Majeure Data Science 2020-2021

Régression Linéaire

(UP 2 : Apprentissage statistique et automatique)

Déroulement du cours :

- 12H0 cours/TD + TP
- Evaluation = TP (+ examen ?)

Contenu :

0. **Introduction**
1. **Un premier exemple (RLS, cf. TP1)**
2. **Régression linéaire (multiple)**
3. **Analyse des résidus**
4. **Outils de diagnostic**
5. **Extensions**

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 restent encore les techniques de base de tout « data scientist »!
- bien interpréter les résultats
- aller vers de très nombreuses extensions

Soit (X, Y) un couple de v.a. réelles. **Problème** :

prédire/expliquer au mieux Y connaissant X $\Leftrightarrow$ réduire l'incertitude sur Y

Mesure usuelle de l'incertitude : $\mathrm{Var}(Y) = E[\,(Y - EY)^2\,]$ = **variance** de Y

☺ On considère **$< U \mid V > = E(UV)$** qui définit un **produit scalaire** sur l'espace vectoriel $L^2 = \{\, U \text{ v.a.} \mid E(U^2) < +\infty \,\}$. Alors :

$$\mathbf{Cov(U, V) = < U - EU \mid V - EV >}$$
$$\mathbf{Var(U) = \| U - EU \|^2}$$

Idée : déterminer la fonction $\varphi : \mathbb{R} \to \mathbb{R}$ telle que $\varphi(X) \in L^2$ et

$$\| Y - \varphi(X) \|^2 = E[\,(Y - \varphi(X))^2\,] \ \ \textbf{minimale}$$

☛ C'est un problème d'approximation

En notant $L^2(X)$ le sous-espace vectoriel formé par toutes les v.a. de la forme $\varphi(X) \in L^2$, la fonction qui approche le mieux est

$$\mathbf{P_{L^2(X)}(Y) := \text{projection orthogonale de } Y \text{ sur } L^2(X)}$$

Considérons le problème d'approximation plus simple de déterminer la **projection orthogonale $P_F(Y)$ de Y sur F := $\{\beta_0 + \beta_1 X : \beta_0, \beta_1$ réels $\}$** le sous-espace vectoriel de dimension 2 des fonctions affines de X.

En écrivant que $P_F(Y) = \boldsymbol{\beta_0 + \beta_1 X \in F}$ **et** $\boldsymbol{Y} - P_F(Y) \perp 1$ et X, on obtient (résolution d'un système linéaire à 2 inconnues) :

$$\mathbf{P_F(Y) = E(Y) + \frac{Cov(X, Y)}{Var(X)}(X - EX) = \beta_0 + \beta_1 X}$$

avec $$\mathbf{\beta_0 = E(Y) - \frac{Cov(X, Y)}{Var(X)} EX \; ; \; \beta_1 = \frac{Cov(X, Y)}{Var(X)}}$$

Formule d'analyse ou de décomposition de la variance (de Y) :

$$Y = P_F(Y) + \varepsilon \text{ où } \varepsilon := Y - P_F(Y) \perp 1 \text{ et } X$$

En particulier : $E(\varepsilon) = 0 \Rightarrow E(Y) = E[P_F(Y)]$ et $\mathrm{Cov}(\varepsilon, X) = 0$

(ANOVA) $\mathbf{Y - EY = P_F(Y) - EY + \varepsilon \text{ où } \varepsilon \perp P_F(Y) - EY}$

$\Rightarrow$ $\mathbf{Var(Y) = Var}(P_F(Y)) + \mathbf{Var(\varepsilon)} = \rho^2 \times \mathbf{Var(Y) + Var(\varepsilon)}$

où $\rho = \mathrm{Corr}(X, Y) = \dfrac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y}$ **est le fameux coefficient de corrélation linéaire.**

% de variance de Y expliquée par X : $100 \times \dfrac{\mathrm{Var}(P_F(Y))}{\mathrm{Var}(Y)} = 100 \times \rho^2\ \%$

A.N. Avec $\rho \approx \pm 0.87$, % de variance expliquée ≈ 75 %

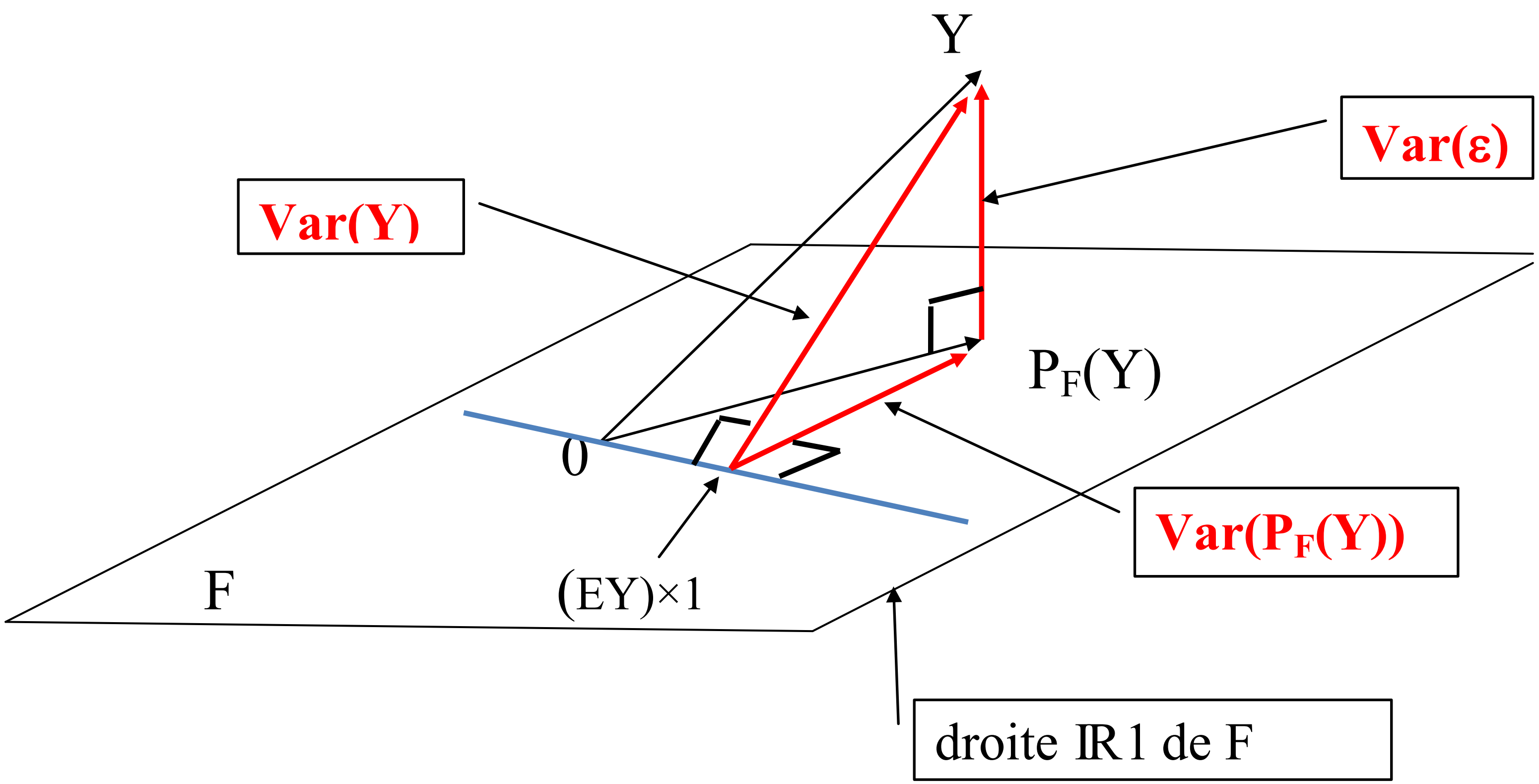


Illustration de : $\| Y - EY \|^2 = \| P_F(Y) - EY \|^2 + \| \varepsilon \|^2$

$\Leftrightarrow \qquad \mathrm{Var}(Y) = \mathrm{Var}(P_F(Y)) + \mathrm{Var}(\varepsilon)$

