Modélisation Statistique
Chapitre 1: Régression linéaire simple

Mohamed Essaied Hamrita

ISMAI, Université Kairouan. Tunisie
mhamrita@gmail.com
http://hamrita.e-monsite.com/

Février 2016

Plan du chapitre

Présentation du modèle

Le **modèle linéaire simple** permet d'expliquer une variable endogène (à expliquée ou dépendante) en fonction d'une variable exogène (explicative).

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, \ldots, N \tag{1}$$

avec:
y: la variable endogène, elle est déterminée par le modèle.
x: la variable explicative supposée exogène.
ε: le terme d'erreur qui est aléatoire.
N: Le nombre d'observations.
A partir des données observées, il est possible d'estimer la relation (1).
On note $\hat{\beta}_0$ et $\hat{\beta}_1$ les estimateurs des paramètres β_0 et β_1, la droite de régression de l'échantillon est donnée par:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, \qquad i = 1, \ldots, N \tag{2}$$

et e_i le résidu d'estimation, $e_i = y_i - \hat{y}_i$, mesure pour chaque individu l'écart entre la valeur observée y_i et la valeur estimée $\hat{y}_i$.

Exemples

Exemple 1 Un exemple traditionnel en économie est l'étude de la relation entre la consommation des ménages (variable dépendante) et leur revenu (variable explicative).

$$C_i = c_0 + cR_i$$

Exemple 2 Quel est le lien entre le salaire d'un individu (variable dépendante) et son niveau d'études (variable explicative)?
Exemple 3 L'étude de la relation entre la quantité produite d'un bien (variable dépendant) et son prix (variable explicative).

$$Q = a + bp$$

Hypothèses sur le modèle

La méthode d'estimation des paramètres du modèle (1) est la méthode des Moindres Carrés Ordinaires (MCO). Avant de présenter cette denière technique, on commence par l'exposé des hypothèses sur le modèle de régression linéaire simple.

Les hypothèses:

H1: L'espérance des aléas est nulle, $E(\varepsilon_i) = 0$, $i = 1, \ldots, N$.

H2: Les alés sont homoscédastiques (de variance constante) et non auto-corrélés (leur covariance est nulle).

$$V(\varepsilon_i) = \sigma^2, \quad i = 1, \ldots, N$$

$$cov(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0, \quad i, j = 1, \ldots, N, \quad i \neq j.$$

H3: La variable x est supposée certaine (exogène).

H4: La variance empirique de la variable x est non nulle et donc les observations x_i ne sont pas toutes identiques.

Estimateurs des MCO

L'estimateur des MCO est déduit de la minimisation de la somme des carrés des erreurs (alés).

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{N} \varepsilon_i^2 = \min_{\beta_0,\beta_1} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2 = \min_{\beta_0,\beta_1} S(\beta_0, \beta_1) \qquad (3)$$

Ce programme donne comme estimateurs:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \text{ et } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad (4)$$

La droite de la régression de l'échantillon est alors donnée par

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \qquad i = 1, \ldots, N$$

Caractéristique de la solution

La solution du programme possède les propriétés suivantes:
- La droite de régression passe par le point moyen $(\bar{x}, \bar{y})$.

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

- La somme des résidus est nulle; $\sum_{i=1}^{N} e_i = 0$, par conséquent $\bar{y} = \bar{\hat{y}}$.
- Le vecteur des résidus et de la variable explicative sont orthogonaux; $\sum_{i=1}^{N} e_i x_i = 0$.

