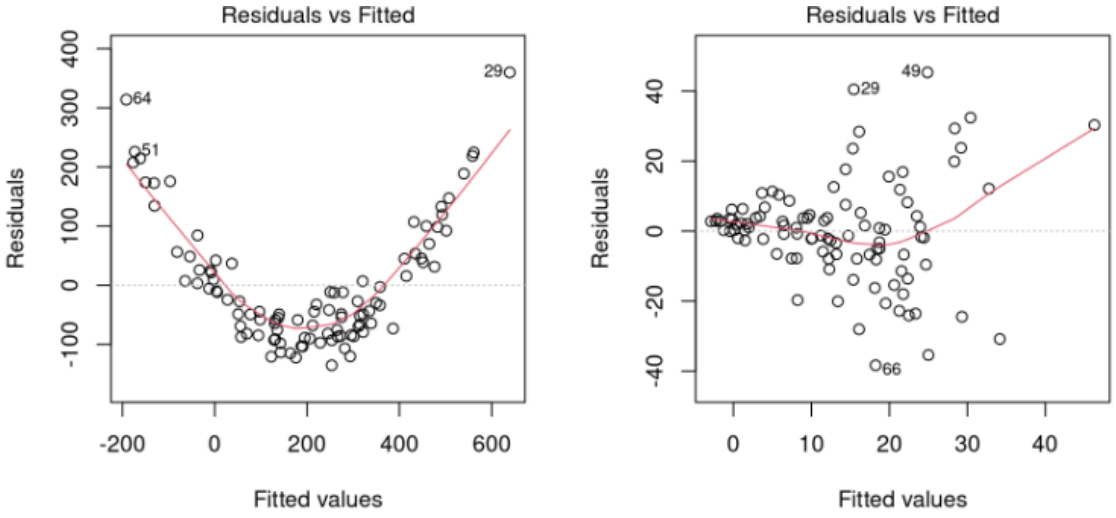

FIGURE 1.5 – À gauche le modèle linéaire ne semble pas être adapté. À droite l'hypothèse d'hétéroscédasticité semble plus raisonnable. Graphiques tirées de [2].

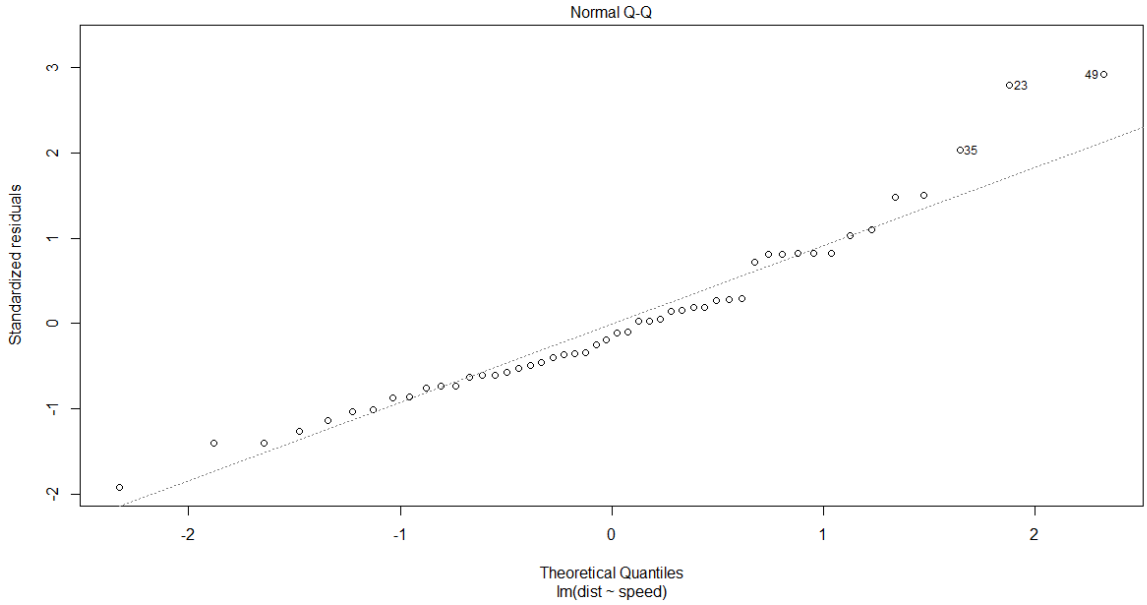

FIGURE 1.6 – QQplot : quantiles empiriques en fonction des quantiles théoriques.