Pour aller plus loin dans l'analyse, supposons que le vecteur aléatoire (X, Y) soit **gaussien**. Alors :

- X et ε sont en réalité **indépendantes** (pas seulement non corrélées)
- $\mathbf{P_{L^2(X)}(Y) = P_F(Y)} = \beta_0 + \beta_1 X$
  - Pour tout x, la loi conditionnelle de Y sachant X = x est la loi normale

$$\mathbf{N(\beta_0 + \beta_1 x, Var(\varepsilon)) = N(\beta_0 + \beta_1 x, (1 - \rho^2) \times \sigma_Y^2)}$$

A.N. : $\rho \approx \pm\, 0.87 \Rightarrow 1 - \rho^2 \approx 0.25 = 0.5^2 \Rightarrow \sigma_\varepsilon = \sqrt{1-\rho^2}\, \sigma_Y \approx \frac{\sigma_Y}{2}$

On appelle $\mathbf{m : x \to \beta_0 + \beta_1 x}$ **fonction de régression** de sorte que

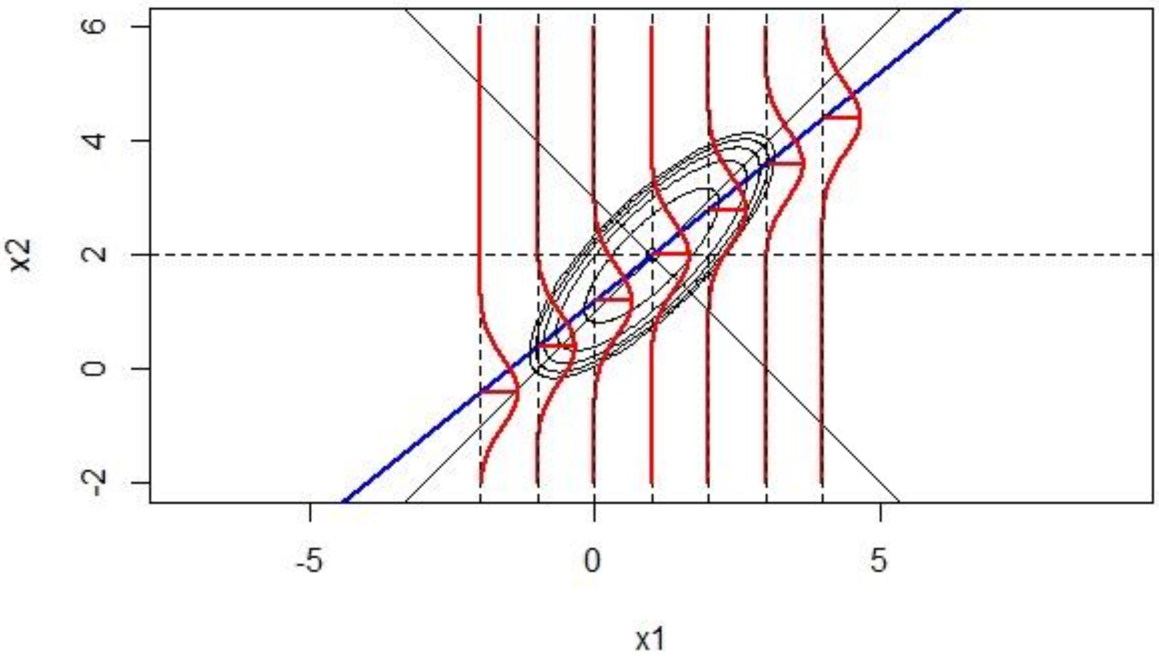
$$\mathbf{m(x) = E(Y \mid X = x)} = \textbf{espérance de la loi de } \mathbf{Y \mid X = x}$$

Comme $P_{L^2(X)}(Y) = P_F(Y) = m(X)$, on écrit encore que

$\mathbf{P_{L^2(X)}(Y) = E(Y \mid X)}$ **espérance de Y sachant X**

$\mathbf{P_F(Y) = E_L(Y \mid X)}$ **espérance linéaire de Y sachant X**

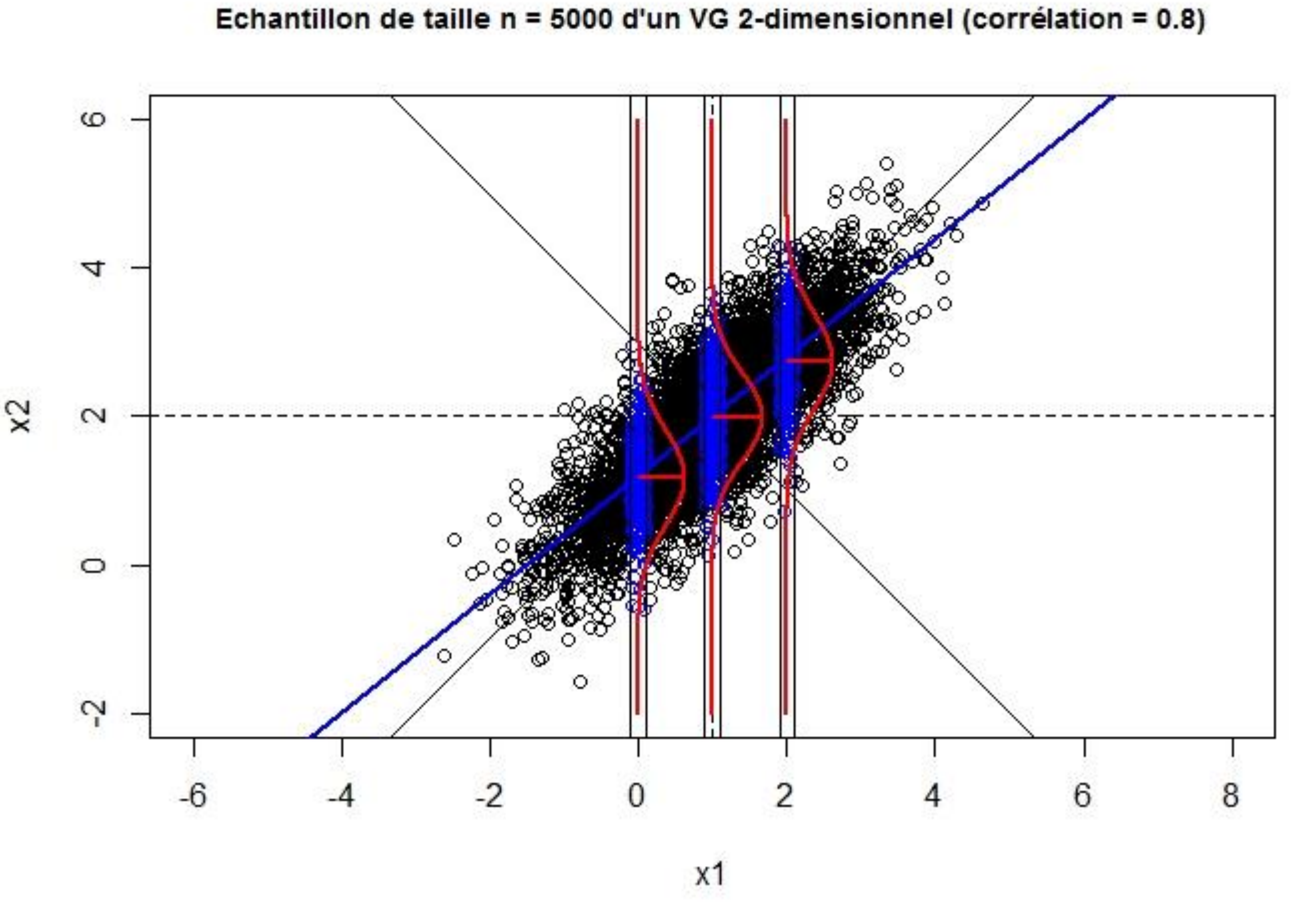
Illustration - lois conditionnelles dans le cas gaussien



