Chapitre I : Régression linéaire simple

M. Champion

Université de Paris

2020-2021

I. Le modèle de régression linéaire simple : généralités

Un modèle linéaire simple :

- permet de **décrire** et **modéliser** la relation entre une variable aléatoire quantitative continue Y et une variable quantitative contrôlée (non aléatoire) X,

- utilise des observations $(x_i, y_i)_{i=1,\cdots,n}$ d'un échantillon de taille n où :
  - $x_1, \cdots, x_n$: valeurs connues et fixées (non aléatoires) de X,
  - $y_1, \cdots, y_n$: réponses obtenues considérées comme n réalisations de Y.

I. Le modèle de régression linéaire simple :

On appelle **modèle linéaire simple gaussien** un modèle statistique qui peut s'écrire sous la forme :

$$Y = \alpha X + \beta + \varepsilon,$$

où :

- Y est une variable aléatoire que l'on observe et que l'on souhaite expliquer et/ou prédire,
- X est la variable explicative (non aléatoire),
- ε est l'erreur résiduelle aléatoire,
- α et β sont les paramètres réels inconnus du modèle à estimer.

I. Le modèle de régression linéaire simple :

On appelle **modèle linéaire simple gaussien** un modèle statistique qui peut s'écrire sous la forme :

$$Y = \alpha X + \beta + \varepsilon,$$

où :

- Y est une variable aléatoire que l'on observe et que l'on souhaite expliquer et/ou prédire,
- X est la variable explicative (non aléatoire),
- ε est l'erreur résiduelle aléatoire,
- α et β sont les paramètres réels inconnus du modèle à estimer.

Remarque : 1. L'estimation de α et β est basée sur n observations de (X, Y) réalisées sur n individus supposés indépendants.
2. α est la pente de la droite de régression linéaire et β l'ordonnée à l'origine.

I. Le modèle de régression linéaire simple : hypothèses

On appelle **modèle linéaire simple gaussien** un modèle statistique qui peut s'écrire sous la forme :

$$\forall i = 1, \cdots, n \quad Y_i = \alpha x_i + \beta + \varepsilon_i,$$

où les erreurs ε_i sont des observations **indépendantes** d'une variable aléatoire ε distribuée suivant une loi $\mathcal{N}(0, \sigma^2)$.

Hypothèses

- $(\varepsilon_1, ..., \varepsilon_n)$ indépendants
- $\forall i = 1, \cdots, n$, $\mathbb{E}(\varepsilon_i) = 0$ et $Var(\varepsilon_i) = \sigma^2$
- $\forall i = 1, \cdots, n$, $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

I. Le modèle de régression linéaire simple : remarque

Le modèle

$$\forall i = 1, \cdots, n, \quad Y_i = \alpha x_i + \beta + \varepsilon_i, \quad \varepsilon_i \text{ iid}, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

est équivalent à

$$\forall i = 1, \cdots, n, \quad \text{les } Y_i \text{ sont indépendants, } Y_i \sim \mathcal{N}(\alpha x_i + \beta; \sigma^2).$$

Ce qui implique que :

- $\mathbb{E}(Y_i) = \alpha x_i + \beta$; $Var(Y_i) = \sigma^2$,
- X n'influe que sur la moyenne et pas sur la variance de Y,
- Y_i se décompose en
  - une partie fixe : $\alpha x_i + \beta$ expliquée par le modèle,
  - une partie aléatoire : ε_i qui reste non expliquée.

I. Le modèle de régression linéaire simple : dimension

Trois paramètres sont inconnus et à estimer :

- les **paramètres d'espérance** α et β (reliés à l'espérance de Y)
- le **paramètre de variance** σ^2 (relié à la variance de Y)

La dimension du modèle est

- la dimension de l'espace dans lequel vit l'**espérance** des variables aléatoires Y_i,
- le nombre de paramètres d'espérance envisagés dans la modélisation moins le nombre de contraintes d'identifiabilité nécessaires à l'estimation des paramètres.

$\rightarrow$ Le modèle de régression linéaire simple est de dimension 2.

II. Estimation des paramètres : exemple

Pour étudier la relation qui peut exister entre le rendement de blé et la quantité d'engrais utilisée dans une région donnée, des agronomes recueillent sur $n = 7$ parcelles :

- la quantité d'engrais x_i (en quintaux) utilisée sur la $i$ème parcelle,
- le rendement de blé y_i (en kg) correspondant mesuré.