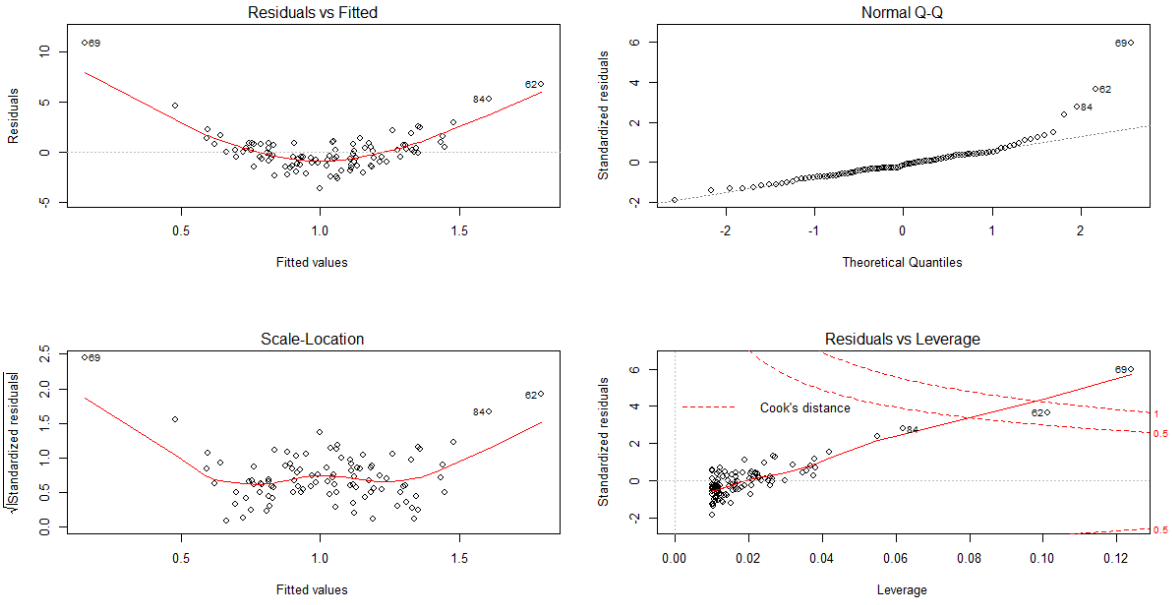

FIGURE 1.7 – Ces graphiques concernent le modèle gaussien et non-linéaire $Y_i = X_i^2 + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 1)$.

