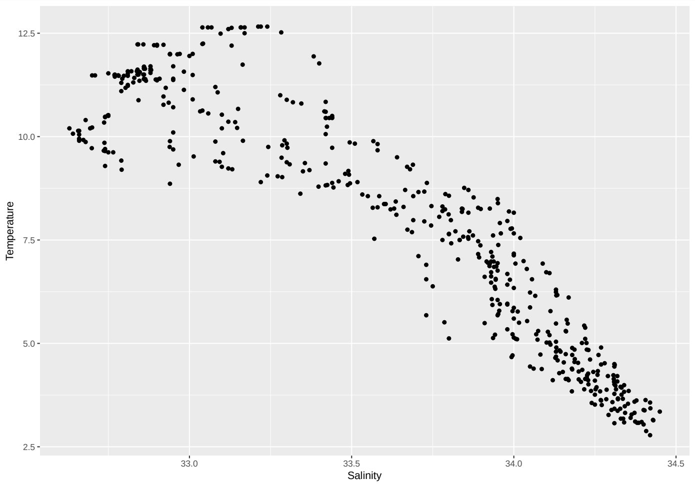
Régression Linéaire Simple

Bruno Pelletier

Outline

Introduction

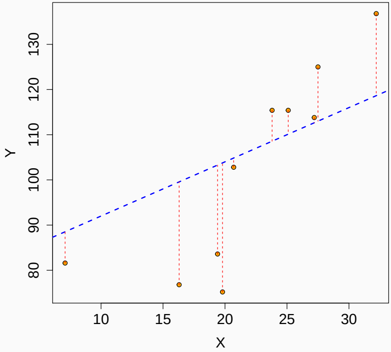
Exemple : Salinité et Température CalCOFI



Objectif : Expliquer la Température (y) à partir de la Salinité (x) de sorte que $y_i \approx \beta_0 + \beta_1 x_i$.

Ajustement par moindres carrés

- On dispose de n observations $(x_1, y_1), \dots, (x_n, y_n)$.
- On cherche la meilleure droite d'équation $y = ax + b$ permettant de modéliser les observations de sorte que $y_i \approx ax_i + b$.



Pour une droite candidate d'équation $y = ax + b$, la i^{eme} erreur d'ajustement est $y_i - (ax_i + b)$.

Ajustement par moindres carrés

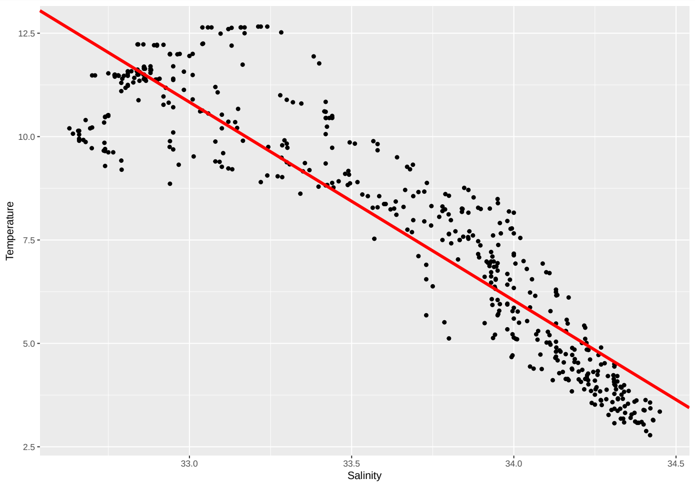
- La qualité de l'ajustement est mesurée par le critère des moindres carrés :

$$\mathcal{L}(a,b) = \sum_{i=1}^{n} (y_i - a x_i - b)^2 .$$

- Sous conditions, $\mathcal{L}$ admet un unique minimum :

$$\begin{aligned} \hat{a} &= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\ \hat{b} &= \bar{y} - \hat{a}\bar{x}. \end{aligned}$$

Exemple : Salinité et Température CalCOFI



Equation de la droite de régression : $y = 169.1178 - 4.79646x$.

Objectifs de la régression

- Etude du lien entre deux variables (sur l'exemple, Salinité et Température).
- Compréhension/Interprétation du phénomène étudié : au travers de l'étude des coefficients du modèle.
- Prévision : quelle valeur de Température peut-on prédire pour une nouvelle observation $x^\star$ de la Salinité ?

$\longrightarrow$ Modélisation statistique des observations

Modélisation statistique

Modélisation

L'étude porte sur deux variables :

- X : variable explicative,
- Y : variable à expliquer.

Deux points de vue quant à la nature des variables :

1. X déterministe et Y aléatoire, ou bien
2. X et Y aléatoires.

Dans tout le cours, on considèrera X déterministe (que l'on notera x). Le cas X aléatoire se traite de manière équivalente en écrivant le modèle conditionnellement à X.

Définition du modèle de régression linéaire simple

- On dispose d'observations (x_i, y_i), pour $i = 1, \dots, n$, issues du modèle suivant :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \text{pour tout } i = 1, \dots, n.$$

- Chaque ε_i est une variable aléatoire appelée le bruit.

