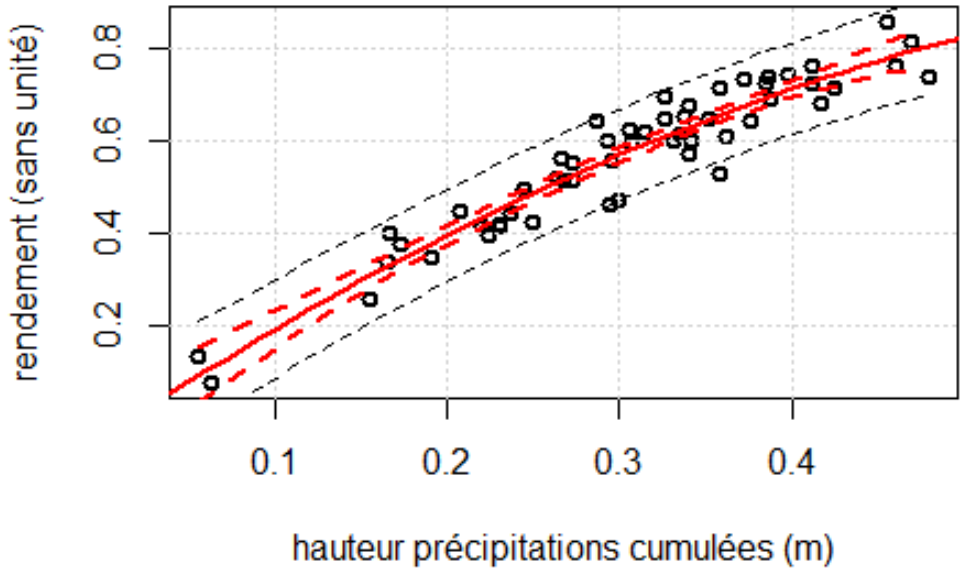
**Majeure Data Science 2025-2026 – UP1 Fondements Probabilistes**

# Régression Linéaire (Multiple)

**(9h00)**




**Prédiction de rendement de parcelles de blé**

**Objectifs du cours**

- Apporter les compétences de base minimales pour mettre en œuvre des techniques de régression linéaire et analyser les résultats obtenus

- Donner une certaine pratique à l'aide du logiciel R sur quelques exemples simples

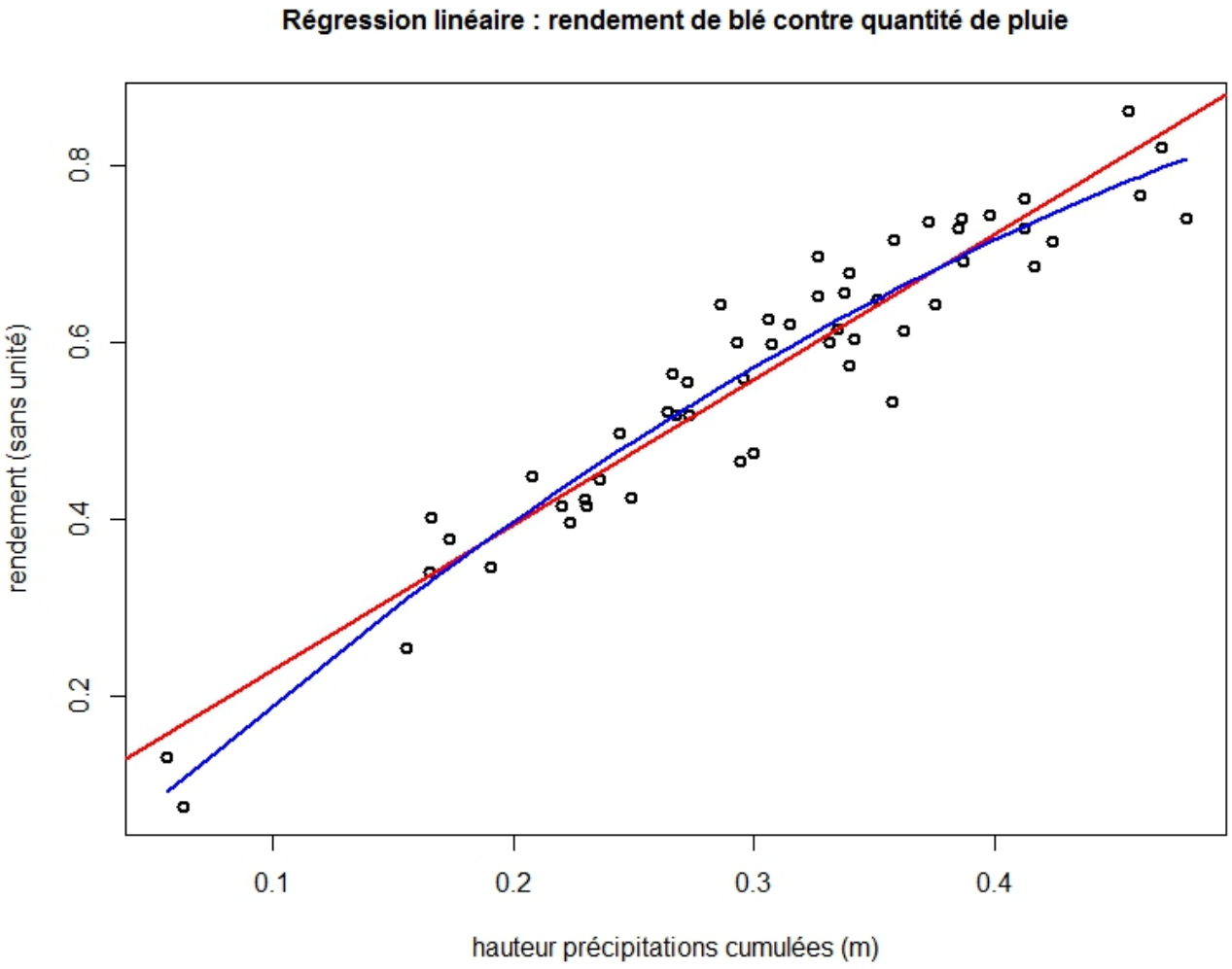
☞ important de bien comprendre les mathématiques qui sont à la base des techniques de régression linéaire de manière à pouvoir :

- Utiliser aux mieux ces techniques (en fonction des objectifs) qui sont encore les techniques de base de tout « data scientist »!

- Savoir bien interpréter les résultats

- Pouvoir envisager les très nombreuses extensions (Ridge, LASSO, …)

# Exemple : rendement de parcelles de blé contre précipitations annuelles




**Exemple 1 - « fil rouge » du support de cours**

## Ce que vous allez apprendre

- ☞ **estimer un modèle de régression à partir de données pour apprendre la relation de dépendance entre une variable d'intérêt (réponse) et des variables explicatives ou prédictives**

- ☞ **tester la significativité de la relation de dépendance**

- ☞ **valider le modèle par analyse des résidus**

- ☞ **tester des hypothèses**

- ☞ **estimer la réponse moyenne pour de nouvelles valeurs des prédicteurs avec intervalle de confiance**

- ☞ **faire des prédictions et construire des intervalles de prédiction**

**…**

## Suite exemple 1 - résultats obtenus sous R pour l'ajustement d'un modèle avec terme quadratique : rendement = $\beta_0 + \beta_1 \times$ pluie $+ \beta_2 \times$ pluie$^2$ + erreur

```
Call: lm(formula = rend ~ pluie + pluie2)
Residuals:
     Min       1Q   Median       3Q      Max
-0.125759 -0.031894 -0.000287  0.035384  0.093880
Coefficients:
             Estimate   Std. Error t value   Pr(>|t|)
(Intercept) -0.04246    0.04512    -0.941    0.35109
pluie        2.50171    0.31868    7.850     2.49e-10 ***
pluie2      -1.52278    0.54701   -2.784     0.00752 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04893 on 51 degrees of freedom
Multiple R-squared:  0.914,          Adjusted R-squared:  0.9106
F-statistic:   271 on 2 and 51 DF,  p-value: < 2.2e-16
```

## Contenu

| **Objectif central de la Science des Données** | = | **Prédiction (classification supervisée, prévision)** |
|---|---|---|

**Techniques classiques** : Régression Linéaire, Régression Logistique, Arbres de Décision et Forêts Aléatoires, SVM (Séparateurs à Vastes Marges ou Support Vector Machines), Réseaux de Neurones, modèles SARIMA (pour la prévision)…

Considérons le problème le plus simple qui soit de prédiction : prédire une variable $y$ (variable à expliquer ou réponse) à partir d'une variable $x$ (variable explicative ou prédictive, prédicteur)

**Formalisme probabiliste** : (X,Y) vecteur aléatoire de loi conjointe décrite par une densité de probabilité, il faut déterminer les lois conditionnelles de Y sachant X = $x$ « pour toutes les valeurs possibles de X »

**Point délicat** de la modélisation : dans la distribution jointe des deux variables X et Y, traduire entièrement et correctement la **dépendance** (si l'on souhaite faire de « bonnes » prédictions en termes de précision et de justesse)

**Cas le plus simple pour tenir compte de dépendance** : via le coefficient de **corrélation linéaire** $\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

Supposons donc (X,Y) **vecteur gaussien** de moyenne $\mu$ et de covariance $\Gamma$ :

$$\mu = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} ; \Gamma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix} \text{ avec } \sigma_X, \sigma_Y > 0 \text{ et } |\rho| < 1$$

On va considérer les deux approches :

**Approche théorique** : les paramètres sont connus (pas besoin de données)

**Approche pratique** : les paramètres sont inconnus, il faut les estimer (apprendre) à partir de données

## Prédiction dans le cas d'un modèle gaussien entièrement spécifié

La loi de Y sachant X = $x$ est une loi normale de la forme $\mathbf{N(\beta_0+\beta_1 x, \sigma^2)}$ avec :

- $\beta_1 = \dfrac{Cov(X,Y)}{\sigma_X^2} = \rho \times \dfrac{\sigma_Y}{\sigma_X}$ et $\beta_0 = \mu_Y - \beta_1 \mu_X$
- $\sigma^2 = \sigma_Y^2 \times (1 - \rho^2)$

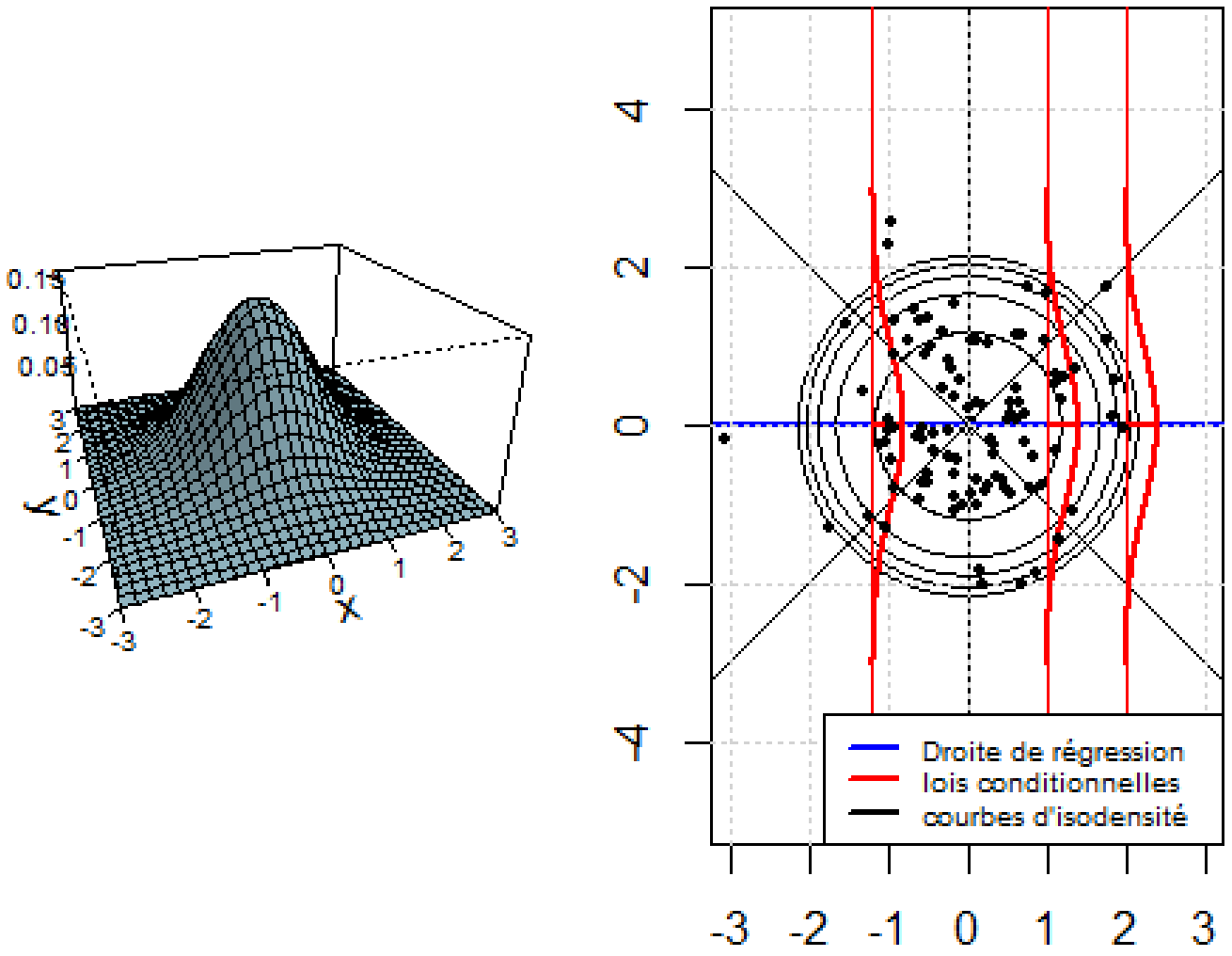
☞ Seule la moyenne conditionnelle $E(Y|X=x) = \beta_0+\beta_1 x$ dépend de $x$
☞ La variance conditionnelle $Var(Y|X=x) = \sigma^2$ est la même quelle que soit la valeur $x$ du prédicteur X

