Oniris
VetAgroBio Nantes
ÉCOLE NATIONALE

EC 651 : Modélisation Statistique

Guillaume Franchi

Cursus Ingénieur 1ère année

Partie 1 : La régression linéaire

Pré-requis

- Connaissances sur le logiciel R.
- Statistiques descriptives.
- Tests d'hypothèses et estimation.
- Un peu d'algèbre linéaire...

Bibiliographie

- R pour la statistique et la science des données, *François Husson et al.*, Presses Universitaires de Rennes.
- Probabilités, Analyse des données et Statistique, *Gilbert Saporta*, Editions Technip.
- Le polycopié d'*Arnaud Guyader*.

◎ Objectifs

- Expliquer un lien possible entre différentes variables décrivant des phénomènes physiques, biologiques, chimiques, etc...
- Prédire la valeur d'une variable réponse en fonction de variables explicatives.

Enseignement

- Partie 1 : Régression linaire (G. Franchi - 1 CM, 6 TDs).
- Partie 2 : Analyse de variance (ANOVA) (B.Mahieu - 1CM, 6 TDs).

Modalités d'évaluation

- 1 évaluation écrite pour la partie 2 : Analyse de variance.
- 1 projet pour la partie 1 : Régression linéaire.

1. La corrélation linéaire

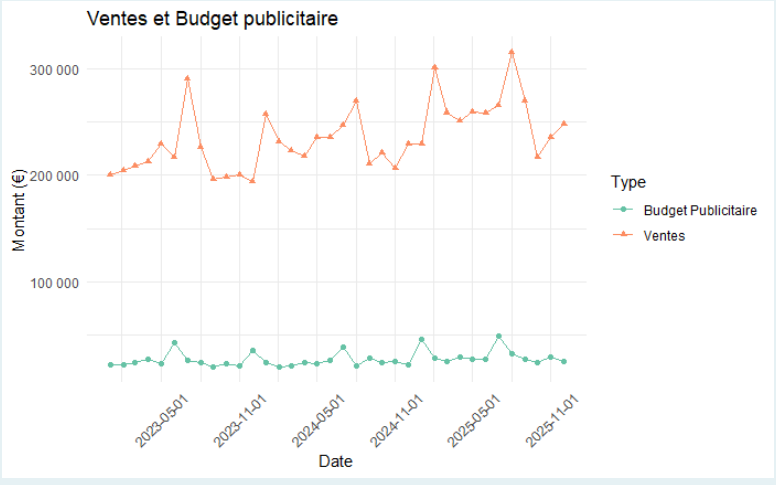
Exemple (*Investissement publicitaire*)

On a recensé les ventes d'un produit cosmétique ainsi que le budget publicitaire accordé à ce produit sur une durée de 36 mois.

Mois	Budget Pub (€)	Budget Pub mois précédent (€)	Ventes (€)
1	22 742	-	200 392
2	22 133	22 742	204 989
3	24 875	22 133	208 374
4	27 493	24 875	212 876
5	23 128	27 493	229 645
6	43 434	23 128	216 858

Exemple (*Investissement publicitaire*)

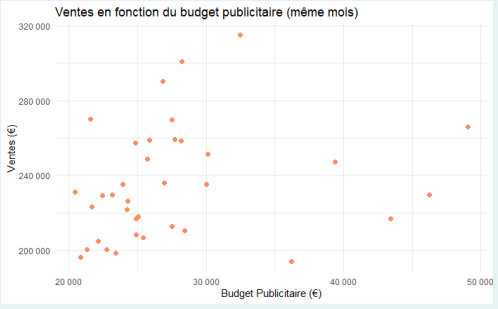
Ci-dessous, on a représenté ces valeurs sous forme de *séries temporelles*.



Exemple (*Investissement publicitaire*)

Question : Existe-t-il un lien entre le budget publicitaire et le montant des ventes réalisées ?

- Le nuage de points présentant les ventes réalisées en fonction du budget publicitaire ne montre pas de lien clair...



- Le coefficient de corrélation linéaire calculé pour ces deux variables est d'ailleurs $\rho \approx 0,21$.

Définition

Soient X et Y deux variables, dont on possède n observations $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$.

Le **coefficient de corrélation linéaire** entre les variables X et Y est

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2} \times \sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

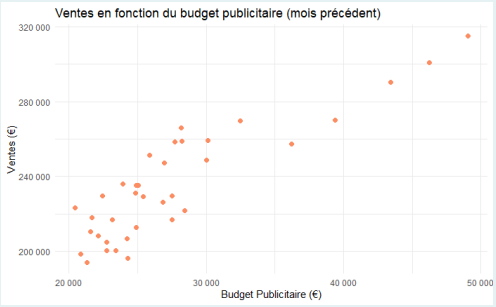
où $\overline{x}$ et $\overline{y}$ désignent les moyennes des observations des variables X et Y.

Remarques

- ρ est toujours compris entre -1 et 1 : $-1 \leqslant \rho \leqslant 1$.
- ρ mesure à quel point l'égalité $Y = aX + b$ est vraie sur les observations :
  - Si $\rho = 1$, on a $Y = aX + b$ avec $a > 0$.
  - Si $\rho = -1$, on a $Y = aX + b$ avec $a < 0$.
  - Si $\rho = 0$, les observations sont orthogonales (*elles sont non corrélées*).

Exemple (*Investissement publicitaire*)

- On représente maintenant les ventes réalisées chaque mois en fonction du budget publicitaire du mois précédent



- Le coefficient de corrélation linéaire calculé pour ces deux variables est $\rho \approx 0,88$.

⚙ La corrélation ρ calculée entre X et Y est-elle significativement non nulle, ou cela provient-il d'un effet d'échantillonnage ?

⚙ On se retrouve dans le cadre du test statistique :

$$H_0 : Cor(X, Y) = 0 \quad \text{contre} \quad H_1 : Cor(X, Y) \neq 0.$$

Propriété

Si le vecteur aléatoire (X, Y) est gaussien, on a sous l'hypothèse H_0

$$\frac{\rho}{\sqrt{\frac{1-\rho^2}{n-2}}} \sim \mathcal{T}_{n-2}$$

où $\mathcal{T}_{n-2}$ désigne la loi de Student à $n-2$ d.d.l.

Remarque

On rejettera donc l'hypothèse nulle, au niveau $\alpha > 0$, si

$$\frac{|\rho|}{\sqrt{\frac{1-\rho^2}{n-2}}} > t_{n-2}(1-\alpha/2).$$

Exemple (*Investissement publicitaire*)

On a les sorties R suivantes.

