Rappel : Intervalles de confiance & tests d'hypothèse

Qu'est-ce qu'une régression ?

- Etude de la dépendance d'une variable par rapport à une ou plusieurs autre(s) variables.
- Objectif = estimer ou prédire la valeur moyenne (population) à partir de valeurs connues ou fixes prises par des variables explicatives.

$\Longrightarrow$

Objectif = estimer la fonction de régression population à partir de la fonction de regression échantillon.

FRP $\equiv$ $\boxed{Y_i = \beta_1 + \beta_2 X_i + u_i}$

avec $E(Y_i/X_i) = \beta_1 + \beta_2 X_i$

FRE $\equiv$ $\boxed{Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i}$

avec $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

Rem : $\hat{Y}_i$ estimateur de $E(Y/X_i)$.

$\hat{\beta}_1$ estimateur de β_1.

$\hat{\beta}_2$ estimateur de β_2.

Les moindres carrés ordinaires (MCO)

On choisit la fonction de régression échantillon qui minimise la somme des carrés des résidus :

$$\underset{\hat{\beta}_1, \hat{\beta}_2}{\text{Min}} \sum \hat{u}_i^2 \equiv \underset{\hat{\beta}_1, \hat{\beta}_2}{\text{Min}} \sum (Y_i - \hat{Y}_i)^2$$

$$\equiv \underset{\hat{\beta}_1, \hat{\beta}_2}{\text{Min}} \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

$$\boxed{\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}} \qquad \boxed{\hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}}$$

où $\bar{X}$ et $\bar{Y}$ sont les moyennes échantillon de X et Y.

$x_i = (X_i - \bar{X})$

$y_i = (Y_i - \bar{Y})$

Les valeurs prises par les estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ dépendent des données de l'échantillon. Comme les données varient d'un échantillon à l'autre, les valeurs prises par les estimateurs varient d'un échantillon à l'autre.

> Comment juge-t-on de la précision d'une estimation ?
> Comment peut-on savoir si la valeur de l'estimation est proche de celle du paramètre population ?

La précision d'un estimateur se mesure par son "erreur standard" ("standard error"), càd par l'écart-type de la distribution d'échantillonage de l'estimateur.

- Sous les hypothèses du modèle de régression linéaire classique :

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \Rightarrow \boxed{\text{s.e.}(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}}}$$

$$\text{var}(\hat{\beta}_1) = \frac{\sum X_i^2}{n \sum x_i^2} \cdot \sigma^2 \Rightarrow \boxed{\text{s.e.}(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n \sum x_i^2} \cdot \sigma^2}}$$

où
- var = variance, $x_i = X_i - \bar{X}$, $y_i = Y_i - \bar{Y}$
- s.e = standard error
- σ^2 = variance homoscédastique du terme d'erreur u_i ($= \text{var}(u_i / X_i)$)

L'estimateur des MCO de σ^2 (paramètre pop. inconnu) est donné par :

$$\boxed{\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2}}$$

où
- $(n-2)$ = nbre de degrés de liberté (df)
- $\sum \hat{u}_i^2$ = somme des carrés des résidus.

df = nbre d'obs − nbre de paramètres à estimer (nbre de paramètres sur lequel l'indicateur est basé).

- Pour faire de l'inférence statistique, il est nécessaire de spécifier la distribution de probabilité du terme d'erreur (car estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ sont des fction linéaire de Y ... qui dépend linéairement de u_i)

Autrement dit, pour faire de l'inférence statistique, il faut qu'on connaisse la distribution d'échantillonage exacte des estimateurs.

- Si terme d'erreur suit distribution normale (+ les hyp. du modèle de régression linéaire classique) :

a) $\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$

b) $\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2})$

(car fct° linéaire d'une v.a. suivant loi normale = v.a. suivant loi normale)

c) $(n-2)(\hat{\sigma}^2/\sigma^2)$ suit une distr. de χ^2 à $(n-2)$ degrés de liberté

ou encore :

a) $\dfrac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim N(0,1)$

b) $\dfrac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} \sim N(0,1)$

c) $\hat{\sigma}^2$ suit distrib. de χ^2

$\Rightarrow$ Les propriétés peuvent être utilisées pour juger de la précision d'une estimation, calculer des intervalles de confiance et faire des tests d'hyp.

3.1. Les intervalles de confiance

La précision d'une estimation peut (aussi) être déterminée en calculant un intervalle de confiance pour le paramètre inconnu.

