

Compléments sur la régression linéaire simple

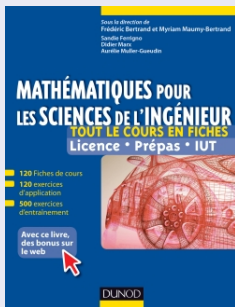
Anova et inférence sur les paramètres

Myriam Maumy-Bertrand¹

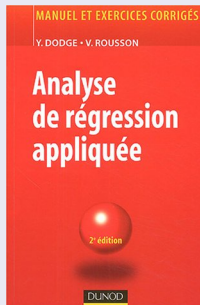
¹IRMA, Université de Strasbourg, France

02-02-2016

Ce chapitre s'appuie sur le livre :
« Mathématiques pour les Sciences de l'Ingénieur »,
de F. Bertrand, S. Ferrigno, D. Marx, M. Maumy-Bertrand, A.
Muller-Gueudin, Éditions Dunod, 2013.



Ce chapitre s'appuie également sur ce livre :
« Analyse de régression appliquée »,
de Y. Dodge et V. Rousson, Éditions Dunod, 2004.



Sommaire

- 1 Test et analyse de variance de la régression
- 2 Distribution des paramètres
 - Modèle de régression linéaire simple
 - Distribution de la pente du modèle
 - Distribution de l'ordonnée à l'origine
- 3 Tests et intervalles de confiance sur les paramètres
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - Test sur l'ordonnée à l'origine
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

Sommaire

- 1 Test et analyse de variance de la régression
- 2 Distribution des paramètres
 - Modèle de régression linéaire simple
 - Distribution de la pente du modèle
 - Distribution de l'ordonnée à l'origine
- 3 Tests et intervalles de confiance sur les paramètres
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - Test sur l'ordonnée à l'origine
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

- Il existe plusieurs démarches pour tester la validité de la linéarité d'une régression linéaire simple.
- Nous montrons l'équivalence de ces différents tests.
- Conséquence : Cela revient à faire **le test du coefficient de corrélation linéaire**, appelé aussi le coefficient de Bravais-Pearson.

Remarque

Nous renvoyons le lecteur à un cours sur le coefficient de corrélation linéaire.

Problème

Nous souhaitons tester l'hypothèse nulle :

$$\mathcal{H}_0 : \rho(X, Y) = 0$$

contre l'hypothèse alternative :

$$\mathcal{H}_1 : \rho(X, Y) \neq 0$$

où

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X] \text{Var}[Y]}},$$

avec

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \text{Cov}(Y, X).$$

Solution

La méthode que nous allons employer ici est :

la méthode de l'ANOVA

utilisée par les logiciels de statistique.

Remarques

- 1 ANOVA pour ANalysis Of VAriance ou encore analyse de la variance.
- 2 Nous renvoyons le lecteur à un cours sur le test du coefficient de corrélation linéaire.

Remarque

Nous avons établi dans le cours numéro 5 :

**Somme des Carrés Totale = Somme des Carrés Expliquée
+ Somme des Carrés Résiduelle**

ce qui s'écrit mathématiquement par :

$$\sum_{i=1}^n (Y_i - \bar{Y}_n)^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y}_n)^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2.$$

À chaque somme de carrés est associé son nombre de degrés de liberté (*ddl*). Ces *ddl* sont présents dans le tableau de l'ANOVA.

Tableau de l'ANOVA

Source de variation	sc	ddl	cm
expliquée sc_{reg}	$\sum_{i=1}^n (\hat{y}_i - \bar{y}_n)^2$	1	$sc_{reg}/1$
résiduelle sc_{res}	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - 2$	$sc_{res}/(n - 2)$
totale sc_{tot}	$\sum_{i=1}^n (y_i - \bar{y}_n)^2$	$n - 1$	

Remarques

1 Le coefficient de détermination

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

mesure le pourcentage d'explication du modèle par la régression linéaire.

2 Le rapport

$$cm_{res} = \frac{SC_{res}}{n - 2}$$

est l'estimation de la variance résiduelle.

À partir du tableau de l'ANOVA, nous effectuons **le test de la linéarité de la régression** en calculant **la statistique de Fisher F** qui suit une loi de Fisher $F(1, n - 2)$.

Cette variable aléatoire F se réalise en :

$$F_{obs} = \frac{SC_{reg}/1}{SC_{res}/(n-2)} = (n-2) \frac{SC_{reg}}{SC_{res}}.$$

Décision

Si

$$F_{obs} \geq F_{1-\alpha}(1, n-2),$$

alors nous décidons de rejeter l'hypothèse nulle \mathcal{H}_0 et par conséquent d'accepter l'hypothèse alternative \mathcal{H}_1 au risque α , c'est-à-dire qu'il existe une liaison linéaire significative entre X et Y .

Si

$$F_{obs} < F_{1-\alpha}(1, n-2),$$

alors nous décidons de ne pas rejeter l'hypothèse nulle \mathcal{H}_0 et par conséquent de l'accepter, c'est-à-dire nous concluons qu'il n'existe pas de liaison linéaire entre X et Y .

Remarque

En effet, si l'hypothèse nulle \mathcal{H}_0 est vérifiée alors cela implique que $\rho(X, Y) = 0$ c'est-à-dire $\text{Cov}(X, Y) = 0$. Donc il n'existe aucune liaison linéaire entre X et Y .

Sommaire

- 1 Test et analyse de variance de la régression
- 2 **Distribution des paramètres**
 - Modèle de régression linéaire simple
 - Distribution de la pente du modèle
 - Distribution de l'ordonnée à l'origine
- 3 Tests et intervalles de confiance sur les paramètres
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - Test sur l'ordonnée à l'origine
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

Sommaire

- 1 Test et analyse de variance de la régression
- 2 **Distribution des paramètres**
 - **Modèle de régression linéaire simple**
 - Distribution de la pente du modèle
 - Distribution de l'ordonnée à l'origine
- 3 Tests et intervalles de confiance sur les paramètres
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - Test sur l'ordonnée à l'origine
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

Modélisation

Le modèle de régression linéaire simple est

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

où les ε_i sont des variables aléatoires inobservables, appelées **les erreurs**.