Bien sûr, si $|\rho| \to 1$, la prédiction par la moyenne $m(x) = \beta_0+\beta_1 x$ est de plus en plus précise puisque $\sigma \downarrow 0$

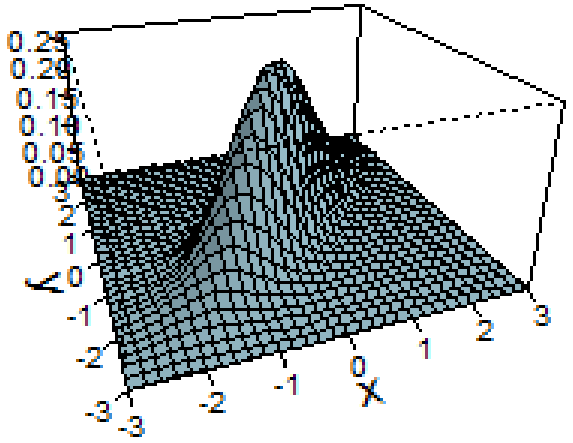
**Intervalle de prédiction à 95%** de la forme $\mathbf{[m(x) - 1.96\times\sigma ; m(x) + 1.96\times\sigma]}$

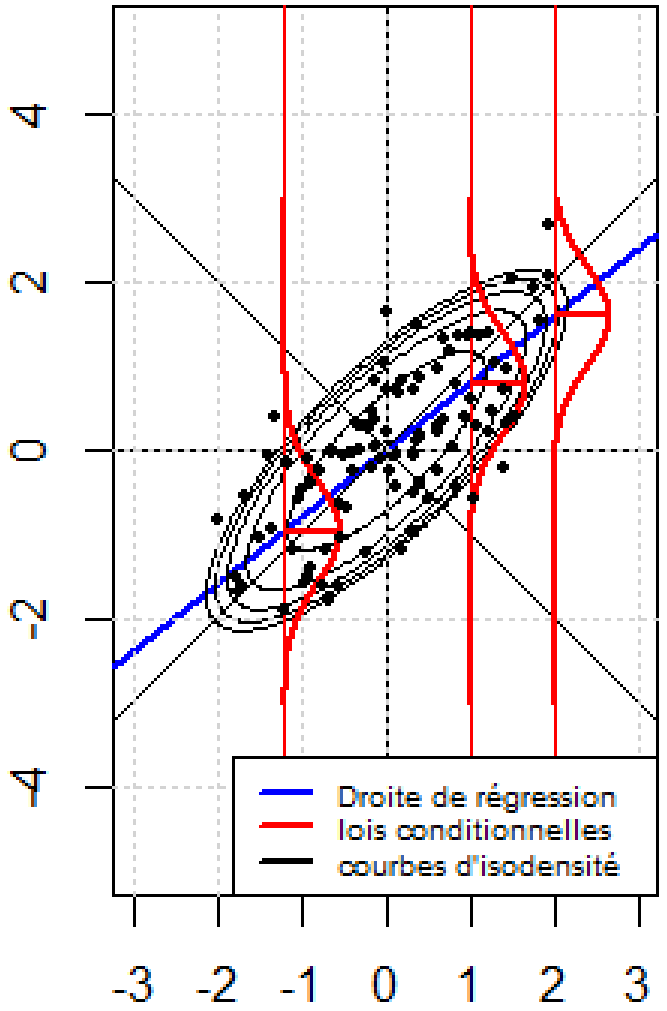
**Application numérique** : $\sigma = \dfrac{\sigma_Y}{2}$ pour $|\rho| \approx 0.87$.

# Illustration - cas de l'indépendance ρ = 0

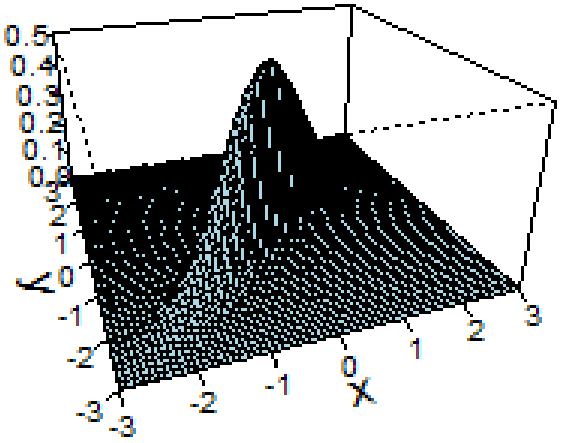


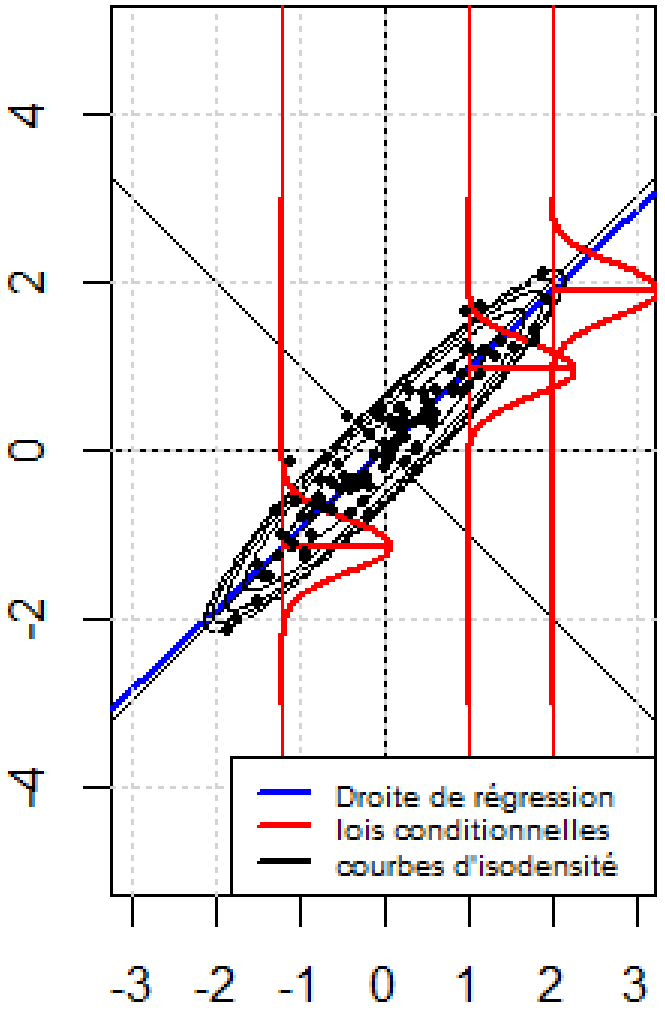
$$\rho = 0.8 \; ; \; \sigma = 0.6 \times \sigma_Y$$



$$\rho = 0.95 \; ; \; \sigma \approx 0.3 \times \sigma_Y$$



**Propriété fondamentale de la fonction de prédiction $x \rightarrow \mathbf{m}(x)$ :**

$$\forall \varphi \text{ fonction : } E[\,(Y - m(X))^2\,] \leq E[\,(Y - \varphi(X))^2\,]$$

**☛ La fonction de prédiction $x \rightarrow \mathbf{m}(x) = \beta_0 + \beta_1 x$ minimise le critère d'Erreur Quadratique Moyenne (EQM ou MSE ou critère $L^2$)**

Soit $L^2$ l'espace vectoriel de toutes les v.a. Z telles que $E(Z^2) < +\infty$ muni du produit scalaire $< U \mid V > = E(UV)$. Notons $L^2(X)$ le sous-espace vectoriel des v.a. de $L^2$ qui sont des fonctions de X, de la forme $\varphi(X)$.

**Vision géométrique ☺ : la v.a. m(X) est la projection orthogonale de Y sur $L^2(X)$**

**Décomposition et ANOVA** : $\mathbf{Y = m(X) + \varepsilon}$ où $m(X) = \beta_0 + \beta_1 X$ ; $\boldsymbol{\varepsilon = Y - m(X)}$

$$\mathbf{Var(Y) = \sigma_Y^2 = Var(m(X)) + Var(\varepsilon)} \text{ avec } Var(m(X)) = \rho^2 \times \sigma_Y^2 \text{ ; } Var(\varepsilon) = \sigma^2$$

**Le prédicteur X et la variable résiduelle ε sont indépendants.**
**De plus, $\varepsilon \sim N(0, \sigma^2)$.**

## Prédiction dans le cadre d'un modèle gaussien avec paramètres inconnus

Cette fois, on dispose de **données** $(x_i , y_i)$, $1 \leq i \leq n$ pour le couple ou vecteur aléatoire (X,Y). Pour l'analyse asymptotique ($n \uparrow +\infty$), on fera l'hypothèse de réalisations **indépendantes**.

☛ on doit donc apprendre le modèle (ses paramètres) avec les données avant de l'utiliser en prédiction

### Approche statistique classique : estimation de la moyenne $\mu$ et de la matrice de covariance $\Gamma$

Méthode d'estimation « naturelle » : **méthode des moments**

Cela conduit aux estimations usuelles $\hat{\beta}_0$ , $\hat{\beta}_1$ de la Régression Linéaire Simple (RLS) et donc à la prédiction ponctuelle $\hat{m}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

**Estimation par Maximum de Vraisemblance (EMV) :** la vraisemblance a une forme compliquée, difficile à maximiser a priori

**Approche par régression linéaire :** on peut écrire que

$Y = \beta_0 + \beta_1 X + \varepsilon$ avec $X \sim N(\mu_X, \sigma_X^2)$ ; $\varepsilon \sim N(0, \sigma^2)$ ; X et ε indépendantes

C'est une autre manière (équivalente) de paramétrer la loi conjointe du couple (X,Y).

Le calcul de l'EMV se simplifie considérablement : ce calcul est équivalent à la minimisation du critère des moindres carrés de la RLS !

**Approche par le critère des moindres carrés** : c'est la déclinaison sur les données de l'approche $L^2$ pour le calcul de la fonction de prédiction $x \rightarrow m(x)$

Dans le cas d'un vecteur gaussien ($X_1$, …, $X_p$, Y) quelconque, on a de même

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

avec $\boldsymbol{\varepsilon \sim N(0, \sigma^2)}$ et **indépendante** du vecteur $X = (X_1, \ldots, X_p)$

En particulier, la loi de Y sachant $X = x$ est la loi normale

$$\mathbf{N(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \sigma^2)}$$

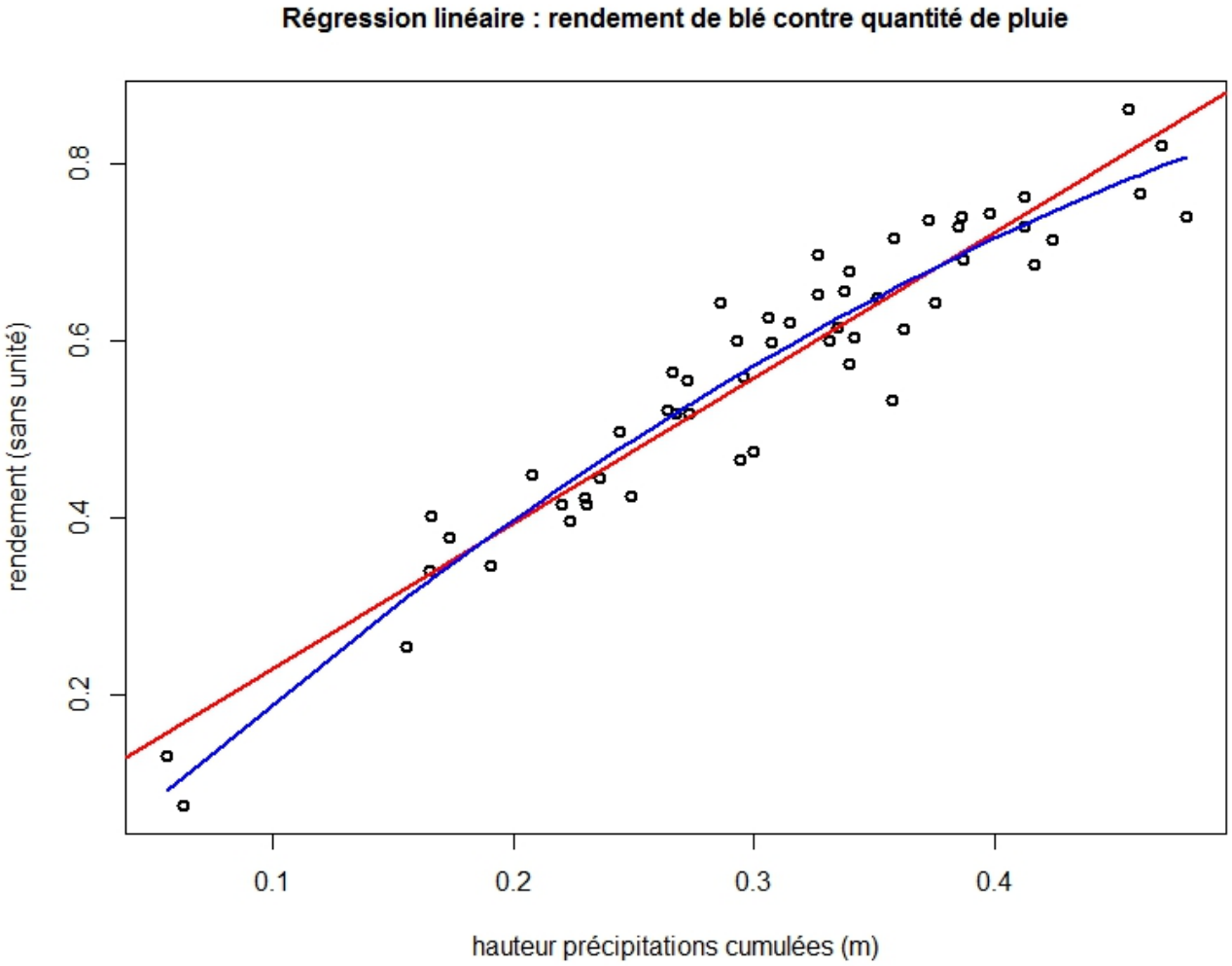
de moyenne

$$\mathbf{m(x) = E(Y \mid X = x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p} x_p$$