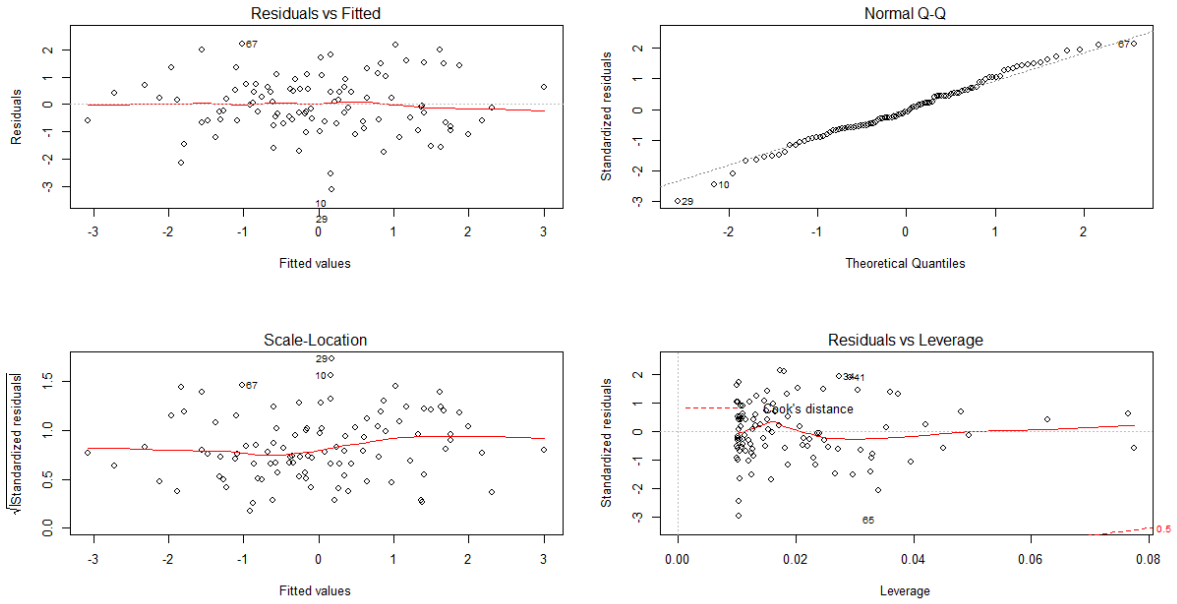

FIGURE 1.8 – Ces graphiques concernent le modèle linéaire gaussien $Y_i = X_i + \epsilon_i,\ \ \epsilon_i \sim \mathcal{N}(0,1)$.

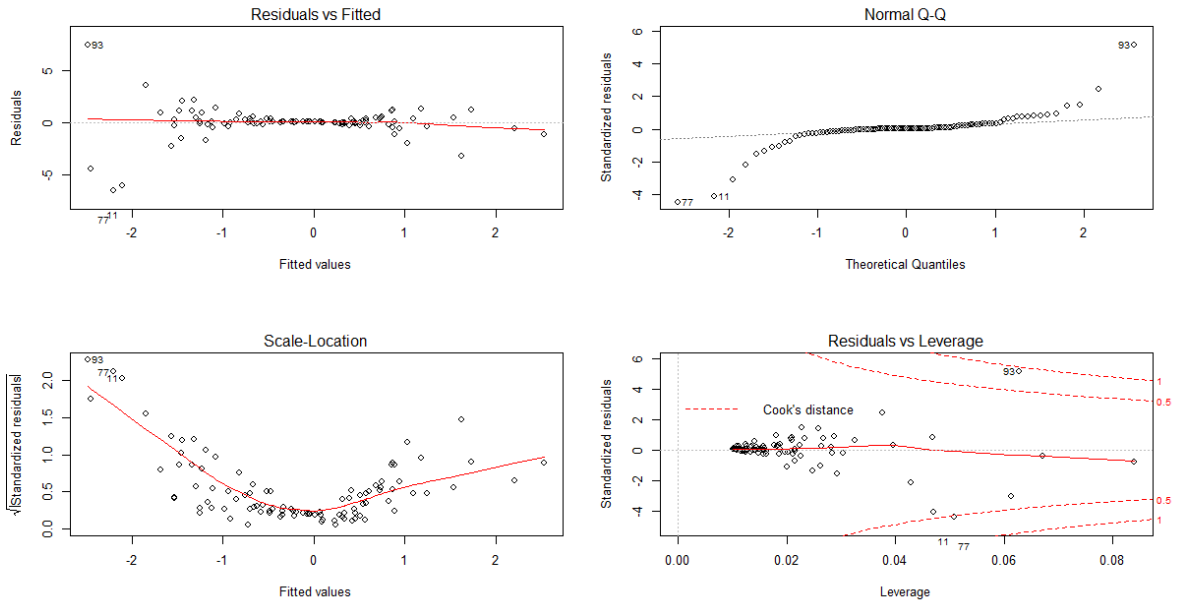

FIGURE 1.9 – Ces graphiques concernent le modèle linéaire gaussien mais dans un cas d'hétéroscédasticité : $Y_i = X_i + \epsilon_i,\ \ \epsilon_i \sim \mathcal{N}(0, X_i^4)$.