Conséquence : Les variables Y_i sont aléatoires.

Première hypothèse : $\mathbb{E}[\varepsilon_i] = 0$.

Conséquence : $\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_i$.

D'autre part, nous avons :

$$\text{Var}[Y_i] = \text{Var}[\varepsilon_i].$$

Les quatre hypothèses indispensables pour construire la théorie :

- 1 Les variables aléatoires ε_i sont indépendantes.
- 2 Les variables aléatoires ε_i sont normalement distribuées.
- 3 L'espérance des variables aléatoires ε_i est égale à 0.
- 4 La variance des variables aléatoires ε_i est égale à σ^2 (inconnue) ne dépendant pas de x_i .

Nous avons donc pour tout $i = 1, \dots, n$:

$$\text{Var}[\varepsilon_i] = \text{Var}[Y_i] = \sigma^2.$$

Cette condition s'appelle d'**homoscédasticité**.

Résumons-nous

Ces quatre hypothèses sont équivalentes à :

les variables aléatoires ε_i sont indépendantes et identiquement distribuées selon une loi normale de moyenne nulle et de variance σ^2 .

Nous notons :

$$\varepsilon_i \text{ i.i.d. } \sim \mathcal{N}(0; \sigma^2).$$

Conséquences importantes :

- 1 La normalité des variables aléatoires ε_i implique la normalité des variables aléatoires Y_i .
- 2 L'indépendance des variables aléatoires ε_i implique l'indépendance des variables aléatoires Y_i .
En effet, nous montrons en calculant que :

$$\begin{aligned} \text{Cov}[Y_i, Y_j] &= \text{Cov}[\beta_0 + \beta_1 x_i + \varepsilon_i, \beta_0 + \beta_1 x_j + \varepsilon_j] \\ &= \text{Cov}[\varepsilon_i, \varepsilon_j] \\ &= 0. \end{aligned}$$

Sommaire

- 1 Test et analyse de variance de la régression
- 2 **Distribution des paramètres**
 - Modèle de régression linéaire simple
 - **Distribution de la pente du modèle**
 - Distribution de l'ordonnée à l'origine
- 3 Tests et intervalles de confiance sur les paramètres
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - Test sur l'ordonnée à l'origine
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

Résultat

Nous avons :

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}_n) Y_i}{\sum (x_i - \bar{x}_n)^2}, \text{ où } \bar{x}_n = \frac{\sum x_i}{n}.$$

Conséquences

Il en résulte que :

- $\hat{\beta}_1$ **est une variable aléatoire** car $\hat{\beta}_1$ dépend des variables Y_i qui sont des variables aléatoires.
- $\hat{\beta}_1$ **est une fonction linéaire des variables aléatoires** Y_i .
- Comme les variables aléatoires Y_i par hypothèse sont normalement distribuées, alors $\hat{\beta}_1$ **est normalement distribuée**.

Il reste donc à calculer ces deux valeurs pour caractériser la loi de l'estimateur $\hat{\beta}_1$:

1 $\mathbb{E} \left[\hat{\beta}_1 \right]$

2 $Var \left[\hat{\beta}_1 \right].$

Calcul de l'espérance de $\hat{\beta}_1$

D'une part, nous calculons l'espérance de $\hat{\beta}_1$ ainsi :

$$\begin{aligned}\mathbb{E}[\hat{\beta}_1] &= \mathbb{E}\left[\frac{\sum(x_i - \bar{x}_n)Y_i}{\sum(x_i - \bar{x}_n)^2}\right] \\&= \frac{\sum(x_i - \bar{x}_n)\mathbb{E}[Y_i]}{\sum(x_i - \bar{x}_n)^2} \\&= \frac{\sum(x_i - \bar{x}_n)(\beta_0 + \beta_1 x_i)}{\sum(x_i - \bar{x}_n)^2} \\&= \frac{\beta_0 \sum(x_i - \bar{x}_n) + \beta_1 \sum(x_i - \bar{x}_n)x_i}{\sum(x_i - \bar{x}_n)^2} \\&= \frac{0 + \beta_1 \sum(x_i - \bar{x}_n)x_i}{\sum(x_i - \bar{x}_n)^2}.\end{aligned}$$

Suite et fin du calcul de l'espérance de $\hat{\beta}_1$

En effet, nous montrons que : $\sum (x_i - \bar{x}_n) = 0$. De plus, comme nous avons :

$$\sum (x_i - \bar{x}_n)^2 = \sum (x_i - \bar{x}_n)x_i$$

alors nous obtenons :

$$\mathbb{E} [\hat{\beta}_1] = \beta_1.$$

Remarque importante

Donc la variable aléatoire $\hat{\beta}_1$ est **un estimateur sans biais** du coefficient β_1 .

Calcul de la variance de $\hat{\beta}_1$

D'autre part, nous calculons la variance de $\hat{\beta}_1$ ainsi :

$$\begin{aligned} \text{Var} [\hat{\beta}_1] &= \text{Var} \left[\frac{\sum (x_i - \bar{x}_n) Y_i}{\sum (x_i - \bar{x}_n)^2} \right] \\ &= \frac{\sum (x_i - \bar{x}_n)^2 \text{Var}[Y_i]}{(\sum (x_i - \bar{x}_n)^2)^2} \\ &= \frac{\sum (x_i - \bar{x}_n)^2 \sigma^2}{(\sum (x_i - \bar{x}_n)^2)^2} \\ \text{Var} [\hat{\beta}_1] &= \frac{\sigma^2}{\sum (x_i - \bar{x}_n)^2} \end{aligned}$$