Remarques :

- $x \mapsto \beta_0 + \beta_1 x$ graphe $\equiv$ droite ;
- Pour chaque i :

$$Y_i = \underbrace{\beta_0 + \beta_1 x_i}_{\text{déterministe}} + \underbrace{\varepsilon_i}_{\text{aléatoire}}.$$

Ainsi Y_i est aléatoire au travers de ε_i.

Hypothèses sur les bruits

1. Bruits centrés : $\mathbb{E}[\varepsilon_i] = 0$ pour tout $i = 1, \ldots, n$.
    - D'où : $\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_i$.
    - Les Y_i fluctuent autour de la droite de régression.
2. Variances égales : $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $\forall i = 1, \ldots, n$.
    - Homoscédasticité
    - D'où $\mathrm{Var}(Y_i) = \sigma^2$.Les fluctuations des Y_i ne dépendent pas de la condition expérimentale x_i.
3. Décorrélation des bruits : $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$, $\forall\ i \neq j$.
    - D'où $\mathrm{Cov}(Y_i, Y_j) = 0\ \forall i \neq j$.
    - Les variables Y_i sont décorrélées.
    - Les bruits ne sont pas supposés indépendants.

Résumé

Définition (Modèle Linéaire Standard)

$$\begin{aligned}
Y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \forall i = 1, \dots, n, \\
\mathbb{E}[\varepsilon_i] &= 0, \\
\mathrm{Var}(\epsilon_i) &= \sigma^2, \quad \forall i, \\
\mathrm{Cov}(\epsilon_i, \epsilon_j) &= 0, \quad \forall i \neq j.
\end{aligned}$$

Il s'agit de l'écriture analytique du modèle.

Le modèle contient :

- deux coefficients : β_0 et β_1,
- un paramètre de variance des bruits : σ^2.

Ecriture matricielle du modèle

L'étude du modèle en est facilitée par une écriture matricielle. On pose :

- $\mathbf{Y} = (Y_1, \dots, Y_n)'$: vecteur aléatoire de $\mathbb{R}^n$.
- $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$: vecteur aléatoire de $\mathbb{R}^n$.
- $\beta = (\beta_0, \beta_1)'$: vecteur des paramètres.
- Matrice $n \times 2$ (du plan d'expérience/de design) déterministe :

$$\mathbb{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

Le modèle s'écrit :

$$\mathbf{Y} = \mathbb{X}\beta + \varepsilon.$$

Hypothèse en écriture matricielle

- Modèle :

$$\mathbf{Y} = \mathbb{X}\beta + \varepsilon.$$

- Hypothèses :
  - Bruits centrés :

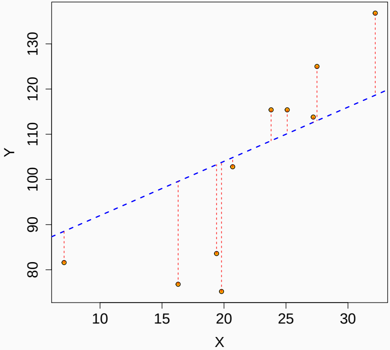
$$\mathbb{E}[\varepsilon] = 0, \qquad (0 \text{ de } \mathbb{R}^n).$$

  - Homoscédasticité + Décorrélation :

$$\mathbb{V}(\varepsilon) = \sigma^2 \mathbf{I}_n = \mathrm{diag}(\sigma^2, \dots, \sigma^2).$$

Estimation des coefficients du modèle

Critère de perte



Pour une droite candidate d'équation $y = \beta_0 + \beta_1 x$, la $i^{\text{ème}}$ erreur d'ajustement est $y_i - (\beta_1 x_i + \beta_0)$. Comment combiner ces n erreurs pour définir un critère pertinent ?

Critère de perte

- Somme des erreurs ?

$\longrightarrow$ non pertinent car les erreurs peuvent se compenser.

- Somme des valeurs absolues des erreurs ?

$$\mathcal{L}(\beta_0, \beta_1) = \sum_{i=1}^{n} |y_i - (\beta_0 + \beta_1 x_i)|$$

$\longrightarrow$ pertinent mais difficulté technique car $\mathcal{L}$ n'est pas différentiable.

Définition (Critère des Moindres Carrés Ordinaires (MCO))

$$\mathcal{L}(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 .$$

$\longrightarrow$ pertinent et différentiable.