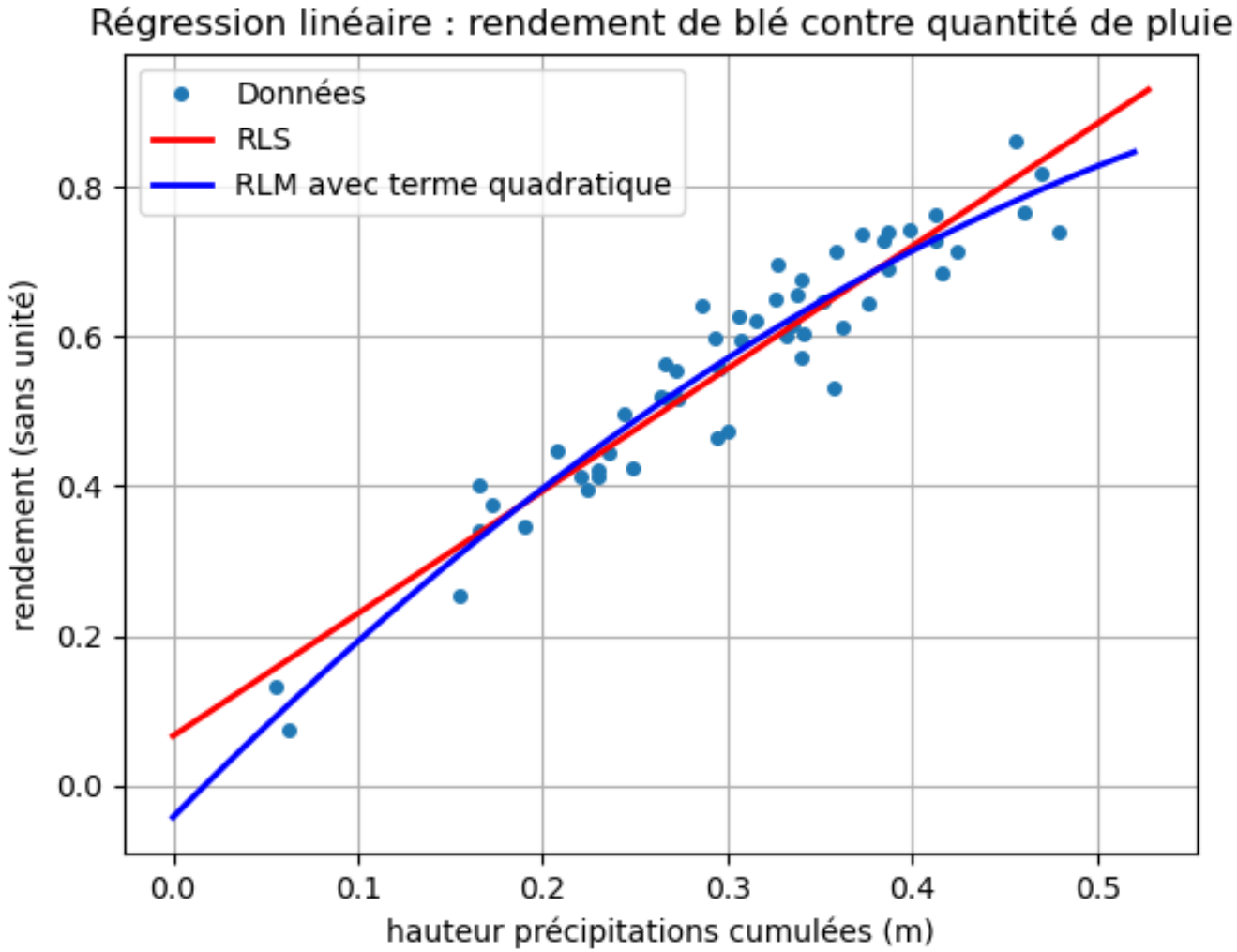
et de variance $\sigma^2$ **constante** ($\sigma^2$ ne dépend pas de $x = (x_1, \ldots, x_p)$).

C'est cette propriété de tout vecteur gaussien qui est à la base du **modèle de Régression Linéaire Multiple** avec des extensions majeures.

# Retour sur l'exemple 1 des parcelles de blé




**Avec R et la fonction lm()**




**Avec Python et la bibliothèque statsmodels (fonction OLS)**

On devine une relation de la forme (droite en rouge)

$$\textbf{rendement} = \beta_0 + \beta_1 \times \textbf{pluie} + \textbf{Erreur}$$

☛ modèle de **régression linéaire simple** avec p = 1 prédicteur : $x$ = pluie

ou, de forme plus complexe,

$$\textbf{rendement} = \beta_0 + \beta_1 \times \textbf{pluie} + \beta_2 \times \textbf{pluie}^{\,2} + \textbf{Erreur}$$

☛ modèle de **régression linéaire multiple** avec p = 2 prédicteurs : $x^{(1)}$ = pluie et $x^{(2)}$ = pluie $^2$

☛ la variable $x$ = pluie est **aléatoire** ou **non contrôlée**. Même dans le cas où ce prédicteur serait **contrôlé**, la réponse Y = « rendement » serait aléatoire compte tenu du terme **Erreur,** la composante résiduelle qui intègre tous les autres facteurs (aléatoires ou non) influençant le rendement…

**Modèle linéaire de régression** : $\mathbf{Y = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_p x^{(p)} + \varepsilon}$

Y = « **réponse** » est de loi connaissant $x = (x^{(1)}, \ldots, x^{(p)})$ une **distribution** de moyenne

$$\mathbf{E(Y \mid x) = \beta_0 + \beta_1 x^{(1)} + \ldots + \beta_p x^{(p)}}$$

et de variance « homogène »

$$\mathbf{Var(Y \mid x) = Var(\varepsilon) = \sigma^2}$$

**Interprétation des coefficients de régression** : $\beta_0$ ordonnée à l'origine (« intercept ») et (si $k \geq 1$) $\beta_k$ = « pente » pour la **réponse espérée** ou **réponse moyenne** $E(Y \mid x)$

**Modèle linéaire de Régression** décliné sur un échantillon de taille n (données)

$$\mathbf{1 \le i \le n,} \quad y_i = \beta_0 + \beta_1 x_i^{(1)} + \ldots + \beta_p x_i^{(p)} + \varepsilon_i \quad \textbf{où } \varepsilon_i \textbf{ résidu (théorique)}$$

Sous forme matricielle : $\mathbf{y = X\beta + \varepsilon}$

- $y = (y_1 \ \ldots \ y_n)'$ vecteur colonne des réponses de taille n
- X matrice de taille $n \times (p+1)$
- $\beta = (\beta_0 \ \ldots \ \beta_p)'$ de taille (p+1) : paramètres pour la **réponse espérée**
- $\varepsilon = (\varepsilon_1 \ \ldots \ \varepsilon_n)'$ vecteur colonne des **résidus théoriques** ou **erreurs** de régression (**bruit**)

**HYPOTHÈSE FORTE : $\varepsilon_1, \ldots, \varepsilon_n$ réalisations de v.a. i. i. d. de loi $N(0, \sigma^2)$**

**Hypothèse plus faible :** $\varepsilon_1, \ldots, \varepsilon_n$ centrées, de même variance $\sigma^2$ et non corrélées

## ☹💣 LE PROBLEME : $\varepsilon_1, \ldots, \varepsilon_n$ ne sont pas observés directement !

Toute l'analyse et la compréhension du modèle linéaire de Régression dans sa version forte repose sur la propriété suivante :

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

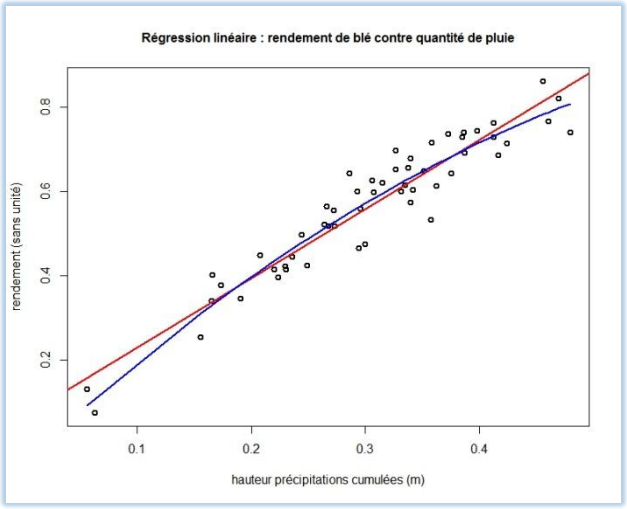
est un **Vecteur Gaussien** (VG) si $\varepsilon$ est un bruit blanc gaussien $N(0, \sigma^2 I_n)$.

A savoir, Y est de loi normale n-dimensionnelle $\mathbf{N(X\beta, \sigma^2 I_n)}$

☛ On considère la matrice **X déterministe** : on raisonne conditionnellement aux valeurs connues ou observées des prédicteurs non contrôlés (aléatoires)

# Modèle avec terme quadratique : rendement ~ pluie + pluie$^2$

```
Call: lm(formula = rend ~ pluie + pluie2)
Residuals:
     Min       1Q   Median       3Q      Max
-0.125759 -0.031894 -0.000287  0.035384  0.093880
Coefficients:
             Estimate   Std. Error t value   Pr(>|t|)
(Intercept) -0.04246    0.04512   -0.941    0.35109
pluie        2.50171    0.31868    7.850    2.49e-10 ***
pluie2      -1.52278    0.54701   -2.784    0.00752 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.04893 on 51 degrees of freedom
Multiple R-squared:  0.914,          Adjusted R-squared:  0.9106
F-statistic:   271 on 2 and 51 DF,  p-value: < 2.2e-16
```

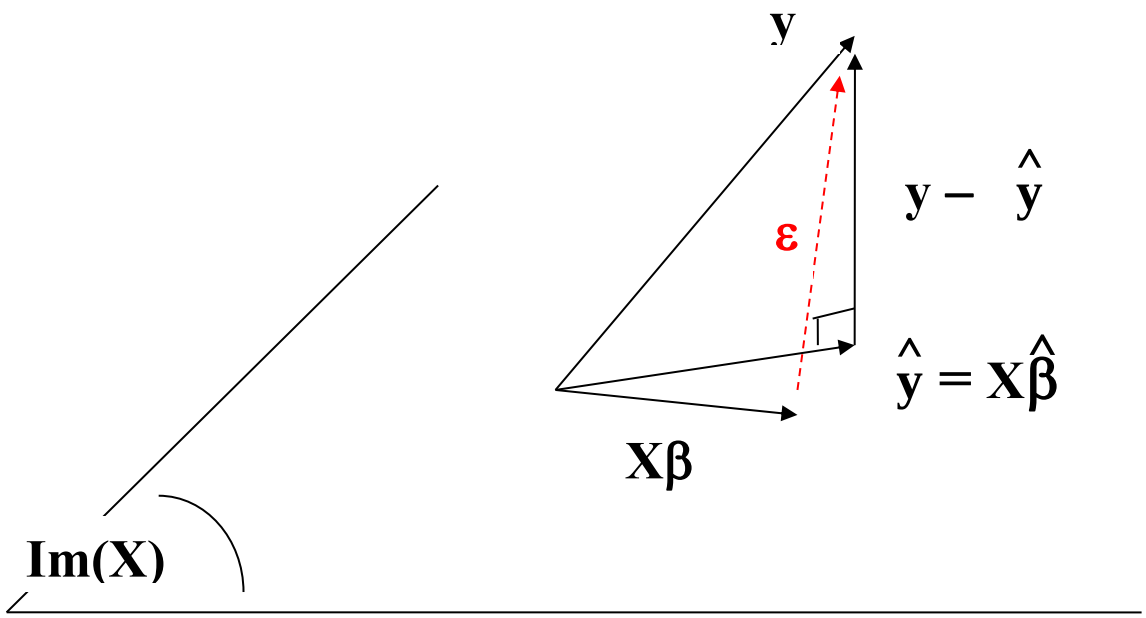



☛ On commence par estimer les **(p+1) paramètres** du modèle et l’étude des estimateurs correspondants.