Chapitre 2

Problématiques

2.1 Problématique I

Les observations, ci-dessous, donnent les vitesses (en mph) de 50 voitures (variable speed) et les distances de freinage (variable dist.) (unité pieds). Source : ici. L'objectif est de comprendre, en se basant sur 50 observations, l'effet de la vitesse sur la distance de freinage. Cela permettra d'avertir, entre autres, sur le danger des excès de vitesse.

(AI) Données et étude empirique

car	speed	dist
1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10
7	10	18
8	10	26
9	10	34
10	11	17
11	11	28
12	12	14
13	12	20
14	12	24
15	12	28
16	13	26
17	13	34
18	13	34
19	13	46
20	14	26

car	speed	dist
21	14	36
22	14	60
23	14	80
24	15	20
25	15	26
26	15	54
27	16	32
28	16	40
29	17	32
30	17	40
31	17	50
32	18	42
33	18	56
34	18	76
35	18	84
36	19	36
37	19	46
38	19	68
39	20	32
40	20	48

car	speed	dist
41	20	52
42	20	56
43	20	64
44	22	66
45	23	54
46	24	70
47	24	92
48	24	93
49	24	120
50	25	85

1. Quelle est la variable explicative ? Quelle est la variable à expliquer ?
2. Calculer, en utilisant un langage informatique de votre choix, les paramètres empiriques de ces observations : les deux moyennes empiriques, les deux variances empiriques, la covariance empirique..
3. Représenter graphiquement ce nuage de ces 50 points : en abscisses la variable *speed*, en ordonnées la variable *dist*. Représenter sur ce même graphe le point moyen de ce nuage de points. [Indication : voir la figure 1.1.]
4. Que constate-t-on ?

(BI) Hypothèses de modélisation et modèle linéaire

On dispose donc des observations $(v_i, d_i)_{1\leq i\leq n}$, $n = 50$. On suppose que ces 50 observations sont des réalisations de v.a. $(V_i, D_i)_{1\leq i\leq n}$ i.i.d. de loi inconnue. On introduit le modèle linéaire suivant : pour $1 \leq i \leq n$

$$D_i = b + \beta V_i + \epsilon_i, \;\; E(\epsilon_i|V_i) = 0, \;\; \mathrm{Var}(\epsilon_i^2|V_i) = \sigma^2 \;\; ps.$$

1. Écrire le critère des moindres carrés permettant d'estimer b et β à l'aide des observations $(V_i, D_i)_{1\leq i\leq n}$.
2. Écrire, en utilisant un logiciel informatique de votre choix, un programme permettant de donner les estimations de b et β selon le critère des moindres carrés.
3. Exécuter le programme et donner l'équation de la droite des moindres carrés sur les données ci-haut introduites. Représenter la graphiquement avec le nuage de points. [indication : voir la figure 1.2.]

(CI) Analyse graphique et hypothèses de modélisations

1. Représenter graphiquement le nuage des points $(\hat{y}_i, \hat{\epsilon}_i)$. Analyser ce graphique. [Indication : La figure 1.4 est un exemple de tel graphique.]
2. Faire un graphique donnant les quantiles des résidus standardisés aux quantiles de la loi normale centrée réduite. Analyser le graphique obtenu.
3. Étudier graphiquement la validité du modèle linéaire gaussien ainsi que la validité des différentes hypothèses de modélisation utilisées. [indication : voir le graphique 1.6].

(DI) Intervalles de confiance et tests d'hypothèse

On supposera que le modèle linéaire est gaussien.

1. Construire des intervalles de confiance des paramètres du modèle (en faisant les calculs et aussi en utilisant le calcul d'un logiciel de votre choix).
2. Étudier la significativité du modèle.

(EI) Conclure

Synthétiser vos résultats et donner une conclusion concrète (en rapport avec le problème posé).