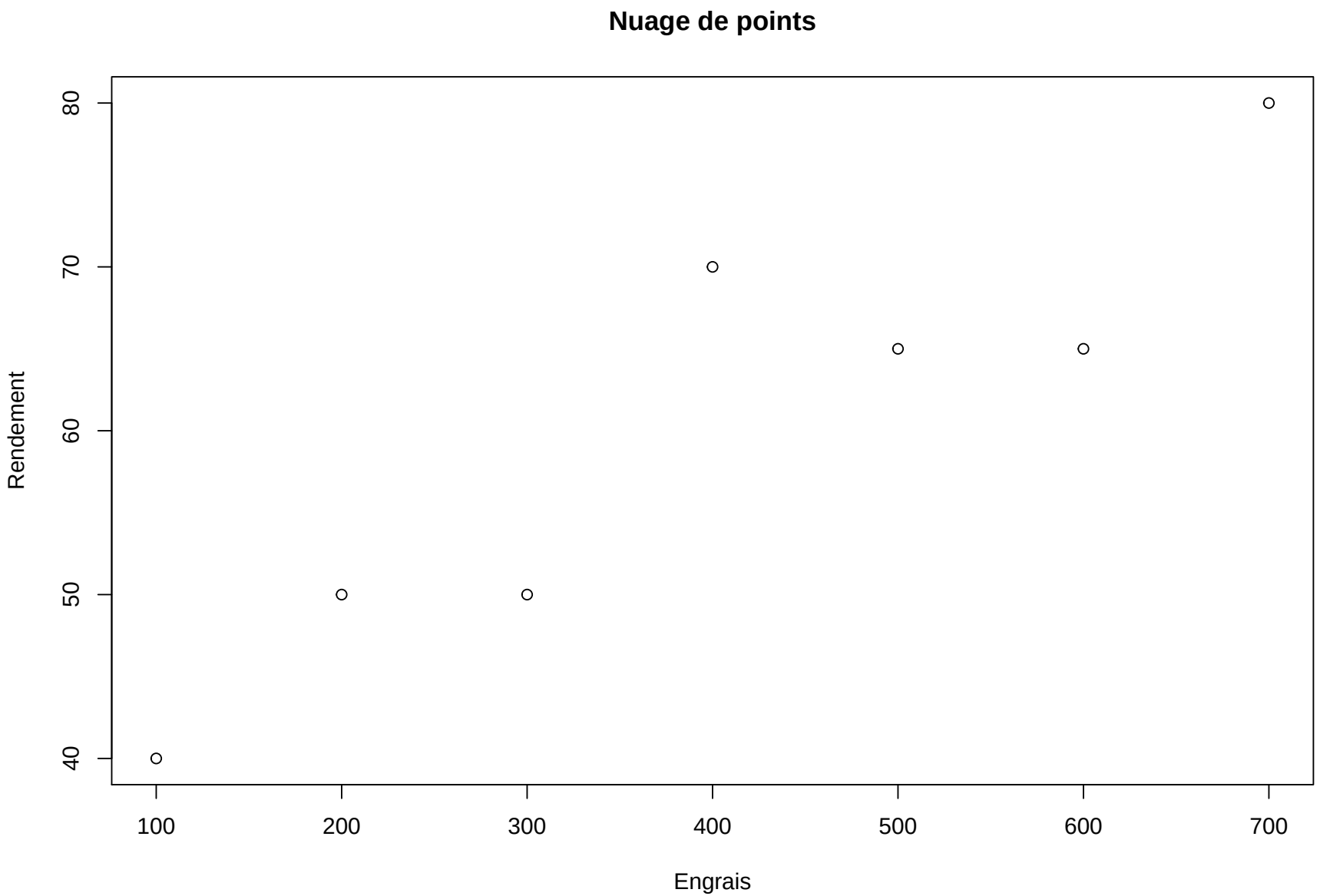
x_i	100	200	300	400	500	600	700
y_i	40	50	50	70	65	65	80

On donne

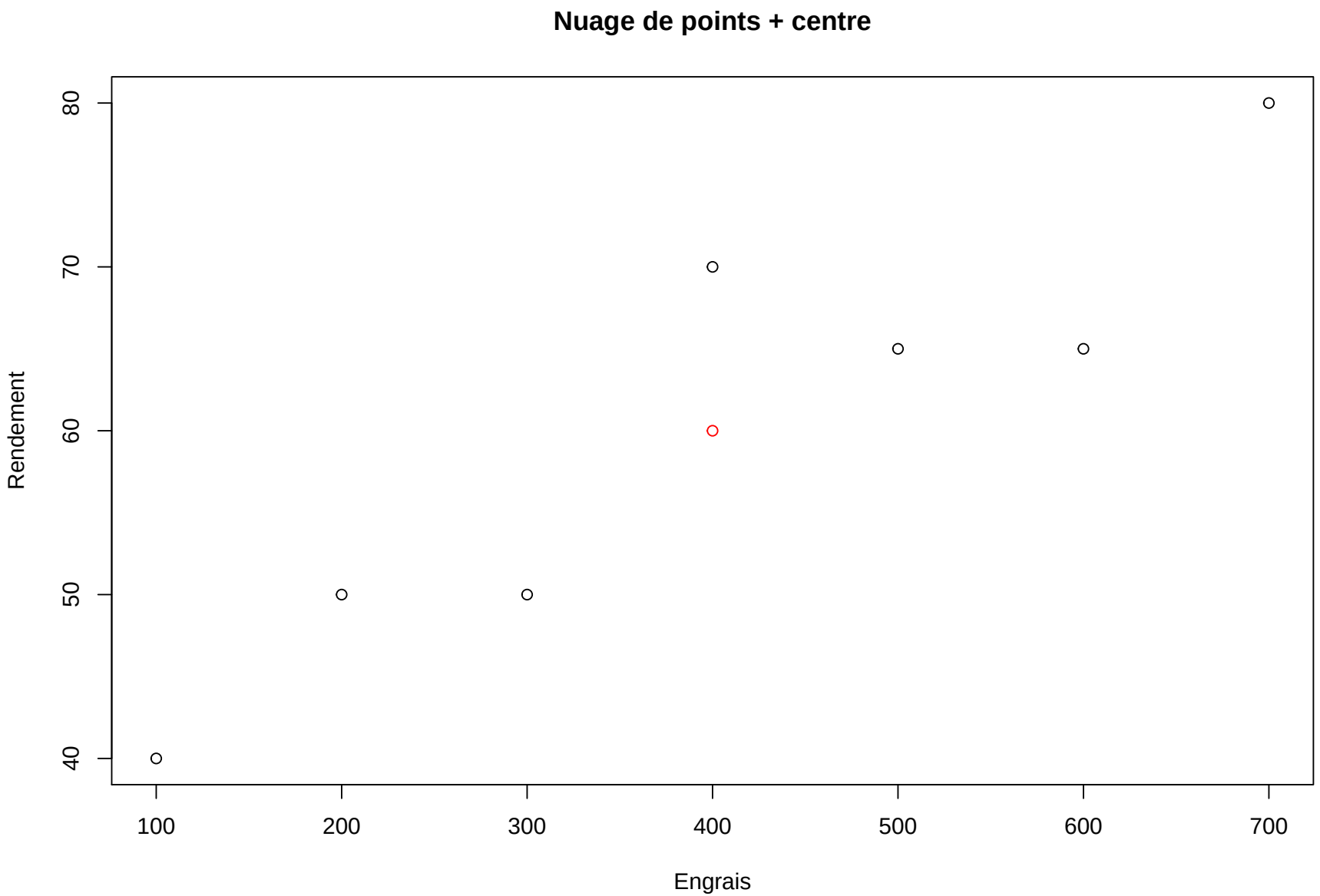
$$\sum_{i=1}^{n} x_i = 2800, \quad \sum_{i=1}^{n} y_i = 420, \quad \sum x_i^2 = 1400000$$