**Estimation des paramètres β par la méthode des moindres carrés (MC ou MCO) :** le calcul explicite (via le calcul matriciel) conduit à

$$\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon$$

et l'interprétation géométrique **dans $\mathbb{R}^n$**



**L'obtention de $\hat{\beta}$ passe par la résolution des équations normales :**

$$\mathbf{X'X\beta = X'y}$$

avec la petite réserve :

**X′X inversible ⇔ X de rang p+1**

**⇔ colonnes de X = famille libre de vecteurs (variables)**

⇔ Gare au phénomène dit de **multicolinéarité** (multicollinearity)

Rappelons l'estimateur obtenu par la méthode des moindres carrés :

$$\hat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'\varepsilon \text{ avec } \varepsilon \sim N(0, \sigma^2 I_n)$$

**RESULTAT :**

$$\hat{\beta} \text{ est un vecteur gaussien de loi } N(\beta, \sigma^2(X'X)^{-1})$$

En particulier : $\forall\, \beta,\ E(\hat{\beta}) = \beta$

☞ **$\hat{\beta}$ estimateur sans biais de $\beta$**

Retenir que $\mathbf{Cov}(\hat{\beta}) = \sigma^2(X'X)^{-1}$

**Estimation de la variance $\sigma^2$ des résidus (bruit) :** il faut estimer les résidus, ce qui conduit à considérer les quantités

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i^{(1)} + \ldots + \hat{\beta}_p x_i^{(p)} \qquad (1 \leq i \leq n)$$

☞ $\hat{y}_i$ = **réponse estimée ou prédite par le modèle pour la i-ème observation**

$\hat{\varepsilon}_i = y_i - \hat{y}_i$ ☞ **i-ème résidu estimé**

On estime alors la variance $\sigma^2$ des résidus par

$$\hat{\sigma}^2 = \frac{1}{n-(p+1)} \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \frac{1}{n-(p+1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

**Résultat :** c'est un estimateur **sans biais** de la variance résiduelle $\sigma^2$

☛**(n – (p+1))** est le nombre de degrés de liberté du vecteur $\hat{\varepsilon} = Y - \hat{Y}$ utilisé pour calculer $\hat{\sigma}^2$

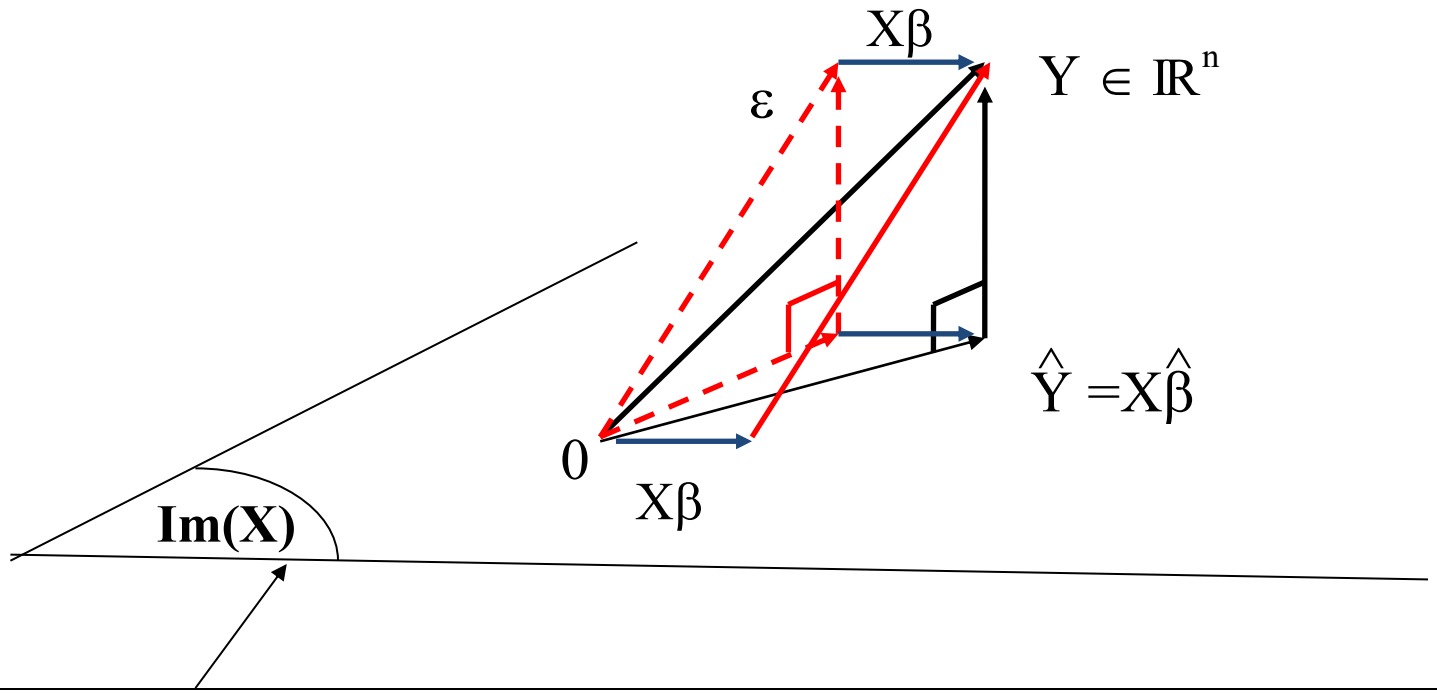
## Loi des estimateurs (sous hypothèse forte sur les résidus)

(i) **(rappel)** Le vecteur $\hat{\beta}$ est **gaussien** $N(\beta, \sigma^2(X'X)^{-1})$

(ii) $\frac{(n-(p+1))\,\hat{\sigma}^2}{\sigma^2} = \frac{\| Y - X\hat{\beta} \|^2}{\sigma^2}$ est de **loi $\chi^2_{n-(p+1)}$** et $\hat{\sigma}$ est **indépendant** de $\hat{\beta}$

(iii) $\frac{\hat{\beta}_j - \beta_j}{\sqrt{c_j}\,\hat{\sigma}}$ est de **loi de Student $t_{n-(p+1)}$** où $c_j$ terme diagonal de la matrice $(X'X)^{-1}$ correspondant à $\beta_j$ (j variant de 0 à p)

**Preuve** : elle utilise l'interprétation géométrique de la régression




sous-espace de **dim (p+1)** engendré par les colonnes de X dans $\mathbb{R}^n$

$\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = HY$ projection orthogonale de Y sur Im(X)

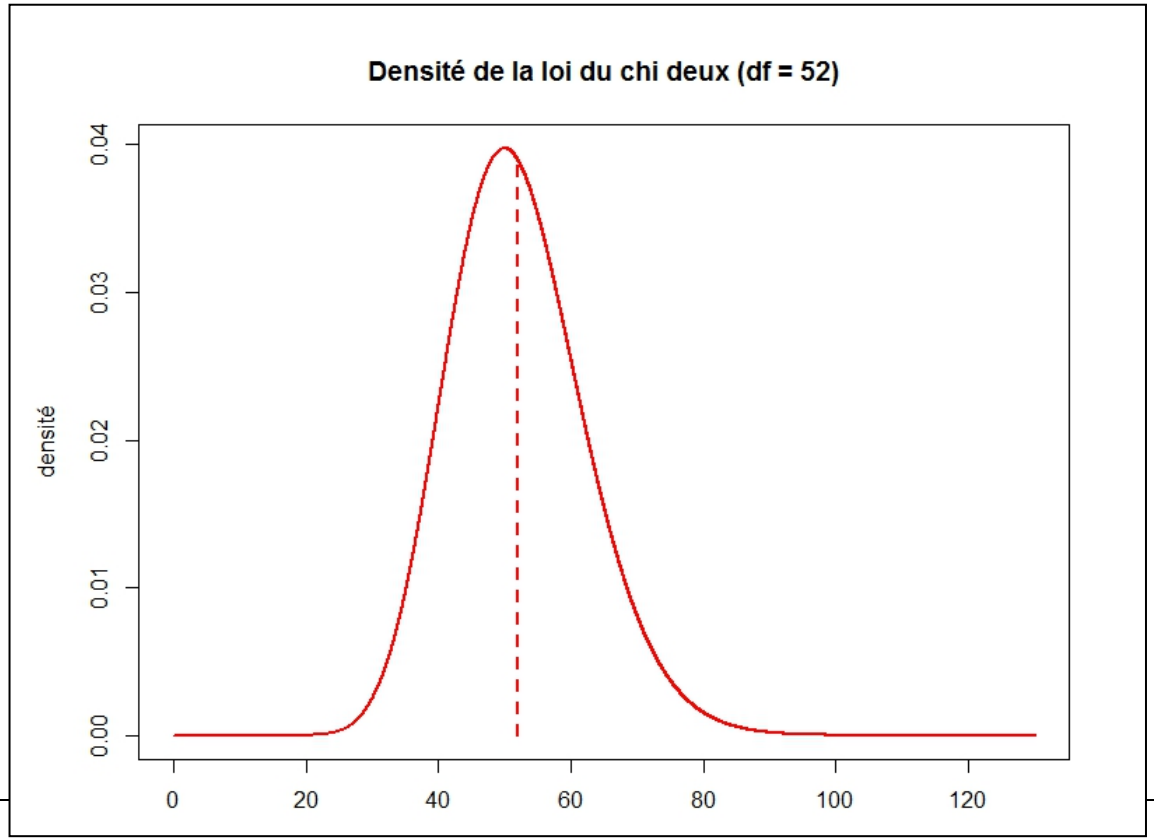
☞ **$H = X(X'X)^{-1}X'$ est la matrice "chapeau" ("hat matrix")**

# Petit rappel sur les lois du khi-deux et de Student (William Gosset, 1908, brasserie Guinness, Dublin)

**Loi du khi-deux à d degrés de liberté (dl) ou $\chi^2_d$** : loi de $\|\varepsilon\|^2 = \varepsilon_1^2 + \ldots + \varepsilon_d^2$ où $\varepsilon \sim N(0, I_d)$

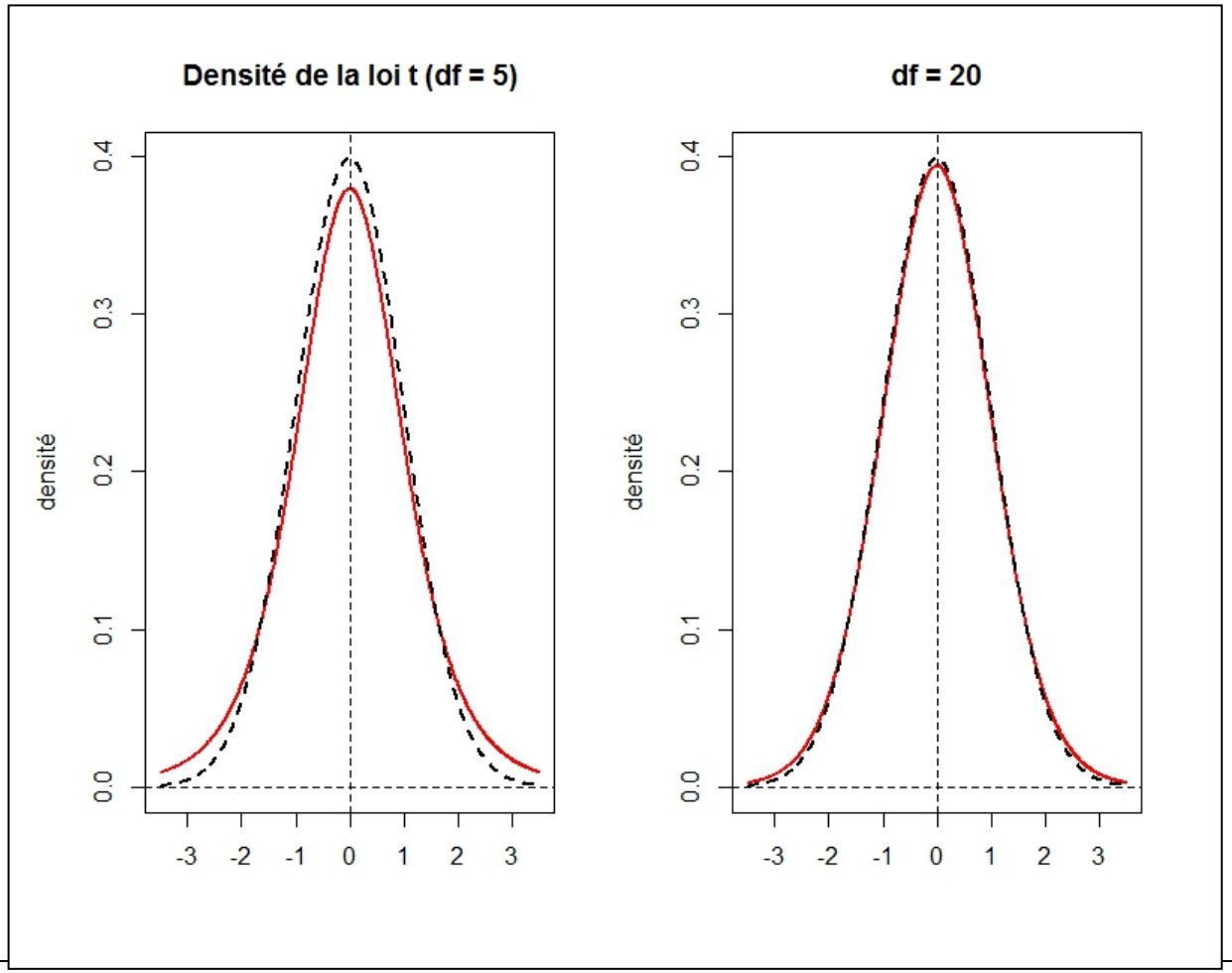
**Densité** : $f(x) = \frac{1}{2^{d/2}\Gamma(d/2)} x^{d/2-1} e^{-x/2} 1_{]0,+\infty[}(x)$ ; $\boxed{E(X) = d}$ ; $\boxed{Var(X) = 2d}$