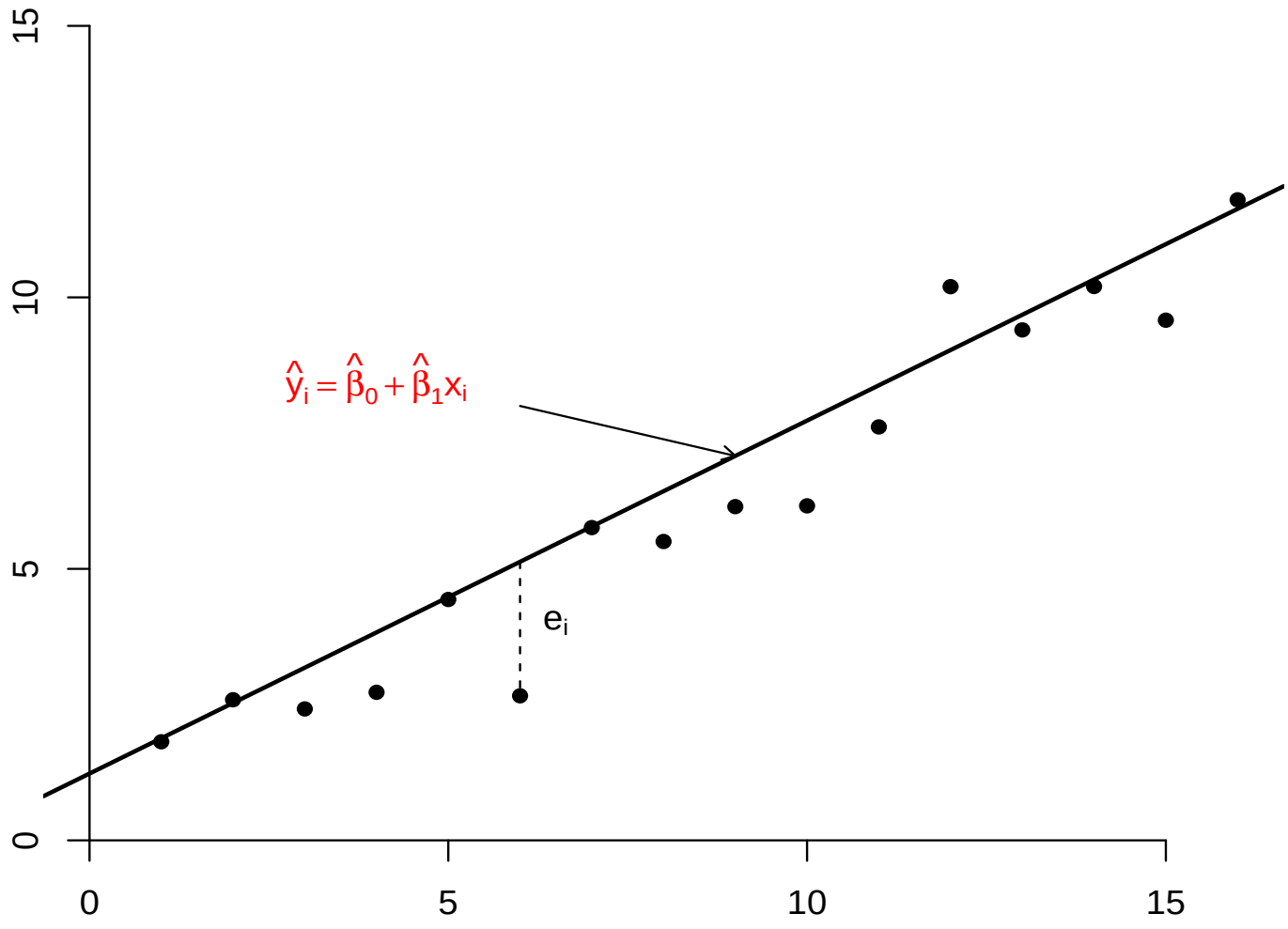

Figure : Droite d'ajustement linéaire

Décomposition de la variance

De ces dernières propriétés algébriques, on déduit l'équation d'analyse de la variance

$$\underbrace{\sum_{i=1}^{N}(y_i - \bar{y})^2}_{SCT} = \underbrace{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}_{SCE} + \underbrace{\sum_{i=1}^{N} e_i^2}_{SCR} \quad (5)$$

avec **SCT**: somme des carrés totale. Elle indique la variabilité totale de Y.
SCE: somme des carrés expliquée. Elle indique la variabilité expliquée par le modèle, c-à-d la variation de Y expliquée par X.
SCR: somme des carrés résiduelle. Elle indique la variabilité non-expliquée par le modèle.
On peut vérifier que:

$$SCE = \hat{\beta}_1^2 S_{xx} = \hat{\beta}_1 S_{xy}$$

On en déduit que:

$$SCR = S_{yy} - \hat{\beta}_1^2 S_{xx} = S_{yy} - \hat{\beta}_1 S_{xy}$$

Qualité d'ajustement

Il est possible de déduire un indicateur synthétique à partir de l'équation d'analyse de variance. C'est le coefficient de détermination R^2.