$$\sum x_i y_i = 184500, \quad \sum y_i^2 = 26350.$$

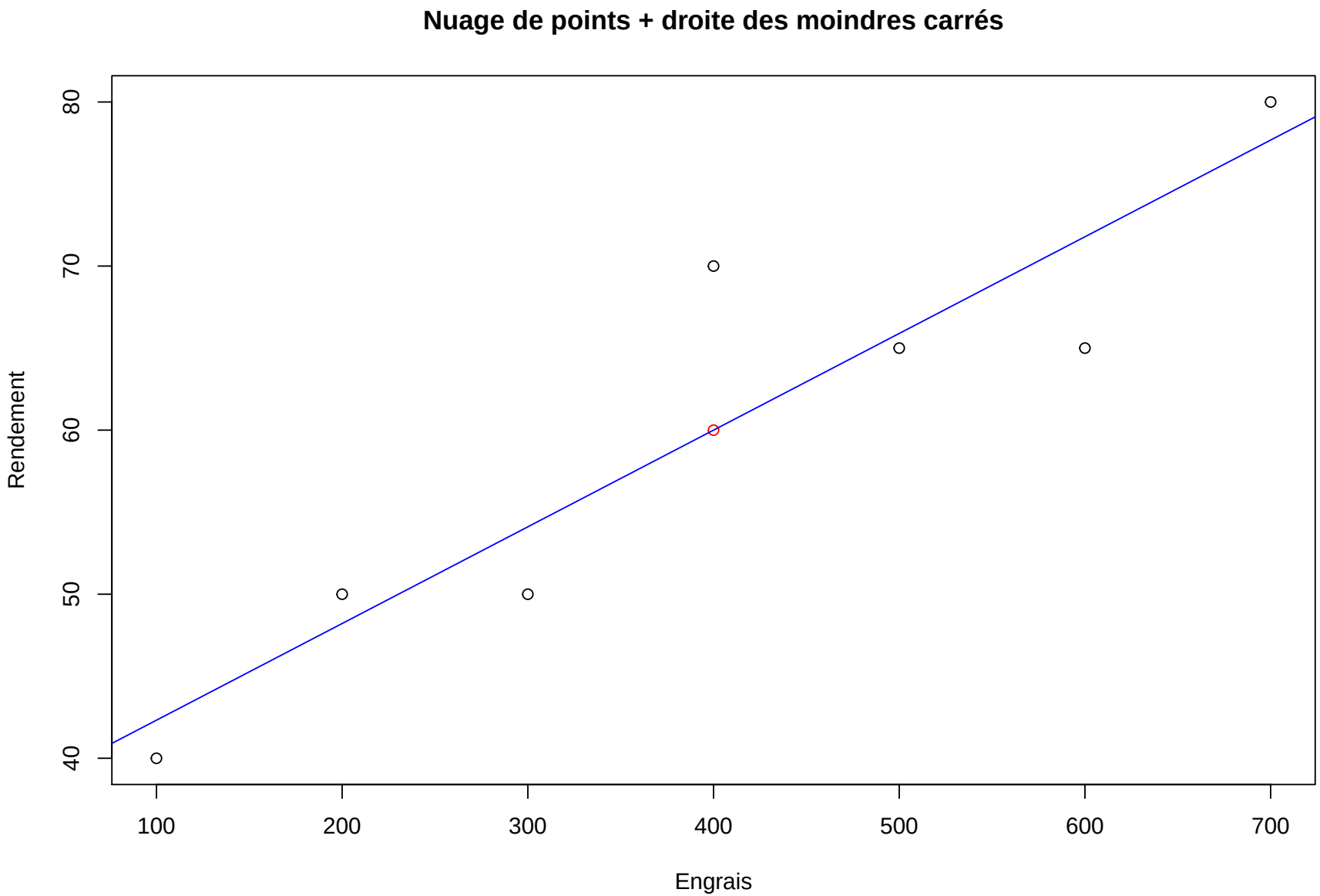
II. Estimation des paramètres : exemple



II. Estimation des paramètres : exemple



II. Estimation des paramètres : exemple



II. Estimation des paramètres : les moindres carrés

La **méthode des moindres carrés** consiste à estimer α et β en minimisant la somme des carrés des résidus ou erreur quadratique :

$$\min_{\alpha,\beta} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha x_i - \beta)^2.$$

Les valeurs de α et β solutions du problème s'expriment alors en fonction des moyennes, variances et covariances :

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

a et b sont les estimateurs des moindres carrés.

II. Estimation des paramètres : les moindres carrés

La **méthode des moindres carrés** consiste à estimer α et β en minimisant la somme des carrés des résidus ou erreur quadratique :

$$\min_{\alpha,\beta} \quad \frac{1}{n}\sum_{i=1}^{n}(Y_i - \alpha x_i - \beta)^2.$$

Les valeurs de α et β solutions du problème s'expriment alors en fonction des moyennes, variances et covariances :

$$a = \frac{S_{xy}}{S_x^2}, \quad \text{avec} \quad S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \quad \text{et} \quad S_x^2 = S_{xx},$$

$$b = \bar{y} - a\bar{x}.$$

a et b sont les estimateurs des moindres carrés.

II. Estimation des paramètres : les moindres carrés (R)

```
modele <- lm(y~x)
modele
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##    36.42857      0.05893
```

```
coef(modele)
```

```
## (Intercept)           x
## 36.42857143  0.05892857
```

II. Estimation des paramètres : variance σ^2

Remarques :

- Les résidus théoriques $\varepsilon_i = Y_i - \alpha x_i - \beta$ sont non observables.
- Les résidus aléatoires $e_i = Y_i - ax_i - b$ sont observables.