Exercice. Supposons que le couple (X, Y) admette une densité de probabilité pas nécessairement gaussienne. Prouver que la fonction m : $x \to E(Y \mid X = x)$ est bien la solution du problème d'approximation :

$$\mathbf{P_{L^2(X)}(Y) = m(X)}$$

Aspect empirique (cf. TP2 du cours de Probabilités – exercice 2).



Soit (x_i, y_i), $1 \le i \le n$, un jeu de données de taille n du vecteur gaussien (X, Y).
On sait que

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

avec ε indépendante de X et de loi normale $N(0, \sigma^2)$.

Ecrivons

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

avec donc $\varepsilon_1, \ldots, \varepsilon_n$ réalisations indépendantes d'une loi $N(0, \sigma^2)$.

Questions.

- Estimation naturelle des coefficients β_0 et β_1 à partir des données ?
- Estimateur de σ^2

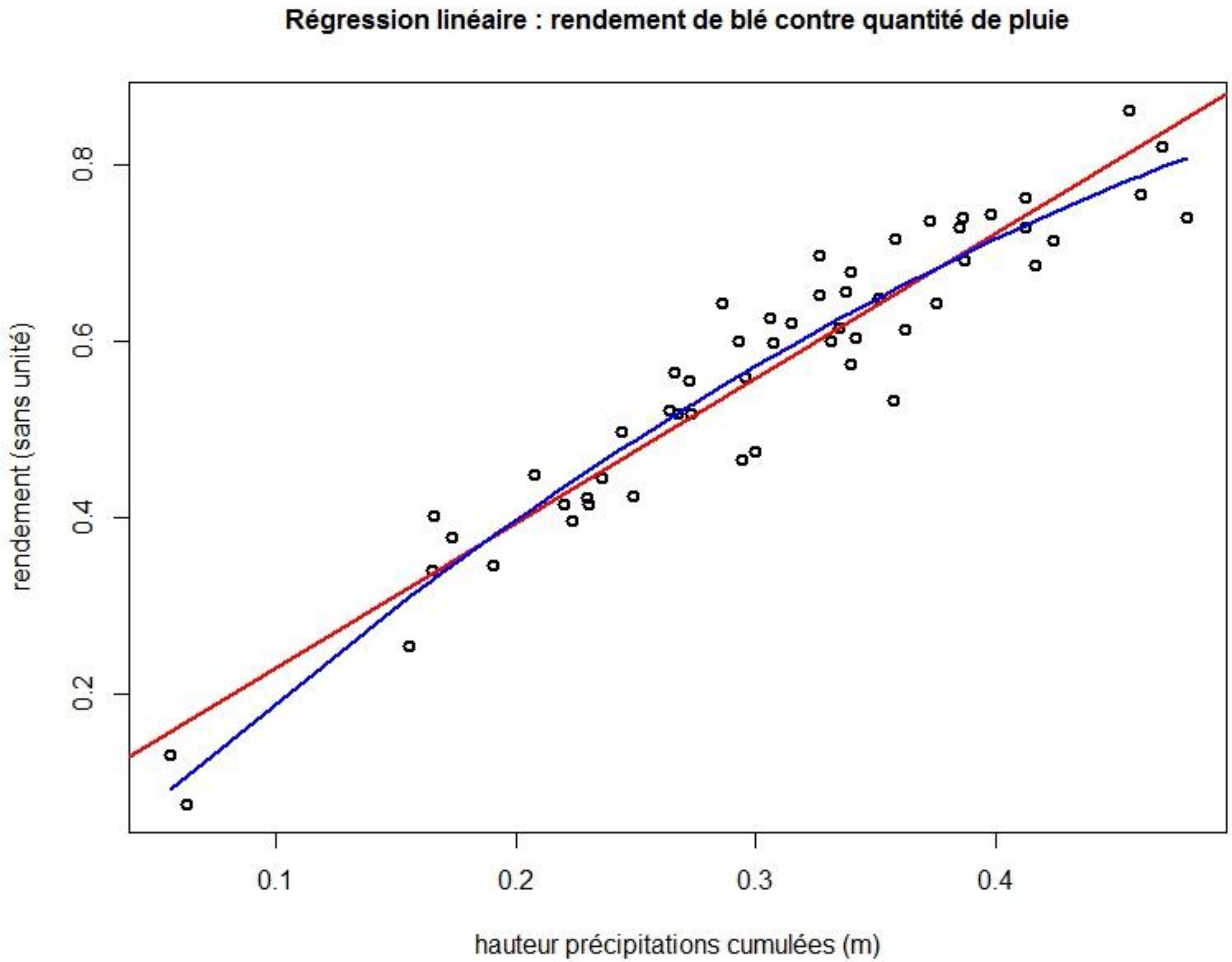
Le modèle de Régression Linéaire – dans le cas d'un vecteur gaussien (X, Y), on a une relation du type

$$Y = \beta_0 + \beta_1 X + \varepsilon \text{ avec } \varepsilon \sim N(0, \sigma^2) \text{ indépendante de X}$$

En particulier, la loi de Y sachant X = x est la loi normale $N(\beta_0 + \beta_1 x, \sigma^2)$ de moyenne $E(Y \mid X = x) = \beta_0 + \beta_1 x$ et de variance σ^2 **constante** (ne dépend pas de x).