- Supposons que nous soyons intéressés par la proximité de $\hat{\beta}_2$ et β_2. Pour cela on pourrait essayer de trouver 2 nombres α et δ, compris de 0 et 1, tq la probabilité que l'intervalle aléatoire $(\hat{\beta}_2 - \delta, \hat{\beta}_2 + \delta)$ contienne la vraie valeur du paramètre β_2 soit égale à $1-\alpha$ (ou $(1-\alpha)\%$).

  $\Rightarrow$ Trouver 2 nombres α et $\delta \in (0,1)$ tq :

$$\boxed{Pr\left(\hat{\beta}_2 - \delta \leq \beta_2 \leq \hat{\beta}_2 + \delta\right) = 1-\alpha} \quad (1)$$

  $\equiv$ Intervalle de confiance au niveau de probabilité $(1-\alpha)$

- <u>Caractéristiques</u> ?

a) Il s'agit d'un intervalle « aléatoire ».

b) La probabilité qui y est associée doit être considérée dans le contexte d'une distribution d'échantillonnage*.

c) Pour un échantillon particulier, l'intervalle de confiance est fixe : la probabilité que l'intervalle recouvre la vraie valeur du paramètre population vaut 0 ou 1.

* L'expression (1) indique que si on procède à des tirages répétés d'échantillons et que pour chacun d'entre eux on construit un intervalle de confiance au niveau de prob. $1-\alpha$, en moyenne $(1-\alpha)\%$ de ces intervalles contiendront la vraie valeur du paramètre population.

- Comment construit-on un intervalle de confiance pour β_2 ?

$$Z = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} = \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\sigma} \sim N(0,1)$$

σ généralement inconnu.

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n-2}}$$

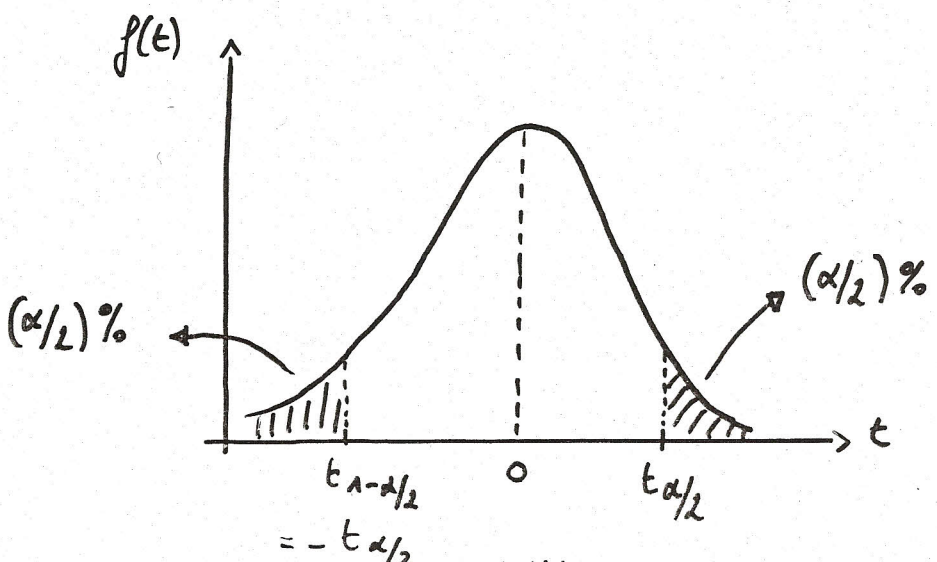
$$\Rightarrow \boxed{t = \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\hat{\sigma}}} \quad (2)$$

Cette statistique t suit une loi de Student à $(n-2)$ degrés de liberté.

Si α est un niveau de probabilité (compris tel 0 et 1) et t est une variable aléatoire suivant une loi de Student, on trouve que :

$$\boxed{Pr\left(t_{1-\alpha/2} \leq t \leq t_{\alpha/2}\right) = 1-\alpha} \quad (3)$$

où $t_{\alpha/2}$ et $1-\alpha/2$ sont les valeurs critiques de la distribution de Student.



