Régression linéaire multiple

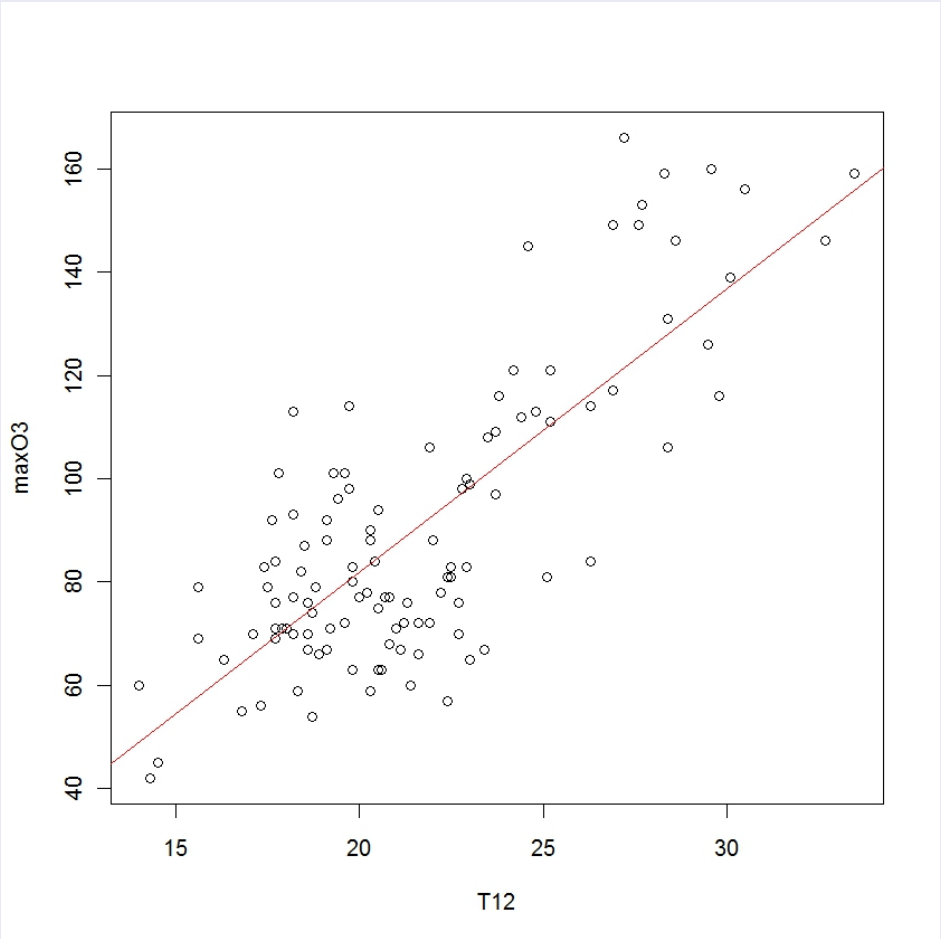
Myriam Maumy-Bertrand et Marie Chion[1]

[1] IRMA, Université de Strasbourg
France

Master 1
2019-2020

Exemple : Issu du livre « Statistiques avec R », P.A. Cornillon, *et al.*, Deuxième édition, 2010

- **Problème :** Étude de la concentration d'ozone dans l'air.
- **Modèle :** La température à 12 heures (v.a. X_1) et la concentration d'ozone (v.a. Y) sont liées de manière linéaire :
$$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$
- **Observations :** $n = 112$ observations de la température à 12 heures et de la concentration d'ozone.
- **But :** Estimer β_0 et β_1 afin de prédire la concentration d'ozone connaissant la température à 12 heures.


maxO3
40
60
80
100
120
140
160
15
20
25
30
T12

Affiner le modèle

Souvent la régression linéaire est trop simpliste. Il faut alors utiliser d'autres modèles plus réalistes mais parfois plus complexes :

- Utiliser d'autres fonctions que les fonctions affines comme les fonctions polynômiales, exponentielles, logarithmiques...
- Considérer plusieurs variables explicatives.
  Exemple : La température à 12 heures **et** la vitesse du vent.