Sommaire

- 1 Test et analyse de variance de la régression
- 2 **Distribution des paramètres**
 - Modèle de régression linéaire simple
 - Distribution de la pente du modèle
 - **Distribution de l'ordonnée à l'origine**
- 3 Tests et intervalles de confiance sur les paramètres
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - Test sur l'ordonnée à l'origine
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

Nous avons :

$$\hat{\beta}_0 = \bar{Y}_n - \hat{\beta}_1 \bar{x}_n$$

où

$$\bar{x}_n = \frac{\sum x_i}{n} \quad \text{et} \quad \bar{Y}_n = \frac{\sum Y_i}{n}.$$

- $\hat{\beta}_0$ **est une variable aléatoire** car $\hat{\beta}_0$ dépend de $\hat{\beta}_1$ qui est une variable aléatoire.
- $\hat{\beta}_0$ **est une fonction linéaire de $\hat{\beta}_1$.**
- Comme $\hat{\beta}_1$ est normalement distribuée, alors $\hat{\beta}_0$ **est normalement distribuée.**

Il reste donc à calculer ces deux valeurs pour caractériser l'estimateur $\hat{\beta}_0$:

1 $\mathbb{E} [\hat{\beta}_0]$

2 $Var [\hat{\beta}_0]$.

Calcul de l'espérance de $\hat{\beta}_0$

D'une part, nous calculons l'espérance de $\hat{\beta}_0$ ainsi :

$$\begin{aligned}\mathbb{E} [\hat{\beta}_0] &= \mathbb{E} [\bar{Y}_n - \hat{\beta}_1 \bar{x}_n] \\ &= \mathbb{E} [\bar{Y}_n] - \bar{x}_n \mathbb{E} [\hat{\beta}_1] \\ &= \mathbb{E} [\bar{Y}_n] - \bar{x}_n \beta_1,\end{aligned}$$

car nous venons de démontrer que $\hat{\beta}_1$ est un estimateur sans biais du coefficient β_1 .

Il reste à calculer la valeur :

$$\mathbb{E} [\bar{Y}_n] .$$

Suite du calcul de l'espérance de $\hat{\beta}_0$

Or nous avons :

$$\begin{aligned}\mathbb{E}[\bar{Y}_n] &= \mathbb{E}\left[\frac{\sum Y_i}{n}\right] \\ &= \frac{\sum \mathbb{E}[Y_i]}{n} \\ &= \frac{\sum (\beta_0 + \beta_1 x_i)}{n} \\ &= \frac{n\beta_0 + \beta_1 \sum x_i}{n} \\ &= \beta_0 + \bar{x}_n \beta_1.\end{aligned}$$

Fin du calcul de l'espérance de $\hat{\beta}_0$

Nous obtenons donc :

$$\begin{aligned}\mathbb{E} [\hat{\beta}_0] &= \mathbb{E} [\bar{Y}_n] - \bar{X}_n \beta_1 \\ &= (\beta_0 + \bar{X}_n \beta_1) - \bar{X}_n \beta_1 \\ &= \beta_0.\end{aligned}$$

Remarque

Donc la variable aléatoire $\hat{\beta}_0$ est **un estimateur sans biais** du coefficient β_0 .

Calcul de la variance de $\hat{\beta}_0$

D'autre part, nous calculons la variance de $\hat{\beta}_0$ ainsi :

$$\begin{aligned} \text{Var} [\hat{\beta}_0] &= \text{Var} [\bar{Y}_n - \hat{\beta}_1 \bar{x}_n] \\ &= \text{Var} [\bar{Y}_n] + \bar{x}_n^2 \text{Var} [\hat{\beta}_1] - 2 \bar{x}_n \text{Cov} [\bar{Y}_n, \hat{\beta}_1]. \end{aligned}$$

Il reste donc à calculer la valeur :

$$\text{Cov} [\bar{Y}_n, \hat{\beta}_1].$$

Suite du calcul de la variance de $\hat{\beta}_0$

Par les calculs, nous montrons que :

$$\begin{aligned}
 \text{Cov} [\bar{Y}_n, \hat{\beta}_1] &= \text{Cov} \left[\frac{\sum Y_i}{n}, \frac{\sum (x_j - \bar{x}_n) Y_j}{\sum (x_i - \bar{x}_n)^2} \right] \\
 &= \frac{\sum_i \sum_j (x_j - \bar{x}_n) \text{Cov}[Y_i, Y_j]}{n \sum (x_i - \bar{x}_n)^2} \\
 &= \frac{\sum_i (x_i - \bar{x}_n) \text{Var}[Y_i]}{n \sum (x_i - \bar{x}_n)^2} \\
 &= \frac{\sigma^2 \sum_i (x_i - \bar{x}_n)}{n \sum (x_i - \bar{x}_n)^2} \\
 &= 0.
 \end{aligned}$$

Suite du calcul de la variance de $\hat{\beta}_0$

Comme nous avons que

$$\text{Var} [\bar{Y}_n] = \frac{\sigma^2}{n},$$

nous obtenons, alors :

$$\begin{aligned}\text{Var} [\hat{\beta}_0] &= \text{Var} [\bar{Y}_n] + \bar{x}_n^2 \text{Var} [\hat{\beta}_1] \\ &= \frac{\sigma^2}{n} + \frac{\bar{x}_n^2 \sigma^2}{\sum (x_i - \bar{x}_n)^2} \\ &= \frac{\sigma^2 \left(\sum (x_i - \bar{x}_n)^2 + n \bar{x}_n^2 \right)}{n \sum (x_i - \bar{x}_n)^2}.\end{aligned}$$

Fin du calcul de la variance de $\hat{\beta}_0$

En rappelant que :

$$\sum (x_i - \bar{x}_n)^2 = \sum x_i^2 - n\bar{x}_n^2,$$

nous avons finalement :

$$\text{Var} [\hat{\beta}_0] = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x}_n)^2}.$$