Comme la distribution de Student est symétrique autour de sa moyenne ($=0$) :

$$\boxed{Pr\left(-t_{\alpha/2} \leq t \leq t_{\alpha/2}\right) = 1-\alpha} \quad (4)$$

En substituant la valeur de la statistique t dans l'expression (4) :

$$\boxed{Pr\left(-t_{\alpha/2} \leq \frac{(\hat{\beta}_2 - \beta_2)}{se(\hat{\beta}_2)} \leq t_{\alpha/2}\right) = 1-\alpha} \quad (5)$$

$\Rightarrow$

$$\boxed{Pr\left(\hat{\beta}_2 - t_{\alpha/2} \cdot se(\hat{\beta}_2) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} \cdot se(\hat{\beta}_2)\right) = 1-\alpha} \quad (6)$$

$\Rightarrow$ $\boxed{\left[\hat{\beta}_2 \pm t_{\alpha/2} \cdot se(\hat{\beta}_2)\right]}$ recouvre la vraie valeur (7)
du paramètre β_2 avec une probabilité égale à $1-\alpha$ (ou $(1-\alpha)\%$)

- <u>Comment construit-on un IC pour β_1 ?</u>

Même raisonnement :

$\Rightarrow$ $\boxed{\left[\hat{\beta}_1 \pm t_{\alpha/2} \cdot se(\hat{\beta}_1)\right]}$ recouvre vraie valeur du (8)
paramètre β_1 avec prob. $(1-\alpha)$

- <u>Remarque</u> :
La largeur de l'IC de β_1 (et β_2) est proportionnelle à l'erreur standard de l'estimateur. Lorsque l'erreur standard $\uparrow$, l'incertitude // valeur du paramètre pop. $\uparrow$.

Table 3.2. Dépenses de cons. et revenus en $ de 10 ménages

Dép. de cons. Y	Revenus X
70	80
65	100
90	120
95	140
110	160
115	180
120	200
140	220
155	240
150	260

- Résultats par MCO :

$\hat{\beta}_1 = 24,4545$ $\quad \hat{\beta}_2 = 0,5091$

$se(\hat{\beta}_1) = 6,4138$ $\quad se(\hat{\beta}_2) = 0,0357$

$\hat{\sigma}^2 = 42,1591$ $\quad cov(\hat{\beta}_1, \hat{\beta}_2) = -0,2172$

$R^2 = 0,9621$

- IC pour β_1 et β_2 ?

Valeur critique pour une distribution de Student à 8 df pour $\alpha = 0,05$:

$$t_{\alpha/2} = t_{0,025} = 2,306$$

IC pour β_1 : $[\hat{\beta}_1 \pm t_{\alpha/2} \cdot se(\hat{\beta}_1)] = 24,4545 \pm 2,306\,(6,4138$

$= [24,4545 \pm 14,7902]$

$\Rightarrow [9,6643 \leq \beta_1 \leq 39,2448]$

IC pour β_2 : $[\hat{\beta}_2 \pm t_{\alpha/2} \cdot se(\hat{\beta}_2)] = 0,5091 \pm 2,306 \cdot (0,0357$

$= [0,5091 \pm 0,0823]$

$\Rightarrow [0,4268 \leq \beta_2 \leq 0,5914]$

- Comment construit-on un IC pour σ^2 ?

Sous l'hypothèse de normalité du terme d'erreur, la statistique :

$$\boxed{\chi^2 = (n-2)\left(\frac{\hat{\sigma}^2}{\sigma^2}\right)} \qquad (1)$$

suit une distribution de Chi-carré à $(n-2)$ degrés de liberté.

$$\boxed{Pr\left(\chi^2_{1-\alpha/2} \leq \chi^2 \leq \chi^2_{\alpha/2}\right) = 1-\alpha} \qquad (2)$$

où $\chi^2_{1-\alpha/2}$ et $\chi^2_{\alpha/2}$ sont les valeurs critiques de la distribution de Chi-carré.

En substituant la valeur de la statistique χ^2 dans l'expression (2) :

$$Pr\left(\chi^2_{1-\alpha/2} \leq (n-2)\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) \leq \chi^2_{\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow \boxed{Pr\left((n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha}$$

Exemple : 10 familles dont on connaît dép. de cons et revenu.

