Régression linéaire multiple

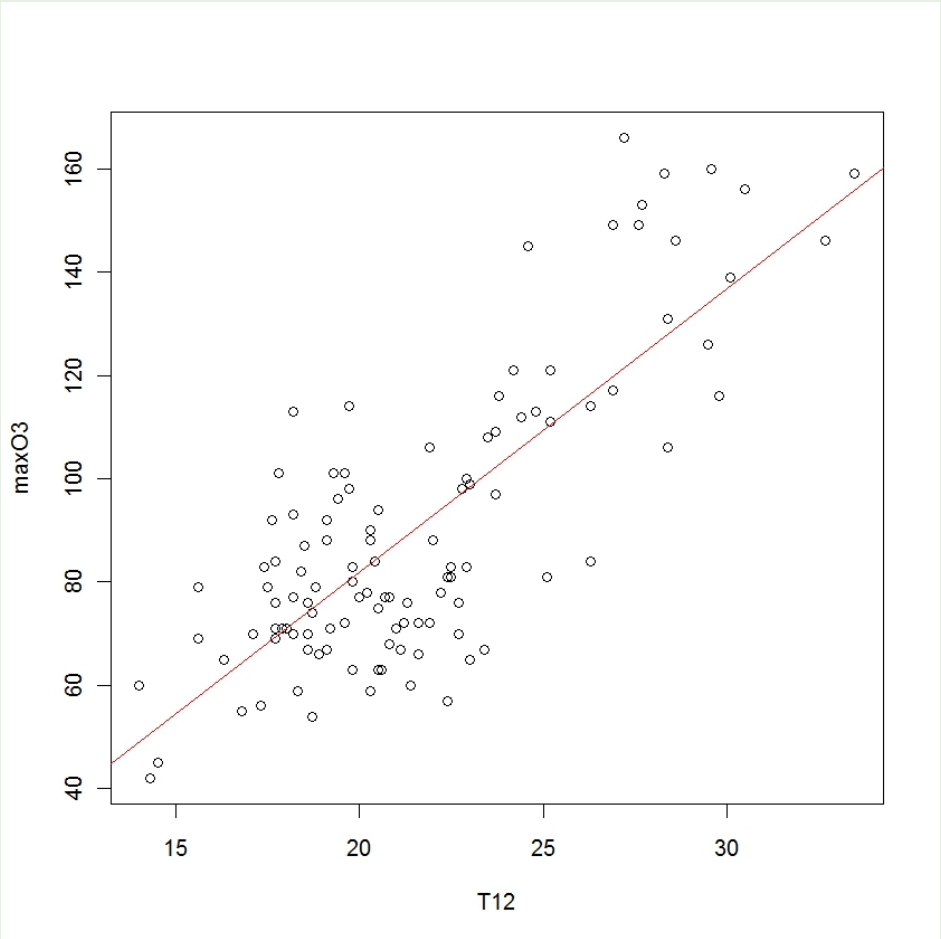
Myriam Maumy-Bertrand[1]

[1] IRMA, Université de Strasbourg
France

ESIEA 4ème Année 02-02-2016

Exemple : Issu du livre « Statistiques avec R », P.A. Cornillon, *et al.*, Troisième édition augmentée, 2012, P.U.R.

- **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.

Retour à l'exemple de la concentration d'ozone

La température à 12 heures **et** la vitesse du vent à 12 heures.

Régression linéaire multiple

Le principe de la régression linéaire multiple est simple :

- Déterminer la variable expliquée Y.
- Déterminer $(p-1)$ variables explicatives $X_1, \ldots, X_{p-1}$.
- Il ne reste plus qu'à appliquer un modèle linéaire :

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

Concentration d'ozone

- La variable expliquée Y : la concentration d'ozone.
- Les $(p-1)$ variables explicatives $X_1, \ldots, X_{p-1}$: X_1 température à 9 heures, X_2 température à 12 heures, X_3 température à 15 heures, X_4 nébulosité à 9 heures, ...
- Il ne reste plus qu'à appliquer un modèle linéaire :

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{11} X_{11} + \varepsilon.$$

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 p 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 la quantité suivante :

$$\|\varepsilon\|^2 = {}^t\varepsilon\varepsilon.$$

Les calculs

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

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

La dérivée de $\|\varepsilon\|^2$ par rapport à β est alors égale à :

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

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 ($p = 2$). D'une part nous avons :

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

D'autre part, en calculant, nous obtenons :

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

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 5 d'esti .