Remarque

Plus généralement, on peut se donner une fonction de perte

$$\ell : (u, v) \mapsto \ell(u, v)$$

et définir le critère de perte suivant :

$$\mathcal{L}(\beta_0, \beta_1) = \sum_{i=1}^{n} \ell\left(y_i, \beta_0 + \beta_1 x_i\right).$$

Le critère des MCO correspond à la perte quadratique :

$$\ell(u, v) = (u - v)^2.$$

Estimateurs des MCO

- L'estimateur des MCO est construit par minimisation du critère des MCO sur $\mathbb{R}^2$.

$$\mathcal{L}(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 .$$

- On recherche les points critiques de $\mathcal{L}$ sur $\mathbb{R}^2$:

$$\left\{ \begin{array}{ccc} \frac{\partial \mathcal{L}}{\partial \beta_0} & = & 0 \\ \frac{\partial \mathcal{L}}{\partial \beta_1} & = & 0 \end{array} \right. \iff \left\{ \begin{array}{ccc} -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) & = & 0 \\ -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) & = & 0 \end{array} \right.$$

Posons :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad \text{et} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i .$$

Il vient :

Estimateurs des MCO

$$\begin{cases} n\bar{y} - n\beta_0 - \beta_1 n\bar{x} = 0 \\ \sum_{i=1}^n x_i y_i - \beta_0 n\bar{x} - \beta_1 \sum_{i=1}^n x_i^2 = 0 \end{cases}$$

$$\begin{cases} \beta_0 = \bar{y} - \beta_1 \bar{x} \\ \sum_{i=1}^n x_i y_i - (\bar{y} - \beta_1 \bar{x}) n\bar{x} - \beta_1 \sum_{i=1}^n x_i^2 = 0 \qquad (\star) \end{cases}$$

$$(\star) \iff \beta_1 \left(n\bar{x}^2 - \sum_{i=1}^n x_i^2 \right) + \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} = 0,$$

$$\beta_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Définition (Estimateurs des MCO)

$$\begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{cases}$$

Remarques :

- $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des variables aléatoires.
- On vérifie que la matrice Hessienne de $\mathcal{L}$ est définie positive en $(\hat{\beta}_0, \hat{\beta}_1)$, de sorte que le minimum est unique.
- Condition ?

$$\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0.$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0 \iff x_i = \bar{x} \quad \forall i = 1, \dots, n$$

D'où :

$$\begin{aligned} \textstyle\sum_{i=1}^{n}(x_i - \bar{x})^2 \neq 0 \quad &\iff \quad \text{au moins deux } x_i \text{ sont différents.} \\ &\iff \quad \mathbb{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \text{est de rang plein.} \end{aligned}$$

On supposera cette condition satisfaite par la suite.

Propriétés des estimateurs des MCO

On s'intéresse aux propriétés statistiques des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$:

- Fluctuent-ils "autour" des paramètres β_0 et β_1 ?
- Quelles sont les amplitudes de ces fluctuations ?

Rappels : Soit $\hat{\theta}$ un estimateur d'un paramètre réel θ.

- Le biais de l'estimateur est $\mathbb{E}[\hat{\theta}] - \theta$.
- $\hat{\theta}$ est dit sans biais si $\mathbb{E}[\hat{\theta}] - \theta = 0$ pour tout θ.
- On mesure les fluctuations de $\hat{\theta}$ par $\mathbb{V}(\hat{\theta})$.
- Erreur quadratique moyenne :

$$EQM = \mathbb{E}[(\hat{\theta} - \theta)^2] = \underbrace{(\mathbb{E}[\hat{\theta}] - \theta)^2}_{\text{biais au carré}} + \underbrace{\mathbb{V}(\hat{\theta})}_{\text{variance}} .$$

Propriétés des estimateurs MCO

Propriété 1 : $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont linéaires en les Y_i.

- $$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x}) Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n w_i Y_i, \quad \text{avec}$$

$$w_i = \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2}.$$

- $$\widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x} = \frac{1}{n} \sum_{i=1}^n Y_i - \bar{x} \sum_{i=1}^n w_i Y_i = \sum_{i=1}^n \left(\frac{1}{n} - \bar{x} w_i \right) Y_i.$$

Propriété 2 : $\hat{\beta}_0$ et $\hat{\beta}_1$ sont sans biais

- $$\mathbb{E}[\hat{\beta}_1] = \mathbb{E}\left[\sum_{i=1}^n w_i Y_i\right] = \sum_{i=1}^n w_i \mathbb{E}[Y_i] = \sum_{i=1}^n w_i(\beta_0 + \beta_1 x_i)$$

$$\mathbb{E}[\hat{\beta}_1] = \beta_1 \sum_{i=1}^n w_i x_i \quad \text{car } \sum_{i=1}^n w_i = 0. \text{ En outre :}$$

$$\sum_{i=1}^n w_i x_i = \sum_{i=1}^n \frac{x_i - \bar{x}}{\sum_{j=1}^n (x_j - \bar{x})^2} x_i = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2} = 1.$$

D'où $\quad \mathbb{E}[\hat{\beta}_1] = \beta_1$.

Conclusion : $\hat{\beta}_1$ est un estimateur sans biais de β_1.