- Corrélation avec le budget publicitaire du même mois :

```
cor.test(df$`Budget Pub ()`,df$`Ventes ()`)
```

```
 Pearson's product-moment correlation

data:  df$`Budget Pub ()` and df$`Ventes ()`
t = 1.2587, df = 34, p-value = 0.2167
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.1262841  0.5045639
sample estimates:
      cor
0.2110061
```

Ici, on ne rejette pas l'hypothèse nulle : les ventes ne semblent pas corrélées au budget publicitaire du même mois (*au seuil de 5%)*.

Exemple (*Investissement publicitaire*)

- Corrélation avec le budget publicitaire du mois précédent :

```
cor.test(df$`Budget Pub mois précédent ()`,df$`Ventes ()`)
```

```
 Pearson's product-moment correlation

data:  df$`Budget Pub mois précédent ()` and df$`Ventes ()`
t = 10.829, df = 34, p-value = 1.46e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7765776 0.9377500
sample estimates:
      cor
0.8804748
```

Ici, on rejette l'hypothèse nulle : les ventes sont significativement corrélées avec le budget publicitaire du mois précédent (*au seuil de 5%)*.

2. La régression linéaire simple

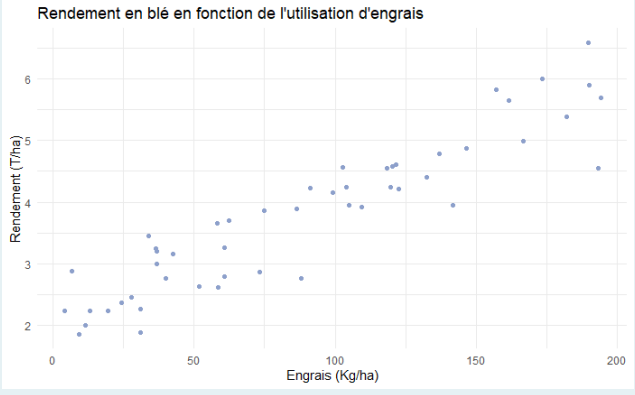
2.1 Contexte et exemple

On souhaite dorénavant **expliquer** le lien linéaire pouvant exister entre deux variables X et Y.

On souhaite également **prédire** la valeur de la variable Y en fonction de la variable X.

Exemple (*Engrais et blé*)

- On s'intéresse au lien entre la quantité d'engrais utilisé dans un champ de blé *(en Kg/ha)* et le rendement de ce champ *(en T/ha)*.
- On a ainsi mesuré ces deux quantités dans 50 champs.



Exemple (*Engrais et blé*)

- Une étude rapide de la corrélation montre un lien linéaire significatif *(au seuil de 5%)* entre ces deux quantités.

```
 Pearson's product-moment correlation

data:  df_wheat$Engrais and df_wheat$Rendement
t = 16.983, df = 48, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8724140 0.9574931
sample estimates:
      cor
0.9259172
```

- On cherche à expliquer davantage le lien linéaire entre le rendement d'un champ (valeurs y_i) et la quantité d'engrais utilisé (valeurs x_i).

2.2 Modélisation

Modèle linéaire simple

- On suppose dans la suite que les observations $y_1, \dots, y_n$ s'expliquent en fonction des $x_1, \dots, x_n$ selon le modèle

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad 1 \leqslant i \leqslant n$$

où $(\beta_0, \beta_1) \in \mathbb{R}^2$ et les ε_i sont des variables aléatoires, appelées bruits.

- On suppose de plus que les bruits $(\varepsilon_i)_{1 \leqslant i \leqslant n}$:
  - sont indépendants ;
  - sont de même loi $\mathcal{N}(0, \sigma^2)$ avec $\sigma > 0$.

⚙ Pour expliquer le lien entre les y_i et les x_i, il faut donc **estimer** les valeurs de β_0, β_1 et σ.

💡 Il s'agit du principe de la **régression linéaire**.

2.3 Moindres Carrés Ordinaires (MCO)

Définition

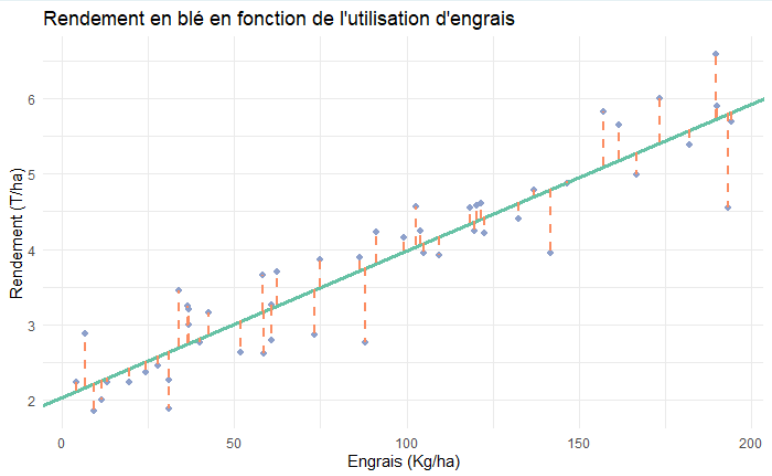
On appelle **estimateurs des moindres carrés ordinaires**, notés $\hat{\beta}_0$ et $\hat{\beta}_1$, les valeurs de β_0 et β_1 minimisant

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

i.e.

$$\left(\hat{\beta}_0, \hat{\beta}_1\right) = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{argmin} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2.$$

Schéma



Les estimateurs des MCO définissent la droite minimisant les écarts entre les observations et celle-ci.

Propriété

Les estimateurs des MCO ont pour expressions

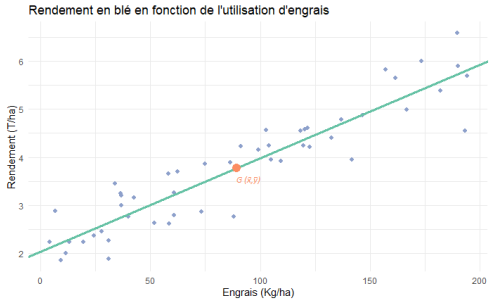
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \quad \text{et} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}.$$

Remarque :

La droite des MCO, donnée par l'équation

$$y = \hat{\beta}_0 + \hat{\beta}_1 x,$$

passe toujours par le centre de gravité $(\overline{x}, \overline{y})$ du nuage de points $((x_i, y_i))_{1 \leqslant i \leqslant n}$.



En R, la régression linéaire par les moindres carrés se fait avec la fonction `lm`.