Régression linéaire multiple

Le principe de la régression linéaire multiple est simple :

- Déterminer la variable expliquée Y.
  Exemple : La concentration d'ozone.
- Déterminer $(p-1)$ variables explicatives $X_1, \ldots, X_{p-1}$
  Exemple : X_1 température à 12 heures, X_2 vitesse du vent, ...
- Il ne reste plus qu'à appliquer un modèle linéaire :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon.$$

Dans un échantillon de n individus, nous mesurons y_i, $x_{i,1}$, ..., $x_{i,p-1}$ pour $i = 1, \ldots, n$.

Observations	Y	X_1	...	X_{p-1}
1	y_1	$x_{1,1}$	...	$x_{1,p-1}$
2	y_2	$x_{2,1}$	...	$x_{2,p-1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_n	$x_{n,1}$	...	$x_{n,p-1}$

Remarque

Les variables $x_{i,j}$ sont fixes tandis que les variables Y_i sont aléatoires.

Problème

Il faut estimer les paramètres $\beta_0, \ldots, \beta_{p-1}$ du modèle de régression et ce de manière optimale.

Solution

Utiliser la méthode des moindres carrés. Cette méthode revient à minimiser la quantité suivante :

$$\min_{\beta_0,\ldots,\beta_{p-1}} \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1}\right)\right)^2 .$$

Le système peut se réécrire :

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

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

Vecteur des résidus : $\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\widehat{\beta}$.

Remarque

Les variables **y** et **X** sont mesurées tandis que l'estimateur $\widehat{\beta}$ est à déterminer.
La méthode des moindres carrés ordinaires consiste à trouver le vecteur $\widehat{\beta}$ qui minimise $\|\varepsilon\|^2 = {}^t\varepsilon\varepsilon$.

Les calculs

$$\begin{aligned} \|\varepsilon\|^2 &= {}^t(\mathbf{y} - \mathbf{X}\widehat{\beta})(\mathbf{y} - \mathbf{X}\widehat{\beta}) \\ &= {}^t\mathbf{y}\mathbf{y} - {}^t\widehat{\beta}{}^t\mathbf{X}\mathbf{y} - {}^t\mathbf{y}\mathbf{X}\widehat{\beta} + {}^t\widehat{\beta}{}^t\mathbf{X}\mathbf{X}\widehat{\beta} \\ &= {}^t\mathbf{y}\mathbf{y} - 2{}^t\widehat{\beta}{}^t\mathbf{X}\mathbf{y} + {}^t\widehat{\beta}{}^t\mathbf{X}\mathbf{X}\widehat{\beta} \end{aligned}$$

car ${}^t\widehat{\beta}{}^t\mathbf{X}\mathbf{y}$ est un scalaire. Donc il est égal à sa transposée.

La dérivée par rapport à $\widehat{\beta}$ est alors égale à :

$$-2{}^t\mathbf{X}\mathbf{y} + 2{}^t\mathbf{X}\mathbf{X}\widehat{\beta}.$$

Problème

Nous cherchons $\widehat{\beta}$ qui annule cette dérivée. Donc nous devons résoudre l'équation suivante :

$$^t\mathbf{X}\mathbf{X}\widehat{\beta} = {}^t\mathbf{X}\mathbf{y}.$$

Solution

Nous trouvons après avoir inversé la matrice $^t\mathbf{X}\mathbf{X}$ (il faut naturellement vérifier que $^t\mathbf{X}\mathbf{X}$ est carrée et inversible c'est-à-dire qu'aucune des colonnes qui compose cette matrice ne soit proportionnelle aux autres colonnes)

$$\widehat{\beta} = (^t\mathbf{X}\mathbf{X})^{-1}\,{}^t\mathbf{X}\mathbf{y}.$$

Remarque

Retrouvons les résultats de la régression linéaire simple ($p = 2$)

$$ {}^t\mathbf{XX} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}; \qquad {}^t\mathbf{Xy} = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \end{pmatrix}. $$

Donc :

$$ \begin{aligned} ({}^t\mathbf{XX})^{-1} &= \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \\ &= \frac{1}{\sum (x_i - \overline{x}_n)^2} \begin{pmatrix} \sum x_i^2/n & -\overline{x}_n \\ -\overline{x}_n & 1 \end{pmatrix}. \end{aligned} $$

Suite et fin de la remarque

Finalement nous retrouvons bien :

$$\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \dfrac{\overline{Y}_n \sum x_i^2 - \overline{x}_n \sum x_i Y_i}{\sum (x_i - \overline{x}_n)^2} \\ \\ \dfrac{\sum x_i Y_i - n\overline{x}_n \overline{Y}_n}{\sum (x_i - \overline{x}_n)^2} \end{pmatrix}$$

ce qui correspond aux estimateurs de la régression linéaire simple que nous avons déjà rencontrés dans le cours 6.