Propriétés des MCO

- $$\mathbb{E}[\widehat{\beta}_0] = \mathbb{E}[\bar{Y} - \widehat{\beta}_1 \bar{x}] = \mathbb{E}[\bar{Y}] - \bar{x}\mathbb{E}[\widehat{\beta}_1] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[Y_i] - \bar{x}\beta_1.$$

$$\mathbb{E}[\widehat{\beta}_0] = \frac{1}{n}\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i) - \bar{x}\beta_1 = \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1.$$

D'où $\quad \mathbb{E}[\widehat{\beta}_0] = \beta_0.$

Conclusion : $\widehat{\beta}_0$ est un estimateur sans biais de β_0.

Propriétés des MCO

Variances de $\hat{\beta}_0$ et $\hat{\beta}_1$
Le calcul donne :

$$\mathrm{Var}(\hat{\beta}_0) = \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2,$$

$$\mathrm{Var}(\hat{\beta}_1) = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2,$$

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \sigma^2.$$

- σ^2 est la variance des bruits.
- Rappel des hypothèses sur les bruits : $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma^2$, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0 \, \forall i \neq j$.
- Il conviendra donc d'estimer la variance des bruits.
- Interprétation ?

Propriété de Gauss-Markov
Parmi tous les estimateurs de β_0 et β_1 linéaires en les Y_i ($\star$) et sans biais (†), les estimateurs des MCO $\hat{\beta}_0$ et $\hat{\beta}_1$ sont de variance minimale. On considère la classe des estimateurs $(\hat{\theta}_0, \hat{\theta}_1)$:

- ($\star$) de la forme $\hat{\theta}_0 = \sum_{i=1}^{n} \alpha_i Y_i$ et $\hat{\theta}_1 = \sum_{i=1}^{n} \gamma_i Y_i$,
- (†) tels que $\mathbb{E}[\hat{\theta}_0] = \beta_0$ et $\mathbb{E}[\hat{\theta}_1] = \beta_1$.

Estimation de la variance

Rappel : Modèle Linéaire Standard

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \forall i = 1, \dots, n.$$

Hypothèses :

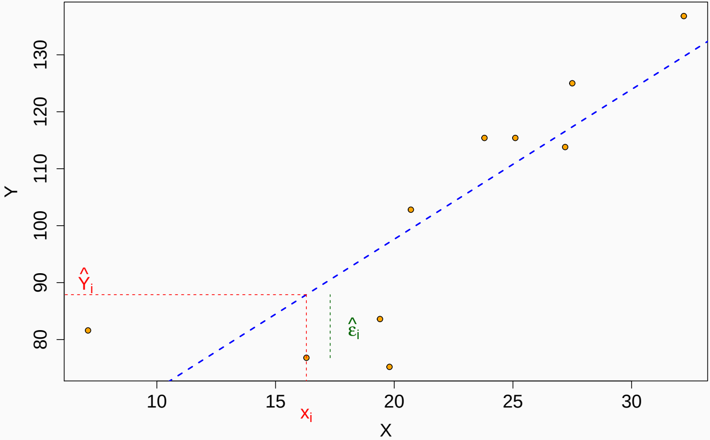
- Centrage : $\mathbb{E}[\varepsilon_i] = 0$,
- Homoscédasticité : $\mathrm{Var}(\epsilon_i) = \sigma^2, \quad \forall i$,
- Décorrélation : $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0, \quad \forall i \neq j$.

On s'intéresse à l'estimation du paramètre σ^2, la variance des bruits.Si les bruits étaient observés, un estimateur naturel serait

$$\frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2.$$

Mais ce n'est pas le cas...

Résidus et valeurs ajustées



- Droite de régression d'équation "$y = \widehat{\beta}_0 + \widehat{\beta}_1 x$".
- $i^{\text{ème}}$ valeur ajustée : $\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$.
- $i^{\text{ème}}$ résidus : $\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$.
- Résidus et valeurs ajustées sont des variables aléatoires.

Propriétés des résidus

Propriété 1

$$\begin{aligned}\mathbb{E}[\hat{\varepsilon}_i] &= \mathbb{E}[Y_i] - \mathbb{E}[\hat{Y}_i] = \beta_0 + \beta_1 x_i - \mathbb{E}[\hat{\beta}_0 + \hat{\beta}_1 x_i] \\ &= \beta_0 + \beta_1 x_i - \mathbb{E}[\hat{\beta}_0] - \mathbb{E}[\hat{\beta}_1] x_i \\ &= 0.\end{aligned}$$

Les résidus sont des variables aléatoires centrées, i.e. à moyennes nulles.

Propriété 2

$$\begin{aligned}\sum_{i=1}^n \hat{\varepsilon}_i &= \sum_{i=1}^n (Y_i - \hat{Y}_i) = \sum_{i=1}^n Y_i - \sum_{i=1}^n \hat{Y}_i \\ &= n\bar{Y} - \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i) = n\bar{Y} - n\hat{\beta}_0 - \hat{\beta}_1 n\bar{x} \\ &= n\bar{Y} - n(\bar{Y} - \hat{\beta}_1 \bar{x}) - \hat{\beta}_1 n\bar{x} = 0.\end{aligned}$$

La somme des résidus est nulle.Par conséquent les résidus ne sont pas indépendants et $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$.