Sommaire

- 1 Test et analyse de variance de la régression
- 2 Distribution des paramètres
 - Modèle de régression linéaire simple
 - Distribution de la pente du modèle
 - Distribution de l'ordonnée à l'origine
- 3 **Tests et intervalles de confiance sur les paramètres**
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - Test sur l'ordonnée à l'origine
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

Sommaire

- 1 Test et analyse de variance de la régression
- 2 Distribution des paramètres
 - Modèle de régression linéaire simple
 - Distribution de la pente du modèle
 - Distribution de l'ordonnée à l'origine
- 3 Tests et intervalles de confiance sur les paramètres
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - Test sur l'ordonnée à l'origine
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

Rappel

Nous rappelons que :

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1; \sigma^2(\hat{\beta}_1))$$

où

$$\sigma^2(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x}_n)^2}.$$

Nous obtenons alors :

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma(\hat{\beta}_1)} \sim \mathcal{N}(0; 1).$$

Problème

Nous ne connaissons pas le paramètre σ^2 , c'est-à-dire la variance des variables aléatoires ε_i .

Que pouvons-nous faire alors pour résoudre ce problème ?

Proposition

Estimer ce paramètre ! Oui, mais comment ?

Solution

- Nous estimons d'abord σ^2 par CM_{res} l'estimateur sans biais de σ^2 :

$$CM_{res} = \frac{\|\varepsilon\|^2}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}.$$

- Nous estimons ensuite $\sigma^2(\hat{\beta}_1)$ par :

$$s^2(\hat{\beta}_1) = \frac{CM_{res}}{\sum (x_i - \bar{x}_n)^2}.$$

Fin de la solution

- Nous montrons alors que :

$$\frac{\hat{\beta}_1 - \beta_1}{s(\hat{\beta}_1)} \sim T_{n-2},$$

où la variable aléatoire T_{n-2} désigne une variable de Student avec $(n - 2)$ degrés de liberté.

Mise en place du test sur la pente

Nous souhaitons tester l'hypothèse nulle :

$$\mathcal{H}_0 : \beta_1 = 0$$

contre l'hypothèse alternative :

$$\mathcal{H}_1 : \beta_1 \neq 0.$$

Nous utilisons alors la statistique de Student suivante :

$$t_{obs} = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)}$$

pour décider de l'acceptation ou du rejet de \mathcal{H}_0 .

Décision

Nous décidons de rejeter l'hypothèse nulle \mathcal{H}_0 et donc d'accepter l'hypothèse alternative \mathcal{H}_1 au seuil de signification α si

$$|t_{obs}| \geq t_{n-2; 1-\alpha/2}$$

où la valeur critique $t_{n-2; 1-\alpha/2}$ est le $(1 - \alpha/2)$ -quantile d'une loi de Student avec $(n - 2)$ ddl.

Dans ce cas, nous disons que la relation linéaire entre X et Y est significative au seuil α .

Décision - Suite et fin

Nous décidons d'accepter l'hypothèse nulle \mathcal{H}_0 au seuil de signification α si

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

où la valeur $t_{n-2; 1-\alpha/2}$ est le $(1 - \alpha/2)$ -quantile d'une loi de Student avec $(n - 2)$ ddl.

Dans ce cas, Y ne dépend pas linéairement de X . Le modèle devient alors :

$$Y_i = \beta_0 + \varepsilon_i$$

Le modèle proposé $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ est inadéquat. Nous testons alors un nouveau modèle.

Sommaire

- 1 Test et analyse de variance de la régression
- 2 Distribution des paramètres
 - Modèle de régression linéaire simple
 - Distribution de la pente du modèle
 - Distribution de l'ordonnée à l'origine
- 3 Tests et intervalles de confiance sur les paramètres
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - Test sur l'ordonnée à l'origine
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

IC pour β_1

Un intervalle de confiance au niveau $(1 - \alpha)$ pour le coefficient inconnu β_1 est défini par

$$\left] \hat{\beta}_1 - t_{n-2;1-\alpha/2} \times s(\hat{\beta}_1) ; \hat{\beta}_1 + t_{n-2;1-\alpha/2} \times s(\hat{\beta}_1) \right[.$$

Cet intervalle de confiance est construit pour contenir, dans $(1 - \alpha)\%$ des cas, la vraie valeur inconnue du coefficient β_1 .

Sommaire

- 1 Test et analyse de variance de la régression
- 2 Distribution des paramètres
 - Modèle de régression linéaire simple
 - Distribution de la pente du modèle
 - Distribution de l'ordonnée à l'origine
- 3 Tests et intervalles de confiance sur les paramètres
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - **Test sur l'ordonnée à l'origine**
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

Rappel

Nous rappelons que :

$$\hat{\beta}_0 \sim \mathcal{N}(\beta_0; \sigma^2(\hat{\beta}_0))$$

où

$$\sigma^2(\hat{\beta}_0) = \frac{\sigma^2 \sum x_i^2}{n \sum (x_i - \bar{x}_n)^2}.$$

Nous obtenons alors :

$$\frac{\hat{\beta}_0 - \beta_0}{\sigma(\hat{\beta}_0)} \sim \mathcal{N}(0; 1).$$

Problème

Nous ne connaissons pas le paramètre σ^2 , c'est-à-dire la variance des variables aléatoires ε_i .

Que pouvons-nous faire alors pour résoudre ce problème ?

Solution

Estimer ce paramètre !

Solution

- Nous estimons d'abord σ^2 par CM_{res} l'estimateur sans biais de σ^2 :

$$CM_{res} = \frac{\|\varepsilon\|^2}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}.$$

- Nous estimons ensuite $\sigma^2(\hat{\beta}_0)$ par :

$$s^2(\hat{\beta}_0) = \frac{CM_{res} \sum x_i^2}{n \sum (x_i - \bar{x}_n)^2}.$$

Suite de la solution

- Nous montrons alors que :

$$\frac{\hat{\beta}_0 - \beta_0}{s(\hat{\beta}_0)} \sim T_{n-2},$$

où T_{n-2} désigne une v.a. de Student avec $(n - 2)$ ddl.