**Exemple** : $\chi^2_{52}$



**Loi de Student (à d dl) ou $t_d$** : loi de $\dfrac{X}{\sqrt{Y/d}}$ où $X \sim N(0, 1)$ et $Y \sim \chi^2_d$ **indépendantes**

**Densité** : $f(x) = \dfrac{1}{\sqrt{d}\, B(1/2, d/2)} \dfrac{1}{(1 + x^2/d)^{(d+1)/2}}$ ; $\boxed{E(X) = 0}$ ; $\boxed{Var(X) = \dfrac{d}{d-2}}$ $(n \geq 3)$



# Exemple rendement ~ pluie : analyse des résultats sous R

```
data1.reg <- lm(rend ~ pluie)
data1.reg.s <- summary(data1.reg)
print(data1.reg.s)
```

```
Call:
lm(formula = rend ~ pluie, data = data1)

Residuals:
      Min         1Q     Median         3Q        Max
-0.119861 -0.034987  0.003603  0.040208  0.108037

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.06620    0.02405   2.752  0.00813 **
pluie        1.63673    0.07526  21.747  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05201 on 52 degrees of freedom
Multiple R-squared:  0.9009, Adjusted R-squared:  0.899
F-statistic: 472.9 on 1 and 52 DF,  p-value: < 2.2e-16
```

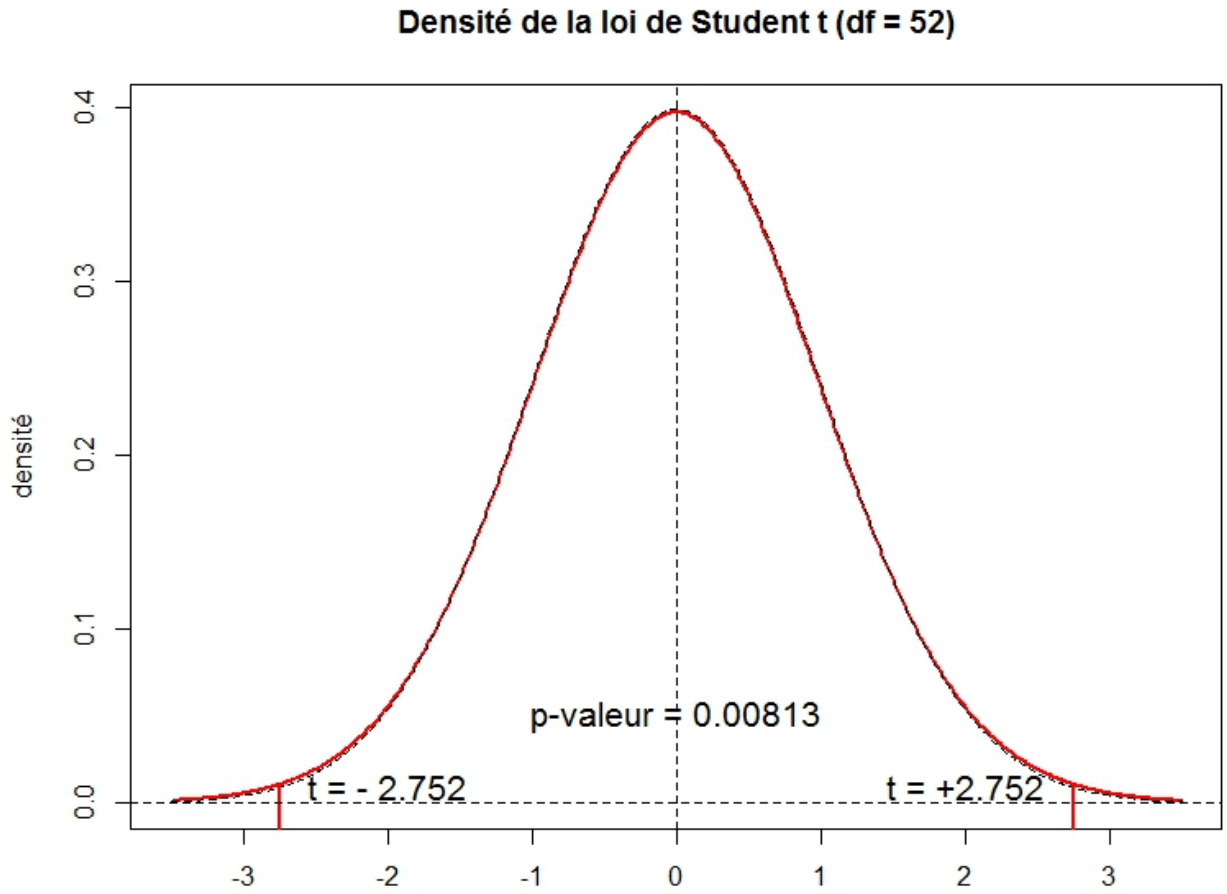
Par exemple, pour le coefficient $\beta_0$ (**intercept**), on lit

$$\hat{\beta}_0 = 0.06620 \text{ et } \sqrt{c_0}\,\hat{\sigma} = 0.02405 \text{ (écart-type estimé de l'estimateur } \hat{\beta}_0 \text{ )}$$

**Sous l'hypothèse $H_0 : \beta_0 = 0$**, la statistique $\dfrac{\hat{\beta}_0 - \beta_0}{\sqrt{c_0}\,\hat{\sigma}} = \dfrac{\hat{\beta}_0}{\sqrt{c_0}\,\hat{\sigma}}$ est de loi de Student $t_{n-(p+1)}$ avec ici p =1, n = 54, soit $\dfrac{\hat{\beta}_0}{\sqrt{c_0}\,\hat{\sigma}} \sim t_{52}$

On lit $\dfrac{\hat{\beta}_0}{\sqrt{c_0}\,\hat{\sigma}} = 2.752$ (t value)

On regarde s'il est « vraisemblable » que cette valeur provienne d'une loi $t_{52}$ :




$P(|t_{52}| \geq 2.752) = 0.00813 \Rightarrow$ on rejette $H_0$ au seuil $\alpha = 5\%$ (risque de première espèce)

Pour analyser la partie

```
Residual standard error: 0.05201 on 52 degrees of freedom
Multiple R-squared:  0.9009, Adjusted R-squared:  0.899
F-statistic: 472.9 on 1 and 52 DF,  p-value: < 2.2e-16
```

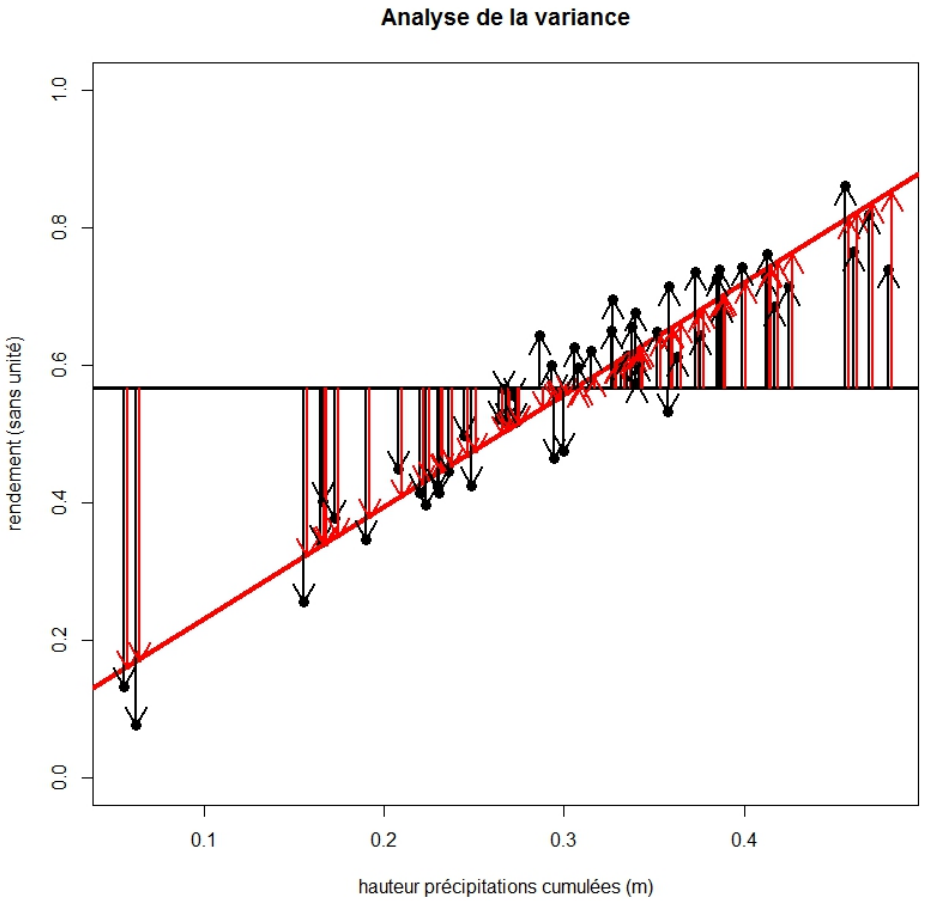
il faut considérer la table d'analyse de la variance (ANOVA) suivante :

```
anova(data1.reg)
```

```
Analysis of Variance Table

Response: rend
           Df  Sum Sq Mean Sq F value    Pr(>F)
pluie       1 1.27917  1.2792  472.92 < 2.2e-16 ***
Residuals 52 0.14065  0.0027
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

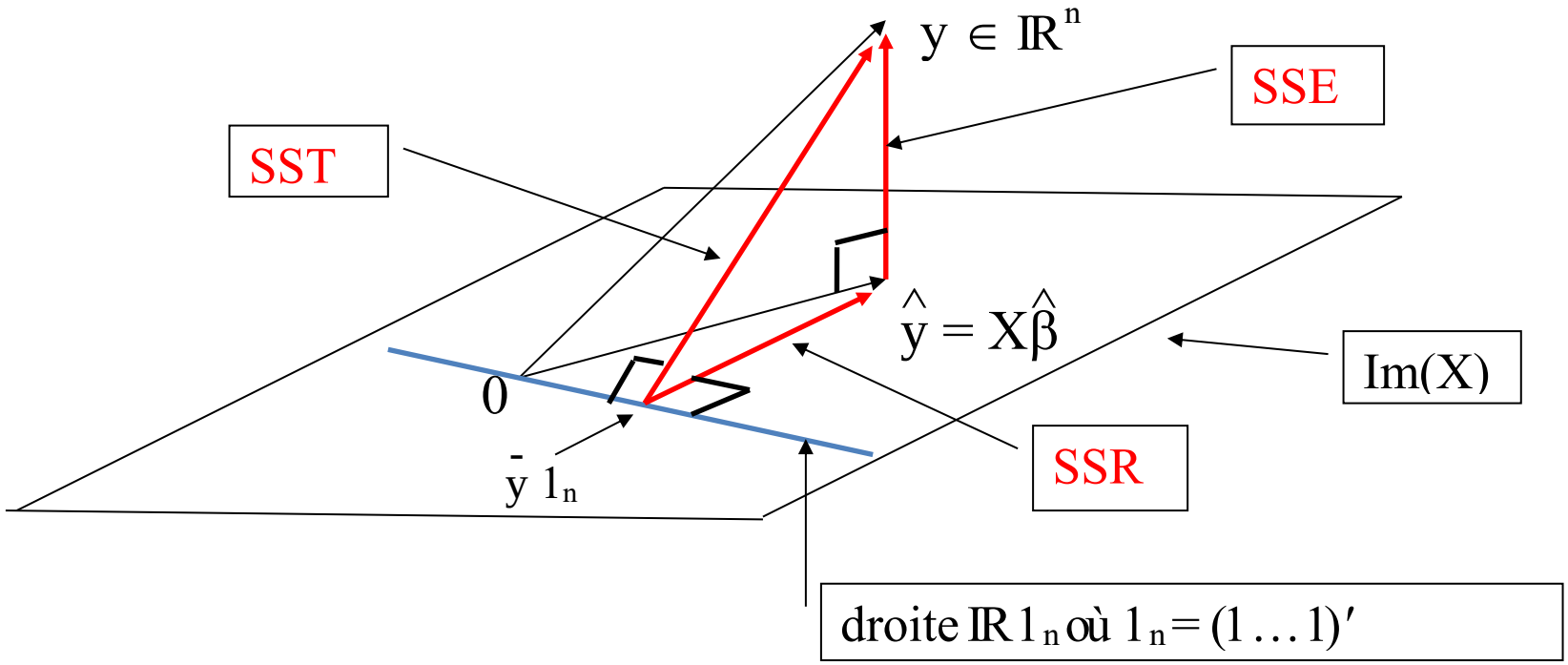
Expliquons le principe de l'analyse de la variance dans le cas p=1 de la régression linéaire simple (droite aux moindres carrés) :




Analyse de la « variabilité » de la réponse :

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y}_i = \hat{y}_i - \bar{y} + \hat{\varepsilon}_i \qquad 1 \leq i \leq n$$

Vision géométrique :




**Formule d'analyse de la variance : SST = SSR + SSE**

**(Total Sum of Squares)** $\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$

☞ somme des carrés des écarts de la variable $y$ à sa moyenne

**(Regression Sum of Squares)** $\mathbf{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$

☞ somme des carrés des écarts lorsque les valeurs $y_i$ sont remplacées par les prévisions obtenues par le modèle de régression ou écarts expliqués par le modèle

**(Error Sum of Squares)** $\mathbf{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}$

☞ somme des carrés des erreurs ou des écarts résiduels

Reprenons la table d'analyse de la variance :