Exemple (*Engrais et blé*)

Reprenons le cas du rendement d'un champ de blé.

```
lm_wheat <- lm(data=df_wheat,formula = Rendement~.)
summary(lm_wheat)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.25014 -0.24539  0.00236  0.34201  0.83999

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.048246   0.121292   16.89   <2e-16 ***
Engrais     0.019442   0.001145   16.98   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.463 on 48 degrees of freedom
Multiple R-squared:  0.8573, Adjusted R-squared:  0.8544
F-statistic: 288.4 on 1 and 48 DF,  p-value: < 2.2e-16
```

Ici, on a $\hat{\beta}_0 \approx 2.05$ et $\hat{\beta}_1 \approx 0.02$.

Lois des estimateurs des MCO

On a

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma_1^2) \quad \text{où} \quad \sigma_1^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

et

$$\hat{\beta}_0 \sim \mathcal{N}(\beta_0, \sigma_0^2) \quad \text{où} \quad \sigma_0^2 = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right).$$

Remarque :
Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont non biaisés.

Théorème (Gauss-Markov)

Les estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ sont de variance minimale parmi les estimateurs sans biais linéaires en $(y_i)_{1 \leqslant i \leqslant n}$.

Problème :
La valeur de σ^2 est inconnue.

2.4 Résidus

Définition

- Pour toute observation $i \in \{1, \dots, n\}$, on note

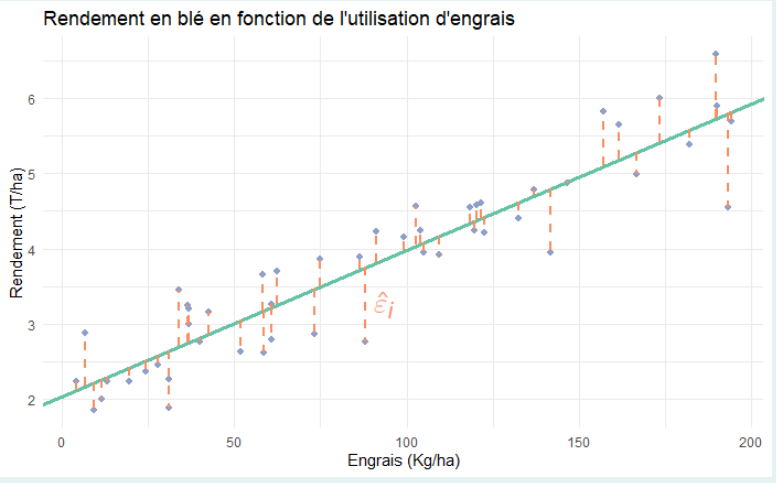
$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

  la **valeur ajustée** du modèle en x_i.
- Les **résidus** de la régression sont définis pour tout $i \in \{1, \dots, n\}$ par

$$\begin{aligned} \hat{\varepsilon}_i &= y_i - \hat{y}_i \\ &= y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right). \end{aligned}$$

Remarque

Les résidus correspondent aux erreurs d'ajustement du modèle.



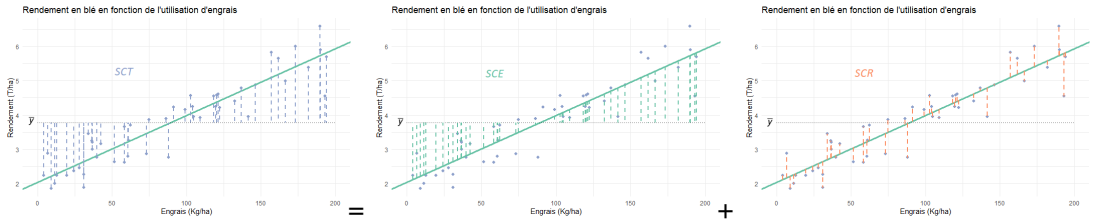
Propriété

On a

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

ou encore

$$SCT = SCE + SCR.$$



Définition

Le **coefficient de détermination** R^2 du modèle est défini par

$$\begin{aligned} R^2 &= \frac{SCE}{SCT} \\ &= 1 - \frac{SCR}{SCT} \\ &= (\text{Dispersion expliquée par le modèle})/(\text{Dispersion totale}). \end{aligned}$$

Remarques :

- C'est un nombre réel compris entre 0 et 1, représentant la proportion de variance expliquée par le modèle.
- Ce coefficient permet de mesurer l'adéquation du modèle aux données. Plus il est proche de 1, plus le modèle est adapté.

Exemple (*Engrais et blé*)

Dans notre cas, on a $R^2 \approx 0.8573$: la proportion de variance expliquée par le modèle est de 85,73%.

```
lm_wheat <- lm(data=df_wheat,formula = Rendement~.)
summary(lm_wheat)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.25014 -0.24539  0.00236  0.34201  0.83999

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.048246   0.121292   16.89   <2e-16 ***
Engrais     0.019442   0.001145   16.98   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.463 on 48 degrees of freedom
Multiple R-squared:  0.8573, Adjusted R-squared:  0.8544
F-statistic: 288.4 on 1 and 48 DF,  p-value: < 2.2e-16
```

2.5 Lois des estimateurs avec variance estimée

Propriétés

- La statistique $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$ est un estimateur non biaisé de σ^2.
- On a $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$.
- $\hat{\sigma}^2$ est indépendante de $\hat{\beta}_0$ et $\hat{\beta}_1$.

Exemple (*Engrais et blé*)

Dans notre cas, on estime $\hat{\sigma} \approx 0.463$.

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.25014 -0.24539  0.00236  0.34201  0.83999

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.048246   0.121292   16.89   <2e-16 ***
Engrais     0.019442   0.001145   16.98   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.463 on 48 degrees of freedom
Multiple R-squared:  0.8573, Adjusted R-squared:  0.8544
F-statistic: 288.4 on 1 and 48 DF,  p-value: < 2.2e-16
```

Lois des estimateurs avec variance estimée

- On note

$$\hat{\sigma}_1^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \quad \text{et} \quad \hat{\sigma}_0^2 = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \right)$$

  les estimateurs de $Var(\hat{\beta}_1)$ et $Var(\hat{\beta}_0)$.

- On a pour $i \in \{0, 1\}$:

$$\frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_i} \sim \mathcal{T}_{n-2}$$

  où $\mathcal{T}_{n-2}$ désigne la loi de Student à $n - 2$ d.d.l.

Remarque :

Ces propriétés permettent de construire des intervalles de confiance pour β_0, β_1 et σ^2.

Propriétés

- Pour $i \in \{0, 1\}$, un intervalle de confiance pour β_i au niveau $1 - \alpha$ est donné par

$$IC(\beta_i) = \left[\hat{\beta}_i - \hat{\sigma}_i \times t_{n-2}(1 - \alpha/2) \; ; \; \hat{\beta}_i + \hat{\sigma}_i \times t_{n-2}(1 - \alpha/2)\right]$$

  où $t_{n-2}(1 - \alpha/2)$ désigne le quantile d'ordre $(1 - \alpha/2)$ de la loi de Student $\mathcal{T}_{n-2}$.

- Un intervalle de confiance pour σ^2 au niveau $1 - \alpha$ est donné par

$$IC(\sigma^2) = \left[\frac{(n-2)\hat{\sigma}^2}{c_{n-2}(1 - \alpha/2)} \; ; \; \frac{(n-2)\hat{\sigma}^2}{c_{n-2}(\alpha/2)}\right]$$

  où $c_{n-2}(\alpha/2)$ et $c_{n-2}(1 - \alpha/2)$ désignent respectivement les quantiles d'ordres $\alpha/2$ et $(1 - \alpha/2)$ de la loi χ^2_{n-2}.

Exemple (Engrais et blé)

- Dans notre exemple, on a $\hat{\beta}_0 \approx 2.05$ et $\hat{\beta}_1 \approx 0.0194$.