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Il indique la proportion de la variabilité de Y expliquée par la variable X.

- Plus il sera proche de la valeur 1, meilleur sera le modèle.
- Lorsque R^2 est proche de 0, cela veut dire que X n'apporte pas d'informations utiles sur Y.

Exemple

Soient les $n = 5$ observations suivantes sur les y_i et les x_i:

y_i	2	4	5	7	10
x_i	1	2	3	4	5

Estimer par MCO les paramètres de l'équation $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ et vérifier numériquement les propriétés algébriques de la solution du programme.
Les estimateyrs par MCO sont donnés par l'équation (4).
Remarque:

$$\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sum_{i=1}^{N} x_i^2 - N\bar{x}^2 \text{ et } \sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}.$$

Exemple

Soient les $n = 5$ observations suivantes sur les y_i et les x_i:

y_i	2	4	5	7	10
x_i	1	2	3	4	5

Estimer par MCO les paramètres de l'équation $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ et vérifier numériquement les propriétés algébriques de la solution du programme.
Les estimateyrs par MCO sont donnés par l'équation (4).
Remarque:
$\sum_{i=1}^{N}(x_i - \bar{x})^2 = \sum_{i=1}^{N} x_i^2 - N\bar{x}^2$ et $\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{N} x_i y_i - N\bar{x}\bar{y}$.
On a $\sum x_i = 15$, $\sum y_i = 28$, $\sum x_i^2 = 55$, $\sum y_i^2 = 194$ et $\sum x_i y_i = 103$.
D'où, $\hat{\beta}_1 = \dfrac{103 - 15 \times 25/5}{55 - 15^2/5} = 1.9$ et $\hat{\beta}_0 = 28/5 - 1.9 \times 15/5 = -0.1$.
$\hat{y}_i = -0.1 + 1.9 x_i$: 1.8 3.7 5.6 7.5 9.4
$\sum e_i = \sum(y_i - \hat{y}_i) = 0$.
$\sum x_i e_i = 0$

Avec R

```
yi=c(2,4,5,7,10); xi=1:5
reg1=lm(yi~xi)
coefficients(reg1)

## (Intercept)          xi
##        -0.1         1.9

residuals(reg1)

##    1    2    3    4    5
##  0.2  0.3 -0.6 -0.5  0.6

round(sum(residuals(reg1)*xi),14)

## [1] 0
```

x_i	y_i	$e_i = y_i - \hat{y}_i$	$x_i e_i$
1.00	2.00	0.20	0.20
2.00	4.00	0.30	0.60
3.00	5.00	-0.60	-1.80
4.00	7.00	-0.50	-2.00
5.00	10.00	0.60	3.00
$\sum x_i = 15$	$\sum y_i = 28$	$\sum e_i = 0$	$\sum x_i e_i = 0$

Propriétés des estimateurs

Théorème de Gauss-Markov

Sous les hypothèse H1, H2, H3 et H4, l'estimateur des MCO est BLUE (Best Linear Unbiaised Estimator), ie; meilleur estimateur linéaire sans biais.