Mise en place du test sur l'ordonnée à l'origine

Nous souhaitons tester l'hypothèse nulle

$$\mathcal{H}_0 : \beta_0 = 0$$

contre l'hypothèse alternative

$$\mathcal{H}_1 : \beta_0 \neq 0.$$

Nous utilisons la statistique de Student suivante :

$$t_{obs} = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)}$$

pour décider de l'acceptation ou du rejet de \mathcal{H}_0 .

Décision

Nous décidons de refuser l'hypothèse nulle \mathcal{H}_0 et d'accepter l'hypothèse alternative \mathcal{H}_1 au seuil de signification α si :

$$|t_{obs}| \geq t_{n-2;1-\alpha/2}$$

où la valeur critique $t_{n-2;1-\alpha/2}$ est le $(1 - \alpha/2)$ -quantile d'une loi de Student avec $(n - 2)$ ddl.

Dans ce cas, le coefficient β_0 du modèle est dit significatif au seuil α .

Décision - Suite et fin

Nous décidons de ne pas refuser et donc d'accepter l'hypothèse nulle \mathcal{H}_0 au seuil de signification α si

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

où la valeur critique $t_{n-2; 1-\alpha/2}$ est le $(1 - \alpha/2)$ -quantile d'une loi de Student avec $(n - 2)$ ddl.

Dans ce cas, l'ordonnée de la droite de régression passe par l'origine :

$$Y_i = \beta_1 x_i + \varepsilon_i.$$

Sommaire

- 1 Test et analyse de variance de la régression
- 2 Distribution des paramètres
 - Modèle de régression linéaire simple
 - Distribution de la pente du modèle
 - Distribution de l'ordonnée à l'origine
- 3 Tests et intervalles de confiance sur les paramètres
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - Test sur l'ordonnée à l'origine
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

IC pour β_0

Un intervalle de confiance au niveau $(1 - \alpha)$ pour le coefficient inconnu β_0 est défini par :

$$\left[\hat{\beta}_0 - t_{n-2;1-\alpha/2} \times s(\hat{\beta}_0) ; \hat{\beta}_0 + t_{n-2;1-\alpha/2} \times s(\hat{\beta}_0) \right].$$

Cet intervalle de confiance est construit pour contenir, dans $(1 - \alpha)\%$ des cas, la vraie valeur inconnue du coefficient β_0 .

Sommaire

- 1 Test et analyse de variance de la régression
- 2 Distribution des paramètres
 - Modèle de régression linéaire simple
 - Distribution de la pente du modèle
 - Distribution de l'ordonnée à l'origine
- 3 Tests et intervalles de confiance sur les paramètres
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - Test sur l'ordonnée à l'origine
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

Construction de l'intervalle de confiance pour une valeur moyenne

Nous allons voir comment trouver un intervalle de confiance pour la valeur moyenne x :

$$\mu_Y(x) = \beta_0 + \beta_1 x,$$

c'est-à-dire pour l'ordonnée du point d'abscisse x se trouvant sur la droite des moindres carrés ordinaire.

L'estimateur de $\beta_0 + \beta_1 x$ est donné par la droite des moindres carrés ordinaire :

$$\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x,$$

où

$$\hat{Y}(x) \sim \mathcal{N}(\beta_0 + \beta_1 x; \sigma^2(\hat{Y}(x)))$$

et

$$\sigma^2(\hat{Y}(x)) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right).$$

Ce qui peut s'écrire aussi :

$$\frac{\hat{Y}(x) - \mu_Y(x)}{\sigma(\hat{Y}(x))} \sim \mathcal{N}(0; 1).$$

Problème

La variance σ^2 est inconnue.

Solution

- Nous estimons d'abord σ^2 par l'estimateur CM_{res} .
- Nous estimons ensuite $\sigma^2(\hat{Y}(x))$ par :

$$s^2(\hat{Y}(x)) = cm_{res} \left(\frac{1}{n} + \frac{(x - \bar{x}_n)^2}{\sum (x_i - \bar{x}_n)^2} \right).$$

- Ainsi nous obtenons :

$$\frac{\hat{Y}(x) - \mu_Y(x)}{s(\hat{Y}(x))} \sim T_{n-2}.$$

Intervalle de confiance pour une valeur moyenne

Il est possible de construire un intervalle de confiance de la valeur moyenne de Y sachant que $X = x_0$. L'estimation ponctuelle pour cette valeur de x_0 est alors égale à $\hat{y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

L'intervalle de confiance de la valeur moyenne prise par la variable Y lorsque $X = x_0$ est égal à :

$$\left[\hat{y}(x_0) - t_{n-2;1-\alpha/2} \times s(\hat{y}(x_0)) ; \hat{y}(x_0) + t_{n-2;1-\alpha/2} \times s(\hat{y}(x_0)) \right].$$

Remarque

Cet intervalle de confiance est construit pour contenir, dans $(1 - \alpha)\%$ des cas, la vraie valeur moyenne inconnue $\mu_Y(x_0)$.

L'ajustement affine peut servir à prévoir une valeur attendue pour la variable Y quand nous fixons $X = x_0$. L'estimation ponctuelle de cette valeur est alors égale à $\hat{y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Intervalle de prévision d'une valeur individuelle

Un intervalle de prévision au niveau $(1 - \alpha)$ pour la variable Y sachant que $X = x_0$ est égal à :

$$\left[\hat{y}(x_0) - t_{n-2;1-\alpha/2} \sqrt{cm_{res} + s^2(\hat{y}(x_0))}; \right. \\ \left. \hat{y}(x_0) + t_{n-2;1-\alpha/2} \sqrt{cm_{res} + s^2(\hat{y}(x_0))} \right].$$

Remarques

- 1 Cet intervalle de prévision est construit pour contenir, dans $(1 - \alpha)\%$ des cas, la vraie valeur individuelle inconnue $Y(x_0)$.
- 2 L'utilisation d'une valeur estimée $\hat{y}(x_0)$ n'est justifiée que si R^2 est proche de 1.
- 3 Notez que l'intervalle de confiance va produire une étendue de valeurs plus petite, parce qu'il s'agit d'une estimation d'un intervalle pour une moyenne plutôt que de l'estimation d'un intervalle pour une seule observation.