Résultats MCO : $\hat{\beta}_1 = 24,4545$, $\hat{\beta}_2 = 0,5091$

$se(\hat{\beta}_1) = 6,4138$, $se(\hat{\beta}_2) = 0,0357$

$\hat{\sigma}^2 = 42,1591$, $cov(\hat{\beta}_1, \hat{\beta}_2) = -0,2172$

$R^2 = 0,9621$

- Soit $\alpha = 0,05$ (coefficient de confiance de 95%).
Pour une chi-carré à 8 degrés de liberte ($8 = 10-2$)
on trouve les valeurs critiques suivantes : $\begin{cases} \chi^2_{0,025} = 17,5346 \\ \chi^2_{0,975} = 2,1797 \end{cases}$

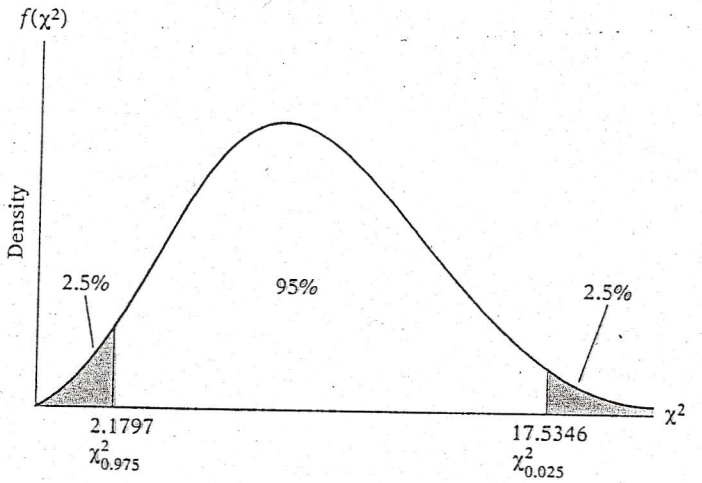


FIGURE 5.1 The 95% confidence interval for χ^2 (8 df).

- IC à 95% pour σ^2 : $\left[(n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} \; ; \; (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}} \right] = 0,95$

$$\Rightarrow \left[8.\left(\frac{42,1591}{17,5346}\right) \; ; \; 8.\left(\frac{42,1591}{2,1797}\right) \right]$$

$$\Rightarrow [19,2347 \; ; \; 154,7336]$$

$$\Rightarrow [19,2347 \leq \sigma^2 \leq 154,7336] = 0,95$$

La probabilité que la vraie valeur du paramètre population σ^2 sous comprise dans cet IC est de 95%.

3.2. Tests d'hypothèses

En pratique, on désire vérifier si un résultat est "compatible" avec une valeur (choisie a priori) pour un paramètre population inconnu.

"Compatible" signifie que le résultat est suffisamment proche de la valeur hypoth. pop, pour que l'hyp. concernant la valeur du paramètre pop. ne soit pas rejetée.

H_0 : hypothèse nulle
H_1 : hypothèse alternative

On appelle l'hypothèse nulle, la proposition qui contredit la proposition de recherche. L'hypothèse nulle doit donc être énoncée de telle manière que son rejet entraîne l'acceptation de l'hypothèse alternative, càd. la proposition de recherche.

Pour décider si une hyp. nulle peut être rejetée ou non, on utilise les tests de significativité.

Deux éléments centraux : a) la statistique de test,
b) la distribution de cette stat. de test sous H_0.

Règle de décision basée sur valeur de la stat. de test (calculée à partir de l'échantillon).

3.2.1. Tests relatifs aux coeff. de régression

- Sous l'hyp. de normalité, $t = \dfrac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} = \dfrac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum x_i^2}}{\hat{\sigma}}$

suit une loi de Student à $(n-2)$ degrés de liberté.

- $Pr\left(-t_{\alpha/2} \leq \dfrac{\hat{\beta}_2 - \beta_2^*}{se(\hat{\beta}_2)} \leq t_{\alpha/2} \right) = 1-\alpha$

où $\begin{cases} \beta_2^* = \text{valeur de } \beta_2 \text{ sous } H_0 \\ -t_{\alpha/2} \text{ et } t_{\alpha/2} = \text{valeurs crit. de Student à } (n-2) \text{ df.} \end{cases}$

$\Rightarrow$ $\boxed{Pr\left(\beta_2^* - t_{\alpha/2} \cdot se(\hat{\beta}_2) \leq \hat{\beta}_2 \leq \beta_2^* + t_{\alpha/2}\, se(\hat{\beta}_2) \right) = 1-\alpha}$

Intervalle dans lequel l'estimation $\hat{\beta}_2$ se situe avec une probabilité de $(1-\alpha)\%$ sous l'hypothèse nulle, càd. lorsque $\beta_2 = \beta_2^*$.