```
lm_wheat <- lm(data=df_wheat,formula = Rendement~.)
lm_wheat$coefficients
```

```
(Intercept)     Engrais
 2.04824563  0.01944231
```

- Des intervalles de confiance de niveau 0.95 pour ces valeurs sont donnés par

$$IC(\beta_0) \approx [1.80\ ;\ 2.29] \quad \text{et} \quad IC(\beta_1) \approx [0.017\ ;\ 0.022].$$

```
confint(lm_wheat,level=0.95)
```

```
                2.5 %    97.5 %
(Intercept) 1.80437222 2.2921190
Engrais     0.01714052 0.0217441
```

2.6 Nouvelles prévisions

Supposons que l'on ait une nouvelle observation x_{n+1} de la variable explicative X, indépendante des observations $x_1, \ldots, x_n$.

On cherche naturellement à prédire la valeur associée y_{n+1} de la variable Y par

$$\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}.$$

Question : Comment mesurer l'incertitude sur notre prévision $\hat{y}_{n+1}$?

Propriété

On note toujours ici : $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

- On a

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}} \sim \mathcal{T}_{n-2}.$$

- On en déduit un intervalle de confiance de niveau $1-\alpha$ pour y_{n+1}

$$IC(y_{n+1}) = \left[\hat{y}_{n+1} \pm t_{n-2}(1-\alpha/2)\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)}\right]$$

où $t_{n-2}(1-\alpha/2)$ désigne le quantile d'ordre $(1-\alpha/2)$ de la loi $\mathcal{T}_{n-2}$.

Exemple (*Engrais et blé*)

- Supposons que dans 3 champs de blés différents, on ait utilisé 75 Kg, 130 Kg et 180 Kg d'engrais par hectare.
- Ci-dessous, on présente les rendements estimés pour ces champs, avec leurs intervalles de confiance au niveau 95%.

```
new_obs <- data.frame(Engrais = c(75,130,180))
predict(lm_wheat,newdata = new_obs,interval = "prediction")
```

```
       fit      lwr      upr
1 3.506419 2.565659 4.447179
2 4.575746 3.630870 5.520621
3 5.547861 4.584710 6.511012
```

2.7 Nullité des coefficients

⚙ Une question statistique d'intérêt :

Les paramètres β_0 et β_1 sont-ils significativement non nuls ?

⚙ En effet, même si on a comme estimations $\hat{\beta}_0 \neq 0$ et $\hat{\beta}_1 \neq 0$, il se peut que l'on ait pour les vraies valeurs $\beta_0 = 0$ ou $\beta_1 = 0$.

Remarque : Si $\beta_1 = 0$, cela signifie que la variable *X* n'explique en fait pas la variable *Y* (*du moins linéairement...)*.

Pour $i \in \{0, 1\}$, on considère le test d'hypothèses :

$$H_0 : \beta_i = 0 \quad \text{contre} \quad H_1 : \beta_i \neq 0.$$

Propriété

- Sous l'hypothèse nulle H_0, on a

$$\frac{\hat{\beta}_i}{\hat{\sigma}_i} \sim \mathcal{T}_{n-2}.$$

- Donc, si $\left|\frac{\hat{\beta}_i}{\hat{\sigma}_i}\right| > t_{n-2}(1 - \alpha/2)$, on rejette H_0 au niveau α.

Exemple (*Engrais et blé*)

Dans notre exemple, on rejette l'hypothèse nulle pour chacune des coefficients. Ceux-ci sont donc significativement non nuls au seuil de 5%.

```
lm_wheat <- lm(data=df_wheat,formula = Rendement~.)
summary(lm_wheat)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.25014 -0.24539  0.00236  0.34201  0.83999