Un estimateur naturel de σ^2 est la variance empirique des résidus, i.e. $\frac{1}{n}\sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2$, avec $\bar{\hat{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i$. Les résidus étant centrés, le terme de recentrage apparaît superflu. On peut donc considérer un estimateur de la forme $\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$.

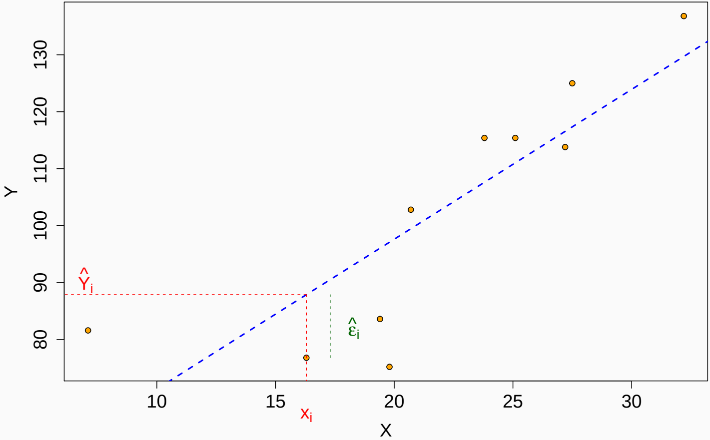
Théorème (Estimateur sans biais de σ^2)
L'estimateur $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ est un estimateur sans biais de σ^2.

Remarques :

- Normalisation par $n-2$: le modèle contient 2 paramètres.
- Comparer avec l'estimation sans biais de la variance d'un échantillon IID.

Coefficient de détermination R^2

Comment mesurer la qualité de l'ajustement ?



- Avant ajustement : $\{Y_i\}$ dispersion $\equiv \sum_{i=1}^n (Y_i - \bar{Y})^2$.
- Après ajustement : $\{\hat{Y}_i\}$ dispersion $\equiv \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$.
- Rappel : $\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i$.

$\longrightarrow$ comparons ces deux quantités.

Equation de décomposition de la variance

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\bar{Y})^2 &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i+\hat{Y}_i-\bar{Y})^2\\ &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2+\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2+\\ &\quad +2\underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\bar{Y})}_{\text{terme croisé}}\end{aligned}$$

On montre que le terme croisé est nul.D'où :

Théorème (Equation de décomposition de la variance)

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{SCT}=\underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}_{SCR}+\underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{SCE}.$$

(S)omme des (C)arrés : (T)otale, (R)ésiduelle, (E)xpliquée.

Définition (Coefficient de détermination R^2)
Le coefficient de détermination R^2 est défini par

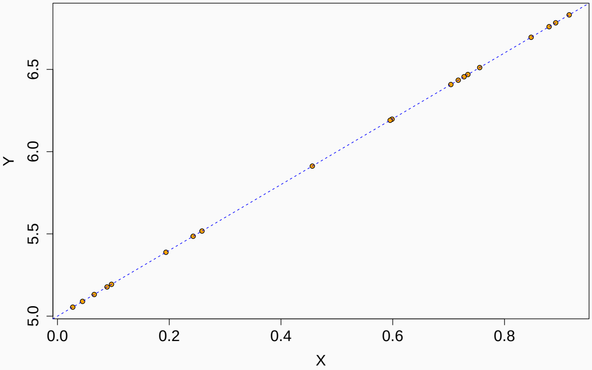
$$R^2 = \frac{\mathsf{SCE}}{\mathsf{SCT}} = 1 - \frac{\mathsf{SCR}}{\mathsf{SCT}}.$$

Remarques :

- Le R^2 représente la proportion de variance expliquée par le modèle.
- Le R^2 est compris entre 0 et 1.
- Deux cas extrêmes : $R^2 = 1$ et $R^2 = 0$.

Cas $R^2 = 1$

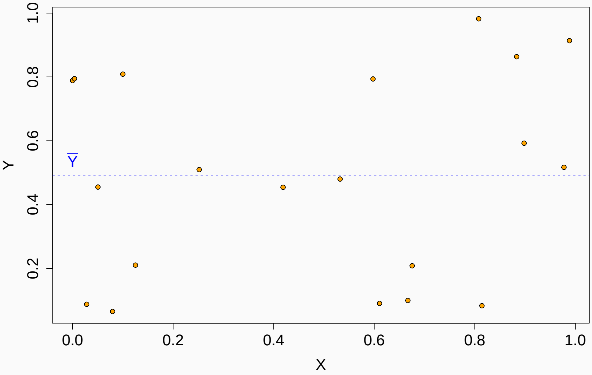
$$R^2 = 1 \iff \frac{\text{SCE}}{\text{SCT}} = 1 \iff \text{SCE} = \text{SCT} \iff \text{SCR} = 0$$
$$\iff \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = 0 \iff Y_i = \hat{Y}_i, \quad \forall i = 1, \ldots, n.$$



Ajustement parfait : tous les couples (x_i, Y_i) sont situés sur la droite de régression.

