# Modèles Linéaires

# Correction TD 1 : Régression linéaire simple
# Licence 3 MIASHS - Informatique

**Guillaume Metzler, Francesco Amato, Alejandro Rivera**
**Institut de Communication (ICOM)**
**Université de Lyon, Université Lumière Lyon 2**
**Laboratoire ERIC UR 3083, Lyon, France**
guillaume.metzler@univ-lyon2.fr ; francesco.amato@univ-lyon2.fr
alejandro.rivera@univ-lyon2.fr

**Résumé**

Cette première séance aborde généralement le modèle linéaire gaussien simple, et plus précisément les points suivants :

- Rappel sur les hypothèses du modèle linéaire gaussien,
- Estimation des paramètres du modèle par MCO, dans le cas modèle linéaire simple,
- Ecriture du modèle sous forme matricielle
- Estimation par maximum de vraisemblance

## Modèle de régression simple

On rappelle que le modèle linéaire gaussien simple s'écrit sous la forme

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

où $Y$ est la variable à expliquer, $X$ est la variable qui explique et $\varepsilon$ est une variable aléatoire représentant les erreurs du modèle.

Dans ce TD, on se concentre sur l'estimation des paramètres du modèle à l'aide de plusieurs méthodes différentes et nous appliquerons ensuite cela sur les données présentées ci-dessous, pour obtenir la droite de régression associée, présentée en Figure 1.

| $Y$ : Score examen 2 | 3.5 | 4 | 5 | 1 | 2 | 1.5 | 2.5 | 5.5 | 6 | 6.5 |
|---|---|---|---|---|---|---|---|---|---|---|
| $X$ : Score examen 1 | 4 | 3 | 3.5 | 1 | 1.5 | 1 | 1.5 | 4 | 3.5 | 4.5 |

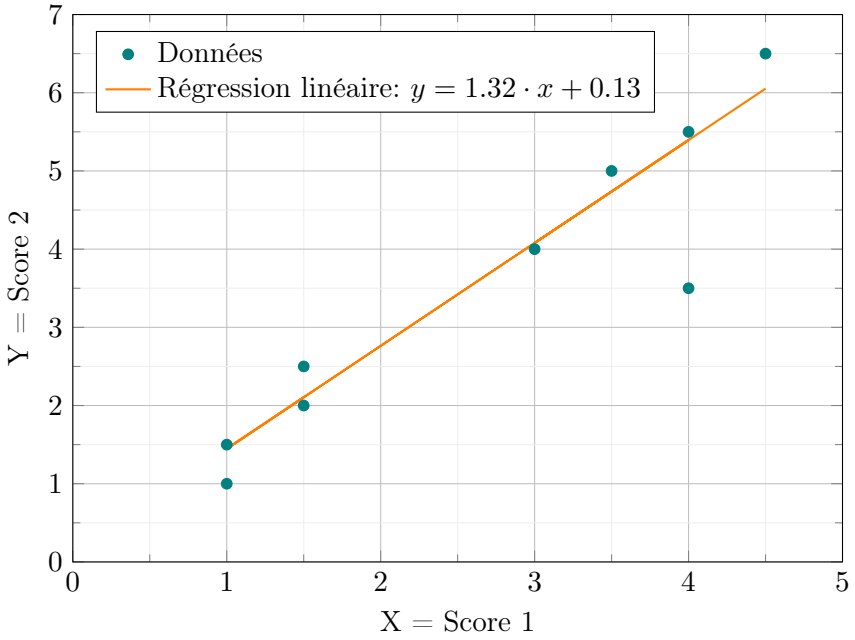


FIGURE 1 – Application de la régression linéaire simple gaussien sur les données présentées dans la table associée. On cherche alors à expliquer le score obtenu au deuxième examen en fonction du score obtenu au premier examen.

1. Rappeler les hypothèses du modèle de régression linéaire gaussien.

On rappelle les hypothèses du modèle linéaire simple gaussien

- $(y_i, x_i)_{i=1}^n$ doivent être *i.i.d*, *i.e.* indépendantes et identiquement distribuées,
- $y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$ : lien entre $y_i, x_i$. On suppose que les $x_i$ sont déterministes.
- $\varepsilon_i \underset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$ : hypothèse d'homoscédasticité des résidus.