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.048246   0.121292   16.89   <2e-16 ***
Engrais     0.019442   0.001145   16.98   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.463 on 48 degrees of freedom
Multiple R-squared:  0.8573, Adjusted R-squared:  0.8544
F-statistic: 288.4 on 1 and 48 DF,  p-value: < 2.2e-16
```

3. La régression linéaire multiple

3.1 Contexte et exemple

La modélisation du rendement d'un champ effectuée précédemment est simpliste : d'autres variables peuvent l'expliquer comme la qualité du sol, la température moyenne, etc...

On va chercher à expliquer une variable Y en fonction de plusieurs variables explicatives $X_1, X_2, \ldots, X_p$.

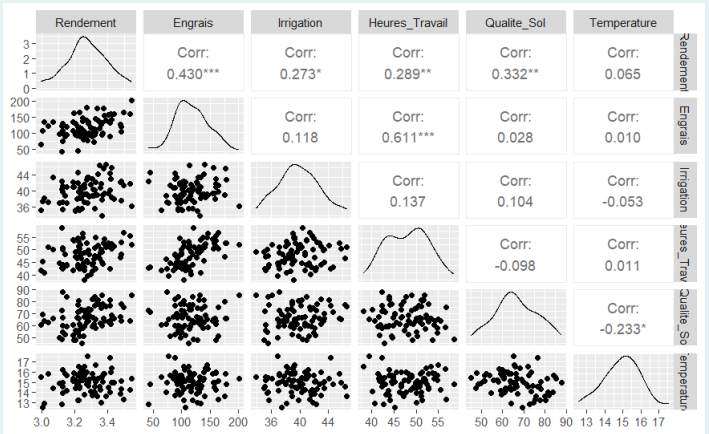
Exemple (*Champs de tomates*)

On a cette fois-ci le rendement de $n = 80$ champs de tomates, supposés indépendants, ainsi que de cinq autres variables explicatives non constantes.

	Y (T/ha)	X1 (Kg/ha)	X2 (mm/ha)	X3 (H/ha)	X4	X5 (°C)
	Rendement	Engrais	Irrigation	Heures_Travail	Qualite_Sol	Temperature
1	3.234626	113.79049	40.01729	52.15813	57.11378	15.23743
2	3.125442	105.39645	41.15584	43.85247	59.97801	16.21811
3	3.337675	176.17417	38.88802	47.21953	79.96061	13.66123
4	3.216521	96.41017	41.93313	54.72312	53.62696	15.66082
5	3.402045	157.58575	39.33854	52.74943	63.20948	14.47709
6	3.151549	119.30130	40.99535	43.89468	84.02362	15.68375

Exemple (*Champs de tomates*)

On représente ci-dessous les corrélations linéaires pouvant exister entre ces variables.



3.2 Modélisation

Notations :

- Y la variable à expliquer : observations $y_1, \ldots, y_n$.
- p le nombre de variables explicatives $X_1, \ldots, X_p$.
- Pour tout $j \in \{1, \ldots, p\}$, les observations de X_j sont :

$$x_{1,j}, x_{2,j}, \ldots, x_{n,j}.$$

- On notera X la matrice de design

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,p} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,p} \\ \vdots & \vdots & & \vdots \\ x_{n,1} & x_{n,2} & \ldots & x_{n,p} \end{bmatrix}$$

X1 Engrais	*X2* Irrigation	*X3* Heures_Travail	*X4* Qualite_Sol	*X5* Temperature
113.79049	40.01729	52.15813	57.11378	15.23743
105.39645	41.15584	43.85247	59.97801	16.21811
176.17417	38.88802	47.21953	79.96061	13.66123
96.41017	41.93313	54.72312	53.62696	15.66082
157.58575	39.33854	52.74943	63.20948	14.47709
119.30130	40.99535	43.89468	84.02362	15.68375

supposée de rang p (*aucune variable ne s'écrit linéairement en fonction des autres*).

Modèle linéaire multiple

- On suppose dans la suite que les observations $y_1, \dots, y_n$ s'expliquent selon le modèle

$$y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p} + \varepsilon_i = \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i, \quad 1 \leqslant i \leqslant n$$

  où les β_j dont des réels et les ε_i des variables aléatoires (*bruits*).
- On suppose toujours que les bruits $(\varepsilon_i)_{1 \leqslant i \leqslant n}$ sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$, où $\sigma > 0$.

Remarque

On a en fait

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

ce qui s'écrit $Y = X \cdot \beta + \varepsilon$, où $\varepsilon \sim \mathcal{N}_{\mathbb{R}^n}(0, \sigma^2 I_n)$.

Remarque

- Dans la plupart des cas, l'équation de la régression contient une constante, correspondant au cas où la variable X_1 est constante et égale à 1 ($x_{i,1} = 1$ *pour tout i)*.
  On a alors
$$y_i = \beta_1 + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}.$$
- La matrice de design s'écrit alors
$$X = \begin{bmatrix} 1 & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n,2} & \dots & x_{n,p} \end{bmatrix}$$

3.3 Moindres Carrés Ordinaires (MCO)

Définition

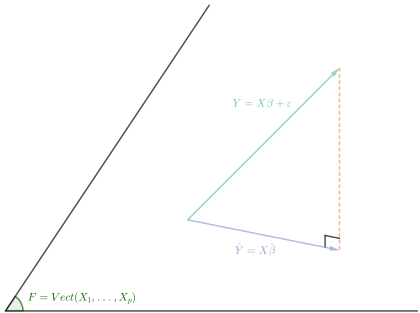
On appelle **estimateur des moindres carrés ordinaires**, noté $\hat{\beta} \in \mathbb{R}^p$

$$\begin{aligned}\hat{\beta} &= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j x_{i,j} \right)^2 \\ &= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \, \|Y - X\beta\|^2.\end{aligned}$$

Remarque : Déterminer

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{argmin} \, \|Y - X\beta\|^2$$

revient à déterminer le projeté orthogonal de Y sur $F = Vect(X_1, \ldots, X_p)$, le sous espace de dimension p de R^n engendré par les vecteurs colonnes de X.



Propriété

L'estimateur des MCO a pour expression

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y.$$

Exemple (*Champs de tomates*)

Dans notre exemple, où l'on cherche à expliquer le rendement *(observations y_i)* en fonction des autres variables, on a :

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.1283588  0.2857600   7.448 1.42e-10 ***
Engrais        0.0013818  0.0004776   2.894 0.005001 **
Irrigation     0.0082729  0.0040953   2.020 0.046994 *
Heures_Travail 0.0021900  0.0031519   0.695 0.489337
Qualite_Sol    0.0041781  0.0011842   3.528 0.000723 ***
Temperature    0.0180394  0.0115351   1.564 0.122114
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1016 on 74 degrees of freedom
Multiple R-squared:  0.3499, Adjusted R-squared:  0.306
F-statistic: 7.965 on 5 and 74 DF,  p-value: 4.767e-06
```

Loi des estimateurs des MCO

- On a

$$\hat{\beta} \sim \mathcal{N}_{\mathbb{R}^p}\left(\beta, \sigma^2 \left(X^T X\right)^{-1}\right).$$

- En particulier, pour tout $j \in \{1, \ldots, p\}$

$$\hat{\beta}_j \sim \mathcal{N}\left(\beta_j, \sigma^2 \left(X^T X\right)^{-1}_{jj}\right).$$

Remarque : L'estimateur $\hat{\beta}$ est donc non biaisé.

Problème : La valeur de σ^2 est inconnue.

3.4 Résidus

Définition

- Le vecteur des **valeurs ajustées** du modèle est donné par

$$\hat{Y} = X \cdot \hat{\beta} \in \mathbb{R}^n$$

On a en fait $\hat{Y} = (\hat{y}_1, \ldots, \hat{y}_n)^T$ où $\hat{y}_i = \sum_{j=1}^{p} \hat{\beta}_j x_{i,j}, \quad 1 \leqslant i \leqslant n.$

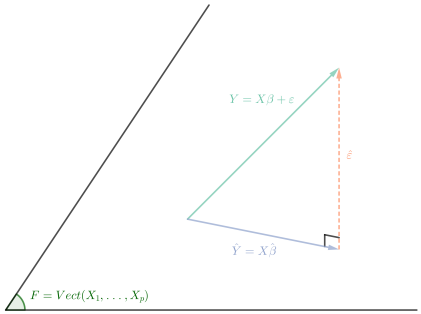
- Le vecteur des **résidus** est donné par

$$\hat{\varepsilon} = Y - \hat{Y}.$$

On a en fait $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^T$ où $\hat{\varepsilon}_i = y_i - \hat{y}_i, \quad 1 \leqslant i \leqslant n.$

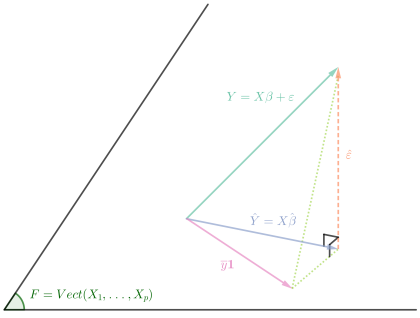
Remarque :

Par définition, le projeté orthogonal de Y sur $F = Vect(X_1, \ldots, X_p)$ est donné par $\hat{Y} = X\hat{\beta}$.