Autrement dit, l'estimateur des MCO est
- une fonction linéaire de y,
- non biaisé: $E(\hat{\beta}_0) = \beta_0$ et $E(\hat{\beta}_1) = \beta_1$,
- efficace: l'estimateur des MCO possède la variance la plus faible parmi les estimateurs linéaires sans biais des paramètres β_0 et β_1. La matrice de variance-covariance des paramètres est donnée par

$$V(\hat{\beta}) = \begin{pmatrix} V(\hat{\beta}_0) & cov(\hat{\beta}_0, \hat{\beta}_1) \\ cov(\hat{\beta}_0, \hat{\beta}_1) & V(\hat{\beta}_1) \end{pmatrix}$$

Demonstration

-$E(\hat{\beta}_1) = \beta_1$:

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum(x_i - \bar{x})y_i - \bar{y}\underbrace{\sum(x_i - \bar{x})}_{=0} = \sum(x_i - \bar{x})y_i$$

D'où

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})\beta_0}{\sum(x_i - \bar{x})^2} + \beta_1 \frac{\sum(x_i - \bar{x})x_i}{\sum(x_i - \bar{x})^2} + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2} \\
&= \beta_1 + \frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}
\end{aligned}$$

Donc $E(\hat{\beta}_1) = \beta_1$ car $E(\varepsilon) = 0$ et x_i est exogène.

Démonstration

- $E(\hat{\beta}_0) = \beta_0$. En effet;

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = \beta_0 + \beta_1 \bar{x} + \bar{\varepsilon} - \hat{\beta}_1 \bar{x} \\ &= \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{\varepsilon} \end{aligned}$$

Donc $E(\hat{\beta}_0) = \beta_0$ car $E(\beta_1 - \hat{\beta}_1) = 0$ et $E(\bar{\varepsilon}) = 0$.

Les variances de $\hat{\beta}_0$ et $\hat{\beta}_1$ sont donnée respectivement par:

$$\boxed{V(\hat{\beta}_0) = \sigma_\varepsilon^2 \left(\frac{1}{N} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})} \right) ; \; V(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{\sum (x_i - \bar{x})}}$$

Calcul des variances

On commence par le calcul de $V(\hat{\beta}_1)$.

$$
\begin{aligned}
V(\hat{\beta}_1) &= E(\hat{\beta}_1 - E(\hat{\beta}_1))^2 = E(\hat{\beta}_1 - \beta_1)^2 \\
&= E\left(\frac{\sum(x_i - \bar{x})\varepsilon_i}{\sum(x_i - \bar{x})^2}\right)^2 \quad \text{on pose } w_i = \frac{(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \\
&= E\left[\left(\sum w_i \varepsilon_i\right)^2\right] = E\left[\sum w_i^2 \varepsilon_i^2 + 2\sum_{i \neq j} w_i w_j \varepsilon_i \varepsilon_j\right] \\
&= \sum w_i^2 E(\varepsilon_i^2) + 2\sum_{i \neq j} w_i w_j E(\varepsilon_i \varepsilon_j) \\
&= \frac{\sigma_\varepsilon^2}{\sum(x_i - \bar{x})^2} \text{ car } E(\varepsilon_i^2) = \sigma_\varepsilon^2 \text{ et } E(\varepsilon_i \varepsilon_j) = 0 \, \forall i \neq j
\end{aligned}
$$

Calcul des variances

$$
\begin{aligned}
V(\hat{\beta}_0) &= E(\hat{\beta}_0 - E(\hat{\beta}_0))^2 = E(\hat{\beta}_0 - \beta_0)^2 \\
&= E\left[\left(\bar{x}(\beta_1 - \hat{\beta}_1) + \bar{\varepsilon}\right)^2\right] \\
&= E\left[\bar{x}^2(\beta_1 - \hat{\beta}_1)^2 + \bar{\varepsilon}^2 + 2\bar{x}\bar{\varepsilon}(\beta_1 - \hat{\beta}_1)\right] \\
&= \bar{x}^2 V(\hat{\beta}_1) + E\left[\left(\frac{1}{N}\sum \varepsilon_i\right)^2\right] + 2\bar{x}\underbrace{E\left[(\beta_1 - \hat{\beta}_1)\bar{\varepsilon}\right]}_{=0} \\
&= \bar{x}^2 \frac{\sigma_\varepsilon^2}{\sum(x_i - \bar{x})^2} + \frac{\sigma_\varepsilon^2}{N} = \sigma_\varepsilon^2\left[\frac{1}{N} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})}\right]
\end{aligned}
$$