C'est cette propriété de tout vecteur gaussien bidimensionnel qui est à la base du modèle linéaire de Régression avec des extensions majeures :

- Prise en compte de plusieurs prédicteurs
- Loi du résidu ε pas nécessairement gaussienne
- Pas d'hypothèse particulière sur la loi de X


Régression linéaire : rendement de blé contre quantité de pluie
rendement (sans unité)
0.2
0.4
0.6
0.8
0.1
0.2
0.3
0.4
hauteur précipitations cumulées (m)

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 »

☛ la variable x 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 théorique de régression linéaire simple (RLS) :

$$\mathbf{y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{x} + \boldsymbol{\varepsilon}$$

y = « **réponse** » est de loi connaissant *x* une **distribution** de moyenne

$$E(y \mid x) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{x}$$

et de variance « homogène » : $\mathbf{Var(y \mid \boldsymbol{x}) = Var(\boldsymbol{\varepsilon}) = \boldsymbol{\sigma}^2}$

Interprétation des coefficients de régression : β_0 ordonnée à l'origine (« intercept ») et β_1 = « pente » pour la **réponse espérée** ou **réponse moyenne** $E(y \mid x) = = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{x}$

☞ **p = 1 prédicteur**

Modèle empirique de RLS (décliné sur la population des n individus) **:**

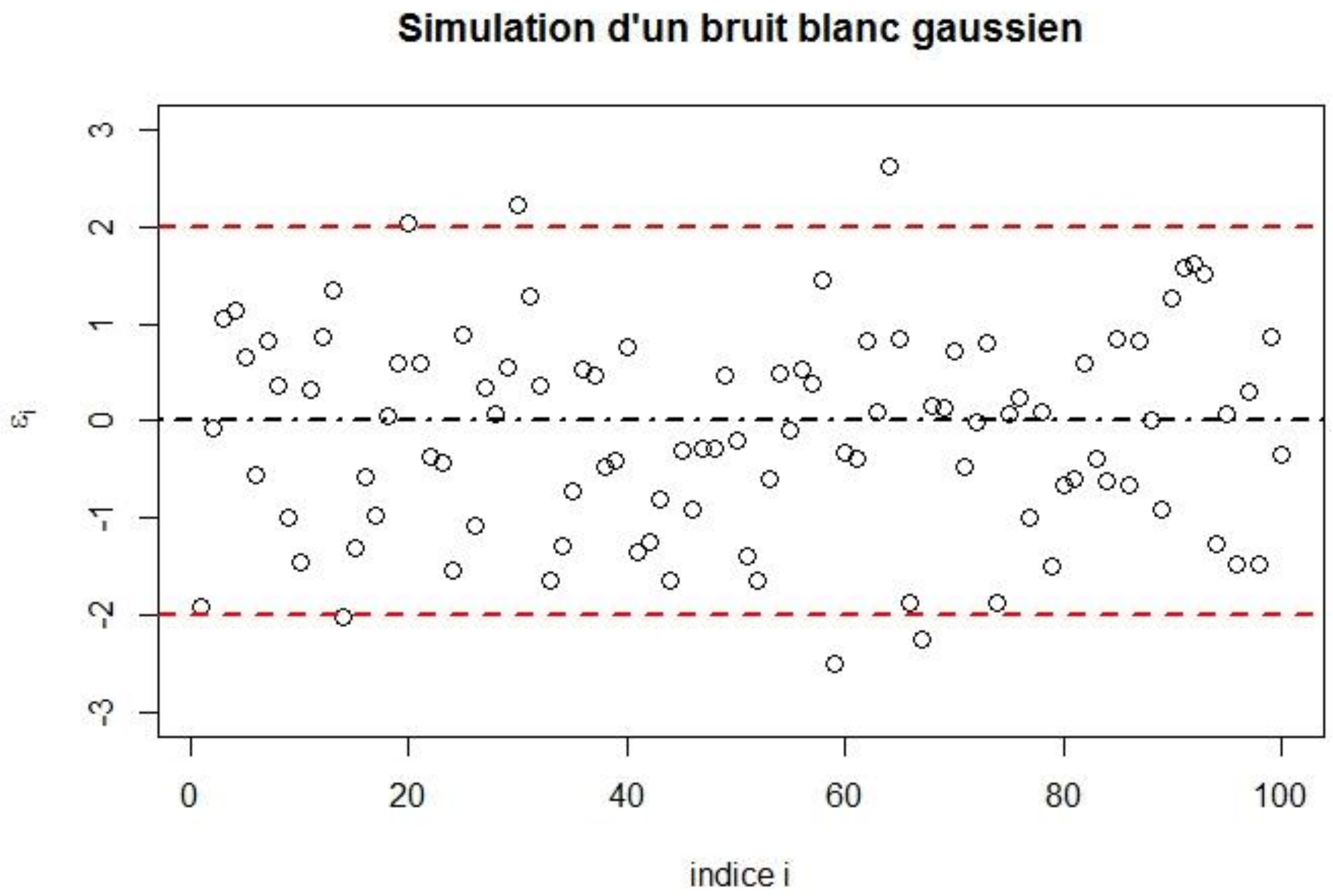
$$1 \leq i \leq n, \quad y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{où } \varepsilon_i \text{ résidu (théorique)}$$

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 $n\times(p+1) = n\times2$
- $\beta = (\beta_0\ \beta_1)'$ de taille $(p+1) = 2$: paramètres pour la **réponse espérée**
- $\varepsilon = (\varepsilon_1 \ldots \varepsilon_n)'$ vecteur colonne des **résidus théoriques** ou **erreurs** de régression (**bruit**)

HYPOTHÈSE FORTE : $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d. de loi $N(0, \sigma^2)$
Hypothèse plus faible : $\varepsilon_1, \ldots, \varepsilon_n$ centrées, de même variance σ^2 et non corrélées

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


Simulation d'un bruit blanc gaussien
ε_i
3
2
1
0
-1
-2
-3
0
20
40
60
80
100
indice i

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

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

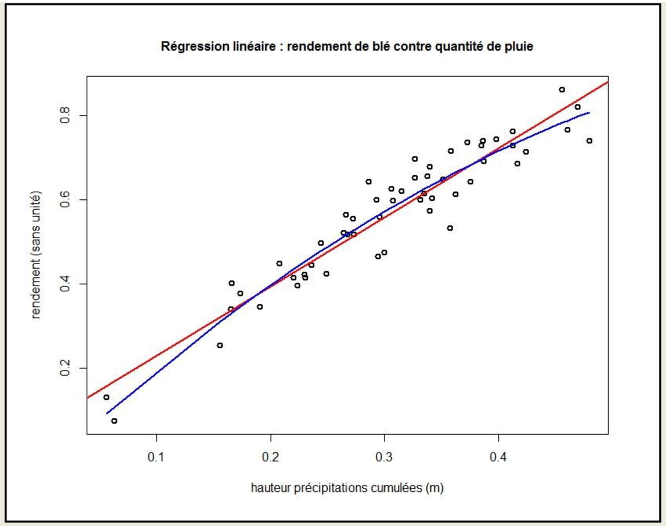
est un **Vecteur Gaussien** (VG) si ε est un bruit blanc gaussien $N(0, \sigma^2 I_n)$.

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

☛ Ainsi, on considère la matrice **X déterministe**, les seuls aléas sur les y_i proviennent des termes d'erreur ε_i pour $1 \leq i \leq n$.