La constante ne fait pas partie du modèle *(X_1 n'est pas constante et égale à 1)*.
Les vecteurs $\hat{Y}$ et $\hat{\varepsilon}$ sont orthogonaux :

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



La constante fait partie du modèle *($X_1 = (1, \ldots, 1)^T = \mathbb{1}$)*.
Les vecteurs $\hat{Y} - \overline{y}\mathbb{1}$ et $\hat{\varepsilon}$ sont orthogonaux :

$$\|Y - \overline{y}\mathbb{1}\|^2 = \|\hat{Y} - \overline{y}\mathbb{1}\|^2 + \|\hat{\varepsilon}\|^2.$$

Dans les deux cas, on note $SCT = SCE + SCR$.

Définition

Le **coefficient de détermination** R^2 est défini par

$$R^2 = (\text{Dispersion expliquée par le modèle})/(\text{Dispersion totale}) = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}.$$

- Si la constante ne fait pas partie du modèle :

$$R^2 = \frac{\|\hat{Y}\|^2}{\|Y\|^2} = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y^2\|}.$$

- Si la constante fait partie du modèle :

$$R^2 = \frac{\|\hat{Y} - \overline{y}\mathbb{1}\|^2}{\|Y - \overline{y}\mathbb{1}\|^2} = 1 - \frac{\|\hat{\varepsilon}\|^2}{\|Y - \overline{y}\mathbb{1}\|^2}.$$

Remarque : Il s'agit toujours d'un nombre réel compris entre 0 et 1, donnant la proportion de variance expliquée par le modèle. Plus il est proche de 1, plus le modèle est adapté aux données.

Exemple (*Champs de tomates*)

Le modèle linéaire explique environ 35% de la variance dans notre étude sur le rendement des champs de tomates.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.1283588  0.2857600   7.448 1.42e-10 ***
Engrais        0.0013818  0.0004776   2.894 0.005001 **
Irrigation     0.0082729  0.0040953   2.020 0.046994 *
Heures_Travail 0.0021900  0.0031519   0.695 0.489337
Qualite_Sol    0.0041781  0.0011842   3.528 0.000723 ***
Temperature    0.0180394  0.0115351   1.564 0.122114
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1016 on 74 degrees of freedom
Multiple R-squared:  0.3499, Adjusted R-squared:  0.306
F-statistic: 7.965 on 5 and 74 DF,  p-value: 4.767e-06
```

Remarque :

Plus il y a de variables à expliquer la variable réponse, plus le R^2 sera proche de 1.

Afin de tenir compte de la dimension de l'espace de projection, on peut considérer la définition suivante.

Définition

Le **coefficient de détermination ajusté**, noté R_a^2 est défini par :

- $R_a^2 = 1 - \frac{n}{n-p} \times \frac{\|\hat{\varepsilon}\|^2}{\|Y\|^2} = 1 - \frac{n}{n-p}(1 - R^2)$ si la constante ne fait pas partie du modèle.
- $R_a^2 = 1 - \frac{n-1}{n-p} \times \frac{\|\hat{\varepsilon}\|^2}{\|Y - \overline{y}\mathbb{1}\|^2} = 1 - \frac{n-1}{n-p}(1 - R^2)$ si la constante fait partie du modèle.

Exemple (*Champs de tomates*)

Le coefficient de détermination ajusté est d'environ 0.306 dans notre modèle.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.1283588  0.2857600   7.448 1.42e-10 ***
Engrais        0.0013818  0.0004776   2.894 0.005001 **
Irrigation     0.0082729  0.0040953   2.020 0.046994 *
Heures_Travail 0.0021900  0.0031519   0.695 0.489337
Qualite_Sol    0.0041781  0.0011842   3.528 0.000723 ***
Temperature    0.0180394  0.0115351   1.564 0.122114
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1016 on 74 degrees of freedom
Multiple R-squared:  0.3499, Adjusted R-squared:  0.306
F-statistic: 7.965 on 5 and 74 DF,  p-value: 4.767e-06
```

3.5 Lois des estimateurs avec variance estimée

Propriétés

- La statistique $\hat{\sigma}^2 = \frac{1}{n-p}\|\hat{\varepsilon}\|^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$ est un estimateur non baisé de σ^2.
- On a $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.
- $\hat{\sigma}^2$ est indépendante de $\hat{\beta}$.

Exemple (*Champs de tomates*)

Dans notre cas, on estime $\hat{\sigma} \approx 0.1016$.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.1283588  0.2857600   7.448 1.42e-10 ***
Engrais      0.0013818  0.0004776   2.894 0.005001 **
Irrigation   0.0082729  0.0040953   2.020 0.046994 *
Heures_Travail 0.0021900  0.0031519   0.695 0.489337
Qualite_Sol  0.0041781  0.0011842   3.528 0.000723 ***
Temperature  0.0180394  0.0115351   1.564 0.122114
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1016 on 74 degrees of freedom
Multiple R-squared:  0.3499, Adjusted R-squared:  0.306
F-statistic: 7.965 on 5 and 74 DF,  p-value: 4.767e-06
```

Lois des estimateurs avec variance estimée

On rappelle que pour tout $j \in \{1, \dots, p\}$

$$\hat{\beta}_j \sim \mathcal{N}\left(0, \sigma^2 \left(X^T X\right)^{-1}_{jj}\right).$$

On note alors $\hat{\sigma}_j^2 = \hat{\sigma}^2 \left(X^T X\right)^{-1}_{jj}$, et on a

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim \mathcal{T}_{n-p}$$

où $\mathcal{T}_{n-p}$ désigne la loi de Student à $n - p$ d.d.l.

Remarque :
Ces propriétés permettent de construire des intervalles de confiance pour les β_j et σ^2.

Propriétés

- Pour tout $j \in \{1, \ldots, p\}$, un intervalle de confiance pour β_j au niveau $1 - \alpha$ est donné par

$$IC(\beta_j) = \left[\hat{\beta}_j - \hat{\sigma}_j \times t_{n-p}(1-\alpha/2) \; ; \; \hat{\beta}_j + \hat{\sigma}_j \times t_{n-p}(1-\alpha/2)\right]$$

où $t_{n-p}(1-\alpha/2)$ désigne le quantile d'ordre $(1-\alpha/2)$ de la loi de Student $\mathcal{T}_{n-p}$.

- Un intervalle de confiance pour σ^2 au niveau $1 - \alpha$ est donné par

$$IC(\sigma^2) = \left[\frac{(n-p)\hat{\sigma}^2}{c_{n-p}(1-\alpha/2)} \; ; \; \frac{(n-p)\hat{\sigma}^2}{c_{n-p}(\alpha/2)}\right]$$

où $c_{n-p}(\alpha/2)$ et $c_{n-p}(1-\alpha/2)$ désignent respectivement les quantiles d'ordres $\alpha/2$ et $(1-\alpha/2)$ de la loi χ^2_{n-p}.

Exemple (*Champs de tomates*)

On obtient ces intervalles de confiance pour les coefficients dans notre modèle avec la fonction `confint` de R.