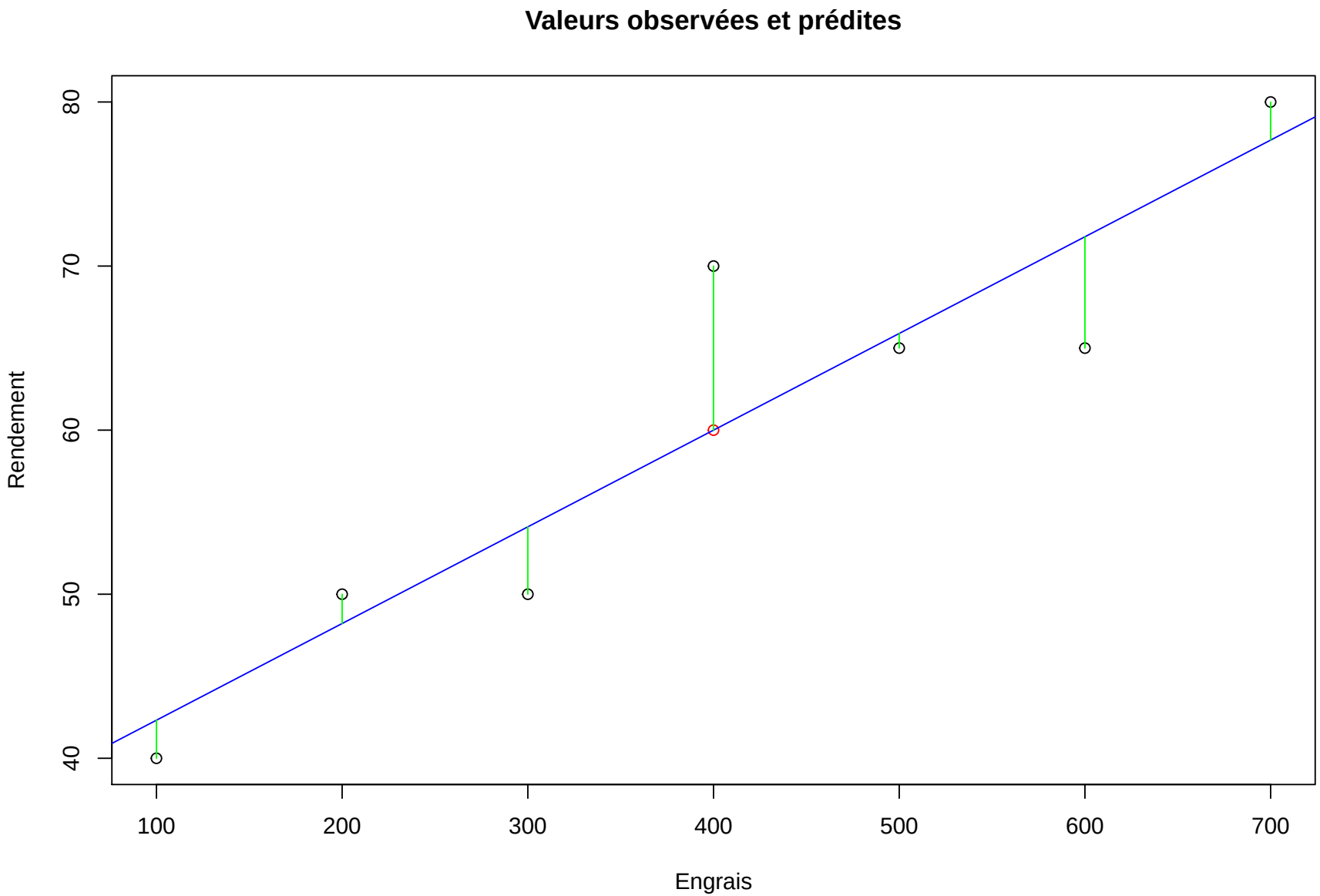
σ^2 est la variance théorique des résidus ε_i. Si on note $\widehat{Y}_i = ax_i + b$, la prévision (aléatoire) de Y_i par le modèle de régression linéaire associée à x_i, les résidus s'écrivent alors :

$$e_i = Y_i - \widehat{Y}_i.$$

Un estimateur s^2 (variance empirique des résidus) de σ^2 est alors :

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - ax_i - b)^2.$$

II. Estimation des paramètres : variance σ^2



II. Estimation des paramètres : variance σ^2 (R)

```
summary(modele)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       1       2       3       4       5       6       7
## -2.3214  1.7857 -4.1071 10.0000 -0.8929 -6.7857  2.3214
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.42857    5.03812   7.231 0.000789 ***
## x            0.05893    0.01127   5.231 0.003379 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.961 on 5 degrees of freedom
## Multiple R-squared:  0.8455, Adjusted R-squared:  0.8146
## F-statistic: 27.36 on 1 and 5 DF,  p-value: 0.003379
```

II. Estimation des paramètres : propriétés et lois

Théorème

a et b sont des estimateurs sans biais et consistants de α et β qui suivent des lois normales d'espérance α et β, et de variance :

$$Var(a) = \frac{\sigma^2}{(n-1)S_x^2} \quad et \quad Var(b) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}\right).$$

Si on remplace σ^2 par s^2 pour obtenir des estimateurs des variances de a et b :

$$s_a^2 = \frac{s^2}{(n-1)S_x^2} \quad et \quad s_b^2 = s^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)S_x^2}\right),$$

on a alors :

$$\frac{a-\alpha}{s_a} \sim St(n-2) \quad et \quad \frac{b-\beta}{s_b} \sim St(n-2).$$

II. Estimation des paramètres : propriétés et lois

Théorème

s^2 est un estimateur sans biais de σ^2 et on a :

$$\frac{(n-2)s^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - ax_i - b)^2}{\sigma^2} \sim \chi^2(n-2).$$

De plus s^2 est indépendant de a, b et $\bar{Y}$.

II. Estimation des paramètres : intervalle de confiance

Les **intervalles de confiance** de niveau de confiance $1-\delta$ sont établis à partir des lois de a et b :

$$IC_{1-\delta}(\alpha) = [a - c_\delta s_a; a + c_\delta s_a]$$

$$IC_{1-\delta}(\beta) = [b - c_\delta s_b; b + c_\delta s_b]$$

où c_δ est le $(1-\delta/2)$-quantile de la distribution de Student $St(n-2)$ satisfaisant :

$$\mathbb{P}\left(St(n-2) \leq c_\delta\right) = 1 - \delta/2.$$

En pratique, c_δ est lu dans table de $St(n-2)$.