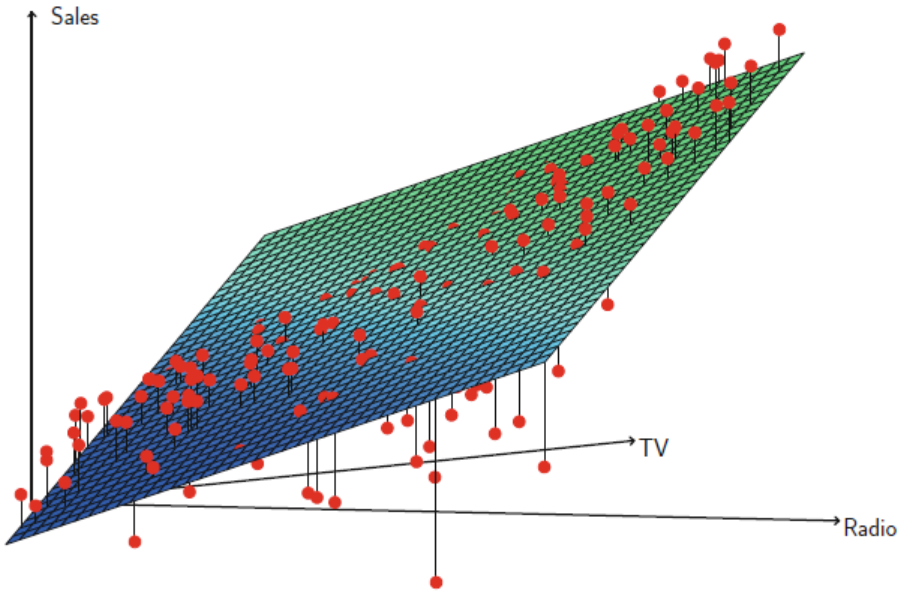

FIGURE 2.1 – Ici $p = 2$. Le graphique, tiré de [4], représente le nuage de points et l'hyperplan obtenu par la méthode MCO. La modélisation linéaire est-elle bien adaptée ?

2.2 Problématique II

On cherche à comprendre la vente Y d'un produit à l'aide des investissements publicitaires à la radio $X^{(1)}$ et à la télévision $X^{(2)}$. Cela permettra de comprendre comment l'investissement publicitaire, par ces deux moyens, expliquera les ventes et aussi de pouvoir prendre des décisions sur les montants à investir, en publicité, à la radio et à la télévision.

Les points rouges de la figure (2.1) représentent les nuages de points $(x_i, y_i)_{1\leq i\leq n}$ avec $x_i = (x_i^{(1)}, x_i^{(2)})^t$, n étant le nombre d'entreprises sur lesquelles portent l'enquête.

Si on suppose que le modèle est linéaire, il s'écrira

$$Y_i = b + \beta^t X_i + \epsilon_i = b + \beta_1 X_i^{(1)} + \beta_2 X_i^{(2)} + \epsilon_i,$$

avec,

$$\mathbb{E}(\epsilon_i|X_i) = 0,\ \mathrm{Var}(\epsilon_i|X_i) = \sigma^2,\ ps.$$

Le critère des moindres carrés ordinaires de la définition 1.2 permet d'estimer les paramètres b, β_1, β_2. Les expressions de ces estimateurs sont données par le théorème 1.1. L'équation de l'hyperplan de la figure (2.1) est donc,

$$z = \hat{b} + \hat{\beta}_1 x + \hat{\beta}_2 y$$

et la prédiction,

$$\hat{Y} = \hat{b} + \hat{\beta}_1 X^{(1)} + \hat{\beta}_2 X^{(2)}.$$

Sur le graphique 2.1, les traits verticaux en noirs sont les résidus $(y_i - \hat{y}_i)_{1 \leq i \leq n}$.

Question. Expliquer en se basant sur le graphique (2.1) qu'un modèle **non linéaire** semble mieux expliquer les données.