Mise en œuvre sous R (cf. tutoriel TP1) :

```
> summary(lm(rend ~ pluie))

Call: lm(formula = rend ~ pluie)
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
```

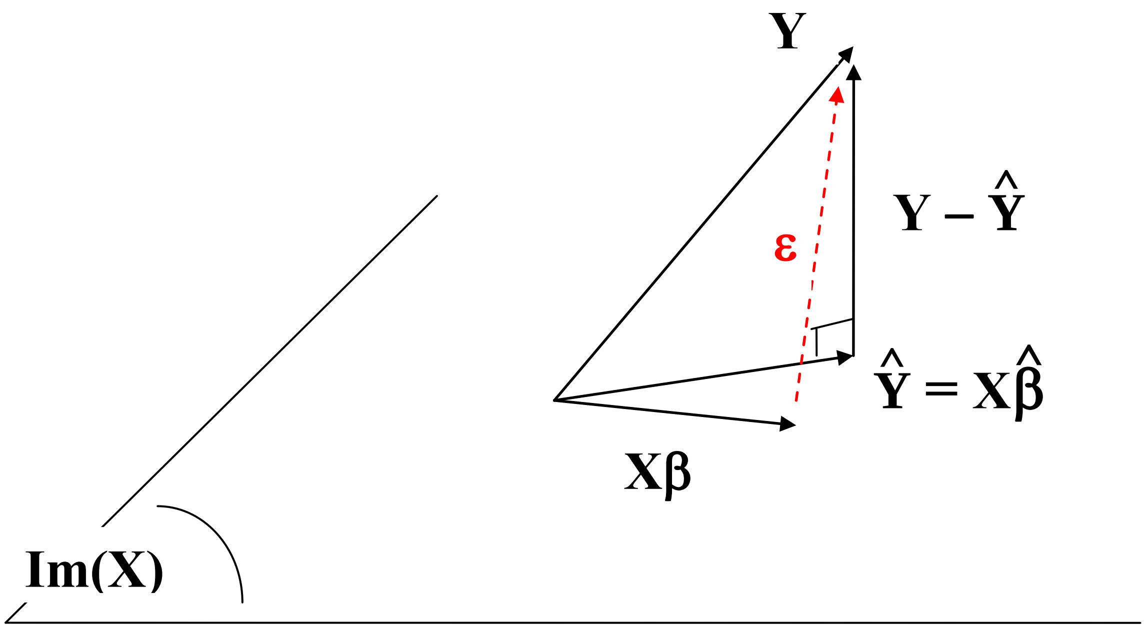


Estimation des paramètres β par moindres carrés (MC ou MCO) :

Le calcul explicite (via le calcul matriciel) :

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

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



L'obtention de β passe par la résolution des équations normales :

$$X'X\beta = X'Y$$

avec la petite réserve :

$X'X$ inversible $\Leftrightarrow$ X de rang $p+1 = 2$

$\Leftrightarrow$ colonnes de X = famille libre de vecteurs (variables)

THÉORÈME DE GAUSS-MARKOV : $\hat{\beta}$ est BLUE (**Best Linear Unbiaised Estimate**), i.e. :

(**Linear**) $\hat{\beta}$ est une fonction linéaire du vecteur des données Y

(**Unbiaised**) $\forall \beta,\ E(\hat{\beta}) = \beta$

(**Best**) $\forall \tilde{\beta}$ estimateur linéaire sans biais de β, $\forall \alpha$, $Var(\alpha'\hat{\beta}) \leq Var(\alpha'\tilde{\beta})$

De plus, $$\mathbf{Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}}$$

Commentaires

- Linear $\Rightarrow \hat{\beta}$ est un vecteur gaussien dans le cas gaussien (hypothèse forte)
- Unbiaised $\Rightarrow \hat{y}(x) = (1\ x)\,\hat{\beta}$ estimateur sans biais de $E(y \mid x)$
- Best $\Rightarrow Var(\hat{\beta}_k) \leq Var(\tilde{\beta}_k)$ pour $1 \leq k \leq p$

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

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

☞ $\hat{y}_i$ = **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 σ^2 des résidus par

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

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

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

Loi des estimateurs et statistiques pivotales (sous hypothèse forte)

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

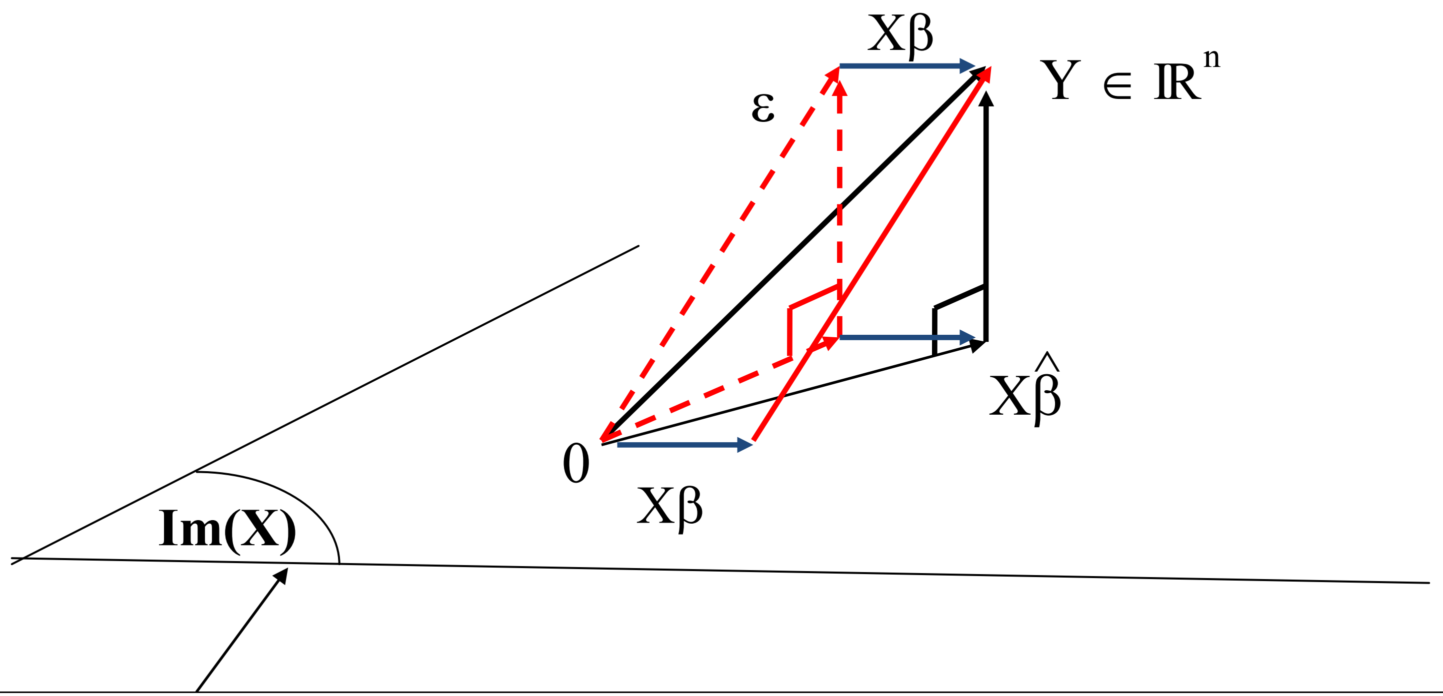
(ii) $\dfrac{(n-(p+1))\,\hat{\sigma}^2}{\sigma^2} = \dfrac{\| Y - X\hat{\beta} \|^2}{\sigma^2}$ est de **loi** $\boldsymbol{\chi^2_{n-2}}$ et $\hat{\sigma}$ est **indépendant** de $\hat{\beta}$