$\equiv$ „Zone d'acceptation" de l'hyp. nulle

Zone qui se situe en dehors de cet inverval = „Région critique"

- <u>Règle de décision pour α fixé :</u>
  - a) si l'estimation $\hat{\beta}_2$ se situe ds la zone d'acceptation, l'hyp. nulle ne peut pas être rejetée au niveau de prob. α.
  - b) si l'estimation $\hat{\beta}_2$ se situe dans la région critique, on rejette l'hyp. nulle au niveau de probabilité α.

Exemple :

Revenus - dépenses de cons. de ménages.

On a que : $\hat{\beta}_2 = 0,5091$, $se(\hat{\beta}_2) = 0,0357$, $df = 8$

Soit $\alpha = 0,05 \longrightarrow t_{\alpha/2} = 2,306$

Soit $\begin{cases} H_0 : \beta_2 = 0,3 \\ H_1 : \beta_2 \neq 0,3 \end{cases}$

Sous l'hyp. nulle, l'intervalle de confiance :

$Pr\ (0,3 - 2,306 \cdot 0,0357 \leq \hat{\beta}_2 \leq 0,3 + 2,306 \cdot 0,0357) = 0,95$

$Pr\ (0,2177 \leq \hat{\beta}_2 \leq 0,3823) = 0,95$

$\hat{\beta}_2 = 0,5091 \longrightarrow$ l'estimation se situe dans la région critique $\longrightarrow$ on rejette H_0

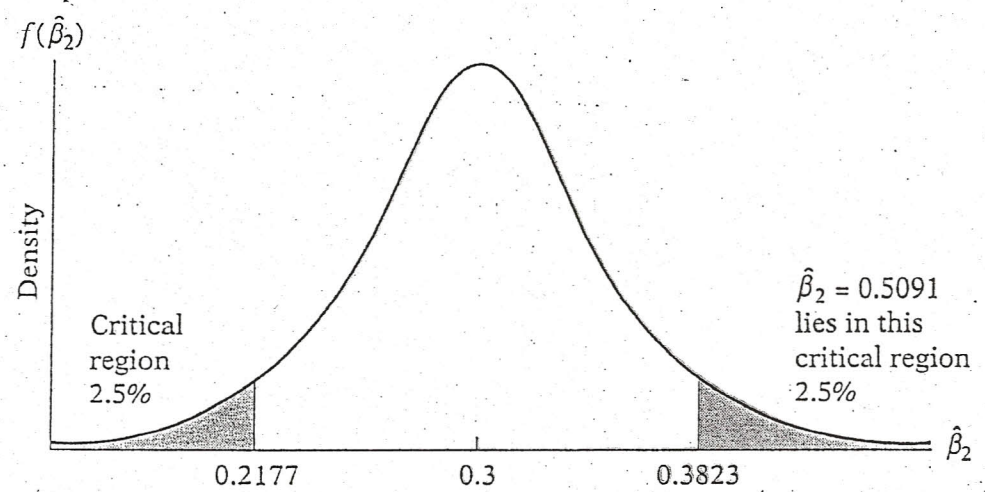


FIGURE 5.3 The 95% confidence interval for $\hat{\beta}_2$ under the hypothesis that $\beta_2 = 0.3$.

- En pratique, il n'est pas nécessaire de calculer explicitmt l'IC sous l'hyp. nulle.

Plus simple : calculer $t = \dfrac{\hat{\beta}_2 - \beta_2^*}{se(\hat{\beta}_2)}$ (où $\beta_2^* = \beta_2$ sous H_0)

et vérifier si elle est comprise ds $-t_{\alpha/2}$ et $t_{\alpha/2}$.

Dans notre exemple :

$t = \dfrac{(0,5091 - 0,3)}{0,0357} = 5,86$ et $t_{\alpha/2} = 2,306$ (pour $\alpha = 0,05$)

Statistique de test se situe dans région critique : RH0

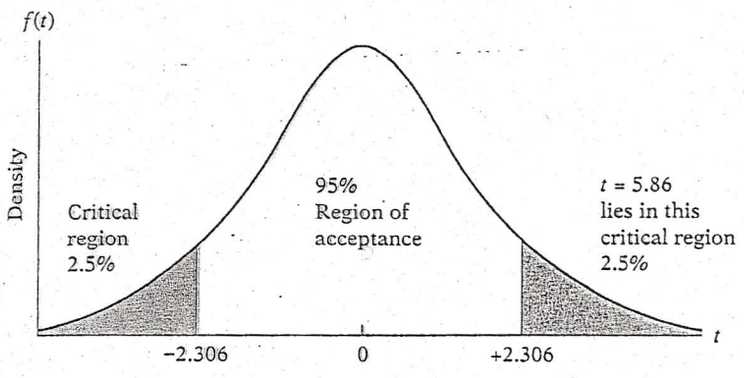


FIGURE 5.4 The 95% confidence interval for t(8 df).