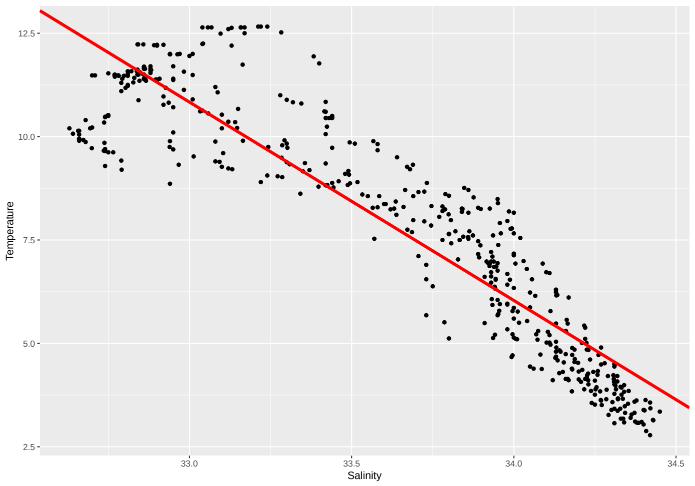
Cas $R^2 = 0$

$$R^2 = 0 \iff \text{SCE} = 0 \iff \sum_{i=1}(\hat{Y}_i - \bar{Y})^2 = 0 \iff \hat{Y}_i = \bar{Y},\ \forall i.$$



Ajustement médiocre : X n'influence pas Y. La droite de régression est horizontale.

Exemple : Salinité et Température CalCOFI



- $R^2 = 0.8517.$
- 85.17% de la variance est expliquée par le modèle.

Une propriété

Théorème (R^2 et coefficient de corrélation)

$$R^2 = \rho_{xy}^2 \quad \text{où :}$$

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right]^{\frac{1}{2}}}.$$

Terme croisé : Preuve de $\sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = 0$

- $$\sum_{i=1}^n (Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) = \sum_{i=1}^n (Y_i - \hat{Y}_i)\hat{Y}_i - \bar{Y} \underbrace{\sum_{i=1}^n (Y_i - \hat{Y}_i)}_{=0} = \sum_{i=1}^n \hat{\varepsilon}_i \hat{Y}_i.$$

$$= \sum_{i=1}^n (\hat{\beta}_0 + \hat{\beta}_1 x_i)\hat{\varepsilon}_i = \hat{\beta}_0 \underbrace{\sum_{i=1}^n \hat{\varepsilon}_i}_{=0} + \hat{\beta}_1 \sum_{i=1}^n x_i \hat{\varepsilon}_i = \hat{\beta}_1 \sum_{i=1}^n x_i \hat{\varepsilon}_i.$$

- $$\sum_{i=1}^n x_i \hat{\varepsilon}_i = \sum_{i=1}^n x_i Y_i - \sum_{i=1}^n x_i \hat{Y}_i = \sum_{i=1}^n x_i Y_i - \sum_{i=1}^n x_i (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

$$= \sum_{i=1}^n x_i Y_i - \bar{Y} \sum_{i=1}^n x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2$$

$$= \sum_{i=1}^n x_i (Y_i - \bar{Y}) - \hat{\beta}_1 \sum_{i=1}^n x_i (x_i - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) - \hat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})^2.$$

Terme croisé : Preuve de $\sum_{i=1}^n (Y_i - \widehat{Y}_i)(\widehat{Y}_i - \bar{Y}) = 0$

En remplaçant $\widehat{\beta}_1$ par son expression, il vient :

$$\underbrace{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}_{(\star)} - \underbrace{\widehat{\beta}_1 \sum_{i=1}^n (x_i - \bar{x})^2}_{=(\star)} = 0.$$

Donc

$$\sum_{i=1}^n x_i \widehat{\varepsilon}_i = 0$$

puis

$$\sum_{i=1}^n (Y_i - \widehat{Y}_i)(\widehat{Y}_i - \bar{Y}) = 0.$$

Interprétation géométrique

Rappel : Ecriture matricielle du modèle

$$\mathbf{Y} = \mathbb{X}\beta + \varepsilon \quad \text{avec :}$$

- $\mathbf{Y} = (Y_1, \dots, Y_n)'$: vecteur aléatoire de $\mathbb{R}^n$.
- $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$: vecteur aléatoire de $\mathbb{R}^n$.
- $\beta = (\beta_0, \beta_1)'$: vecteur des paramètres.
- Matrice $n \times 2$ (du plan d'expérience/de design) déterministe :

$$\mathbb{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}.$$

Hypothèses :

$$\mathbb{E}[\varepsilon] = 0 \quad \text{et} \quad \mathbb{V}(\varepsilon) = \sigma^2 \mathbf{I}_n.$$

Vecteur des résidus

On a montré que :

1. $\sum_{i=1}^{n} \hat{\varepsilon}_i = 0.$
2. $\sum_{i=1}^{n} x_i \hat{\varepsilon}_i = 0.$

Soit le vecteur des résidus $\hat{\varepsilon} = (\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n)'$.Alors

1. $\iff$ $\hat{\varepsilon}$ est orthogonal (dans $\mathbb{R}^n$) à $(1, \dots, 1)'$.
2. $\iff$ $\hat{\varepsilon}$ est orthogonal (dans $\mathbb{R}^n$) à $(x_1, \dots, x_n)'$.