II. Estimation des paramètres : intervalle de confiance (R)

```
confint(modele)
```

```
##                   2.5 %      97.5 %
## (Intercept) 23.47767169 49.37947117
## x            0.02996948  0.08788766
```

```
confint(modele,level=0.90)
```

```
##                    5 %        95 %
## (Intercept) 26.27651593 46.58062692
## x            0.03622789  0.08162926
```

III. Test dans le modèle linéaire simple : principe d'un test

Un test statistique est une procédure de décision entre deux hypothèses. Il s'agit d'une démarche consistant à rejeter ou à ne pas rejeter une hypothèse en fonction d'un échantillon.

Procédure de test :

1. Définir les **hypothèses** nulle (H_0) et alternative (H_1),
2. Définir le **risque** δ du test (souvent défini dans l'énoncé),
3. Définir la **statistique de test** T_n,
4. Trouver le **seuil critique** permettant d'établir la zone de rejet pour T_n au niveau $1-\delta$ (règle de décision),
5. Calculer T_n sur l'échantillon,
6. Conclure.

III. Test dans le modèle linéaire simple : 2 tests

- Test du **caractère significatif de la liaison linéaire**
  - Test de Student de la nullité de la pente de régression

$$(H_0) \ : \ \alpha = 0 \ \text{ contre } \ (H_1) \ : \ \alpha \neq 0$$

  - Test de Fisher de Comparaison de modèles :

$$(H_0) \ : \ \text{modèle } \ M_1 : Y_i = \beta + \varepsilon_i, \quad \varepsilon_i \ \textit{iid}, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

contre l'alternative

$$(H_1) \ : \ \text{modèle } \ M_2 : Y_i = \alpha x_i + \beta + \varepsilon_i, \quad \varepsilon_i \ \textit{iid}, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

- Test d'un **modèle linéaire spécifique**

$$(H_0) \ : \ \alpha = \alpha_0 \ \textit{et} \ \beta = \beta_0 \ \text{ contre } \ (H_1) \ : \ \alpha \neq \alpha_0 \ \textit{ou} \ \beta \neq \beta_0.$$

III. Test dans le modèle linéaire simple : test de Student

Pour le test de Student au niveau de risque $\delta = 5\%$, les hypothèses sont définies par :

$$(H_0) \ : \ \alpha = 0 \ \text{ contre } \ (H_1) \ : \ \alpha \neq 0.$$

- **Statistique de test** $T_n = \frac{a}{s_a}$ dont la loi sous H_0 est connue :

$$T_n = \frac{a}{s_a} \sim_{H_0} St(n-2).$$

- **Règle de décision** : soit c_δ le $(1-\delta/2)$-quantile de la distribution $St(n-2)$,
  - si $|T_n| > c_\delta$, on rejette H_0 (au risque δ),
  - si $|T_n| \leq c_\delta$, on ne rejette pas H_0.

- **Conclusion** : on mesure T_n sur l'échantillon, on lit c_δ dans la table de $St(n-2)$ (avec ordre $1-\delta/2$) et on conclut suivant la règle décision. Si on rejette (H_0), la liaison linéaire est significative.

III. Test dans le modèle linéaire simple : test de Student

La **p-valeur** est utilisée pour quantifier la significativité d'un résultat dans le cadre d'une hypothèse nulle. L'idée est de prouver que l'hypothèse nulle n'est pas vérifiée car dans le cas où elle le serait le résultat observé serait fortement improbable. Un résultat statistiquement significatif est un résultat qui serait improbable si l'hypothèse nulle était vérifiée.

En termes statistiques la p-valeur s'interprète comme la probabilité d'un résultat au moins aussi « extrême » que le résultat observé, « sachant l'hypothèse nulle » :

$$p\text{-valeur} = \mathbb{P}_{H_0}(|T_n| > |t_n|) = 2 * (1 - \mathbb{P}(St(n-2) \leq |t_n|))$$

- si p-valeur $< \delta$, le test de niveau δ est significatif (liaison significative)
- si p-valeur $> \delta$, le test de niveau δ n'est pas significatif (liaison non significative)

III. Test dans le modèle linéaire simple : test de Student

```
summary(modele)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       1       2       3       4       5       6       7
## -2.3214  1.7857 -4.1071 10.0000 -0.8929 -6.7857  2.3214
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.42857    5.03812   7.231 0.000789 ***
## x            0.05893    0.01127   5.231 0.003379 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.961 on 5 degrees of freedom
## Multiple R-squared:  0.8455, Adjusted R-squared:  0.8146
## F-statistic: 27.36 on 1 and 5 DF,  p-value: 0.003379
```

III. Test dans le modèle linéaire simple : test de Fisher

Le test de Fisher est une approche par comparaison de modèles, il :

- compare les modèles M_1 et M_2 (à un et deux paramètres d'espérance) définis par :

$$\begin{aligned} M_1 &: \quad Y_i = \beta + \varepsilon_i, \quad \varepsilon_i \ iid, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \\ M_2 &: \quad Y_i = \alpha x_i + \beta + \varepsilon_i, \quad \varepsilon_i \ iid, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \end{aligned}$$

- revient à tester, au risque δ fixé, l'hypothèse nulle

$$(H_0) : \text{ modèle } M_1$$

contre l'alternative

$$(H_1) : \text{ modèle } M_2$$

III. Test dans le modèle linéaire simple : test de Fisher

On note

- $SCT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ (Somme de Carrés Totale) : variabilité totale de Y
- $SCM = \sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2$ (Somme de Carrés du Modèle) : variabilité de Y expliquée par x
- $SCR = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$ (Somme de Carrés Résiduelle) : variabilité de Y qui reste inexpliquée par la relation linéaire

La variance totale de Y admet pour décomposition :

$$\text{Var}(Y) = \text{Var}(\hat{Y}) + \text{Var}(e),$$

soit $SCT = SCM + SCR$.