Exemple avec le logiciel R

Reprenons l'exemple traité en début de ce chapitre. Introduisons comme variables explicatives la température à 12 heures et la vitesse du vent. Utilisons pour cela R.

```
> modele1 <- lm(max03 ~ T12 + Vx12)
> summary(modele1)

Call:
lm(formula = max03 ~ T12 + Vx12)

Residuals:
Min 1Q Median 3Q Max
-33.08 -12.00 -1.20 10.67 45.67
```

Suite de l'exemple avec le logiciel R

```
Coefficients: Estimate Std. Error t value
Pr(>|t|)
(Intercept) -14.4242 9.3943 -1.535 0.12758
T12 5.0202 0.4140 12.125 < 2e-16 ***
Vx12 2.0742 0.5987 3.465 0.00076 ***
Residual standard error: 16.75 on 109 degrees
of freedom
Multiple R-squared: 0.6533, Adjusted
R-squared: 0.6469
F-statistic: 102.7 on 2 and 109 DF, p-value: <
2.2e-16
```

Test de normalité

Pour lire la p valeur du test de Fisher et celles des tests de Student, il faut s'assurer auparavant que les résidus suivent une loi normale. Pour cela, vous allez réaliser un test de normalité, celui de Shapiro-Wilk avec R.

```
> residus<-residuals(modele1)
> shapiro.test(residus)

Shapiro-Wilk normality test

data: residus
W = 0.9854, p-value = 0.2624
```

Conclusion du test de normalité

La *p*-valeur (*p*-value = 0,2624) du test de Shapiro-Wilk étant strictement supérieure à $\alpha = 5\%$, le test n'est pas significatif. Nous décidons de ne pas rejeter et donc d'accepter $\mathcal{H}_0$ au seuil $\alpha = 5\%$. Le risque d'erreur associé à cette décision est un risque d'erreur de seconde espèce β. Dans le cas présent, vous ne pouvez pas l'évaluer.

Interprétation

Il ne nous reste plus qu'à interpréter la sortie de `summary(modele1)`.

Suite de l'interprétation

La p-valeur (p-value < 2.2e-16) du test de Fisher étant inférieure ou égale à $\alpha = 5\%$, le test est significatif. Nous rejetons $\mathcal{H}_0$ et nous décidons que $\mathcal{H}_1$ est vraie avec un risque de première espèce $\alpha = 5\%$. Donc il y a au moins une des deux variables qui joue le rôle de variable explicative.
La p-valeur (p-value < 2e-16) du test de Student, associée à « T12 » étant inférieure ou égale à $\alpha = 5\%$, le test est significatif. Nous rejetons $\mathcal{H}_0$ et nous décidons que $\mathcal{H}_1$ est vraie avec un risque de première espèce $\alpha = 5\%$.
Nous pouvons faire la même conclusion pour la variable vitesse du vent à 12 heures.