Donc $\hat{\varepsilon}$ est orthogonal à $\mathrm{Im}(\mathbb{X})$, sous-espace vectoriel de $\mathbb{R}^n$ engendré par les colonnes de $\mathbb{X}$.

Soit $\hat{Y} = (\hat{Y}_1, \dots, \hat{Y}_n)'$ le vecteur des valeurs ajustées.

$$\hat{Y} = \hat{\beta}_0 \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} + \hat{\beta}_1 \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \mathbb{X}\hat{\beta}.$$

Donc $\hat{Y}$ est à valeurs dans $\mathrm{Im}(\mathbb{X})$.

Conclusion : $\hat{Y}$ est le projeté orthogonal de Y sur $\mathrm{Im}(\mathbb{X})$

Equation de décomposition de la variance

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{SCT}} = \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{SCR}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{SCE}} .$$

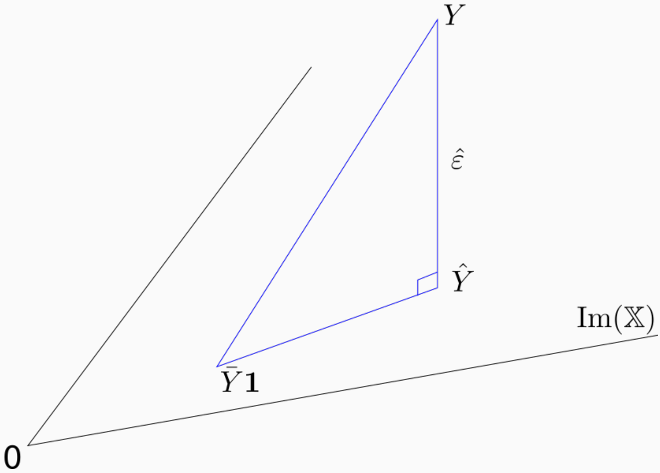
En écriture vectorielle, avec $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^n$:

- SCT $= \|Y - \bar{Y}\mathbf{1}\|^2$.
- SCE $= \|\hat{Y} - \bar{Y}\mathbf{1}\|^2$.
- SCR $= \|\hat{\varepsilon}\|^2 = \|Y - \hat{Y}\|^2$.

Remarques :

- Le vecteur $\bar{Y}\mathbf{1}$ appartient à $\mathrm{Im}(\mathbb{X})$.
- Le vecteur $\hat{Y} - \bar{Y}\mathbf{1}$ appartient à $\mathrm{Im}(\mathbb{X})$.
- $\hat{\varepsilon}$ est orthogonal à $\hat{Y} - \bar{Y}\mathbf{1}$.

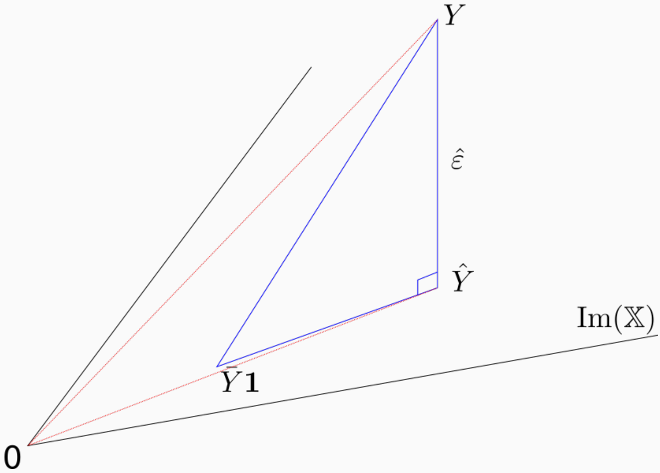
Equation de décomposition de la variance



Equation de décomposition de la variance $\iff$ Pythagore

$$\|Y - \bar{Y}\mathbf{1}\|^2 = \|\hat{Y} - \bar{Y}\mathbf{1}\|^2 + \|\hat{\varepsilon}\|^2$$