$$
\begin{aligned}
E\left[(\beta_1 - \hat{\beta}_1)\bar{\varepsilon}\right] &= -E\left[\left(\sum w_i e_i\right)\frac{1}{N}\sum e_i\right] = -E\left[\frac{1}{N}\left(\sum w_i e_i^2 + \sum_{i\neq j} e_i e_j\right.\right. \\
&= -\frac{1}{N}\sum w_i E(e_i^2) = -\frac{1}{N}\sigma^2 \underbrace{\sum w_i}_{=0} = 0
\end{aligned}
$$

Calcul de la covariance

$$\begin{aligned} Cov(\hat{\beta}_0, \hat{\beta}_1) &= E(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1) \\ &= E\left[(\bar{\varepsilon} - (\hat{\beta}_1 - \beta_1)\bar{x})(\hat{\beta}_1 - \beta_1)\right] \\ &= E\left[\bar{\varepsilon}(\hat{\beta}_1 - \beta_1)\right] - \bar{x}E\left[(\hat{\beta}_1 - \beta_1)^2\right] \\ &= -\frac{\sigma_\varepsilon^2 \bar{x}}{\sum(x_i - \bar{x})^2} \end{aligned}$$

Ainsi, la matrice des variances-covariance des paramètres estimés est donnée par:

$$V(\hat{\beta}) = \begin{pmatrix} \sigma_\varepsilon^2\left(\frac{1}{N} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}\right) & -\frac{\sigma_\varepsilon^2 \bar{x}}{\sum(x_i - \bar{x})^2} \\ -\frac{\sigma_\varepsilon^2 \bar{x}}{\sum(x_i - \bar{x})^2} & \frac{\sigma_\varepsilon^2}{\sum(x_i - \bar{x})^2} \end{pmatrix}$$

Distribution des estimateurs

Afin de déterminer la distribution statistique des paramètres estimés, on introduit une hypothèse supplémentaire

H5: Les aléas sont distribué identiquement et indépendament selon une loi normale

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma_\varepsilon^2)$$

puisque la variable endogène y_i est une fonction linéaire de ε_i, donc est aussi normale

$$y_i \sim N(\beta_0 + \beta_1 x_i, \sigma_\varepsilon^2)$$

Les paramètres estimés sont linéaires par rapport aux y_i. On en déduit

$$\hat{\beta}_0 \sim N(\beta_0, V(\hat{\beta}_0)), \quad \hat{\beta}_1 \sim N(\beta_1, V(\hat{\beta}_1))$$

d'où

$$\frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \sim N(0,1), \quad \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1)$$

Estimateur de la variance des erreurs

La variance de la variable aléatoire e_i est l'un des paramètres inconnus du modèle de régression linéaire et est donnée par

$$V(e_i) = E\left(e_i - E(e_i)\right)^2 = E(e_i^2) = \sigma_\varepsilon^2$$

Un estimateur évident de la variance des erreurs, σ_ε^2, est la moyenne arithemétique du carré des erreurs,

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum e_i^2}{N} = \frac{\sum (y_i - \hat{y}_i)^2}{N}$$

On peut vérifier que cet estimateur est biaisé et on peut déduire l'estimateur sans biais de σ_ε^2:

$$\boxed{\hat{\sigma}_\varepsilon^2 = \frac{\sum e_i^2}{N-2} = \frac{SCR}{N-2}}$$

Et on peut montrer que:

$$\frac{\sum e_i^2}{\sigma_\varepsilon^2} = (N-2)\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \overset{iid}{\sim} \chi^2(N-2)$$

Test de significativité

Il s'agit de tester si les paramètres estimés du modèle sont statistiquement significatifs, ie;

$$H_0 : \ \beta_i = 0 \quad i = 0, 1$$

$$H_1 : \ \beta_i \neq 0$$

On a $\hat{\beta}_i \overset{iid}{\sim} N(\beta_i, V(\hat{\beta}_i))$, donc

$$\frac{\hat{\beta}_i - \beta_i}{\sigma_{\hat{\beta}_i}} \overset{iid}{\sim} N(0, 1)$$