Retour à la concentration d'ozone

Reprenons l'exemple de la concentration d'ozone traité en début de ce cours. Le jeu de données se trouve à l'adresse suivante :

`http://math.agrocampus-ouest.fr/infoglueDeliverLive/enseignement/support2cours/livres/statistiques.avec.R`

Puis téléchargeons et enregistrons sur le bureau par exemple, le jeu de données : `ozone.txt`.

Introduisons deux variables explicatives : la température à 12 heures et la vitesse du vent à 12 heures.

Suite de l'exemple

Tapons la ligne de commande suivante pour charger le jeu de données :

```
> ozone<-read.table(file.choose())
```

Nous pouvons vérifier que le fichier est bien importé en tapant la ligne de commande suivante :

```
> str(ozone)
```

Maintenant, construisons le modèle linéaire en utilisant la commande `lm()` :

```
> modele1<-lm(max03 ~ T12+Vx12,data=ozone)
> summary(modele1)
```

Suite de l'exemple

R renvoie le résultat suivant :

```
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

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Inter.)     -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 ***
--
Signif. codes:  0=***; 0.001=**; 0.01=*;
0.05=.; 0.1= .
```

Suite de l'exemple

```
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
```

Remarque

Nous commenterons tous ces résultats par la suite.

Fin de l'exemple

Le vecteur $\widehat{\beta}$ cherché est égal à :

$$\widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 &= & -14,4242 \\ \widehat{\beta}_1 &= & 5,0202 \\ \widehat{\beta}_2 &= & 2,0742 \end{pmatrix}.$$

Résultats préliminaires

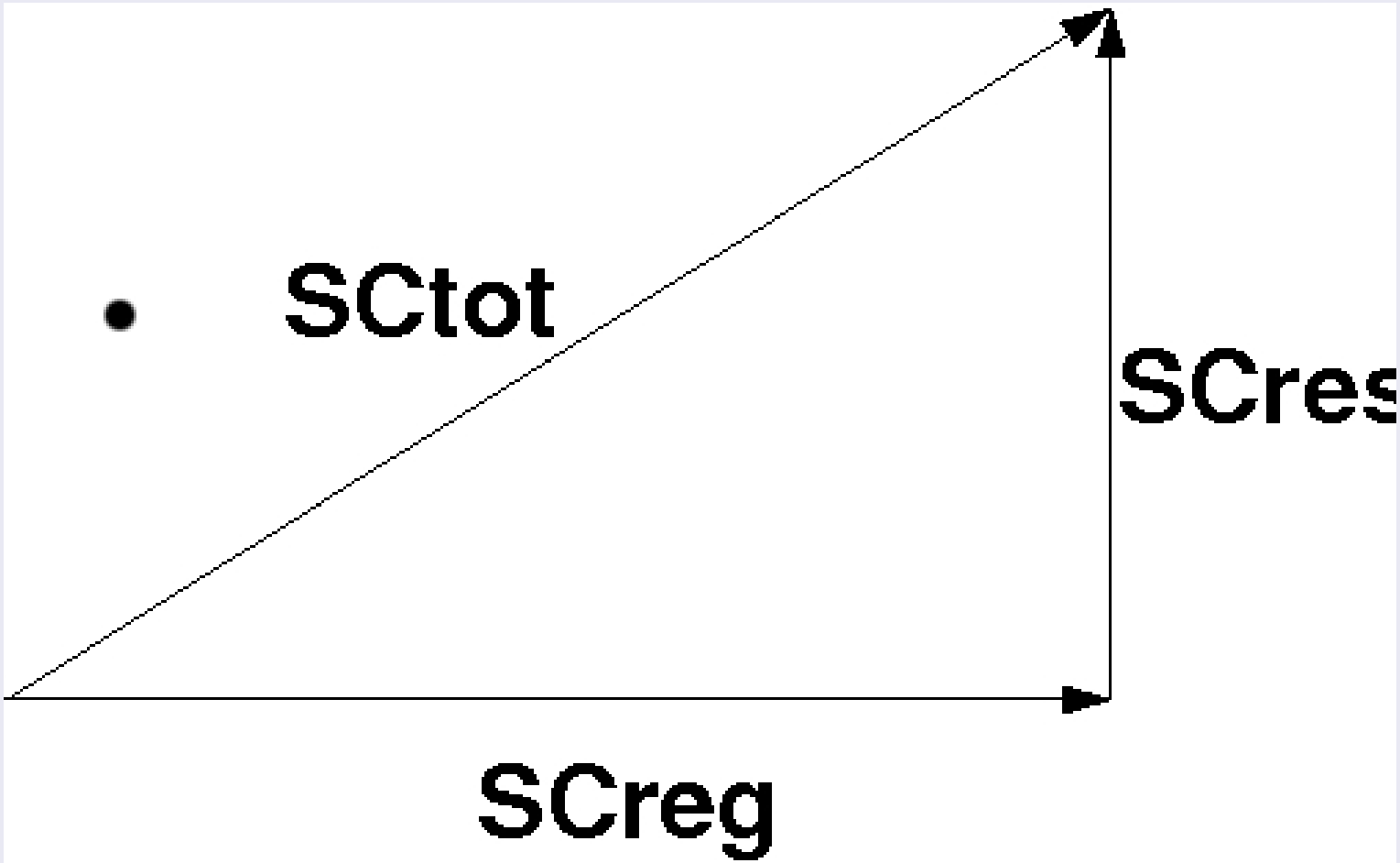
1. $\sum_i \hat{y}_i^2 = \sum_i \hat{y}_i y_i$ ou (forme matricielle) ${}^t\hat{\mathbf{y}}\hat{\mathbf{y}} = {}^t\mathbf{y}\hat{\mathbf{y}}$
2. $\sum_i \hat{y}_i = \sum_i y_i$