Bibliographie

[1] WikiStat cliquer ici et cliquer ici

[2] https://r.qcbs.ca/workshop04/book-fr/r%C3%A9gression-lin%C3%A9aire-avec-r.html

[3] V. Rivoirard, G. Stoltz. Statistique mathématique en action, Master et Agrégation externe de mathématiques (2e édition). Vuibert (2012).

[4] G. James, D. Witten, T. Hastie, R. Tibshirani. An Introduction to Statistical Learning with Applications in R. Springer Texts in Statistics. (2021).

[5] S. Louhichi. Guide pour une leçon de modélisation stochastique. Fonctions de répartition empiriques. Tests de Kolmogorov-Smirnov. Estimation des quantiles. Master. France. Disponible https://hal.science/hal-04365615. (2023).

Index

Coefficient de détermination, 9
Corrélation empirique, 10
Covariance empirique, 6
Critère des moindres carrés ordinaires, 5

Droite des moindres carrés, 6

Estimateur linéaire, 7
Estimateurs des moindres carrés ordinaires, 5

Fonction de régression, 5

Graphique des résidus, 13

Homoscédasticité, 4
Hypothèses de modélisation, 1, 13

Intervalle de prédiction, 13

Matrice chapeau, 8
Modèle linéaire multiple, 3
Modèle linéaire simple, 3
Modèle statistique, 1
Modéliser, 3
Moyenne empirique, 6

Nuage des points, 6

Point moyen du nuage des points, 6
Prédiction, 7

Résidus, 7

Test de Fisher, 12
Test de Student, 12
Théorème de Gauss-Markov, 8

Variance empirique, 6

Table des figures

1.1 Nuage de points et point moyen du nuage. La relation semble être linéaire. On peut la modéliser par un modèle linéaire simple (ici $p = 1$) et étudier ensuite les pertinences de ce modèle. 4
1.2 Nuage de points et droite MCO 7
1.3 Ici $p = 2$. On voudrait expliquer le revenu Y (Income) à l'aide du nombre d'années d'études $X^{(1)}$ (Years of Education) et de l'ancienneté $X^{(2)}$ (seniority). Les points rouges sont le nuage des points. L'hyperplan est celui obtenu, en estimant les paramètres du modèle linéaire multiple par MCO. Il permet de déduire des prédictions d'un revenu pour une ancienneté et un nombre d'années d'études donnés $\hat{Y} = \hat{b} + \hat{\beta}_1 X^{(1)} + \hat{\beta}_2 X^{(2)}$. Les traits verticaux noirs sont les résidus i.e. $Y_i - \hat{Y}_i$. Cet exemple est tiré de [4]. 8
1.4 Graphique des résidus : en abscisses les valeurs $(\hat{y}_i)_i$ et en ordonnées $(\hat{\epsilon}_i)_i$ . . . 15
1.5 À gauche le modèle linéaire ne semble pas être adapté. À droite l'hypothèse d'hétéroscédasticité semble plus raisonnable. Graphiques tirées de [2]. 15
1.6 QQplot : quantiles empiriques en fonction des quantiles théoriques. 16
1.7 Ces graphiques concernent le modèle gaussien et non-linéaire $Y_i = X_i^2 + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 1)$. 16
1.8 Ces graphiques concernent le modèle linéaire gaussien $Y_i = X_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, 1)$. 17
1.9 Ces graphiques concernent le modèle linéaire gaussien mais dans un cas d'hétéroscédasticité : $Y_i = X_i + \epsilon_i$, $\epsilon_i \sim \mathcal{N}(0, X_i^4)$. 17

2.1 Ici $p = 2$. Le graphique, tiré de [4], représente le nuage de points et l'hyperplan obtenu par la méthode MCO. La modélisation linéaire est-elle bien adaptée ? . 21

Notations

BLUE Best Linear Unbiaised Estimator 8

I.I.D. indépendants et identiquements distribuées 3

MCO Critère des moindres carrés ordinaires 5

P.S. Presque sûrement 3

V.A. Variables ou vecteurs aléatoires 3