- Remarque : ⓐ plus la statistique t est grande en valeur absolue, plus la probabilité de rejeter l'hypothèse nulle sera importante.

ⓑ Un test est significatif lorsque la valeur de la stat de test se situe dans la région critique. Dans ce cas l'hyp. nulle est rejetée.

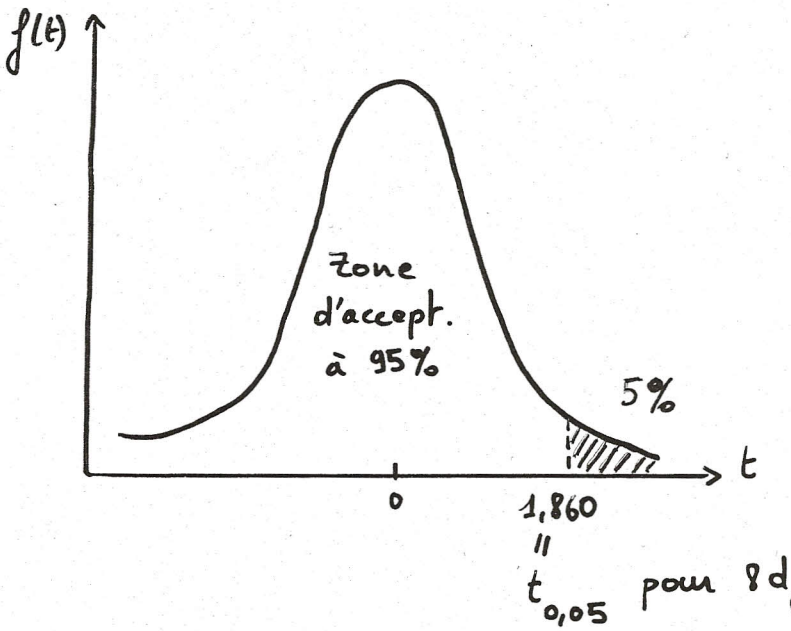
ⓒ test d'hypothèse „bilatéral" car $H_1 : \beta_2 \neq 0,3$
(si H0 est fausse, β_2 peut être > ou < 0,3)

Test unilatéral à droite

Exemple : $\beta_2 > 0,3$?

$\Rightarrow$ $H_0 : \beta_2 \leq 0,3$

$H_1 : \beta_2 > 0,3$



$t_{0,05}$ pour 8 df

Test unilatéral "à droite" : on utilise uniq. la valeur critique supérieure, càd t_α.

Règle de décision : si la t stat est inférieure à t_α on ne rejette pas H_0, sinon RH_0.

Dans notre exemple :

$$t = \frac{0,5091 - 0,3}{0,0357} = 5,86 > t_{0,05} = 1,860$$

$\Rightarrow$ on rejette H_0 à 5%.

Test unilatéral à gauche

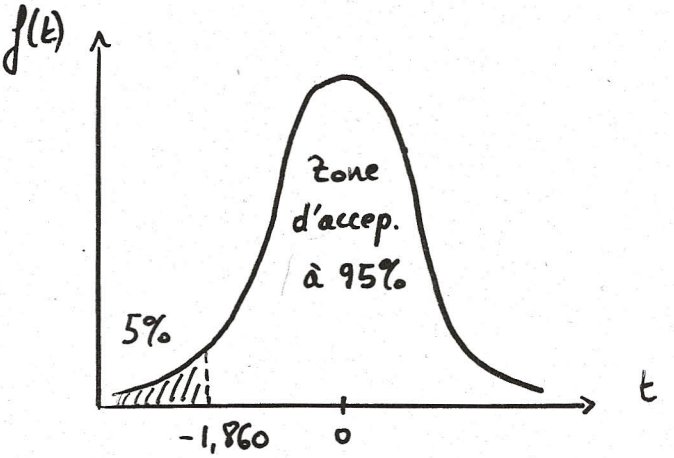
Exemple : $\beta_2 < 0,3$?