```
Analysis of Variance Table

Response: rend
           Df  Sum Sq   Mean Sq  F value    Pr(>F)
pluie       1  1.27917  1.2792   472.92     < 2.2e-16 ***
Residuals 52  0.14065  0.0027
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On lit (colonne `Sum Sq`)

$$SSR = 1.27917$$

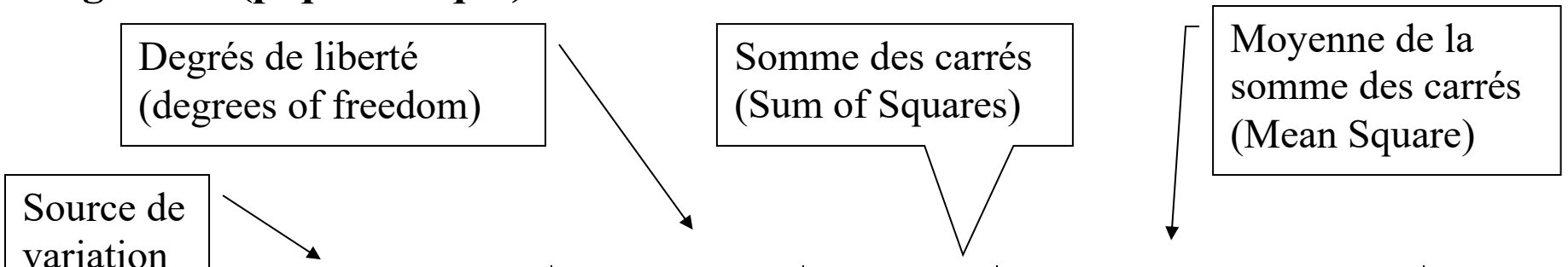
$$SSE = 0.14065$$

et, indirectement, $SST = SSR + SSE = 1.41972$

Pour aller plus loin dans l'analyse, il faut normaliser ces différentes sommes :

$$\frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)}$$ **est de loi de Fisher $F_{1,\,n-2}$ sous l'hypothèse $H_0 : \beta_1 = 0$**

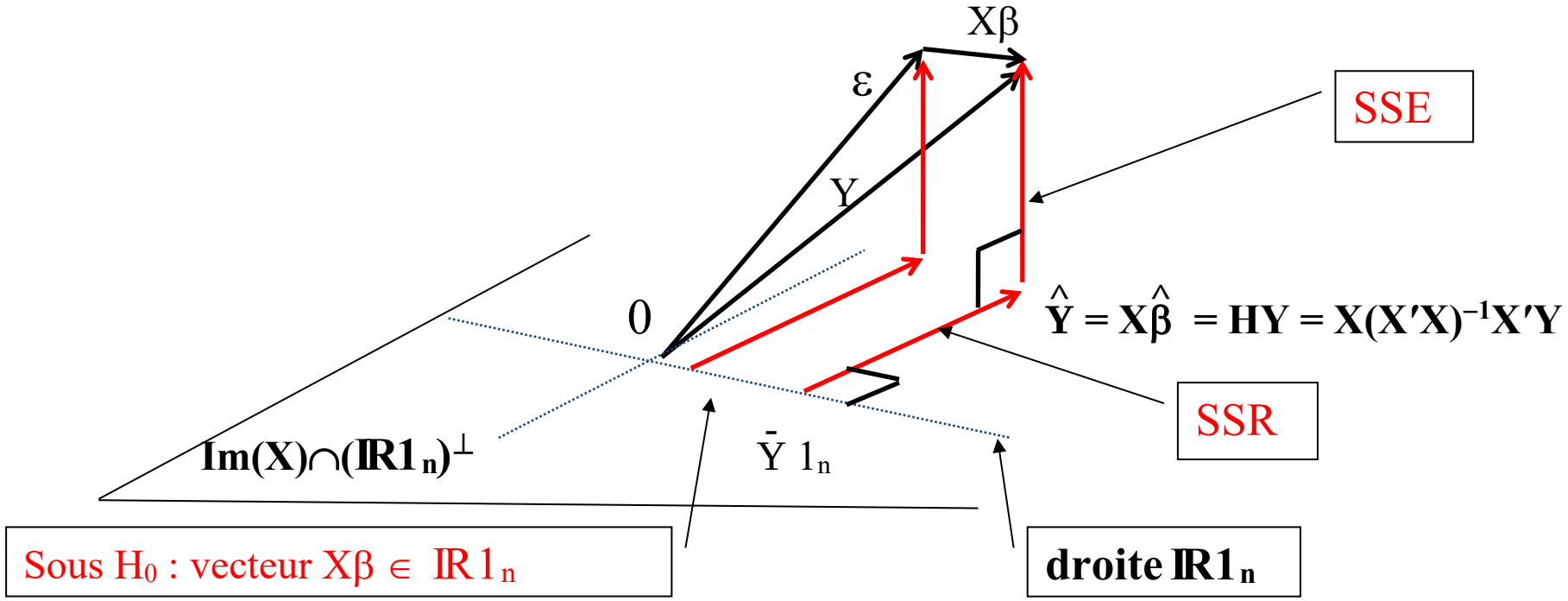
## Cas general (p quelconque)




| Source | DF | SS | MS |
|---|---|---|---|
| Regression | p | SSR | MSR = SSR / p |
| Error | n-p-1 | SSE | MSE = SSE / (n-p-1) |
| Total | n-1 | SST | MST = SST/(n-1) |
| $R^2$ | SSR / SST | | |

DF = nombre de quantités algébriques indépendantes pour calculer les sommes des carrés correspondantes (nombre de carrés indépendants) et MS moyenne des sommes relativement à DF !

Test de l'hypothèse $H_0 : \beta_1 = \ldots = \beta_p = 0$ contre $H_1 = (\text{non } H_0)$




$SSE = \| Y - HY \|^2 = \| \varepsilon - H\varepsilon \|^2 =$ **norme carrée de la projection orthogonale sur** $Im(X)^{\perp}$

$SSR = \| HY - \bar{Y}\,1_n \|^2 = \| H\varepsilon - \bar{\varepsilon}\,1_n \|^2 =$ **norme carrée de la projection** $\perp$ **sur** $Im(X) \cap (\mathbb{R}1_n)^{\perp}$

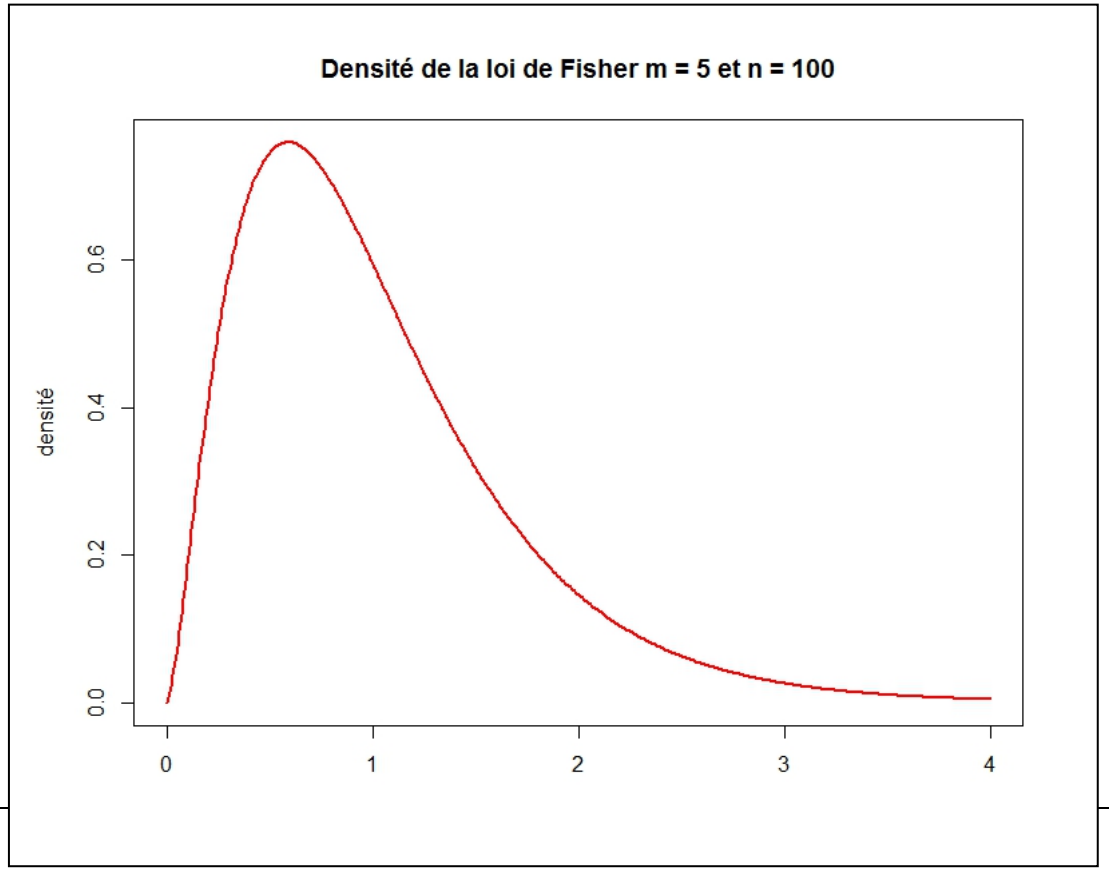
$$\Rightarrow \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-(p+1))} \quad \text{est de loi de Fisher } F_{p,\,n-(p+1)} \text{ sous } H_0$$

**Loi F de Fisher-Snedecor $F_{m,n}$** : loi de $\frac{X/m}{Y/n}$ où $X \sim \chi^2_m$ et $Y \sim \chi^2_n$ **indépendantes**

(Sir Ronald Fisher, biologiste et statisticien, 1890-1962)

**Densité** : $f(x) = \frac{m^{m/2} n^{n/2}}{B(m/2, n/2)} \frac{x^{m/2 - 1}}{(n + mx)^{(m+n)/2}} 1_{]0, +\infty[}(x)$

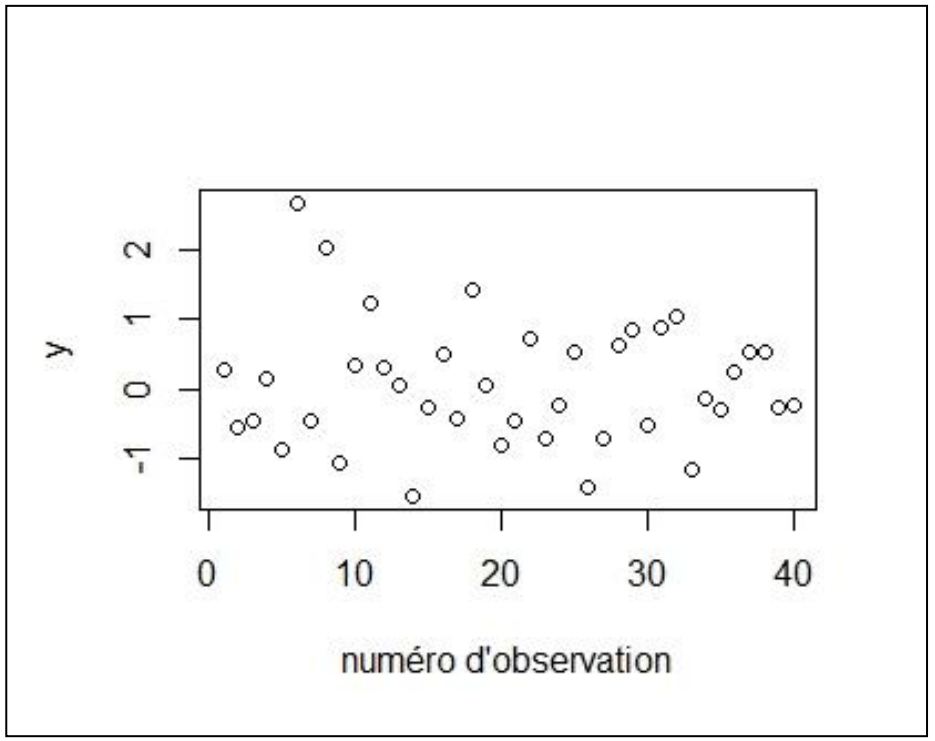
**Exemple** : m = 5 ; n = 100



# Intérêt d'un test global :

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.06144    0.13487   0.456  0.65390
pred1        0.36207    0.60315   0.600  0.55541
pred2       -0.64737    0.60315  -1.073  0.29657
pred3        0.13579    0.60315   0.225  0.82428
pred4        0.19369    0.60315   0.321  0.75161
pred5       -0.70366    0.60315  -1.167  0.25779
pred6        2.04316    0.60315   3.387  0.00309 **
pred7        0.11463    0.60315   0.190  0.85129
pred8        0.69385    0.60315   1.150  0.26426
pred9       -0.95607    0.60315  -1.585  0.12944
pred10       0.42740    0.60315   0.709  0.48717
pred11       0.17216    0.60315   0.285  0.77839
pred12      -0.36186    0.60315  -0.600  0.55563
pred13       0.60755    0.60315   1.007  0.32645
pred14      -0.70361    0.60315  -1.167  0.25783
pred15       0.01004    0.60315   0.017  0.98689
pred16       0.13078    0.60315   0.217  0.83066
pred17      -0.48489    0.60315  -0.804  0.43138
pred18       0.44535    0.60315   0.738  0.46931
pred19       0.16879    0.60315   0.280  0.78262
pred20      -0.28803    0.60315  -0.478  0.63842
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.853 on 19 degrees of freedom
Multiple R-squared:  0.5502,    Adjusted R-squared:  0.07667
F-statistic: 1.162 on 20 and 19 DF,  p-value: 0.3734
```




**Chercher l'erreur**. Indication : $1 - (1 - 0.05)^{20} = 64\%$

# Retour exemple avec terme quadratique : rendement ~ pluie + pluie$^2$