Propriété des moindres carrés ordinaires

$$\sum_i (y_i - \overline{y}_n)^2 = \sum_i (\hat{y}_i - \overline{y}_n)^2 + \sum_i (y_i - \hat{y}_i)^2$$

$$\mathrm{SC}_{tot} = \mathrm{SC}_{reg} + \mathrm{SC}_{res}$$

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.

Retour à l'exemple de la concentration d'ozone

Pour calculer le coefficient de détermination, avec `R`, il existe plusieurs méthodes.
La plus rapide : trouver le coefficient de détermination dans la sortie de la commande `summary()`.
R^2 est en face de `Multiple R-squared:`.
Ici R^2 est égal à 0,6533. Le modèle n'est pas très bon puisque $R^2 < 0,80$. Il faudrait donc envisager un autre modèle pour expliquer la concentration d'ozone.

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 :

- les ε_i sont indépendantes
- $\mathbb{E}(\varepsilon) = 0$
- $\mathbb{V}\text{ar}(\varepsilon) = \sigma^2 \mathbf{I}_n$,

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 :

- $\mathbb{E}(\mathbf{y}) = \mathbf{X}\beta$
- $\mathbb{V}\text{ar}(\mathbf{y}) = \sigma^2\mathbf{I}_n$.

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.
- $\mathbb{V}\text{ar}\left(\widehat{\beta}\right) = \sigma^2({}^t\mathbf{X}\mathbf{X})^{-1}$.

Problème

La variance σ^2 de la population est inconnue. Donc il faut estimer la variance σ^2 !
Comment ?
En construisant un estimateur !

Construction d'un estimateur de la variance 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	$sc_{tot} = \sum_{i=1}^{n} (y_i - \overline{y})^2$	$n-1$		

Retour à la concentration d'ozone

Nous obtenons le tableau d'analyse de la variance en tapant les deux lignes de commande suivantes :

```
> modele0<-lm(maxO3~1,data=ozone)
> anova(modele0,modele1)
```

Source de variation	*sc*	*ddl*	*cm*	F_{obs}
Régression	57611	2	$28805,5$	$102,6749$
Résiduelle	30580	109	$280,5505$	
Totale	88191	111		

Méthode du test de Fisher

1. Calculer la statistique du test de Fisher :

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

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 la statistique F suit une loi de Fisher avec $(p-1)$ et $(n-p)$ degrés de liberté.
3. Comparer la statistique (la valeur observée) à la valeur critique $F_{1-\alpha,p-1,n-p}$.

Règle de décision pour le test de Fisher

- Si $F_{obs} \geqslant F_{1-\alpha,p-1,n-p}$, alors le test est significatif. Nous rejetons l'hypothèse nulle $\mathcal{H}_0$ et décidons d'accepter l'hypothèse alternative $\mathcal{H}_1$. Le risque d'erreur associé à cette décision est un risque de première espèce $\alpha = 5\%$.
- Si $F_{obs} < F_{1-\alpha,p-1,n-p}$, alors le test n'est pas significatif. Nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ et donc de l'accepter, au seuil $\alpha = 5\%$. Le risque d'erreur associé à cette décision est un risque de deuxième espèce β qu'il faudrait évaluer.

Test de normalité sur les résidus

Pour réaliser le test de Fisher, il faut s'assurer auparavant que les erreurs suivent une loi normale.
Pour cela, nous devons réaliser un test de normalité de Shapiro-Wilk sur les résidus auparavant.
Il faut donc commencer par calculer les résidus. Pour cela, il faut utiliser `R` et en particulier la commande `residuals()`.