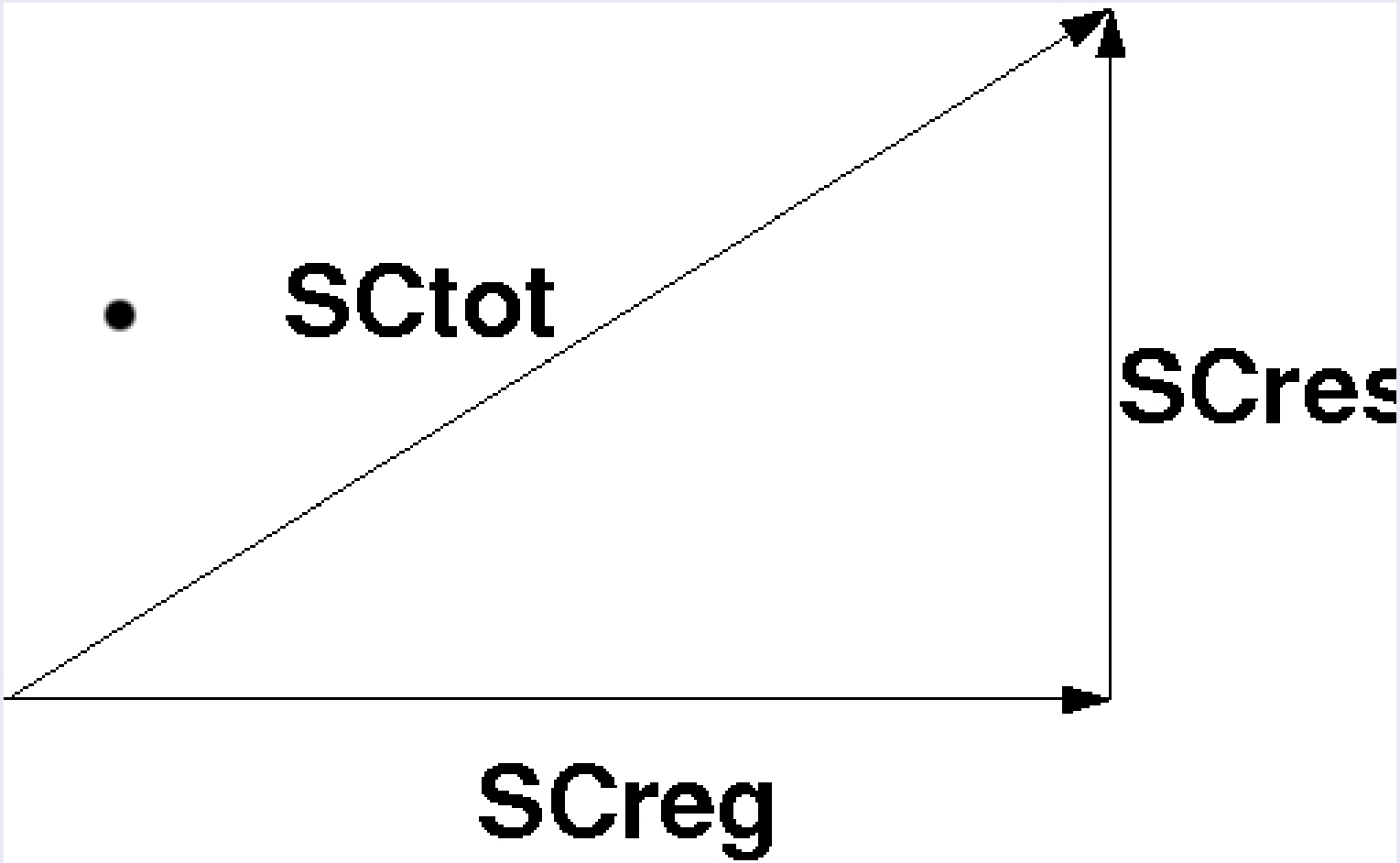
Résultats préliminaires

1. $\sum \hat{y}_i^2 = \sum \hat{y}_i y_i$ ou (forme matricielle) ${}^t\hat{\mathbf{y}}\hat{\mathbf{y}} = {}^t\mathbf{y}\hat{\mathbf{y}}$
2. $\sum \hat{y}_i = \sum y_i$

Propriété des moindres carrés ordinaires

$$\begin{array}{ccccc} \sum(y_i - \overline{y}_n)^2 & = & \sum(\hat{y}_i - \overline{y}_n)^2 & + & \sum(y_i - \hat{y}_i)^2 \\ \mathrm{SC}_{tot} & = & \mathrm{SC}_{reg} & + & \mathrm{SC}_{res} \end{array}$$

Représentation graphique de la relation fondamentale



Rappel sur le coefficient de détermination

Nous rappelons que le **coefficient de détermination** est défini par :

$$R^2 = \frac{SC_{reg}}{SC_{tot}}.$$

Intuitivement ce coefficient de détermination quantifie la capacité du modèle à expliquer les variations de Y.

- Si R^2 est proche de 1 alors le modèle est proche de la réalité.
- Si R^2 est proche de 0 alors le modèle explique très mal la réalité. Il faut alors trouver un meilleur modèle.

Les hypothèses indispensables pour réaliser les tests

Nous faisons les hypothèses suivantes :

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

où le vecteur aléatoire ε suit une loi *multinormale* qui vérifie les hypothèses suivantes :

- $\mathbb{E}[\varepsilon] = 0$
- $Var[\varepsilon] = \sigma^2 \mathbf{I}_n$,

où σ^2 est la variance de la population et $\mathbf{I}_n$ est la matrice identité de taille n.