```
Call: lm(formula = rend ~ pluie + pluie2)
Residuals:
     Min       1Q   Median       3Q      Max
-0.125759 -0.031894 -0.000287  0.035384  0.093880
Coefficients:
             Estimate   Std. Error t value   Pr(>|t|)
(Intercept) -0.04246    0.04512    -0.941    0.35109
pluie        2.50171    0.31868    7.850     2.49e-10 ***
pluie2      -1.52278    0.54701   -2.784     0.00752 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.04893 on 51 degrees of freedom
Multiple R-squared:  0.914,      Adjusted R-squared:  0.9106
F-statistic:   271 on 2 and 51 DF,  p-value: < 2.2e-16
```

## Coefficient de détermination $R^2$ et $R^2$ ajusté

$$\mathbf{R^2 = \frac{SSR}{SST} \in [0, 1]} \text{ ou plutôt } \mathbf{R^2 \times 100\ (\%)}$$

💣 ☠ Attention à l'utilisation de $R^2$ ! (mauvaise idée pour valider un modèle ou comparer des modèles)

**Coefficient de détermination ajusté ou $R^2_{ajusté}$ :** $\mathbf{R^2_{ajusté} = 1 - \frac{MSE}{MST}}$

On a : $1 - R^2_{ajusté} = \frac{MSE}{MST} = \frac{SSE/(n-p-1)}{SST/(n-1)} = \frac{n-1}{n-p-1} \times \frac{SSE}{SST} = \frac{n-1}{n-p-1} \times (1 - R^2)$

D'où $\mathbf{R^2_{ajusté} = 1 - \frac{n-1}{n-p-1} \times (1 - R^2) \leq R^2}$ !

Facteur de pénalisation

# Prédiction avec un modèle de régression linéaire

- **Intervalle de confiance pour la réponse espérée $x_{new}\beta$**

Un **intervalle de confiance** de risque $\alpha$ pour $x_{new}\beta$ (où $x_{new} = (x_{new}{}^{(0)}\ x_{new}{}^{(1)}\ \dots\ x_{new}{}^{(p)})$ ) est

$$\left[\ x_{new}\hat{\beta}\ - s_1(x_{new})t^{-1}_{n-(p+1)}(1-\alpha/2)\ ;\ x_{new}\hat{\beta}\ + s_1(x_{new})\ t^{-1}_{n-(p+1)}(1-\alpha/2)\ \right]$$

où $s_1(x_{new}) = \hat{\sigma}\sqrt{x_{new}(X'X)^{-1}x_{new}'}$ ; $t^{-1}_{n-(p+1)}(1-\alpha/2)$ **quantile de niveau** $(1-\alpha/2)\times 100\%$ **d'une loi** $t_{n-(p+1)}$

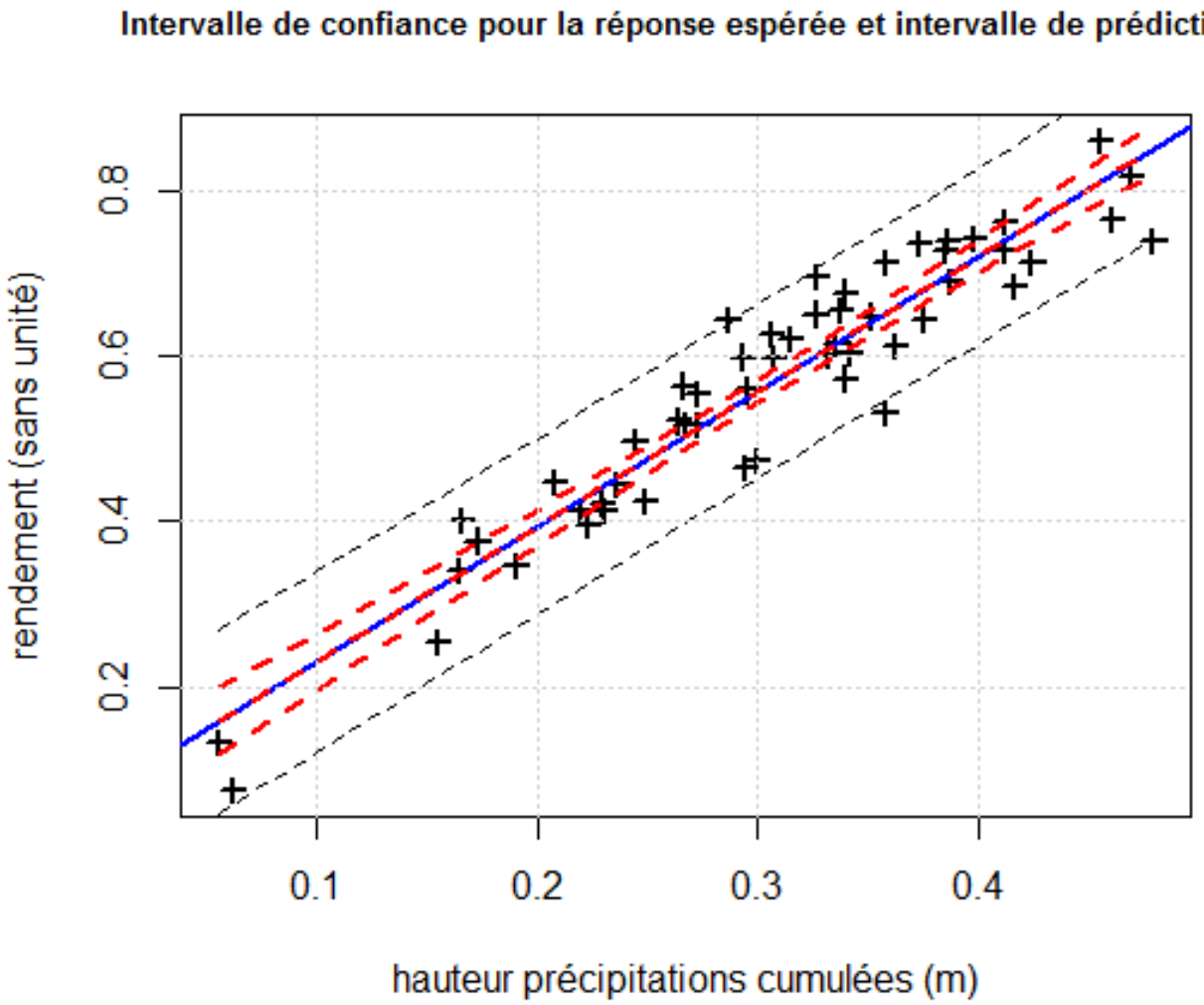
- **Intervalle de prédiction pour la réponse $Y_{new}$**

Un **intervalle de prédiction** de risque $\alpha$ pour la réponse $Y_{new}$ lorsque $x = x_{new}$ est

$$\left[\ x_{new}\hat{\beta}\ - s_2(x_{new})\ t^{-1}_{n-(p+1)}(1-\alpha/2)\ ;\ x_{new}\hat{\beta}\ + s_2(x_{new})\ t^{-1}_{n-(p+1)}(1-\alpha/2)\ \right]$$

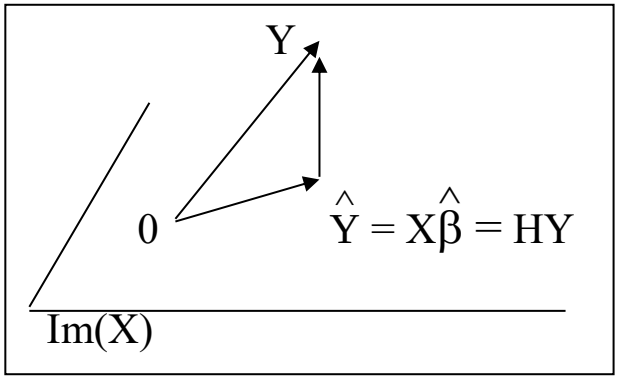
où $s_2(x_{new}) = \hat{\sigma}\sqrt{1 + x_{new}(X'X)^{-1}x_{new}'}$

# Retour exemple : rendement ~ pluie



On note H la matrice $H = X(X'X)^{-1}X'$ qui correspond à la projection orthogonale sur v(X) (sous-espace des prédicteurs dans $\mathbb{R}^n$) de sorte que

$$\hat{Y} = HY$$




On a par contre $\hat{\varepsilon} = Y - \hat{Y} = (I_n - H)Y = (I_n - H)\varepsilon$

**Loi des résidus bruts :** $\hat{\varepsilon}$ est de loi $N(0, \sigma^2(I_n - H))$

**☹ les $\hat{\varepsilon}_i$ sont centrées mais pas de même variance…**

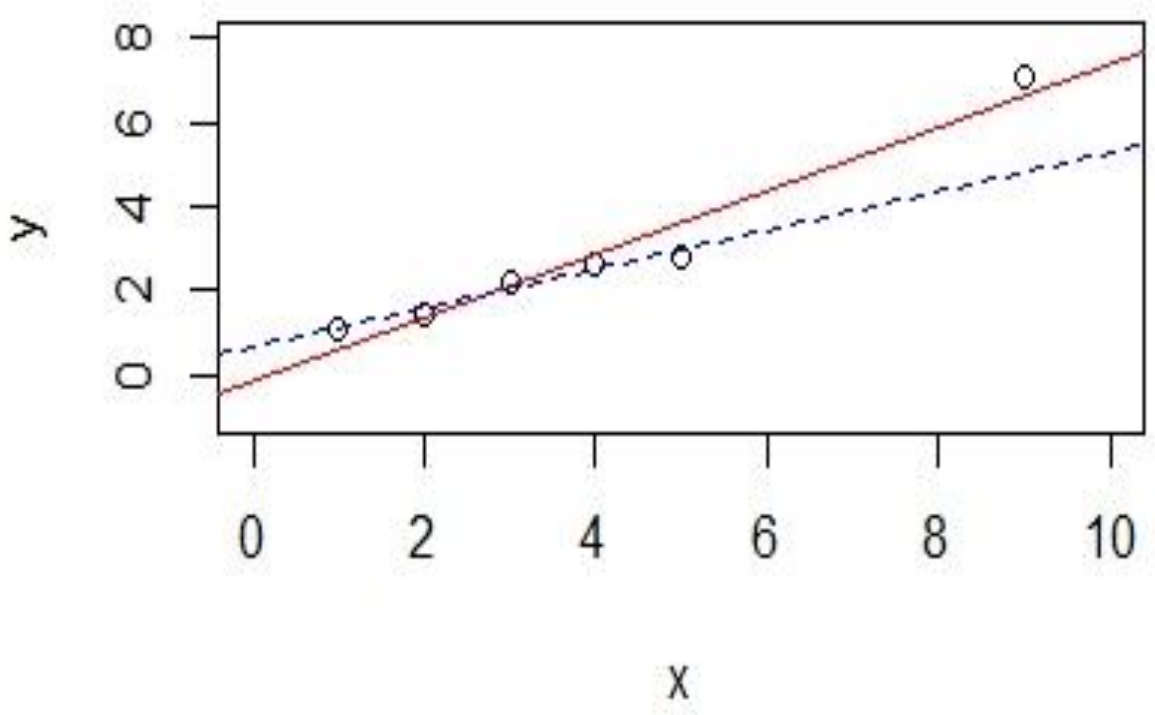
**Résidus standardisés :** $\mathbf{T_i = \dfrac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}}$ où $h_{ii}$ i-ème terme diagonal de H

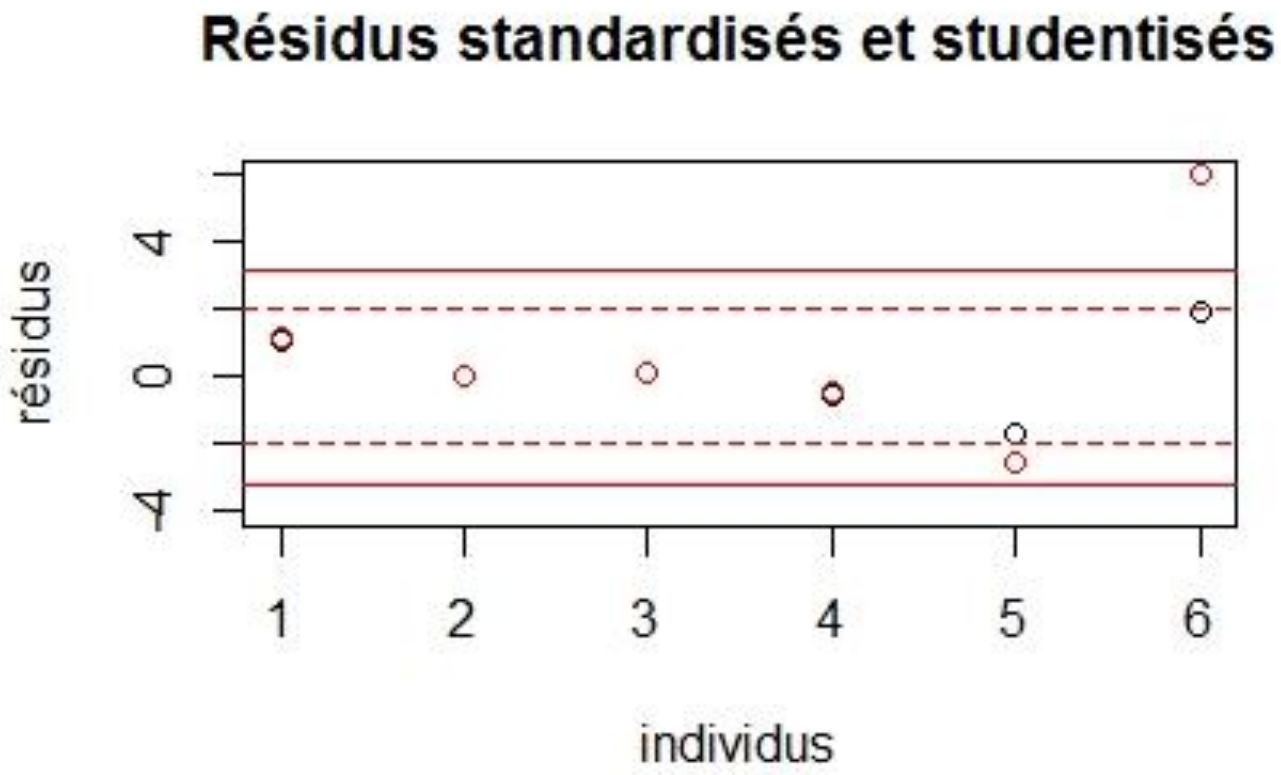
😐 on a des résidus à peu près de loi N(0, 1) d'autant mieux que n est grand. Par contre, on ne peut rien dire sur la loi exacte des $T_i$ car $\hat{\varepsilon}_i$ et $\hat{\sigma}$ ne sont pas des variables indépendantes…

**Résidus studentisés** : $T_i^* = \dfrac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}$ où $\hat{\sigma}_{(i)}$ est l'écart-type résiduel estimé à partir du même modèle linéaire mais sans la i-ème donnée

☺ on parle encore de résidus obtenus par **validation croisée**. **Les $T_i^*$ sont de loi de Student à (n – (p + 2)) degrés de liberté**

# Exemple fictif :




Résidus standardisés et studentisés
résidus
4
0
-4
1
2
3
4
5
6
individus

## Estimation d'une régression en ôtant une observation

On note $X_{(i)}$ , $\hat{\beta}_{(i)}$ , $\hat{\sigma}_{(i)}$ les quantités lorsque la i-ème observation est supprimée.

**(i)** $$\hat{\beta}_{(i)} = \hat{\beta} - (X'X)^{-1} x_i' \frac{\hat{\varepsilon}_i}{1 - h_{ii}}$$

$$(n-p-2)\,\hat{\sigma}_{(i)}^{\,2} = (n-p-1)\,\hat{\sigma}^{\,2} - \frac{\hat{\varepsilon}_i^{\,2}}{1 - h_{ii}}$$