## Estimation par méthode des moindres carrés

La méthode des moindres carrés ordinaires consiste à minimiser le carré des résidus du modèle de régression, *i.e.* minimiser l'écart quadratique entre la prédiction effectuée par le modèle $\hat{\beta}_0 + \hat{\beta}_1 x_i$ et la valeur observée $y_i$.
Pour cela on disposera d'un échantillon de données $S = \{(x_i, y_i)\}_{i=1}^m$ et on s'intéresse au problème de minimisation :

$$\min_{\beta_0,\beta_1 \in \mathbb{R}} \sum_{i=1}^n \varepsilon_i^2 = \min_{\beta_0,\beta_1 \in \mathbb{R}} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \min_{\beta_0,\beta_1 \in \mathbb{R}} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{1}$$

2. Dans la résolution de ce problème, l'hypothèse de normalité sur les résidus $\varepsilon_i$ est-elle importante ? colorblue

   *Démonstration.* Non, c'est pas importante. □

3. Montrer que les solutions du problème 1 sont données par

$$\hat{\beta}_1 = \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[X]} \quad \text{et} \quad \hat{\beta}_0 = \mathbb{E}[Y] - \hat{\beta}_1\,\mathbb{E}[X].$$

   Notons $L$ la fonction que l'on cherche à optimiser, définie par :

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

   Cette dernière est convexe en les variables $\beta_1$ et $\beta_0$, elle admet donc une unique solution. Cette solution est obtenue en résolvant l'équation d'*Euler* se présentant sous la forme d'un système linéaire

$$\frac{\partial L}{\partial \beta_1} = 0 \iff -2\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)x_i = 0, \tag{2}$$

$$\frac{\partial L}{\partial \beta_0} = 0 \iff -2\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0) = 0. \tag{3}$$

   On se concentre sur l'équation (3) pour le moment, on va la développer ce qui nous permet d'écrire

$$\begin{aligned}
& -2\textstyle\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0) && = 0 \\
\iff\ & \textstyle\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i - n\beta_0 && = 0 \\
\iff\ & \frac{1}{n}\textstyle\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i && = \beta_0 \\
\iff\ & \bar{y} - \beta_1 \bar{x} && = \beta_0
\end{aligned}$$

   On obtient une expression de l'estimateur $\hat{\beta}_0$ de $\beta_0$, elle dépend de $a$ dont on va pouvoir déterminer l'expression en injectant la valeur de $\beta_0$ dans l'équation (2)

$$\begin{aligned}
& -2\textstyle\sum_{i=1}^{n}(y_i - \beta_1 x_i - \beta_0)x_i && = 0 \\
\iff\ & \textstyle\sum_{i=1}^{n} y_i x_i - \beta_1 \sum_{i=1}^{n} x_i^2 - \beta_0 \sum_{i=1}^{n} x_i && = 0
\end{aligned}$$

$$
\begin{aligned}
\iff \quad & \textstyle\sum_{i=1}^n y_i x_i - \beta_1 \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \left( \frac{1}{n} \sum_{i=1}^n y_i - \beta_1 \frac{1}{n} \sum_{i=1}^n x_i \right) && = 0 \\
\iff \quad & \textstyle\sum_{i=1}^n y_i x_i - \frac{1}{n} \left(\sum_{i=1}^n x_i\right) \left(\sum_{i=1}^n y_i\right) - \beta_1 \left( \sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2 \right) && = 0 \\
\iff \quad & \frac{\sum_{i=1}^n y_i x_i - \frac{1}{n} \left(\sum_{i=1}^n x_i\right) \left(\sum_{i=1}^n y_i\right)}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left(\sum_{i=1}^n x_i\right)^2} && = \beta_1 \\
\iff \quad & \frac{Cov[x,y]}{Var[x]} && = \beta_1
\end{aligned}
$$

**Rappel** Considérons $X$ et $Y$ des variances aléatoires qui admettent des moments d'ordre 2, *i.e.* qui admettent une variance. Alors,

(i) concernant la variance d'une variable aléatoire, nous avons la **Formule de Koenig-Huygens** :

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 .$$

(ii) la covariance entre deux variables aléatoires est aussi égale à

$$Cov(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y].$$

Ce qui achève la démonstration.