Conséquences

Les hypothèses précédentes impliquent

- $\mathbb{E}[\mathbf{y}] = \mathbf{X}\beta$
- $Var[\mathbf{y}] = \sigma^2 \mathbf{I}_n$.

Nous pouvons alors démontrer, **sous ces hypothèses** :

- $\mathbb{E}\left[\widehat{\beta}\right] = \beta$. Ce qui signifie que le vecteur $\widehat{\beta}$ est un estimateur sans biais de β.
- $Var\left[\widehat{\beta}\right] = \sigma^2({}^t\mathbf{X}\mathbf{X})^{-1}$.

Problème

La variance σ^2 est inconnue. Donc il faut estimer σ^2 !

Construction d'un estimateur de σ^2

Un estimateur sans biais de la variance σ^2 est défini par :

$$CM_{res} = \frac{\sum(Y_i - \widehat{Y}_i)^2}{n-p} = \frac{SC_{res}}{n-p} = \frac{SC_{tot} - SC_{reg}}{n-p}$$

où

- n est le nombre d'individus/d'observations,
- p est le nombre de variables explicatives.

Nous rappelons que la quantité $(n-p)$ est **le nombre de degrés de liberté associé à SC_{res}.**

Test de Fisher

Tester l'hypothèse nulle :

$$\mathcal{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

contre l'hypothèse alternative :

$\mathcal{H}_1 : \exists$ au moins un j pour lequel $\beta_j \neq 0$ où j varie de 1 à $p-1$.

Remarque

Si l'hypothèse nulle $\mathcal{H}_0$ est vérifiée alors le modèle s'écrit :

$$Y_i = \beta_0 + \varepsilon_i.$$

Tableau de l'analyse de la variance

Source de variation	sc	ddl	cm	F_{obs}
Régression	$sc_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$	$p-1$	$\frac{sc_{reg}}{p-1}$	$\frac{cm_{reg}}{cm_{res}}$
Résiduelle	$sc_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$n-p$	$\frac{sc_{res}}{n-p}$	
Totale	$sc_{tot} = \sum_{i=1}^{n} (y_i - \overline{y})^2$	$n-1$		

Méthode

1. Calculer la statistique

$$F_{obs} = \frac{CM_{reg}}{CM_{res}}.$$

2. Lire la valeur critique $F_{1-\alpha,p-1,n-p}$ où $F_{1-\alpha,p-1,n-p}$ est le $(1-\alpha)$-quantile d'une loi de Fisher avec $(p-1)$ et $(n-p)$ degrés de liberté, car si l'hypothèse nulle $\mathcal{H}_0$ est vraie, alors F_{obs} suit une loi de Fisher avec $(p-1)$ et $(n-p)$ degrés de liberté.
3. Comparer la statistique c'est-à-dire la valeur observée à la valeur critique.

Règle de décision

- Nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et d'accepter l'hypothèse alternative $\mathcal{H}_1$, au seuil $\alpha = 5\%$, si

$$|F_{obs}| \geqslant F_{1-\alpha,p-1,n-p}.$$

- Nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ et donc de l'accepter si

$$|F_{obs}| < F_{1-\alpha,p-1,n-p}.$$

Tests de Student

Tester l'hypothèse nulle

$$\mathcal{H}_0 : \beta_j = b_j \quad \text{pour } j = 0, \ldots, p-1$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \beta_j \neq b_j \quad \text{pour un certain } j \text{ entre 0 et } p-1.$$

Méthode

1. Calculer la statistique

$$t_{obs} = \frac{\widehat{\beta}_j - b_j}{s(\widehat{\beta}_j)}$$

où $s^2(\widehat{\beta}_j)$ est l'élément diagonal d'indice j de $CM_{res}({}^t\mathbf{XX})^{-1}$.

2. Lire la valeur critique $t_{n-p;1-\alpha/2}$ où $t_{n-2;1-\alpha/2}$ est le $(1-\alpha/2)$-quantile d'une loi de Student avec $(n-p)$ degrés de liberté, car si l'hypothèse nulle $\mathcal{H}_0$ est vraie, alors t_{obs} suit une loi de Student avec $(n-p)$ degrés de liberté.
3. Comparer la statistique c'est-à-dire la valeur observée à la valeur critique.