$H_0 : \beta_2 \geq 0,3$

$H_1 : \beta_2 < 0,3$

On utilise uniq. la valeur critique supérieure $-t_\alpha$.

Règle de décision : si la t stat est supérieure à la valeur critique on ne rejette pas H_0, sinon RH0



$-t_{0,05}$ pour $df = 8$

Dans notre cas :

$t = 5,86 > -1,860$

$\Rightarrow$ on ne rejette pas H_0.

En résumé :

Le test „t" de significativité

Type d'hyp.	H_0	H_1	Règle de décision : RH0 si
Bilatéral	$\beta_2 = \beta_2^*$	$\beta_2 \neq \beta_2^*$	$\lvert t \rvert > t_{\alpha/2, df}$
Unilatéral à droite	$\beta_2 \leq \beta_2^*$	$\beta_2 > \beta_2^*$	$t > t_{\alpha, df}$
Unilatéral à gauche	$\beta_2 \geq \beta_2^*$	$\beta_2 < \beta_2^*$	$t < -t_{\alpha, df}$

3.2.2. Tests relatifs à la variance des erreurs σ^2

- Sous l'hypothèse de normalité du terme d'erreur, la variable :

$$\chi^2 = (n-2)\left(\frac{\hat{\sigma}^2}{\sigma^2}\right)$$

suit une distribution de chi-carré à $(n-2)$ degrés de liberté.

- $Pr\left(\chi^2_{1-\alpha/2} \leq \chi^2 \leq \chi^2_{\alpha/2}\right) = 1-\alpha$

$$\Rightarrow \boxed{Pr\left(\chi^2_{1-\alpha/2} \leq (n-2)\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) \leq \chi^2_{\alpha/2}\right) = 1-\alpha}$$

- Exemple : revenus - dép. de cons. des ménages

$\hat{\sigma}^2 = 42,1591$, $df = 8$

Soit $\alpha = 0,05 \longrightarrow \begin{cases} \chi^2_{1-\alpha/2} = \chi^2_{0,975} = 2,1797 \\ \chi^2_{\alpha/2} = \chi^2_{0,025} = 17,5346 \end{cases}$

Soit $\begin{cases} H_0 : \sigma^2 = 85 \\ H_1 : \sigma^2 \neq 85 \end{cases}$

- Règle de décision : si la stat. de test $\left(\chi^2 = (n-2)\frac{\hat{\sigma}^2}{\sigma^{2*}}\right)$ sous l'hyp. nulle (càd pour $\sigma^2 = 85$) est comprise tel les 2 valeurs critiques de la chi-carré, on ne rejette pas H_0, sinon on RH$_0$.

- Dans notre exemple :

$$\chi^2 = 8.\left(\frac{42,1591}{85}\right) = 3,97 \quad > \chi^2_{0,975} = 2,1797$$

$$< \chi^2_{0,025} = 17,5346$$

$\Rightarrow$ on ne rejette pas H_0 (illustrat. test "bilatéral")

Test bilatéral car $H_1 : \sigma^2 \neq 85$ (on admet que si H_0 est fausse, σ^2 peut être supérieur ou inférieur à 85).

En résumé : Le test de χ^2

Type d'hyp.	H_0	H_1	Règle de décision : RH_0 si
Bilatéral	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$df.(\hat{\sigma}^2/\sigma_0^2) > \chi^2_{\alpha/2, df}$ ou $df.(\hat{\sigma}^2/\sigma_0^2) < \chi^2_{(1-\alpha/2), df}$
Unilatéral à droite	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$df(\hat{\sigma}^2/\sigma_0^2) > \chi^2_{\alpha, df}$
Unilatéral à gauche	$\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$df(\hat{\sigma}^2/\sigma_0^2) < \chi^2_{1-\alpha, df}$

où df = nbre de degrés de liberté

3.2.3. Tests d'hypothèses : quelques aspects pratiques

a) La signification d'accepter ou de rejeter une hypothèse

Si on ne rejette pas H_0, cela ne signifie pas que H_0 est vraie.

Pourquoi ?