Retour à l'exemple de la concentration d'ozone

Nous commençons par calculer les résidus en tapant la ligne de commande suivante :

```
> residus<-residuals(modele1)
```

Puis, nous réalisons le test de Shapiro-Wilk sur les résidus en tapant la ligne de commande suivante :

```
> shapiro.test(residus)
```

Suite de l'exemple

`R` renvoie le résultat suivant :

```
Shapiro-Wilk normality test
data: residus
W = 0.9854, p-value = 0.2624
```

Suite de l'exemple

La *p*-valeur (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 l'hypothèse nulle $\mathcal{H}_0$ et donc de l'accepter, au seuil $\alpha = 5\%$. Le risque d'erreur associé à cette décision est un risque d'erreur de deuxième espèce β. Dans le cas d'un test de Shapiro-Wilk, nous ne pouvons pas l'évaluer.

Fin de l'exemple

Déterminons la valeur critique en utilisant R et en tapant la ligne de commande suivante :

```
> qf(0.95,2,109)
```

R renvoie le résultat suivant :

```
[1] 3.079596
```

Comme $F_{obs} = 102,6749 > F_{0,95,2,109} = 3,079596$, le test est significatif. Nous décidons de rejeter l'hypothèse nulle $\mathcal{H}_0$ et décidons d'accepter l'hypothèse alternative $\mathcal{H}_1$, au seuil $\alpha = 5\%$. Le risque d'erreur associé à cette décision est un risque d'erreur de première espèce $\alpha = 5\%$.

Donc nous pouvons conclure, au seuil $\alpha = 5\%$, qu'il y a au moins une des deux variables qui joue le rôle de variable explicative dans le modèle.

Remarque

Nous pouvons aussi raisonner avec la *p*-valeur donnée par `R` dans la sortie donnée par la ligne de commande :

```
> summary(modele1)
```

Il faut aussi s'assurer d'avoir la normalité des résidus pour interpréter cette *p*-valeur.

Retour à la concentration d'ozone

Rappelons que `R` nous a renvoyé le résultat suivant :

```
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
```

Fin de l'exemple

La p-valeur ($< 2,2 \times 10^{-16}$) du test de Fisher étant inférieure à $\alpha = 5\%$, le test est significatif. Nous rejetons l'hypothèse nulle $\mathcal{H}_0$ et décidons d'accepter l'hypothèse alternative $\mathcal{H}_1$, au seuil $\alpha = 5\%$. Le risque d'erreur associé à cette décision est un risque de première espèce $\alpha = 5\%$
Donc nous pouvons conclure, au seuil $\alpha = 5\%$, qu'il y a au moins une des deux variables qui joue le rôle de variable explicative dans le modèle.

Remarque

Nous retrouvons la même conclusion que précédemment lorsque nous avons raisonné avec la statistique du test !

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 du test de Student

1. Calculer la statistique du test de Student :

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

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 (la valeur observée) à la valeur critique.

Règle de décision pour le test de Student

- Si $|t_{obs}| \geqslant t_{n-p;1-\alpha/2}$, alors le test est significatif. Nous rejetons l'hypothèse nulle $\mathcal{H}_0$ et décidons d'accepter l'hypothèse alternative $\mathcal{H}_1$. Le risque d'erreur associé à cette décision est un risque de première espèce $\alpha = 5\%$.
- Si $|t_{obs}| < t_{n-p;1-\alpha/2}$, alors le test n'est pas significatif. Nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ et donc de l'accepter, au seuil $\alpha = 5\%$. Le risque d'erreur associé à cette décision est un risque de deuxième espèce β qu'il faudrait évaluer.

Cas particulier du test de Student

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 du test de Student

1. Calculer la statistique du test de Student :

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

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 statistique (la valeur observée) à la valeur critique.

Règle de décision pour le test de Student

- Si $|t_{obs}| \geqslant t_{n-p;1-\alpha/2}$, alors le test est significatif. Nous rejetons l'hypothèse nulle $\mathcal{H}_0$ et décidons d'accepter l'hypothèse alternative $\mathcal{H}_1$. Le risque d'erreur associé à cette décision est un risque de première espèce $\alpha = 5\%$.
- Si $|t_{obs}| < t_{n-p;1-\alpha/2}$, alors le test n'est pas significatif. Nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ et donc de l'accepter, au seuil $\alpha = 5\%$. Le risque d'erreur associé à cette décision est un risque de deuxième espèce β qu'il faudrait évaluer.