4. Effectuer l'application numérique avec les données de l'énoncé.

```
### DONNEES
y = c(3.5, 4, 5, 1, 2, 1.5, 2.5, 5.5, 6, 6.5)
x = c(4, 3, 3.5, 1, 1.5, 1, 1.5, 4, 3.5, 4.5)

### METHODE DES MOINDRES CARRES: FORME UNIVARIEE

beta_1 <- cov(x,y)/var(x)
beta_0 <- mean(y) - beta_1*mean(x)
print(paste("Le coefficient directeur de ma droite est ",
round(beta_1,2)))
```

```
## [1] "Le coefficient directeur de ma droite est  1.32"
```

```
print(paste("L'ordonnée à l'origine de ma droite est ",
round(beta_0,2)))
```

```
## [1] "L'ordonnée à l'origine de ma droite est  0.13"
```

Cette façon de résoudre le problème convient très bien lorsque la modélisation implique une seule variable explicative mais est moins pratique lorsque l'on dispose de plusieurs descripteurs. Il convient alors de passer sous une écriture matricielle de ce même problème de minimisation.

5. Montrer que le problème d'optimisation 1 peut se réécrire sous la forme

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^2} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2,$$

où $\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}$ sont des objets dont on explicitera les définition et dimension.

Pour $n$ observations, on peut écrire le modèle de régression linéaire multiple sous la forme :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{pour } i = 1, \ldots, n$$

On peut condenser la notation en écrivant le modèle matriciellement de la manière suivante :

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

où

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{n-1} \\ 1 & x_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{et} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{pmatrix},$$

et $\boldsymbol{\beta} = (\beta_0, \beta_1) \in \mathbb{R}^2$ est notre vecteur des paramètres du modèle. Le vecteur $\mathbf{y} \in \mathbb{R}^n$ est le vecteur dont on cherche à expliquer les valeurs, la matrice $\mathbf{X} \in \mathbb{R}^{n\times 2}$ est la matrice explicative. $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ est le vecteur des erreurs associées à chaque exemple qui sont supposées gaussiennes.

On peut noter que

$$\varepsilon = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}.$$

Rappelant que la norme 2 (ou norme euclidienne) d'un vecteur $\varepsilon$, qui se note à l'aide de deux barres, est défini comme

$$\|\varepsilon\|_2 = \sqrt{\varepsilon_1^2 + \ldots + \varepsilon_n^2}, \quad \text{et donc} \quad \|\varepsilon\|_2^2 = \varepsilon_1^2 + \ldots + \varepsilon_n^2$$

alors le problème à minimiser

$$\min_{\beta\in\mathbb{R}^2}\sum_{i=1}^{n}\varepsilon_i^2 = \min_{\beta\in\mathbb{R}^2}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \min_{\beta\in\mathbb{R}^2}\sum_{i=1}^{n}(y_i-(\beta_0+\beta_1 x_i))^2$$

revient à écrire

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^2}\|\varepsilon\|_2^2 = \min_{\boldsymbol{\beta}\in\mathbb{R}^2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2,$$

6. Montrer que ce problème d'optimisation est convexe.

   Avant de trouver un estimateur, il faut expliquer pourquoi ce problème a une unique solution.
   On a

$$\begin{aligned}
\min_{\boldsymbol{\beta}\in\mathbb{R}^2}\|Y-\hat{Y}\|_2^2 &= \min_{\boldsymbol{\beta}\in\mathbb{R}^2}\|Y-(X\boldsymbol{\beta})\|_2^2,\\
&\quad\downarrow \text{ on se rappelle que } \hat{Y}=\boldsymbol{\beta}X\\
&= \min_{\boldsymbol{\beta}\in\mathbb{R}^2}\|Y-(\beta_0+\boldsymbol{\beta}_1\mathbf{x})\|_2^2,\\
&\quad\downarrow \text{ on se rappelle que pour tout vecteur } \mathbf{x},\ \langle\mathbf{x},\mathbf{x}\rangle=\mathbf{x}^T\mathbf{x}=\langle\mathbf{x}\mid\mathbf{x}\rangle=\|\mathbf{x}\|_2^2\\
&= \min_{\boldsymbol{\beta}\in\mathbb{R}^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2.
\end{aligned}$$

   On voit bien que cette fonction est convexe. En effet, on se rappelle qu'une fonction $x\mapsto(a-x)^2$ est une fonction *quadratique* donc convexe, et la somme de fonctions convexes reste convexe.