Sommaire

- 1 Test et analyse de variance de la régression
- 2 Distribution des paramètres
 - Modèle de régression linéaire simple
 - Distribution de la pente du modèle
 - Distribution de l'ordonnée à l'origine
- 3 Tests et intervalles de confiance sur les paramètres
 - Test sur la pente
 - Intervalle de confiance pour la pente
 - Test sur l'ordonnée à l'origine
 - Intervalle de confiance pour l'ordonnée à l'origine
- 4 Intervalle de confiance pour une valeur moyenne ou une prévision
- 5 Exemple

Exemple : le tableau de données. D'après Birkes et Dodge (1993)

Pays	Taux d'urbanisation x_i	Taux de natalité y_i
Canada	55,0	16,2
Costa Rica	27,3	30,5
Cuba	33,3	16,9
E.U.	56,5	16,0
El Salvador	11,5	40,2
Guatemala	14,2	38,4
Haïti	13,9	41,3

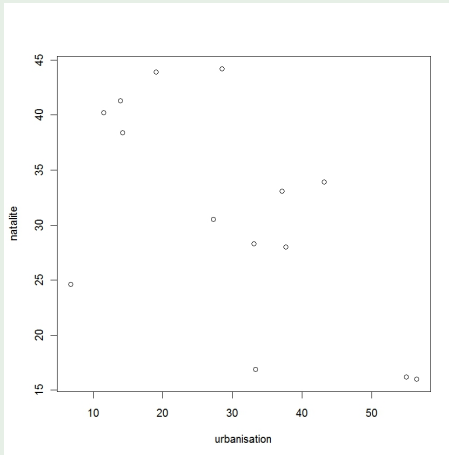
Suite des données

Pays	Taux d'urbanisation x_i	Taux de natalité y_i
Honduras	19,0	43,9
Jamaïque	33,1	28,3
Mexique	43,2	33,9
Nicaragua	28,5	44,2
Trinité-et-Tobago	6,8	24,6
Panama	37,7	28,0
Rép. Dom.	37,1	33,1

Enregistrement des données et nuage de points

```
> urbanisation<-c(55,27.3,33.3,56.5,11.5,14.2,  
+13.9,19,33.1,43.2,28.5,6.8,37.7,37.1)  
> natalite<-c(16.2,30.5,16.9,16.0,40.2,38.4,  
+41.3,43.9,28.3,33.9,44.2,24.6,28,33.1)  
> plot(natalite~urbanisation)
```

Nuage de points



Analyse : calcul du coefficient de corrélation linéaire

Nous souhaitons modéliser la relation entre le taux de natalité et le taux d'urbanisation.

La première question à se poser est : « existe-t-il une relation linéaire entre les deux variables ? »

Pour y répondre, calculons le coefficient de corrélation linéaire de Bravais-Pearson à l'aide de R.

```
> cor(natalite,urbanisation)
[1] -0.6211854
```

Comment interprétons-nous cette valeur ? Il semblerait qu'il puisse exister une relation linéaire entre les deux variables. Il reste donc à réaliser le test du coefficient de corrélation linéaire.

Suite de l'analyse : test du coefficient de corrélation linéaire

Mais pour cela, il faut savoir si le couple (X, Y) suit une loi normale bivariée. Utilisons R.

```
> exemple<-data.frame(urbanisation,natalite)
> transpose<-t(exemple)
> mshapiro.test(transpose)
Shapiro-Wilk normality test
data:  Z
W = 0.927, p-value = 0.2771
```

La p -valeur (p -value = 0,2771) étant supérieure à $\alpha = 5\%$, le test est significatif au seuil de signification $\alpha = 5\%$. Nous décidons donc de ne pas rejeter l'hypothèse nulle \mathcal{H}_0 et par conséquent de l'accepter au seuil de signification $\alpha = 5\%$.

Suite de l'analyse : test du coefficient de corrélation linéaire

Maintenant que l'hypothèse fondamentale est vérifiée, nous pouvons réaliser le test du coefficient de corrélation linéaire.

```
> cor.test(urbanisation,natalite)
Pearson's product-moment correlation
data: urbanisation and natalite
t = -2.7459, df = 12, p-value = 0.01774
alternative hypothesis: true correlation is
not equal to 0
95 percent confidence interval:
-0.8662568 -0.1351496
sample estimates:
cor
-0.6211854
```


Suite et fin de l'analyse : test du coefficient de corrélation linéaire

La p -valeur (p -value = 0,01774) étant inférieure à $\alpha = 5\%$, le test est significatif au seuil de signification $\alpha = 5\%$. Nous décidons donc de rejeter l'hypothèse nulle \mathcal{H}_0 et par conséquent d'accepter l'hypothèse alternative \mathcal{H}_1 au seuil de signification $\alpha = 5\%$. Nous commettons un risque d'erreur de première espèce égal à $\alpha = 5\%$.

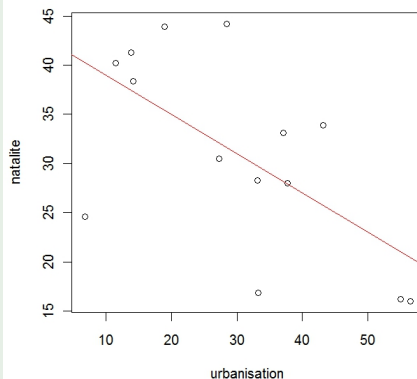
Il existe donc une relation linéaire entre les deux variables, à savoir entre le taux d'urbanisation et le taux de natalité.