(iii) $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{c_j}\,\hat{\sigma}}$ est de **loi de Student** $\boldsymbol{t_{n-2}}$ où c_j terme diagonal de la matrice $(X'X)^{-1}$ correspondant à β_j

Preuve :

(i) Résulte de $\hat{\beta} = \beta + (X'X)^{-1}X'\varepsilon$ et ε gaussien $\sim N(0, \sigma^2 I_n)$

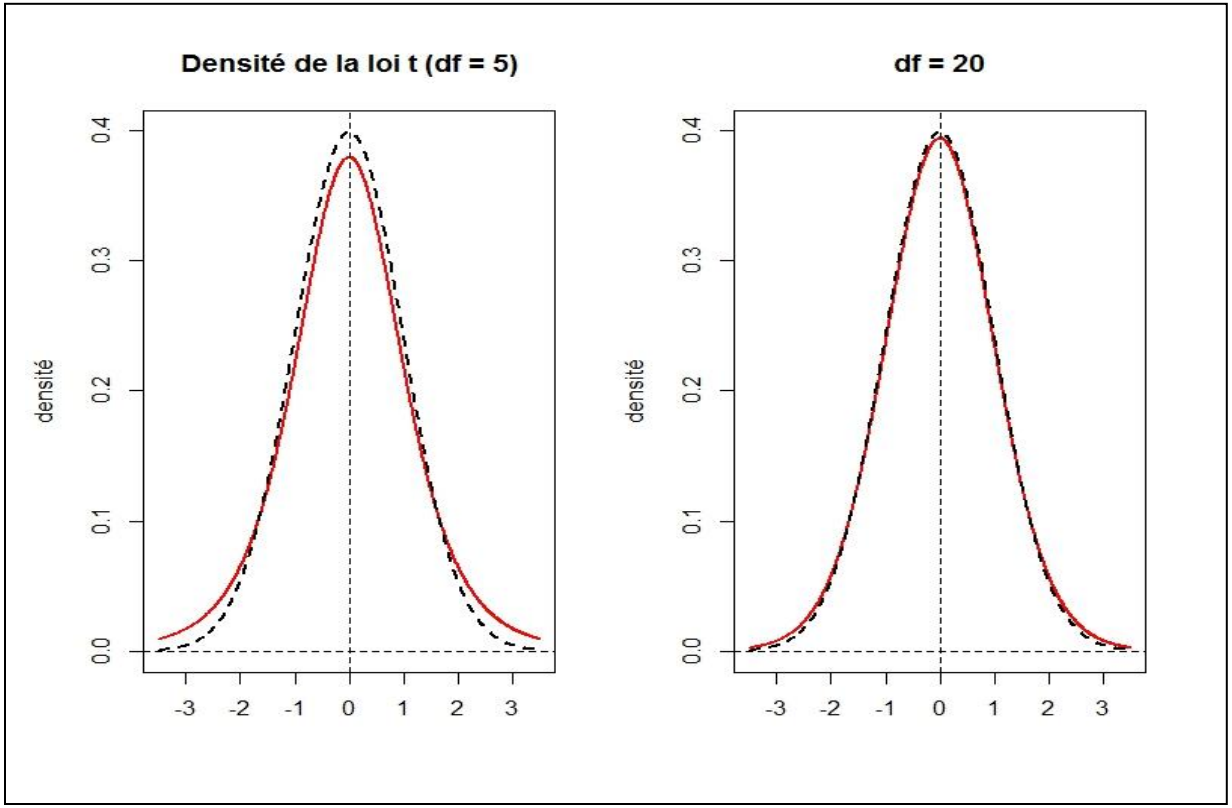
(ii) On utilise l'interprétation géométrique suivante de la régression



sous-espace de dim 2 engendré par les colonnes de X dans $\mathbb{R}^n$

Exercice. Construire les tests de nullité de β_0 et β_1. On utilisera la **statistique de Student** (à d = n – 2 dl ou df) : loi de $\dfrac{X}{\sqrt{Y/d}}$ où $X \sim N(0, 1)$ et $Y \sim \chi^2_d$ indépendantes

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



On peut donc maintenant analyser une partie du résultat retourné par la fonction **lm** sur l'exemple 1 :

```
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
```

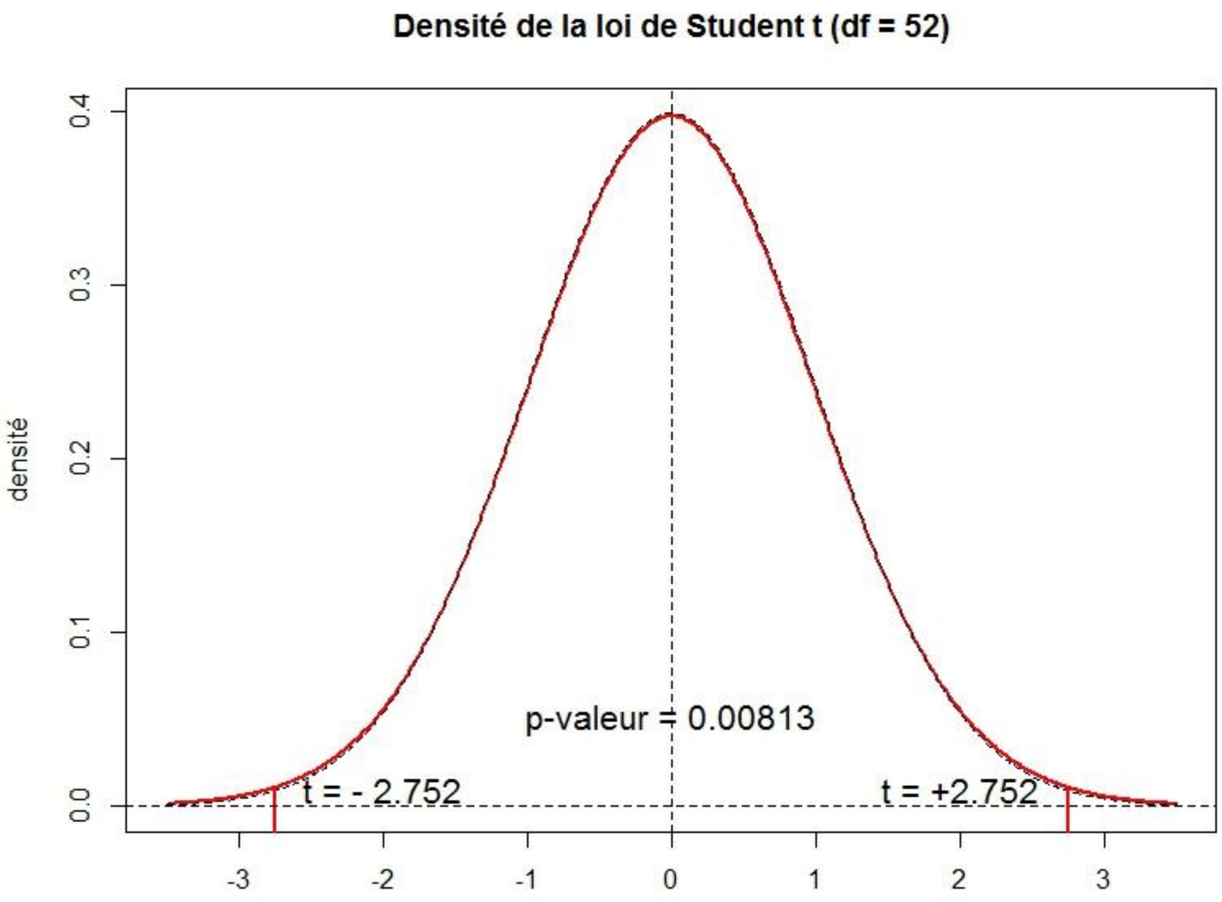
Par exemple, pour le coefficient β_0 (**intercept**), on lit