III. Test dans le modèle linéaire simple : test de Fisher

- **Statistique de test**

$$T_n = \frac{SCM/1}{SCR/(n-2)} \sim_{H_0} F(1, n-2)$$

- **Règle de décision** : soit c_δ le $(1-\delta)$-quantile de la loi $F(1, n-2)$,
  - si $t_n > c_\delta$, on rejette H_0 (au risque δ),
  - si $t_n \leq c_\delta$, on ne rejette pas H_0.

- **Conclusion** : on mesure T_n sur l'échantillon, on lit c_δ dans la table de $F(1, n-2)$ (avec ordre $1-\delta$) et on conclut suivant la règle de décision. Si on rejette (H_0), on conserve le modèle complet, la liaison linéaire est significative.

- ***p*-valeur** $= \mathbb{P}_{H_0}(T_n > t_n) = 1 - P(F(1, n-2) < t_n)$ pour mesurer la significativité du test.

III. Test dans le modèle linéaire simple : test de Fisher

Table d'analyse de variance

Source	ddl	Somme des carrés	Carrés Moyens	stat de test	$p-$value
Modèle	1	SCM	$CMM = SCM/1$	$t_n = \frac{CMM}{CMR}$	$P(F(1, n-2) > t_n)$
Résidu	$n-2$	SCR	$CMR = SCR/(n-2)$		
Total	$n-1$	SCT	$CMT = SCT/(n-1)$		

```
anova(modele)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x          1 972.32  972.32  27.362 0.003379 **
## Residuals  5 177.68   35.54
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

III. Test dans le modèle linéaire simple : test de Fisher

Interprétation

On teste :

$$(H_0) : \text{ modèle } M_1 : Y_i = \beta + \varepsilon_i, \quad \varepsilon_i \text{ iid}, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

contre l'alternative :

$$(H_1) : \text{ modèle } M_2 : Y_i = \alpha x_i + \beta + \varepsilon_i, \quad \varepsilon_i \text{ iid}, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Pour le modèle M_1 :
- Y_i iid, $Y_i \sim \mathcal{N}(\beta, \sigma^2)$,
- $b = \bar{y}$,
- $\hat{y}_i = \bar{y}$
- $s^2 = \frac{1}{n-2} \sum_i (y_i - \bar{y})^2 = \frac{1}{n-2} SCT$.

$SCR(M_1) = SCT$

Pour le modèle M_2 :
- Y_i indépendants, $Y_i \sim \mathcal{N}(\alpha x_i + \beta, \sigma^2)$,
- deux estimateurs d'espérance a et b,
- $\hat{y}_i = a x_i + b$,
- $s^2 = \frac{1}{n-2} \sum_i (\hat{y}_i - y_i)^2 = \frac{1}{n-2} SCR$.

$SCR(M_2) = SCR$

III. Test dans le modèle linéaire simple : test de Fisher

Interprétation

On en déduit que :

$$SCM = SCR(M_1) - SCR(M_2),$$

ce qui signifie que le SCM mesure la réduction d'erreur quand on passe du modèle M_1 au modèle M_2.

Le test répond à la question suivante : la droite des moindres carrés $y = ax + b$ (modèle estimé M_2) explique mieux le nuage de points que la droite horizontale $y = b$ (modèle estimé M_1) mais le gain est-il significatif? On n'abandonnera M_1 pour M_2 que si la réduction d'erreur en passant de M_1 à M_2 est significative.

La statistique de test de Fisher se réécrit :

$$T_n = \frac{(SCR(M_1) - SCR(M_2))/(2-1)}{SCR(M_2)/(n-2)}.$$

III. Test dans le modèle linéaire simple : test de Fisher (R)

```r
modele1 <- lm(y~1)
anova(modele,modele1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ 1
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1      5  177.68
## 2      6 1150.00 -1   -972.32 27.362 0.003379 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

III. Test dans le modèle linéaire simple : modèle spécifique

- **Hypothèses** pour α_0 et β_0 fixés

$$(H_0) : \alpha = \alpha_0 \text{ et } \beta = \beta_0 \text{ contre } (H_1) : \alpha \neq \alpha_0 \text{ ou } \beta \neq \beta_0.$$

  a et b ne sont pas indépendants, on ne peut donc pas faire le test de $\alpha = \alpha_0$ puis celui de $\beta = \beta_0$.

- **Statistique de test** et loi sous H_0

$$T_n = \frac{\sum_{i=1}^n ((a-\alpha_0)x_i + (b-\beta_0))^2 /2}{\sum_{i=1}^n (y_i - ax_i - b)^2 /(n-2)} \sim_{H_0} F(2, n-2)$$

- **Règle de décision** : soit c_δ le $(1-\delta)$-quantile de la loi $F(2, n-2)$,
  - si $T_n > c_\delta$ on rejette H_0 (au risque δ),
  - si $T_n \leq c_\delta$, on ne rejette pas H_0.
- **Conclusion** : on mesure T_n sur l'échantillon, on lit c_δ dans la table de $F(2, n-2)$ (avec ordre $1-\delta$) et on conclut suivant la règle de décision.
- ***p*-valeur** $= \mathbb{P}_{H_0}(T_n > t_n) = 1 - \mathbb{P}(F(2, n-2) < t_n)$ où $F \sim F(2, n-2)$ pour mesurer la significativité du test.

IV. Prévision : problématique

Modèle linéaire simple :

$$\forall i = 1, \cdots, n \quad Y_i = \alpha x_i + \beta + \varepsilon_i, \quad \varepsilon_i \ \ iid, \ \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

Etant donnée une valeur x_{n+1} de x pour laquelle on n'a pas observé de y_{n+1} :

- Comment construire une prévision de ce y_{n+1} non disponible?
- Peut-on intuitivement prédire y_{n+1} par $\widehat{y}_{n+1} = ax_{n+1} + b$?
- Quel sens donner à cette prédiction? Quelle en est sa qualité ?