Suite de l'analyse du modèle de régression

Maintenant, déterminons les coefficients de la droite des moindres carrés ordinaires avec R et traçons-la.

```
> modele<-lm(natalite~urbanisation)
> coef(modele)
(Intercept)  urbanisation
 42.9905457   -0.3988675
> abline(coef(modele),col="red")
```

Nuage des points et droite des MCO



Calcul des résidus

Pour réaliser les tests sur la pente de la droite et sur l'ordonnée à l'origine (appelé `Intercept` sous R), il faut vérifier que les erreurs suivent une loi normale. Nous allons d'abord calculer les résidus avec la fonction `residuals()` du logiciel R.

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

et les placer dans le tableau de données initial.

Tableau de données avec résidus

Pays	Taux d'urbanisation x_i	Taux de natalité y_i	Valeurs estimées \hat{y}_i	Résidus e_i
Canada	55,0	16,2	21,05	-4,85
Costa Rica	27,3	30,5	32,10	-1,60
Cuba	33,3	16,9	29,71	-12,81
E.U.	56,5	16,0	20,45	-4,45
El Salvador	11,5	40,2	38,40	1,80
Guatemala	14,2	38,4	37,33	1,07
Haïti	13,9	41,3	37,45	3,85

Suite des données avec résidus

Pays	Taux d'urbanisation x_i	Taux de natalité y_i	Valeurs estimées \hat{y}_i	Résidus e_i
Honduras	19,0	43,9	35,41	8,49
Jamaïque	33,1	28,3	29,79	-1,49
Mexique	43,2	33,9	25,76	8,14
Nicaragua	28,5	44,2	31,62	12,58
Trinité-et-Tobago	6,8	24,6	40,28	-15,68
Panama	37,7	28,0	27,95	0,05
Rép. Dom.	37,1	33,1	28,19	4,91

Vérification de la normalité des erreurs

Réalisons donc un test de normalité sur les erreurs, par exemple le test de Shapiro-Wilk avec la fonction `shapiro.test()` de R.

```
> shapiro.test(residus)
      Shapiro-Wilk normality test
data:  residus
W = 0.9635, p-value = 0.7797
```

Vérification de la normalité des erreurs

La p -valeur (p -value = 0,7797) étant supérieure à $\alpha = 5\%$, le test n'est pas significatif au seuil de signification $\alpha = 5\%$. Nous décidons donc de ne pas rejeter l'hypothèse nulle \mathcal{H}_0 et par conséquent de l'accepter au seuil de signification $\alpha = 5\%$. Nous commettons un risque d'erreur de deuxième espèce β qu'il faudrait évaluer.

Test sur la pente de la droite β_1 , « à la main »

Nous testons

$$\mathcal{H}_0 : \beta_1 = 0$$

contre

$$\mathcal{H}_1 : \beta_1 \neq 0.$$

Nous calculons la statistique de test de Student :

$$t_{12,obs} = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} = \frac{-0,3989}{0,1453} = -2,746.$$

Remarque

Pour calculer cette statistique à la main, il a fallu récupérer deux valeurs, à savoir celle du numérateur et du dénominateur. Pour cela, vous vous êtes servis de la sortie de R qui donne ces deux valeurs, en utilisant la commande `summary(modele)`.

Sortie de la fonction `summary(modele)`

```
> summary(modele)
Call:
lm(formula = natalite ~ urbanisation)

Residuals:
    Min       1Q   Median       3Q      Max
-15.6782  -3.7413   0.5601   4.6440  12.5772

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   42.9905     4.8454   8.872 1.28e-06 ***
urbanisation  -0.3989     0.1453  -2.746  0.0177 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.154 on 12 degrees of freedom
Multiple R-squared:  0.3859,    Adjusted R-squared:  0.3347
F-statistic:  7.54 on 1 and 12 DF,  p-value: 0.01774
```

Test sur la pente de la droite β_1 , « à la main »

Or la valeur critique de Student à 12 ddl est égale, au seuil de signification $\alpha = 5\%$, à :

$$t_{(12;0,975)} \simeq 2,179.$$

Remarque

Cette dernière valeur est obtenue avec la fonction `qt ()` de R.

```
> qt (0.975, 12)
[1] 2.178813
```

Décision du test de Student « à la main »

Comme

$$|t_{12,obs}| > t_{12;0,975},$$

le test est significatif au seuil de signification $\alpha = 5\%$. Nous décidons donc de refuser l'hypothèse nulle \mathcal{H}_0 . Par conséquent, nous décidons d'accepter l'hypothèse alternative \mathcal{H}_1 , au seuil de signification $\alpha = 5\%$. Nous commettons une erreur de première espèce qui vaut $\alpha = 5\%$.

En conclusion, la relation linéaire entre le taux de natalité et le taux d'urbanisation est significative au seuil $\alpha = 5\%$.

Remarque : test de Student réalisé avec la p -valeur

En se servant de la sortie de la fonction `summary(modele)` du logiciel \mathbb{R} (voir le slide suivant) et en raisonnant en terme de p -valeur, nous retrouvons la même conclusion.

En effet, comme la p -valeur est égale à 0,0177, le test est significatif au seuil de signification $\alpha = 5\%$. Nous décidons donc de refuser l'hypothèse nulle \mathcal{H}_0 .

Par conséquent, nous décidons d'accepter l'hypothèse alternative \mathcal{H}_1 , au seuil de signification $\alpha = 5\%$.

Nous commettons une erreur de première espèce qui vaut $\alpha = 5\%$.

En conclusion, la relation linéaire entre le taux de natalité et le taux d'urbanisation est significative au seuil $\alpha = 5\%$.