7. Déterminer les solutions du problème et préciser à quelle condition cette solution existe.

   La fonction étant convexe, ce que l'on peut vérifier en calculant la matrice hessienne, les minima de cette fonction sont données en cherchant l'endroit où la dérivée de la fonction $\boldsymbol{\beta}\mapsto\|Y-X\boldsymbol{\beta}\|_2^2$ s'annule. On va donc chercher les valeurs de $\beta$ telles que

$$\frac{\partial}{\partial\boldsymbol{\beta}}\|Y-X\boldsymbol{\beta}\|_2^2=0 \iff -2X^T(Y-X\boldsymbol{\beta})=0 \tag{4}$$

   En dérivant à nouveau fonction, on trouve

$$\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\|Y-X\boldsymbol{\beta}\|_2^2 = 2X^TX \succ 0,$$

*i.e.* la matrice hessienne est définie positive, ce qui est le cas ici car il s'agit de la matrice de variance-covariance des données, cette convexité nous permettra de dire que la vecteur $\beta$ vérifiant l'équation (4) est bien solution de notre problème de minimisation. Or

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}}\|Y - X\boldsymbol{\beta}\|_2^2 = 0 &\iff -2X^T(Y - X\boldsymbol{\beta}) = 0, \\
&\downarrow \text{ on peut diviser par } -2 \\
&\iff X^T(Y - X\boldsymbol{\beta}) = 0, \\
&\iff X^TY - X^TX\boldsymbol{\beta} = 0, \\
&\iff X^TX\boldsymbol{\beta} = X^TY, \\
&\downarrow \text{ à condition que la matrice } X^TX \text{ soit inversible} \\
&\iff \boldsymbol{\beta} = (X^TX)^{-1}X^TY.
\end{aligned}$$

8. Effectuer l'application numérique avec les données de l'énoncé.

```
### METHODE DES MOINDRES CARRES: FORME MATRICIELLE

# On peut calculer les coefficients en utilisant la formule matricielle.
# Avant de faire ça, il faut créer la matrice X:

X = cbind(rep(1,length(x)),x)

head(X)
```

```
##          x
## [1,] 1 4.0
## [2,] 1 3.0
## [3,] 1 3.5
## [4,] 1 1.0
## [5,] 1 1.5
## [6,] 1 1.0
```

```
# En applicant la fomule
beta = solve(t(X)%*%X)%*%t(X)%*%y

print(paste("Le coefficient directeur de ma droite est ",
round(beta[2,],2)))
```

```
## [1] "Le coefficient directeur de ma droite est  1.32"
```

```
print(paste("L'ordonnée à l'origine de ma droite est ",
round(beta[1,] ,2)))
```

```
## [1] "L'ordonnée à l'origine de ma droite est  0.13"
```

Etant donnée l'hypothèse formulée sur la distribution de la variable aléatoire $Y$, nous pouvons également estimer les paramètres du modèle par maximum de vraisemblance.

### Estimation par maximum de vraisemblance

Les données $y_i$ suivent une distribution de normale de moyenne $\beta_0 + \beta_1 x_i$ et de variance inconnue $\sigma^2$.

1. Donner la valeur de la densité de la variable aléatoire $Y$.

   En sachant que :

   $$Y = \beta_0 + \beta_1 X + \varepsilon = X\boldsymbol{\beta} + \varepsilon$$

   où $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ est une variable normale avec une moyenne nulle et une variance de $\sigma^2$.

   On rappelle (c'est une des hypothèses du modèle linéaire gaussien) que les données $y_i$ sont distribuées selon une loi gaussienne, *i.e.* $y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. La densité d'une loi gaussienne est donnée par :

   $$f(t; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e \left( -\frac{(t-\mu)^2}{2\sigma^2} \right).$$

   Donc, pour $y_i$ la densité est :

   $$f(y_i; \beta_0 + \beta_1 x_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e \left( -\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2} \right).$$

   Qui en utilisant l'écriture matricielle peut être écrite :

   $$f(y_i; \mathbf{x}_i^T\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e \left( -\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2} \right).$$