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

Sous l'hypothèse H$_0$: β_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)

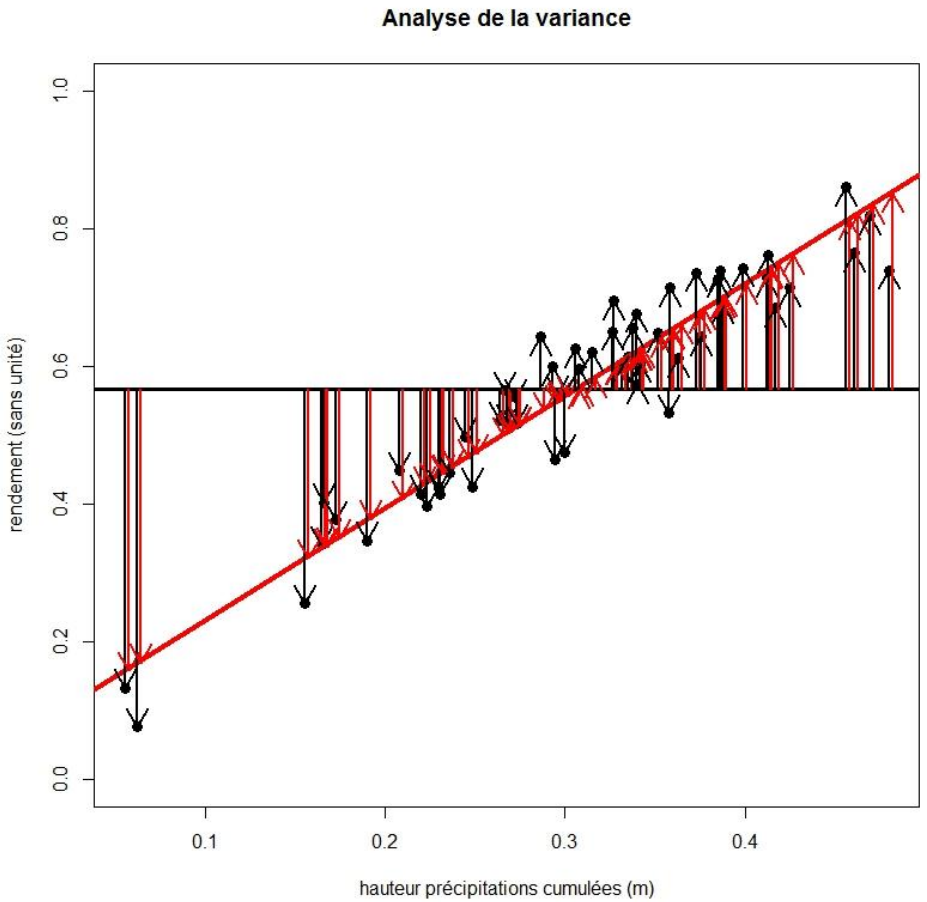
Reste à analyser la dernière partie

```
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
```

ce qui va se faire avec la table d'analyse de la variance (ANOVA)

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

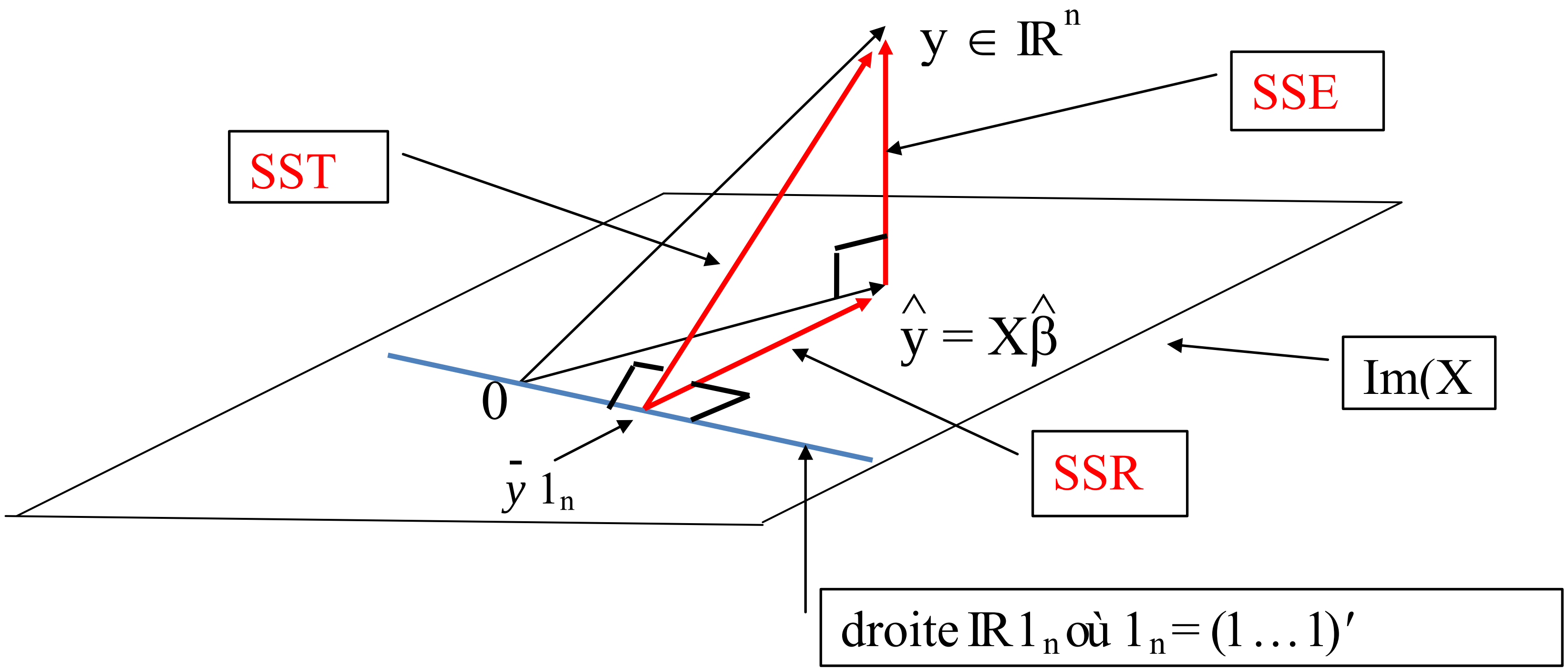
```
Analysis of Variance Table

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
```



Analyse de la « variabilité » de la réponse :

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

Vision géométrique :



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

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

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

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

☞ 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

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

☞ somme des carrés des erreurs ou des écarts résiduels

Considérons 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 :