Règle de décision

- Nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et d'accepter l'hypothèse alternative $\mathcal{H}_1$, au seuil $\alpha = 5\%$, si

$$|t_{obs}| \geqslant t_{n-p;1-\alpha/2}.$$

- Nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ et donc de l'accepter si

$$|t_{obs}| < t_{n-p;1-\alpha/2}.$$

Cas particulier

Tester l'hypothèse nulle

$$\mathcal{H}_0 : \beta_j = 0 \quad \text{pour } j = 0, \ldots, p-1$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \beta_j \neq 0 \text{ pour un certain } j \text{ entre 0 et } p-1.$$

Méthode

1. Calculer la statistique

$$t_{obs} = \frac{\hat{\beta}_j}{s(\hat{\beta}_j)}.$$

2. Lire la valeur critique $t_{n-p;1-\alpha/2}$ où $t_{n-p;1-\alpha/2}$ est le $(1-\alpha/2)$-quantile d'une loi de Student avec $(n-p)$ degrés de liberté, car si l'hypothèse nulle $\mathcal{H}_0$ est vraie, alors t_{obs} suit une loi de Student avec $(n-p)$ degrés de liberté.
3. Comparer la valeur observée et la valeur critique.

Règle de décision

- Nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et d'accepter l'hypothèse alternative $\mathcal{H}_1$, au seuil $\alpha = 5\%$, si

$$|t_{obs}| \geqslant t_{n-p;1-\alpha/2}.$$

- Nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ et donc de l'accepter si

$$|t_{obs}| < t_{n-p;1-\alpha/2}.$$

IC pour β_j

Un intervalle de confiance au niveau $(1-\alpha)$ où α est la probabilité d'erreur pour β_j est défini par

$$\left]\widehat{\beta}_j - t_{n-p;1-\alpha/2} \times s(\widehat{\beta}_j); \widehat{\beta}_j + t_{n-p;1-\alpha/2} \times s(\widehat{\beta}_j)\right[.$$

Remarque

Cet intervalle de confiance est construit de telle sorte qu'il contienne le paramètre inconnu β_j avec une probabilité de $(1-\alpha)$.

Test de Fisher partiel

La nullité d'un certain nombre r de paramètres dans un modèle de p paramètres.
l'hypothèse nulle $\mathcal{H}_0$: *modèle réduit* avec $(p - r)$ paramètres contre
l'hypothèse alternative $\mathcal{H}_1$: *modèle complet* avec p paramètres.

Exemples

- Tester la nullité d'un paramètre, par exemple : β_1.
  $\mathcal{H}_0 : Y_i = \beta_0 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$ contre
  $\mathcal{H}_1 : Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$.
- Tester la nullité de plusieurs paramètres, par exemple les pairs : β_{2j}.
  $\mathcal{H}_0 : Y_i = \beta_1 x_{i1} + \beta_3 x_{i3} + \cdots + \beta_{p-1} x_{ip-1} + \varepsilon_i$ contre
  $\mathcal{H}_1 : Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$ avec p pair.

Méthode

1. Calculer les valeurs estimées $\hat{y}_i$ en utilisant la méthode des moindres carrés pour chacun des 2 modèles définis par $\mathcal{H}_0$ et $\mathcal{H}_1$, notées : $\hat{y}_i(\mathcal{H}_0)$ et $\hat{y}_i(\mathcal{H}_1)$.
2. Calculer ensuite $SC_{res}(\mathcal{H}_0)$ et $SC_{res}(\mathcal{H}_1)$.
3. Calculer la statistique

$$F_{obs} = \frac{SC_{res}(\mathcal{H}_0) - SC_{res}(\mathcal{H}_1)}{SC_{res}(\mathcal{H}_1)} \times \frac{n-p}{r}.$$

Méthode - Suite et fin

4. Lire la valeur critique $F_{1-\alpha,r,n-p}$ où $F_{1-\alpha,r,n-p}$ est le $(1-\alpha)$-quantile d'une loi de Fisher avec r et $(n-p)$ degrés de liberté, car si l'hypothèse nulle $\mathcal{H}_0$ est vraie, alors F_{obs} suit une loi de Fisher avec r et $(n-p)$ degrés de liberté.
5. Comparer la valeur observée et la valeur critique.

Règle de décision

- Nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et par conséquent d'accepter l'hypothèse alternative $\mathcal{H}_1$, au seuil $\alpha = 5\%$, si
$$F_{obs} \geqslant F_{1-\alpha,r,n-p}.$$
- Nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ et donc de l'accepter si
$$F_{obs} < F_{1-\alpha,r,n-p}.$$