```
lm_tomato <- lm(data = df_tomato,formula = Rendement ~.)
confint(lm_tomato)
```

```
                       2.5 %      97.5 %
(Intercept)     1.5589695200 2.697747984
Engrais         0.0004302768 0.002333366
Irrigation      0.0001128166 0.016432983
Heures_Travail -0.0040902134 0.008470244
Qualite_Sol     0.0018185737 0.006537700
Temperature    -0.0049447805 0.041023670
```

3.6 Nouvelles prévisions

⚙ Supposons que l'on ait une nouvelle observation $x_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})$, indépendante des observations $x_1, \ldots, x_n$.

💡 On prédit naturellement la valeur associée y_{n+1} de la variable réponse Y par

$$\hat{y}_{n+1} = \sum_{j=1}^{p} \hat{\beta}_j x_{n+1,j} = x_{n+1} \cdot \hat{\beta}.$$

⚙ **Question :** Comment mesure l'incertitude sur la prévision $\hat{y}_{n+1}$?

Propriété

- On a

$$\frac{y_{n+1} - \hat{y}_{n+1}}{\sqrt{\hat{\sigma}^2 \left(1 + x_{n+1} \left(X^T X\right)^{-1} x_{n+1}^T\right)}} \sim \mathcal{T}_{n-p}.$$

- On en déduit un intervalle de confiance de niveau $1 - \alpha$ pour y_{n+1}

$$IC(y_{n+1} = \left[\hat{y}_{n+1} \pm t_{n-p}(1 - \alpha/2) \sqrt{\hat{\sigma}^2 \left(1 + x_{n+1} \left(X^T X\right)^{-1} x_{n+1}^T\right)}\right]$$

où $t_{n-p}(1 - \alpha/2)$ désigne le quantile d'ordre $(1 - \alpha/2)$ de la loi $\mathcal{T}_{n-p}$.

Exemple (*Champs de tomates*)

- Supposons que l'on dispose des 3 champs différents suivants.

```
new_obs <- data.frame(Engrais = c(70,100,150),
                      Irrigation = c(35,40,45),
                      Heures_Travail = c(40,50,45),
                      Qualite_Sol = c(80,60,75),
                      Temperature = c(15,13,16))
```

```
  Engrais Irrigation Heures_Travail Qualite_Sol Temperature
1      70         35             40          80          15
2     100         40             50          60          13
3     150         45             45          75          16
```

- On obtient les rendements estimés pour ces champs, avec leurs intervalles de confiance au niveau 95%

```
predict(lm_tomato,newdata = new_obs,interval = "prediction")
```

```
       fit      lwr      upr
1 6.813248 6.302925 7.323571
2 6.737884 6.200317 7.275452
3 7.115340 6.604831 7.625849
```

3.7 Nullité des coefficients

⚙ Question statistique d'intérêt : Les coefficients $(\beta_j)_{1 \leqslant j \leqslant p}$ sont-il significativement non nuls ?

⚙ En effet, même si on a $\hat{\beta}_j \neq 0$ pour un certain $j \in \{1, \ldots, p\}$, il se peut que la vraie valeur vérifie $\beta_i = 0$.

Remarques :

- Si $\beta_j = 0$, cela signifie que la variable X_j n'explique pas *(linéairement)* la variable Y.
- On peut tester si un ou plusieurs coefficients son simultanément nuls.

Nullité d'un coefficient

Pour $j \in \{1, \dots, p\}$, on considère le test d'hypothèses :

$$H_0 : \beta_j = 0 \quad \text{contre } H_1 : \beta_j \neq 0.$$

Propriété

Sous l'hypothèse nulle H_0, on a

$$\frac{\hat{\beta}_j}{\hat{\sigma}_j} \sim \mathcal{T}_{n-p}.$$

Donc, si $\left| \frac{\hat{\beta}_j}{\hat{\sigma}_j} \right| > t_{n-p}(1 - \alpha/2)$, on rejette H_0 au niveau α.

Exemple (*Champs de tomates*)

- Dans notre exemple, on ne rejette pas, au seuil de 5%, l'hypothèse $\beta_6 = 0$. En l'occurrence, la température ne semble pas jouer de rôle significatif dans notre modèle.
- En revanche, on rejette l'hypothèse $\beta_3 = 0$. L'impact de l'irrigation est donc siginificativement non nul, au seuil de 5%.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.1283588  0.2857600   7.448 1.42e-10 ***
Engrais        0.0013818  0.0004776   2.894 0.005001 **
Irrigation     0.0082729  0.0040953   2.020 0.046994 *
Heures_Travail 0.0021900  0.0031519   0.695 0.489337
Qualite_Sol    0.0041781  0.0011842   3.528 0.000723 ***
Temperature    0.0180394  0.0115351   1.564 0.122114
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1016 on 74 degrees of freedom
Multiple R-squared:  0.3499, Adjusted R-squared:  0.306
F-statistic: 7.965 on 5 and 74 DF,  p-value: 4.767e-06
```

Remarques :

- Dans l'exemple précédent, si on souhaite par exemple tester

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad \text{contre} \quad H_1 : \exists j \in \{1, 2, 3\},\ \beta_j \neq 0,$$

  utiliser les statistiques de test précédentes n'est pas très pertinent...
- En effet, si on prend comme zone de rejet

$$\mathcal{R} = \bigcup_{j=1}^{3} \left\{ \left| \frac{\hat{\beta}_j}{\hat{\sigma}_j} \right| > t_{n-p}(1 - \alpha/2) \right\},$$

  le risque de 1ère espèce n'est pas nécessairement majoré par α (*il est majoré par* 3α).
- « Plus on cherche un coefficient significatif, plus on a de chances de trouver... »
- On peut appliquer un critère de correction (*Bonferoni par exemple*)...
- ... ou utiliser une zone de rejet beaucoup plus restrictive comme simplement

$$\mathcal{R}' = \left\{ \left| \frac{\hat{\beta}_1}{\hat{\sigma}_1} \right| > t_{n-p}(1 - \alpha/2) \right\},$$

  mais la puissance du test est alors faible.