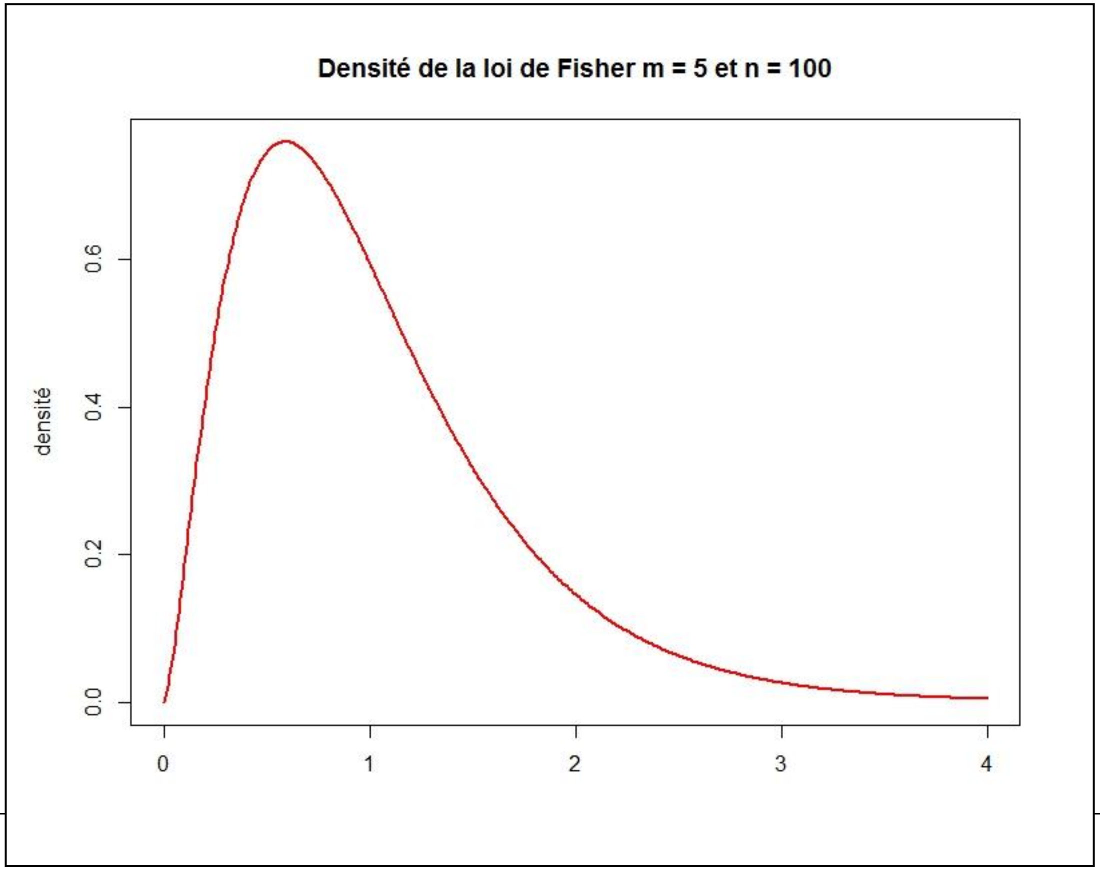
$$\frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)} \quad \textbf{est de loi de Fisher } F_{1,\, n-2}$$

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

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

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

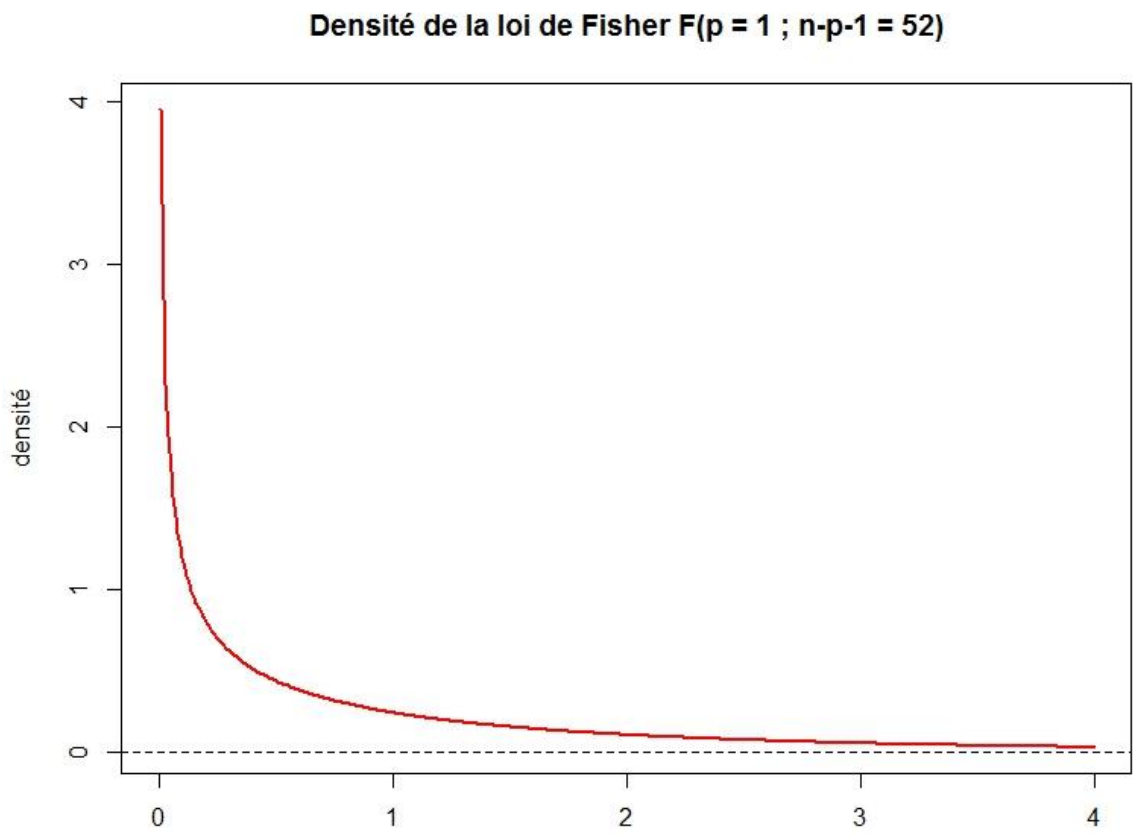
Exemple : m = 5 ; n = 100



On peut maintenant reprendre l'analyse

```
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
```



p-valeur < 2.2 10^{-16} $\Rightarrow$ « partie régression » très significative !

Reste à analyser la ligne :

```
Multiple R-squared:  0.9009, Adjusted R-squared:  0.899
```

Coefficient de détermination ou R^2 : $\boxed{R^2 = \frac{\mathbf{SSR}}{\mathbf{SST}}}$

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

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

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

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

Facteur de pénalisation

Exercice - Inférence avec un modèle de régression linéaire simple (p = 1)

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

Un intervalle de confiance de niveau α pour $x_{new}\beta$ (où $x_{new} = (x_{new}{}^{(0)}, x_{new}{}^{(1)}, \ldots, x_{new}{}^{(p)})$) 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 prévision pour la réponse**

Un intervalle de prévision de niveau α 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 $s_2(x_{new}) = \hat{\sigma}\sqrt{1 + x_{new}(X'X)^{-1}x_{new}'}$