Retour à l'exemple de la concentration d'ozone

Rappelons que `R` a renvoyé le résultat suivant :

```
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 ***
```

Suite de l'exemple

La *p*-valeur (0,12758) du test de Student, associée à la constante étant strictement supérieure à $\alpha = 5\%$, le test n'est pas significatif. Nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ et donc de l'accepter, au seuil $\alpha = 5\%$. Le risque d'erreur associé à cette décision est un risque de deuxième espèce β qu'il faudrait évaluer.
Donc nous pouvons conclure, au seuil $\alpha = 5\%$, que la constante n'intervient pas dans le modèle.

Suite de l'exemple

La p-valeur (p-value $< 2 \times 10^{-16}$) du test de Student, associée à la variable « T12 » étant inférieure ou égale à $\alpha = 5\%$, le test est significatif. Nous rejetons l'hypothèse nulle $\mathcal{H}_0$ et décidons d'accepter l'hypothèse alternative $\mathcal{H}_1$. Le risque d'erreur associé à cette décision est un risque de première espèce $\alpha = 5\%$.

Donc nous pouvons conclure, au seuil $\alpha = 5\%$, que la variable température à 12 heures joue bien un rôle de variable explicative dans le modèle.

Fin de l'exemple

La *p*-valeur (0,00076) du test de Student, associée à la variable « Vx12 » étant inférieure ou égale à $\alpha = 5\%$, le test est significatif. Nous rejetons l'hypothèse nulle $\mathcal{H}_0$ et décidons d'accepter l'hypothèse alternative $\mathcal{H}_1$. Le risque d'erreur associé à cette décision est un risque de première espèce $\alpha = 5\%$.

Donc nous pouvons conclure, au seuil $\alpha = 5\%$, que la variable vitesse du vent à 12 heures joue bien un rôle de variable explicative dans le modèle.

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 $(1-\alpha)$.

Retour à l'exemple de la concentration d'ozone

Pour calculer les intervalle de confiance pour les paramètres du modèle dans lequel il y a la température à 12 heures et la vitesse du vent à 12 heures, nous utilisons la commande `confint()`. Donc, en tapant la ligne de commande suivante :

```
> confint(modeleTV)
```

`R` renvoie le résultat suivant :

```
                2.5 %      97.5 %
(Intercept) -33.0434740 4.195007
T12          4.1995967 5.840850
Vx12         0.8875903 3.260712
```

Test de Fisher partiel

La nullité d'un certain nombre r de paramètres dans un modèle de p paramètres.
Tester l'hypothèse nulle

$$\mathcal{H}_0 : \text{modèle réduit avec } (p-r) \text{ paramètres}$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \text{modèle complet avec } p \text{ paramètres.}$$

Deux exemples

1. Tester la nullité d'un paramètre. Par exemple : β_1.

$$\mathcal{H}_0 : Y_i = \beta_0 + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i(p-1)} + \varepsilon_i$$

contre

$$\mathcal{H}_1 : Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i(p-1)} + \varepsilon_i.$$

Suite des deux exemples

2. Tester la nullité de plusieurs paramètres. Par exemple les coefficients pairs : β_{2j}.

$$\mathcal{H}_0 : Y_i = \beta_1 x_{i1} + \beta_3 x_{i3} + \cdots + \beta_{2p-1} x_{i2p-1} + \varepsilon_i$$

contre

$$\mathcal{H}_1 : Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{2p} x_{i2p} + \varepsilon_i.$$

Méthode du test de Fisher partiel

1. Calculer les valeurs estimées $\hat{y}_i$ en utilisant la méthode des moindres carrés pour chacun des deux 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 du test de Fisher :

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

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 statistique (la valeur observée) à la valeur critique.

Règle de décision

- Si $F_{obs} \geqslant F_{1-\alpha,r,n-p}$, alors le test est significatif. Nous rejetons l'hypothèse nulle $\mathcal{H}_0$ et décidons d'accepter l'hypothèse alternative $\mathcal{H}_1$, au seuil $\alpha = 5\%$. Le risque d'erreur associé à cette décision est un risque de première espèce $\alpha = 5\%$.
- Si $F_{obs} < F_{1-\alpha,r,n-p}$, alors le test n'est pas significatif. Nous décidons de ne pas rejeter l'hypothèse nulle $\mathcal{H}_0$ et donc de l'accepter, au seuil $\alpha = 5\%$. Le risque d'erreur associé à cette décision est un risque de deuxième espèce β qu'il faudrait évaluer.

Un autre exemple

Un exemple d'utilisation du test de Fisher partiel se trouve dans la feuille de T.D. numéro 2, exercice 1.