$$T_i^* = T_i \sqrt{\frac{n-p-2}{(n-p-1-T_i^2)}}$$

**(ii)** $T_i^*$ est de **loi de Student** $t_{n-p-2}$

**(iii)** $T_i^*$ est l'erreur standardisée correspondant à la prévision pour $Y_i$ faite sans la i-ème donnée :

$$T_i^* = \frac{y_i - x_i \hat{\beta}_{(i)}}{\hat{\sigma}_{(i)} \sqrt{1 + x_i (X_{(i)}' X_{(i)})^{-1} x_i'}}$$

**Exercice.** Montrer que $\hat{\beta}_{(i)} = \hat{\beta} - (X'X)^{-1} x_i' \dfrac{\hat{\varepsilon}_i}{1 - h_{ii}}$

en utilisant les propriétés suivantes :

- $\hat{\beta}_{(i)} = (X_{(i)}'X_{(i)})^{-1} X_{(i)}'y_{(i)}$
- $(X_{(i)}'X_{(i)})^{-1} = (X'X)^{-1} + \dfrac{(X'X)^{-1} x_i' x_i (X'X)^{-1}}{1 - x_i (X'X)^{-1} x_i'}$ (résulte de $X'X = X_{(i)}'X_{(i)} + x_i' x_i$)
- $X_{(i)}'y_{(i)} = X'y - y_i x_i'$

**Exercice.** Etablir la relation $\mathbf{(n-p-2)\, \hat{\sigma}_{(i)}^{\,2} = (n-p-1)\, \hat{\sigma}^{\,2} - \dfrac{\hat{\varepsilon}_i^{\,2}}{1 - h_{ii}}}$

en écrivant que

$$(n-p-1)\, \hat{\sigma}^{\,2} = \| y - X\hat{\beta} \|^2 = y'y - (X\hat{\beta})'y = y'y - \hat{\beta}'\, X'y$$

et en utilisant les relations $\hat{\beta} = \hat{\beta}_{(i)} + (X'X)^{-1} x_i' \dfrac{\hat{\varepsilon}_i}{1 - h_{ii}}$ et $X'y = X_{(i)}'y_{(i)} + y_i x_i'$

En déduire que $\mathbf{T_i^* = T_i \sqrt{\dfrac{n-p-2}{(n-p-1-T_i^2)}}}$

**Exercice.** Montrer que $T_i^*$ est l'erreur standardisée correspondant à la prévision pour $Y_i$ faite sans la i-ème donnée

$$\mathbf{T_i^* = \frac{y_i - x_i \hat{\beta}_{(i)}}{\hat{\sigma}_{(i)} \sqrt{1 + x_i (X_{(i)}' X_{(i)})^{-1} x_i'}}}$$

en utilisant l'expression déjà rencontrée

$$(X_{(i)}' X_{(i)})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1} x_i' x_i (X'X)^{-1}}{1 - x_i (X'X)^{-1} x_i'}$$

En déduire que $T_i^*$ est de loi de Student $t_{n-p-2}$

## Observations influentes et aberrantes

- Une observation est dite **aberrante** si elle n'est pas en accord avec le modèle ajusté (voir les résidus studentisés calculés par validation croisée)
- Une observation est dite **influente** si elle influence fortement l'ajustement du modèle

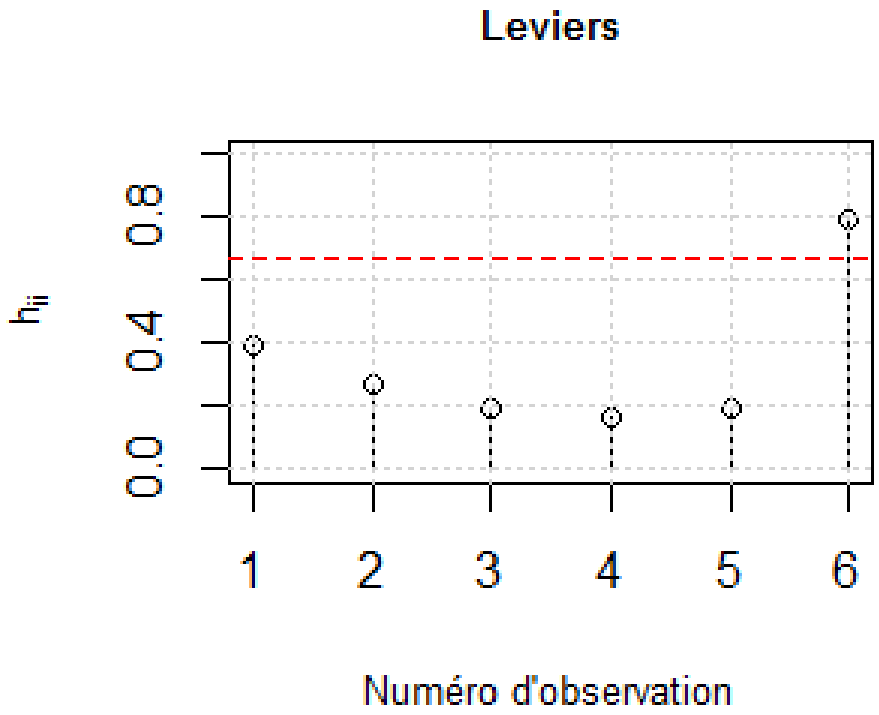
**Levier (leverage).** Le coefficient $\mathbf{h_{ii}}$ est appelé levier de l'observation n° i et apparaît comme une mesure de l'influence de la i-ème observation sur sa propre prédiction :

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij}\, y_j$$

sachant que $\sum_{j=1}^{n} h_{ij}^2 \leq 1$.

Comme $1/n \leq h_{ii} \leq 1$ et $\sum_{i=1}^{n} h_{ii} = p+1$, on utilise parfois le seuil $2\times(p+1)/n$ pour détecter des observations influentes.

## Illustration sur l'exemple fictif




L'observation n° 6 est détectée influente, mais pas nécessairement aberrante !

☞ **Le levier ne dépend que de la matrice X, i.e. des niveaux des prédicteurs.**

## Autre mesure de l'influence de l'observation n° i (qui tient compte des réponses)

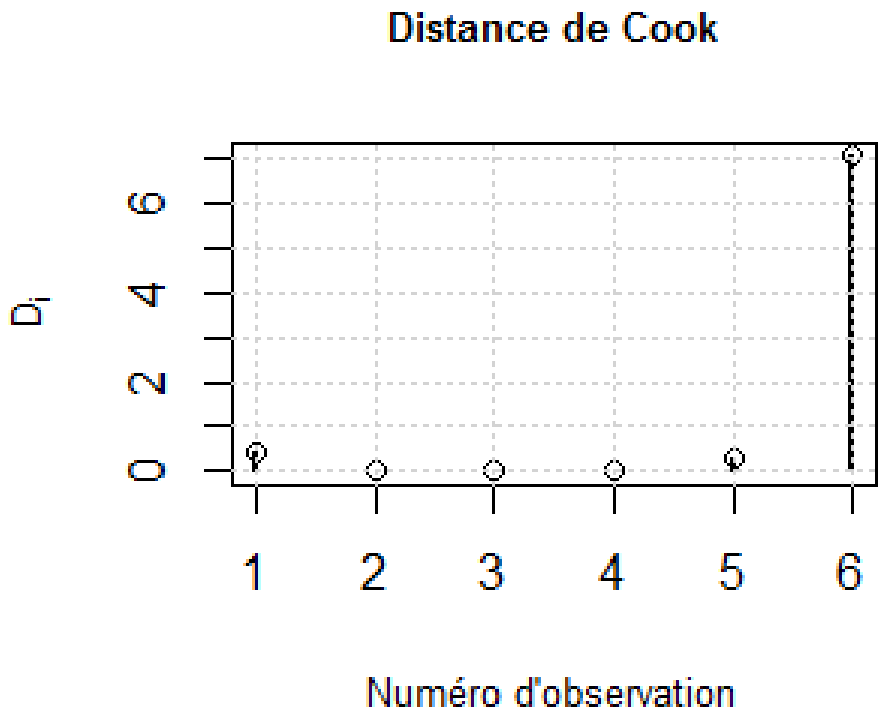
$$D_i = \frac{\| \hat{y} - \hat{y}_{(i)} \|^2}{(p+1)\times\hat{\sigma}^2} = \frac{\| X\hat{\beta} - X\hat{\beta}_{(i)} \|^2}{(p+1)\times\hat{\sigma}^2}$$

☞ **distance de Cook**

## Autre expression (exercice) :

$$D_i = \frac{T_i^2}{p+1} \times \frac{h_{ii}}{1 - h_{ii}}$$

☺ Ainsi, une observation est influente si le **résidu standardisé** est important ainsi que le **levier**




On retrouve le fait que l'observation n°6 est suspecte et qu'il convient de l'examiner de près…

## Colinéarité des régresseurs (= variables prédictives).

Si des régresseurs sont fortement corrélés entre eux, la matrice (X′X) est difficile à inverser, les coefficients β du modèle sont mal estimés…

Pour détecter une colinéarité entre un régresseur $x^{(j)}$ et tous les autres, le **facteur d'inflation de la variance VIF** (de l'anglais *variance inflation factor*) a été introduit

$$\mathbf{1 \leq j \leq p : VIF_j = \frac{1}{1 - R_j^2}}$$

où $R_j^2$ est le coefficient de détermination de la régression de $x^{(j)}$ sur tous les autres régresseurs. On montre que plus $VIF_j$ est élevé, plus la variance de l'estimateur de $\beta_j$ est élevée.

**Cas extrêmes** : $VIF_j = 1$ et $VIF_j = +\infty$