IV. Prévision : remarque

- $\widehat{y}_{n+1}$ est une réalisation de la variable aléatoire $\widehat{Y}_{n+1}$ définie par $\widehat{Y}_{n+1} = ax_{n+1} + b$
  - $\mathbb{E}(\widehat{Y}_{n+1}) = \alpha x_{n+1} + \beta$: $\boxed{\widehat{Y}_{n+1} \text{ estimateur sans biais de } \alpha x_{n+1} + \beta}$

- De plus, si y_{n+1} était disponible, on lui associerait une v.a. Y_{n+1} définie par

$$Y_{n+1} = \alpha x_{n+1} + \beta + \varepsilon_{n+1}, \quad \varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2),$$

  avec $\varepsilon_1, \cdots, \varepsilon_n, \varepsilon_{n+1}$ indépendantes.
  - $\mathbb{E}(Y_{n+1}) = \alpha x_{n+1} + \beta$: $\boxed{\widehat{Y}_{n+1} \text{ est un estimateur sans biais de } \mathbb{E}(Y_{n+1})}$

$\rightarrow \widehat{y}_{n+1}$ est donc à la fois une estimation de $\mathbb{E}(Y_{n+1})$ et une prévision de y_{n+1}.

IV. Prévision : objectifs

On en déduit les deux problématiques suivantes :

- $\widehat{y}_{n+1}$ est une estimation de $\mathbb{E}(Y_{n+1})$: **construire un intervalle de confiance** pour le paramètre $\mathbb{E}(Y_{n+1})$.

- $\widehat{y}_{n+1}$ est une prévision de y_{n+1} : **construire un intervalle de prédiction** (intervalle de pari) pour Y_{n+1}.

IV. Prévision : intervalle de confiance de $\mathbb{E}(Y_{n+1})$

Théorème

$\widehat{Y}_{n+1} = ax_{n+1} + b$ est un estimateur sans biais de $\mathbb{E}(Y_{n+1}) = \alpha x_{n+1} + \beta$, de variance

$$Var(\widehat{Y}_{n+1}) = \sigma^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right),$$

estimée par

$$s^2_{n+1} = s^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right).$$

De plus,

$$\frac{(\widehat{Y}_{n+1} - \mathbb{E}(Y_{n+1}))}{s_{n+1}} \sim St(n-2)$$

IV. Prévision : intervalle de confiance de $\mathbb{E}(Y_{n+1})$

- **Intervalle de confiance** de $\mathbb{E}(Y_{n+1})$ au niveau de confiance $1-\delta$:

$$\begin{aligned} IC_{1-\delta}(\mathbb{E}(Y_{n+1})) &= [\widehat{y}_{n+1} - c_\delta \ s_{n+1}; \widehat{y}_{n+1} + c_\delta \ s_{n+1}] \\ &= \left[\widehat{y}_{n+1} - c_\delta \sqrt{s^2\left(\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)}; \right. \\ &\qquad \left. \widehat{y}_{n+1} + c_\delta \sqrt{s^2\left(\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)}\right] \end{aligned}$$

où c_δ est le $1-\delta/2$-quantile de $St(n-2)$ tq $\mathbb{P}\left(St(n-2) \leq c_\delta\right) = 1-\delta/2$.

- **Intervalle de confiance** de la droite de régression
  - en faisant varier x_{n+1}, les IC définissent deux hyperboles qui sont l'IC de la droite de régression,
  - plus on s'éloigne du point moyen $(\bar{x}, \bar{y})$, moins l'estimation est précise.

IV. Prévision : intervalle de prévision de Y_{n+1}

On montre que

$$\frac{(\widehat{Y}_{n+1} - Y_{n+1})}{\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}} \sim St(n-2)$$

- **Intervalle de prédiction** de Y_{n+1} de niveau $1-\delta$:

$$IP_{1-\delta}(Y_{n+1}) = \left[\widehat{y}_{n+1} - c_\delta\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}; \right.$$
$$\left. \widehat{y}_{n+1} + c_\delta\sqrt{s^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)}\right]$$

où c_δ est le $1-\delta/2$-quantile de $St(n-2)$ tq $\mathbb{P}(St(n-2) \leq c_\delta) = 1-\delta/2$.

- $IC_{1-\delta}(\mathbb{E}(Y_{n+1})) \subset IP_{1-\delta}(Y_{n+1})$

IV. Prévision : intervalles confiance/prédiction - résumé

L'intervalle de **confiance d'une prévision** définit les limites dans lesquelles se situe probablement une valeur individuelle lue sur la droite de régression. Lorsqu'on construit un modèle qui se présente sous la forme d'une droite de régression, l'intervalle de confiance en question dit que, pour une valeur donnée x_{n+1}, la vraie valeur de la variable y devrait se situer au sein de cet intervalle de confiance.

L'intervalle de **prédiction d'une prévision** définit les limites dans lesquelles tombera vraisemblablement une nouvelle observation de y si elle fait partie de la même population statistique que l'échantillon.