Et puisque les variances des paramètres du modèle dépendent de celle des erreurs, qui est estimée, alors on déduit un estimateur des variances des paramètres $\hat{\beta}_i$ en remplaçant σ_ε^2 par $\hat{\sigma}_\varepsilon^2$. Donc on aura

$$\hat{t}_i = \frac{\hat{\beta}_i - \beta_i}{\hat{\sigma}_{\hat{\beta}_i}} \overset{iid}{\sim} St(N - 2)$$

Test de significativité

Règle de décision

On compare le ratio de Student empirique $\hat{t}_i$ sous H_0 à la valeur de $t_{(\alpha/2,N-2)}$ de Student lue dans la table à $N-2$ degrés de liberté et pour un seuil de probabilité égal à α %.
Si $|\hat{t}_i| > t_{(\alpha/2,N-2)}$, on décide de rejeter l'hypothèse nulle. Dans ce cas, on conclut que le paramètre β_i est statistiquement significatif au seuil α %.

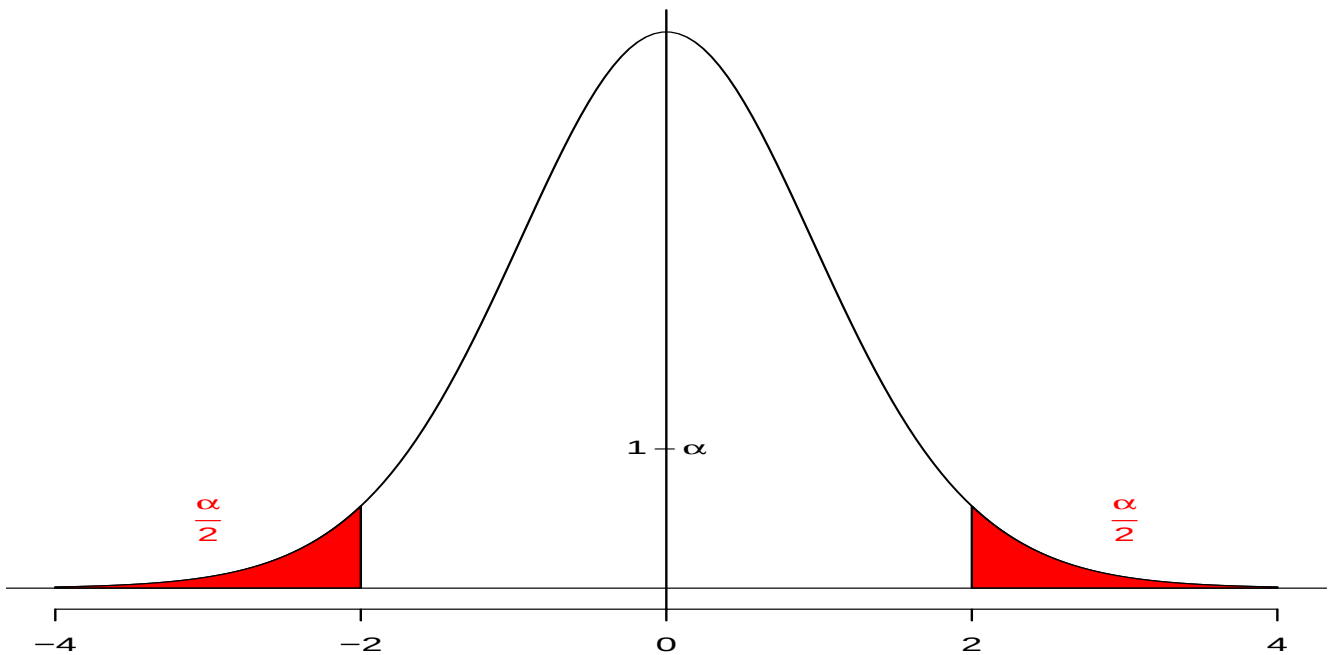

Figure : Distribution de Student

Remarques

1) Le test de significativité individuelle est un cas particulier du test $H_0 : \hat{\beta}_i = \bar{\beta}_i$ où $\bar{\beta}_i = 0$. Dans le cas général, on compare la valeur

$$\hat{t}_i = \left| \frac{\hat{\beta}_i - \bar{\beta}_i}{\hat{\sigma}_{\hat{\beta}_i}} \right|$$

à la valeur tabulée.