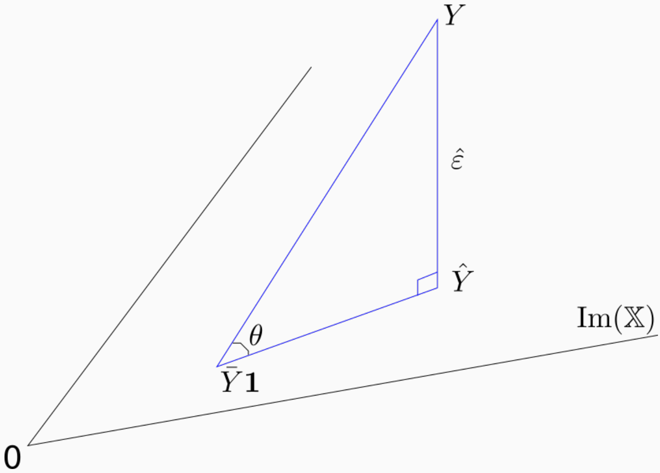
Pythagore



On a aussi

$$\|Y\|^2 = \|\widehat{Y}\|^2 + \|\widehat{\varepsilon}\|^2.$$

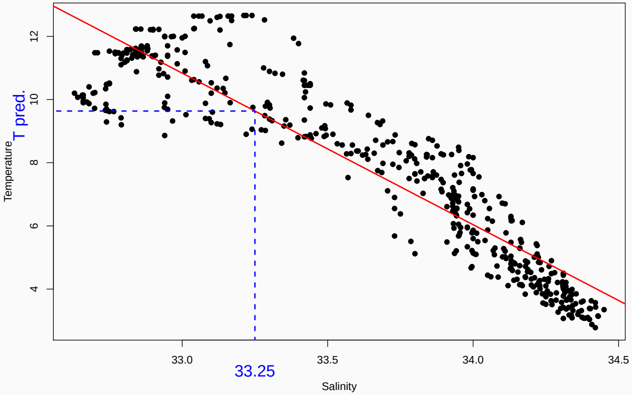
Coefficient de détermination



$$R^2 = \frac{\mathsf{SCE}}{\mathsf{SCT}} = \frac{\|\hat{Y} - \bar{Y}\mathbf{1}\|^2}{\|Y - \bar{Y}\mathbf{1}\|^2} = \cos^2(\theta).$$

Prévision

Exemple : Salinité et Température CalCOFI



- Quelle valeur de température prédire pour une nouvelle observation de salinité égale à 33.25 ?
- L'utilisation du modèle est-elle pertinente ?

Modélisation pour une nouvelle observation.

Rappel : Modèle linéaire standard

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \forall i = 1, \dots, n, \\
\mathbb{E}[\varepsilon_i] &= 0, \\
\mathrm{Var}(\epsilon_i) &= \sigma^2, \quad \forall i, \\
\mathrm{Cov}(\epsilon_i, \epsilon_j) &= 0, \quad \forall i \neq j.
\end{aligned}
$$

Nouvelle observation x_{n+1} :

- Modèle : $Y_{n+1} = \beta_0 + \beta_1 x_{n+1} + \varepsilon_{n+1}$.
- Y_{n+1} n'est pas observée : on souhaite la prédire.
- Hypothèses sur les bruits :

$$\mathbb{E}[\varepsilon_{n+1}] = 0, \quad \mathrm{Var}(\varepsilon_{n+1}) = \sigma^2, \quad \mathrm{Cov}(\varepsilon_{n+1}, \varepsilon_i) = 0, \ \forall i = 1, \dots, n.$$

Valeur prédite et erreur de prévision

- Utilisant le modèle, la prévision naturelle est :

$$Y_{n+1}^{(p)} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{n+1}.$$

- à laquelle est asssociée l'erreur de prévision :

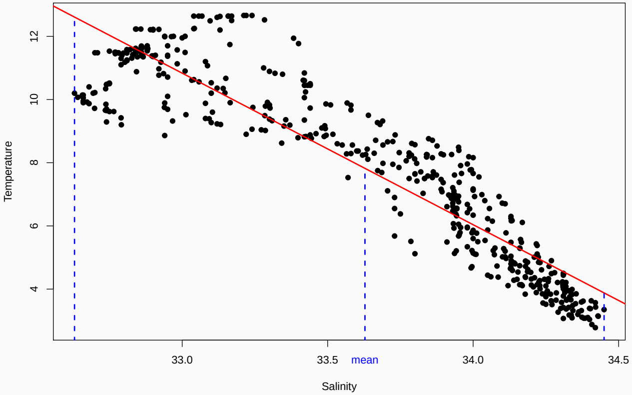
$$\widehat{\varepsilon}_{n+1}^{(p)} = Y_{n+1} - \widehat{Y}_{n+1}^{(p)}.$$

L'erreur de prévision satisfait :

- $\mathbb{E}[\widehat{\varepsilon}_{n+1}^{(p)}] = 0.$
- $$\mathrm{Var}(\widehat{\varepsilon}_{n+1}^{(p)}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).$$

Remarque : La variance augmente lorsque (i) σ^2 augmente, et (ii) lorsque x_{n+1} s'éloigne de $\bar{x}$.

Exemple : Salinité et Température CalCOFI



- $n = 493$
- Nouvelles observations : minimum, moyenne, et maximum.