IV. Prévision : intervalles confiance/prédiction - R

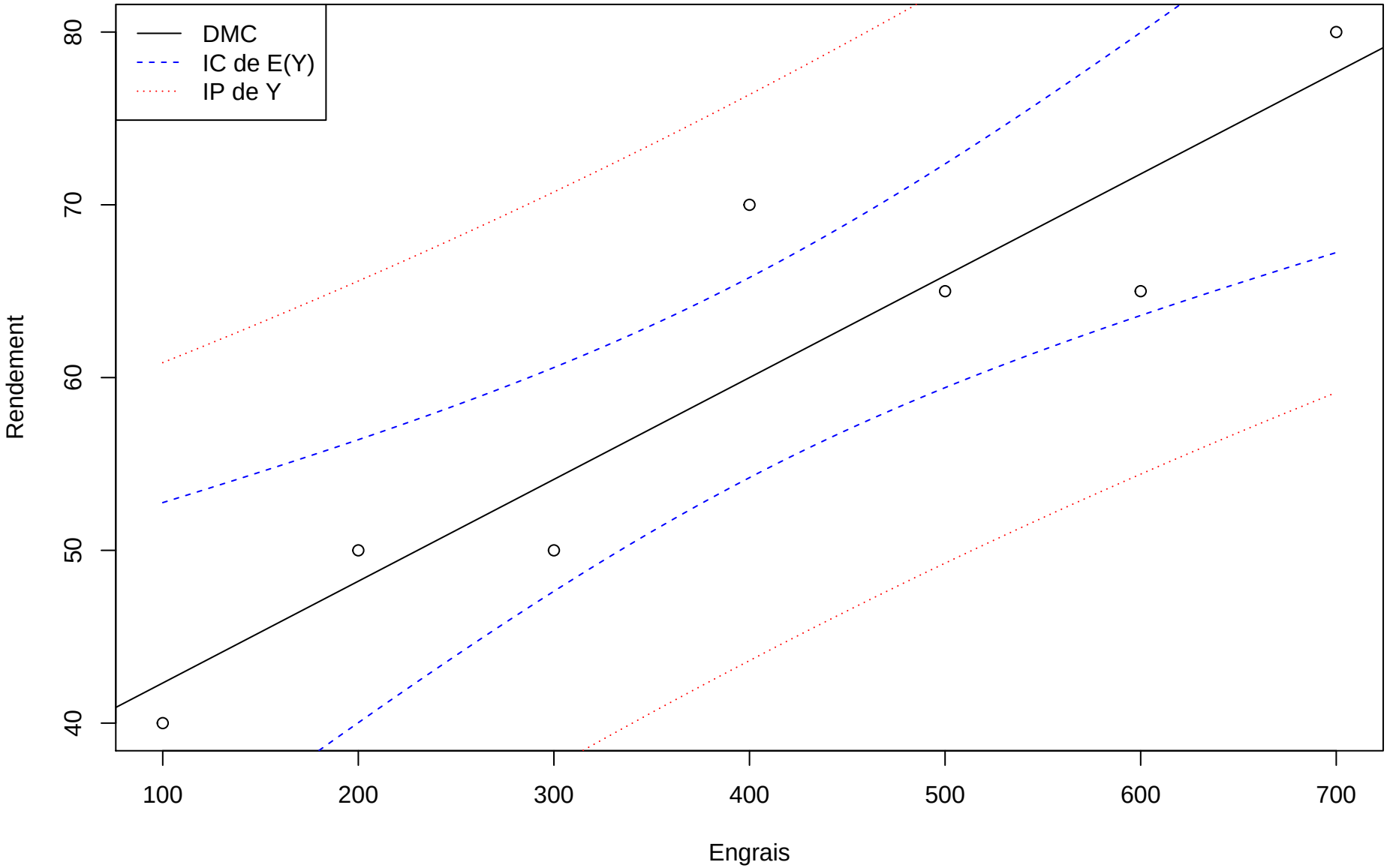
```
newdata <- data.frame(x=450)
predict(modele,newdata,interval="confidence")
```

```
##        fit      lwr     upr
## 1 62.94643 56.97636 68.9165
```

```
predict(modele,newdata,interval="prediction")
```

```
##        fit      lwr      upr
## 1 62.94643 46.50083 79.39203
```

IV. Prévision : intervalles confiance/prédiction - R



V. Validation du modèle : qualité d'ajustement

La qualité d'ajustement du modèle est mesurée par le **coefficient de détermination** :

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCR}{SCT} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}.$$

C'est un nombre compris entre 0 et 1 qui détermine à quel point l'équation de régression est adaptée pour décrire la distribution des points. Plus il est proche de 1, meilleur est l'ajustement.

Remarque : En règle général, on utilise le R^2 ajusté :

$$R^2_{\text{ajusté}} = 1 - \frac{SCR/(n-2)}{SCT/(n-1)}.$$

Le **coefficient de corrélation linéaire** (compris entre -1 et 1) permet également de justifier l'utilisation d'un modèle linéaire.

V. Validation du modèle : qualité d'ajustement (R)

```
summary(modele)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       1       2       3       4       5       6       7
## -2.3214  1.7857 -4.1071 10.0000 -0.8929 -6.7857  2.3214
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.42857    5.03812   7.231 0.000789 ***
## x            0.05893    0.01127   5.231 0.003379 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.961 on 5 degrees of freedom
## Multiple R-squared:  0.8455, Adjusted R-squared:  0.8146
## F-statistic: 27.36 on 1 and 5 DF,  p-value: 0.003379
```

V. Validation du modèle : hypothèses

On considère le modèle linéaire simple :

$$\forall i = 1, \cdots, n \quad Y_i = \alpha x_i + \beta + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

sous les hypothèses suivantes :

- adéquation : $\forall i, \mathbb{E}(\varepsilon_i) = 0$,
- homoscédasticité : $Var(\varepsilon_i) = \sigma^2$, $\forall i$,
- indépendance des erreurs résiduelles,
- normalité des erreurs résiduelles.

$\rightarrow$ Pour la régression simple, le nuage de points (x, y) ou bien (x, e) suffit presque à vérifier ces hypothèses!

V. Validation du modèle : hypothèses

- Hypothèse d'**adéquation**
  - nuage de points (x, y) et/ou (x, e) pour vérifier qu'il n'y a pas de tendance (points répartis autour de l'axe des abscisses)

- Hypothèse d'**homoscédasticité**
  - nuage de points $(\hat{y}, e)$ pour vérifier qu'il n'y a pas de tendance (pas de cône, vague,...)

- Hypothèse d'**indépendance** des erreurs résiduelles
  - hypothèse fondamentale mais difficile à vérifier à votre niveau
  - nuage de point (x_i, e) pour détecter des écarts éventuels dûs à l'apparition de tendances cycliques, une répartition non aléatoire des résidus,...
  - test de Durbin-Watson

- Hypothèse de **normalité** des erreurs résiduelles
  - résultats asymptotiques si n est grand
  - normalité à vérifier dans le cas de petits échantillons
  - histogramme et QQ plot des résidus standardisés

V. Validation du modèle : hypothèses (R)

```
par(mfrow=c(2,2))
plot(model)
```