2) On peut aussi déterminer l'intervalle de confiance du paramètre $\hat{\beta}_i$ pour décider si H_0 doit être accepter ou non au seuil α.

$$IC(\beta_i) = \left[\hat{\beta}_i \pm \hat{t}_{(\alpha/2, N-2)} \hat{\sigma}_{\hat{\beta}_i} \right]$$

Si $\bar{\beta}_i \in IC(\beta_i)$, on accepte H_0 au seuil α.

3) La sortie du logiciel `R` (comme toute sortie) donne la $p-$value. Si cette valeur est inférieur à α, on rejette l'hypothèse nulle en faveur de l'hypothèse alternative.

Intervalle de confiance de σ^2

On peut construire un intervalle de confiance pour σ^2 en utilisant la distribution suivante:

$$(N-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(N-2)$$

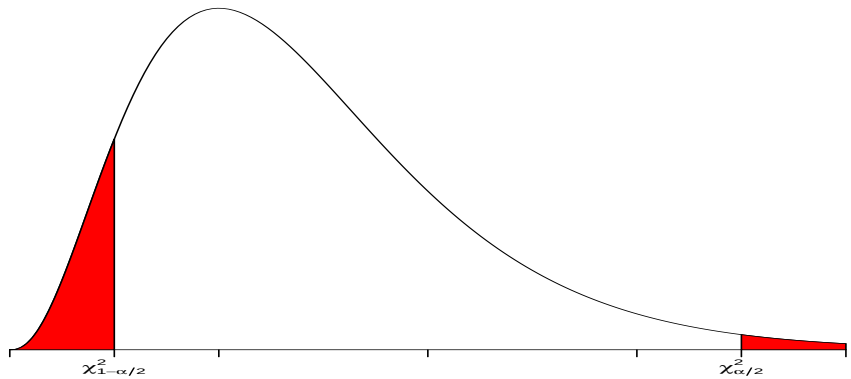

Figure : Distribution de Chi-deux

L'intervalle de confiance de σ^2 est donné par:

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

$\chi^2_{\alpha/2}$ et $\chi^2_{1-\alpha/2}$ sont les deux valeurs critiques tirées à partir d'une table de la loi de χ^2 de degrés de liberté $N-2$.

Exemple 2:

On reprend l'exemple précédent. Le modèle estimé est:

$$y_i = \underset{(0.635)}{-0.10} + \underset{(0.191)}{1.90} x_i + e_i, \quad i = 1, \ldots, 5.$$

Les valeurs entre parenthèses désigneent les écart-types des paramètres estimés.

Tester la significativité individuelle des paramètres du modèle au seui $\alpha = 5\%$.

Exemple 2:

On reprend l'exemple précédent. Le modèle estimé est:

$$y_i = \underset{(0.635)}{-0.10} + \underset{(0.191)}{1.90} x_i + e_i, \quad i = 1, \ldots, 5.$$

Les valeurs entre parenthèses désigneent les écart-types des paramètres estimés.

Tester la significativité individuelle des paramètres du modèle au seui $\alpha = 5\%$.

Sous H_0, la valeur de la statistique est $|\hat{t}_0| = 0.1/0.635$. D'après la table statistique, $t_{0.025,3} = 3.182$. On a $\hat{t}_0 < t_{0.025,3}$, ce qui nous permet de rejeter H_0. Le paramètre β_0 est statistiquement significatif au seuil $\alpha = 5\%$.