2. Si on considère à nouveau notre échantillon $S$, donner l'expression de la vraisemblance de cet échantillon.

Considérons un échantillon *i.i.d.* $S = \{(y_i, \mathbf{x}_i)\}_{i=1}^n$ de $n$ exemples.
La vraisemblance d'un échantillon *i.i.d* est défini comme le produit des valeurs de la densité en chaque valeurs de l'échantillon et la densité d'une loi normale de moyenne $\mu$ et de variance $\sigma^2$ est donnée par Ainsi la vraisemblance $L$ de notre échantillon $S$ est donnée par

$$
\begin{aligned}
L(S, \boldsymbol{\beta}, \sigma) &= \prod_{i=1}^{n} f(S, \boldsymbol{\beta}, \sigma), \\
&\quad \downarrow \text{ densité de la loi gaussienne} \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right)}, \\
&\quad \downarrow \text{ propriété de l'exponentielle} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\left(-\sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}\right)}, \\
&\quad \downarrow \text{ or } \sum_{i=1}^{n} \frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2} = \frac{\|Y - X\boldsymbol{\beta}\|_2^2}{2\sigma^2} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{\left(-\frac{\|Y - X\boldsymbol{\beta}\|_2^2}{2\sigma^2}\right)},
\end{aligned}
$$

où les notations $X$ et $Y$ sont celles employées dans la partie précédente.

3. Estimer les valeurs de $\beta_0$ et de $\beta_1$ par maximum de vraisemblance.

   On se rappelle que les meilleurs paramètres sont ceux qui permettent de maximiser la vraisemblance de nos données. Mais cette expression est bien trop complexe à manipuler. Donc au lieu de maximiser la vraisemblance $L$ on va chercher à maximiser la log-vraisemblance $\ell$, définie par $L = \ln(L)$ soit

$$\ell(S, \boldsymbol{\beta}, \sigma) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{\|Y - X\boldsymbol{\beta}\|_2^2}{2\sigma^2}.$$

   Les valeurs de $\sigma^2$ et $\boldsymbol{\beta}$ qui maximisent la vraisemblance sont données en résolvant le système défini par

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}}(S, \sigma^2, \boldsymbol{\beta}) = 0 \quad \text{et} \quad \frac{\partial \ell}{\partial \sigma^2}(S, \sigma^2, \boldsymbol{\beta}) = 0.$$

   Concentrons nous sur la première équation, ce qui nous donne

$$
\begin{aligned}
&\frac{\partial \ell}{\partial \boldsymbol{\beta}}(S, \sigma^2, \boldsymbol{\beta}) = 0, \\
&\quad \downarrow \text{ on dérive notre norme} \\
\iff &-\frac{X^T(Y - X\boldsymbol{\beta})}{2\sigma^2} = 0, \\
\iff &-X^TY - X^TX\boldsymbol{\beta} = 0, \\
&\quad \downarrow \text{ on isole le vecteur } \boldsymbol{\beta} \\
\iff &\boldsymbol{\beta} = \left(X^TX\right)^{-1} X^TY.
\end{aligned}
$$

4. Estimer la variance des résidus $\sigma^2$ à l'aide de cette même méthode.

   Faisons de même avec la deuxième équation en utilisant l'estimateur $\hat{\boldsymbol{\beta}}$ précédemment obtenu et en notant que

   $$\|Y - X\hat{\boldsymbol{\beta}}\|_2^2 = \|Y - \hat{Y}\|_2^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = RSS.$$

   où $RSS$ signifie *Residual Sum of Squares*. On a ainsi,

   $$
   \begin{aligned}
   &\frac{\partial \ell}{\partial \sigma^2}(S, \sigma^2, \boldsymbol{\beta}) = 0, \\
   &\quad \downarrow \text{ on en utilisant les notations précédentes} \\
   \iff &-\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{RSS}{(\sigma^2)^2} = 0, \\
   &\quad \downarrow \text{ en multipliant par } 2\sigma^2 \\
   \iff &-n\sigma^2 + RSS = 0, \\
   &\quad \downarrow \text{ on isole } \sigma^2 \\
   \iff &\sigma^2 = \frac{RSS}{n}.
   \end{aligned}
   $$

   Ainsi les estimateurs obtenus par maximum de vraisemblance sont donnés par

   $$\hat{\boldsymbol{\beta}} = \left(X^TX\right)^{-1} X^TY \quad \text{et} \quad \hat{\sigma^2} = \frac{1}{n}\sum_{i=1}(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}}) = \frac{RSS}{n}.$$

   **Remarque** L'estimateur de $\sigma^2$ ainsi obtenu est biaisé ! Il faudrait corriger cet estimateur pour le rendre non biaisé.