Sortie de la fonction `summary(modele)`

```
> summary(modele)
Call:
lm(formula = natalite ~ urbanisation)

Residuals:
    Min       1Q   Median       3Q      Max
-15.6782  -3.7413   0.5601   4.6440  12.5772

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.9905     4.8454   8.872 1.28e-06 ***
urbanisation  -0.3989     0.1453  -2.746  0.0177 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.154 on 12 degrees of freedom
Multiple R-squared:  0.3859,    Adjusted R-squared:  0.3347
F-statistic:  7.54 on 1 and 12 DF,  p-value: 0.01774
```

IC pour β_1 « à la main »

Un intervalle de confiance pour le coefficient inconnu β_1 au niveau $(1 - \alpha) = 0,95$ s'obtient en calculant :

$$\hat{\beta}_1 \pm t_{12;0,975} \times s(\hat{\beta}_1) = -0,3989 \pm 2,179 \times 0,1453.$$

Nous avons donc après simplification :

$$]-0,716; -0,082[$$

qui contient la vraie valeur de la pente de la droite inconnue β_1 avec une probabilité de 0,95.

Remarque

Nous remarquons que 0 n'est pas compris dans l'intervalle.

Remarque

En se servant de la sortie de la fonction `confint(modele)` du logiciel R, nous retrouvons le même résultat mais avec plus de précision.

```
> confint(modele1)
                2.5%          97.5%
(Intercept) 32.4332596 53.54783182
xi          -0.7153622 -0.08237281
```

Test sur l'ordonnée β_0

$$\mathcal{H}_0 : \beta_0 = 0$$

contre

$$\mathcal{H}_1 : \beta_0 \neq 0.$$

Nous calculons

$$t_{12,obs} = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)} = \frac{42,9905}{4,8454} = 8,872.$$

Remarque

Pour calculer cette statistique à la main, il a fallu récupérer deux valeurs, à savoir celle du numérateur et du dénominateur. Pour cela, vous vous êtes servis de la sortie de R qui donne ces deux valeurs, en utilisant la commande `summary(modele)`.

Sortie de la fonction `summary(modele)`

```
> summary(modele)
Call:
lm(formula = natalite ~ urbanisation)

Residuals:
    Min       1Q   Median       3Q      Max
-15.6782  -3.7413   0.5601   4.6440  12.5772

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.9905     4.8454   8.872 1.28e-06 ***
urbanisation  -0.3989     0.1453  -2.746  0.0177 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.154 on 12 degrees of freedom
Multiple R-squared:  0.3859,    Adjusted R-squared:  0.3347
F-statistic: 7.54 on 1 and 12 DF,  p-value: 0.01774
```

Test sur l'ordonnée β_0

Or la valeur critique est égale à pour un seuil $\alpha = 5\%$:

$$t_{0,975;12} \simeq 2,179.$$

Remarque

Cette dernière valeur est obtenue avec la fonction `qt ()` de R.

```
> qt (0.975, 12)  
[1] 2.178813
```

Décision

Comme

$$|t_{12,obs}| > t_{12;0,975},$$

le test est significatif au seuil de signification $\alpha = 5\%$. Nous décidons de refuser l'hypothèse nulle \mathcal{H}_0 .

Par conséquent nous décidons d'accepter l'hypothèse alternative \mathcal{H}_1 , au seuil de signification $\alpha = 5\%$.

Nous commettons une erreur de première espèce qui vaut $\alpha = 5\%$.

En conclusion, la droite des moindres carrés ordinaire ne passe pas par l'ordonnée à l'origine au seuil de signification $\alpha = 5\%$.

Remarque

En se servant de la sortie de la fonction `summary(modele)` du logiciel `R` et en raisonnant en terme de p -valeur, nous retrouvons cette conclusion. En effet, comme la p -valeur est égale à $1,28e - 06$ (voir le slide suivant), le test est significatif au seuil de signification $\alpha = 5\%$. Nous décidons donc de refuser l'hypothèse nulle \mathcal{H}_0 .

Par conséquent, nous décidons d'accepter l'hypothèse alternative \mathcal{H}_1 , au seuil de signification $\alpha = 5\%$.

Nous commençons une erreur de première espèce qui vaut $\alpha = 5\%$.

En conclusion, la droite des moindres carrés ordinaire ne passe pas par l'ordonnée à l'origine au seuil de signification $\alpha = 5\%$.

Sortie de la fonction `summary(modele)`

```
> summary(modele)
Call:
lm(formula = natalite ~ urbanisation)

Residuals:
    Min       1Q   Median       3Q      Max
-15.6782  -3.7413   0.5601   4.6440  12.5772

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.9905     4.8454   8.872 1.28e-06 ***
urbanisation  -0.3989     0.1453  -2.746  0.0177 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.154 on 12 degrees of freedom
Multiple R-squared:  0.3859,    Adjusted R-squared:  0.3347
F-statistic:  7.54 on 1 and 12 DF,  p-value: 0.01774
```


IC pour β_0 « à la main »

Un intervalle de confiance pour le coefficient inconnu β_0 au niveau $(1 - \alpha) = 0,95$ s'obtient en calculant :

$$\hat{\beta}_0 \pm t_{n-2;1-\alpha/2} \times s(\hat{\beta}_0) = 42,9905 \pm 2,179 \times 4,8454.$$

Nous avons donc après simplification :

$$]32,433; 53,548[$$

qui contient la vraie valeur du coefficient inconnu β_0 avec une probabilité de 0,95.

Remarque

Nous remarquons que 0 n'est pas compris dans l'intervalle.

Remarque

En se servant de la sortie de la fonction `confint(modele)` du logiciel R, nous retrouvons le même résultat mais avec plus de précision.

```
>confint(modele)
```

	2.5%	97.5%
(Intercept)	32.4332596	53.54783182
xi	-0.7153622	-0.08237281

Calcul du coefficient de détermination

Une dernière information qu'il ne faut pas oublier, est la lecture du coefficient de détermination `Multiple R-squared` est égal à 38,59%. Ce modèle n'est donc pas très bon. Il faudrait donc introduire de nouvelles variables explicatives comme par exemple le nombre d'habitants ou la superficie du pays pour expliquer le taux de natalité.

Maintenant c'est à vous de vous entraîner en faisant le TD
numéro 5 !
Bonne chance.