De même, on a $\hat{t}_0 = 1.9/0.191 = 9.947 > t_{0.025,3}$. Donc on accepte H_0. Le paramètre $\hat{\beta}_1$ est statistiquement non significatif au seuil $\alpha = 5\%$.

Avec R

```
coef(summary(reg1))

##             Estimate Std. Error    t value    Pr(>|t|)
## (Intercept)     -0.1  0.6350853 -0.1574592 0.884883980
## xi               1.9  0.1914854  9.9224264 0.002177513

qt(0.025,3)

## [1] -3.182446
```

Tableau d'analyse de la variance

Il s'agit de tester:

$$\begin{cases} H_0 : & \beta_i = 0,\ \forall i = 0, 1 \\ H_1 : & \exists \text{ au moins un } \beta_i \neq 0 \end{cases}$$

Pour cela, on fait recours à l'équation de la décomposition de la variance. À partir de cette équation, on construit le tableau d'analyse de variance (ANOVA) qui permet de tester la significativité globale du modèle (la linéarité du modèle). Ce tableau est définit comme suit:

Source de variation	Somme des carrés	ddl	Carrés moyens
Régression	$SCE = \hat{\beta}_1 S_{xy} = \hat{\beta}_1^2 S_{xx}$	1	$MCE = SCE/1$
Résidu	$SCR = S_{yy} - SCE$	N-2	$MCR = SCR/(N-2)$
Totale	$SCT = S_{yy}$	N-1	

Le théorème de Cochrane montre que la statistique F est distribuée selon la loi de Ficher de degrés de liberté 1 et $N-2$. Elle s'exprime selon la relation suivante:

$$F = \frac{MCE}{MCR} = \frac{SCE/1}{SCR/(N-2)} \sim F(1, N-2)$$

Test de significativité globale

La règle de décision dépend de la position de la statistique estimée $\hat{F}$ à partir de l'échantillon, par rapport à la valeur critique tabulée $F^{\alpha}_{1,N-2}$.
Si $\hat{F} \leq F^{\alpha}_{1,N-2}$, on accepte l'hypothèse nulle selon laquelle la régression est globalement non significative au seuil α. Autrement dit, la présence de la variable explicative x n'apporte aucune contribution significative à l'explication du comportement de la variable dépendante y.
On peut également effectuer ce dernier test en utilisant le coefficient de détermination R^2. La statistique du test équivalente est donnée par:

$$F = (N-2)\frac{R^2}{1-R^2} \sim F(1, N-2)$$

Remarque:

On peut vérifier que $\hat{t}_1^2 = \hat{F} = (N-2)\frac{R^2}{1-R^2}$

Prévision

Cette étape consiste à prévoir le comportement de la variable dépendante y sur la base de valeurs hypothétiques, notées x_θ attribuées à la variable explicative. Le prédicteur linéaire est donc donné par $\hat{y}_\theta = \hat{\beta}_0 + \hat{\beta}_1 x_\theta$.

Proposition

Dans le modèle de régression linéaire, le prédicteur linéaire a une erreur de prévision $\hat{Y}_\theta - Y_\theta$ telle que :

$$E(\hat{Y}_\theta - Y_\theta) = 0$$

$$V(\hat{Y}_\theta - Y_\theta) = \sigma^2 \left[1 + \frac{1}{N} + \frac{(X_\theta - \bar{X})^2}{S_{xx}} = \hat{\sigma}^2_{prev}\right]$$

Dans le modèle normal, nous avons le résultat

$$\frac{\hat{Y}_\theta - Y_\theta}{\sqrt{V(\hat{Y}_\theta - Y_\theta)}} \sim N(0,1)$$

Prévision

En remplaçant la variance de l'erreur de prévision par son estimateur, nous dérivons

$$\frac{\hat{Y}_\theta - Y_\theta}{\hat{\sigma}_{prev}} \sim t_{(N-2)}$$

Un intervalle de confiance pour Y_θ est fourni comme suit:

$$IC(Y_\theta) = \left[\hat{Y}_\theta \pm \hat{\sigma}_{prev}\, t_{(\alpha/2,N-2)}\right]$$