Exemple (*Champs de tomates*)

Dans notre exemple, on peut tester

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_5 = 0 \quad \text{contre} \quad H_1 : \exists j \in \{1, 2, 3, 5\},\ \beta_j \neq 0.$$

Si on additionne les *p*-values

$$1.42\mathrm{e}^{-10} + 0.005001 + 0.0046994 + 0.000723 \leqslant 0.052718,$$

On ne garantit plus un risque de 1ère espèce inférieur à 5%.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.1283588  0.2857600   7.448 1.42e-10 ***
Engrais        0.0013818  0.0004776   2.894 0.005001 **
Irrigation     0.0082729  0.0040953   2.020 0.046994 *
Heures_Travail 0.0021900  0.0031519   0.695 0.489337
Qualite_Sol    0.0041781  0.0011842   3.528 0.000723 ***
Temperature    0.0180394  0.0115351   1.564 0.122114
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1016 on 74 degrees of freedom
Multiple R-squared:  0.3499, Adjusted R-squared:  0.306
F-statistic: 7.965 on 5 and 74 DF,  p-value: 4.767e-06
```

« Plus on cherche, plus on a de chances de trouver... »

Nullité simultanée de plusieurs coefficients

On veut tester la nullité simultanée des $q = p - p_0$ derniers coefficients du modèle :

$$H_0 : \beta_{p_0+1} = \cdots = \beta_p = 0 \quad \text{contre} \quad H_1 : \exists j \in \{p_0 + 1, \ldots, p\},\ \beta_j \neq 0.$$

On note $\hat{Y}_0$ le vecteur des valeurs ajustées du modèle en ne gardant que les p_0 premières variables, et $\hat{\varepsilon}_0$ le vecteur des résidus associés.

Propriété

Sous H_0, on a

$$F = \frac{n-p}{q} \times \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\|Y - \hat{Y}\|^2} = \frac{n-p}{q} \times \frac{SCR_0 - SCR}{SCR} \sim \mathcal{F}^q_{n-p}$$

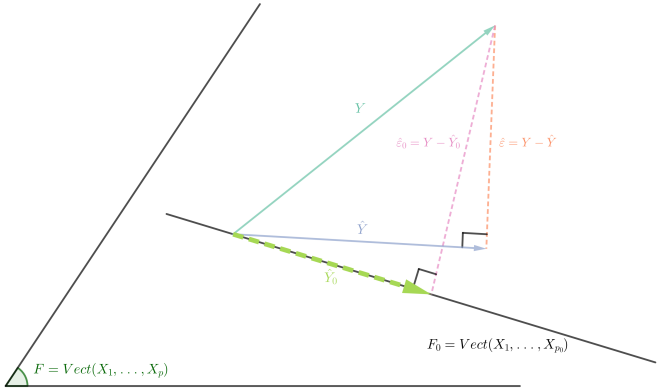
où :

- SCR_0 est la somme des carrés résiduels en ne prenant que les p_0 premières variables ;
- $\mathcal{F}^q_{n-p}$ désigne la loi de Fisher à q et $n - p$ d.d.l.

Donc, si $F > f^q_{n-p}(1-\alpha)$, le quantile d'ordre $(1-\alpha)$ de la loi $\mathcal{F}^q_{n-p}$, on rejette H_0 au niveau α.

Remarque :

L'idée du test est de considérer que si $\hat{Y}_0$ est proche de $\hat{Y}$, alors il semble raisonnable de conserver l'hypothèse nulle, car les q dernières variables n'apportent pas grand chose.



Exemple (*Champs de tomates*)

Dans R, on peut utiliser la fonction `anova` pour effectuer un tel test. On cherche ci-dessous à tester simultanément la significativité des coefficients β_4 et β_5, associés aux variables `Heures_Travail` et `Qualite_Sol`.

```
# Modèle complet
lm_complet <- lm(data = df_tomato,formula = Rendement ~.)
# Modèle réduit
lm_reduit <- lm (data=df_tomato,formula = Rendement ~ Engrais + Irrigation + Temperature)
# Test
anova(lm_complet,lm_reduit)
```

```
Model 1: Rendement ~ Engrais + Irrigation + Heures_Travail + Qualite_Sol +
    Temperature
Model 2: Rendement ~ Engrais + Irrigation + Temperature
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     74 0.76409
2     76 0.89284 -2  -0.12875 6.2347 0.003145 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Ici, on rejette H_0.

⚙ **Remarque :** Ce test, appelé **test entre modèles emboîtés**, est en fait équivalent au test de rapport de vraisemblance maximale.

Cas particulier : test de Fisher global

On peut vouloir tester si tous les coefficients sont nuls, exceptée la constante. Dans ce cas, $\hat{Y}_0 = \overline{y}\mathbb{1}$ et on a comme statistique de test

$$F = \frac{n-p}{p-1} \times \frac{\|\hat{Y} - \hat{Y}_0\|^2}{\|Y - \hat{Y}\|^2} = \frac{n-p}{p-1} \times \frac{R^2}{1-R^2} \sim \mathcal{F}_{n-p}^{p-1}.$$

⚙ **Remarque :** S'il n'y a qu'une seule variable en dehors de la constante, ce test est alors équivalent au test de Student vu pour la régression linéaire simple.

Exemple (*Champs de tomates*)

Dans le cas du test de Fisher global, le résultat est immédiatement fourni par la fonction `summary` de R.

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.1283588  0.2857600   7.448 1.42e-10 ***
Engrais        0.0013818  0.0004776   2.894 0.005001 **
Irrigation     0.0082729  0.0040953   2.020 0.046994 *
Heures_Travail 0.0021900  0.0031519   0.695 0.489337
Qualite_Sol    0.0041781  0.0011842   3.528 0.000723 ***
Temperature    0.0180394  0.0115351   1.564 0.122114
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.1016 on 74 degrees of freedom
Multiple R-squared:  0.3499, Adjusted R-squared:  0.306
F-statistic: 7.965 on 5 and 74 DF,  p-value: 4.767e-06
```

Ici, on rejette au niveau 5% l'hypothèse comme quoi tous les coefficients seraient nuls en dehors de la constante.