Exemple revenus – dép. de cons. des ménages.

i) $H_0 : \beta_2 = 0,5$

$H_1 : \beta_2 \neq 0,5$

$\hat{\beta}_2 = 0,5091$, $se(\hat{\beta}_2) = 0,0357$, $df = 8$

$$t = \frac{0,5091 - 0,5}{0,0357} = 0,25$$

La t stat est non significative au niveau de prob $\alpha = 0,05$.
En effet : $|t| = 0,25 < t_{0,025} = 2,306$

$\Rightarrow$ on ne peut pas rejeter l'hyp. que $\beta_2 = 0,5$.

ii) $H_0 : \beta_2 = 0,48$
$H_1 : \beta_2 \neq 0,48$

$$t = \frac{0,5091 - 0,48}{0,0357} = 0,82 \quad (< 2,306)$$

La t stat est non significative.
$\Rightarrow$ on ne peut pas rejeter l'hyp. que $\beta_2 = 0,48$.

b) <u>L'hypothèse nulle „zéro"</u>

En pratique, on teste sur si un coeff. de régression est significativement $\neq$ de zéro.

Par exemple, $H_0 : \beta_2 = 0$.
Lorsque $df \geq 20$ et $\alpha = 0,05$, l'hyp. nulle $\beta_2 = 0$ peut être rejetée si $|t| > 2$ (où $t = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)}$)

c) <u>Choisir le degré de significativité α</u>

La décision de rejeter H_0 dépend de la valeur de α
α = risque de première espèce
= prob. rejeter H_0 alors qu'elle est vraie.
Lorsqu'on $\downarrow$ le risque première espèce, on $\uparrow$ le risque de seconde espèce (β), où
β = prob. d'accepter une hyp. fausse.

⇒ Il faut faire un arbitrage (el α et β.

En pratique, on fixe la valeur de α et on choisit une stat. de test qui minimise le risque de 2nd espèce (β).

$(1-\beta)$ = puissance du test
⇒ min β ≡ max $(1-\beta)$: on maximise la prob. de rejeter une hypothèse fausse.

Problématique concernant choix de α peut être évitée en utilisant la „p-value".

d) <u>Le degré de significativité exact : la „p-value"</u>

„p-value" = prob. exacte que la distrib. d'éch. de la stat. de test prenne une valeur sup. ou égale à la valeur de la stat. de test obtenue pour l'éch.
„p-value" = prob. exacte de commettre une erreur de première espèce.

Exemple : revenus - dép. de cons. des ménages.
Nous savons que $t = \frac{\hat{\beta}_2 - \beta_2}{se(\hat{\beta}_2)} = 5,86$ sous l'hypothèse nulle que $\beta_2 = \beta_2^* = 0,3$.

Pour $df = 8$ ⟶ p-value associée à 5,86 = 0,000189
⟶ prob. de RHo alors que H_0 est vraie est de +/- 2 pour 10000.

3.3. Analyse de la variance (ANOVA)

- L'hypothèse nulle „ zéro" peut être abordé comme un problème d'analyse de la variance.

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum \hat{u}_i^2 = \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{u}_i^2$$

$$\Longrightarrow \quad SCT = SCE + SCR$$

- Nbre de degrés de liberté ? (= nbre de valeurs qu'on peut choisir arbitrairement

i) SCT $\longrightarrow$ $(n-1)$ df

Pourquoi ?

On perd 1 df en calculant $\bar{Y}$

On peut choisir librement $(n-1)$ valeurs librement mais la n-ème est déterminée par la condition suivante : $\sum y_i = 0$

En effet, $\sum y_i = \sum (Y_i - \bar{Y}) = \sum Y_i - n\bar{Y}$

$$= n\left(\frac{1}{n}\sum Y_i\right) - n\bar{Y}$$

$$= n\bar{Y} - n\bar{Y}$$

$$= 0$$

ii) SCR $\longrightarrow$ $(n-2)$ df

Pourquoi ?

L'utilisation des MCO impose 2 cotions aux résidus $(\hat{u}_i)$:

$\sum \hat{u}_i = 0$ et $\sum \hat{u}_i X_i = 0$

iii) SCE $\longrightarrow$ 1 df

Pourquoi ?

$\sum x_i^2$ est connue, SCE dépend uniq. de $\hat{\beta}_2$

Table 5.3. Tableau ANOVA pour un modèle de régression à deux variables

Source de la variation	Somme des carrés [1]	Somme des carrés moyens [2]	df [3]
Due à la * régression (SCE)	$\sum \hat{y}_i^2 = \hat{\beta}_2^2 \sum x_i^2$	$(\hat{\beta}_2^2 \sum x_i^2)/1$	1
Due aux résidus (SCR)	$\sum \hat{u}_i^2$	$(\sum \hat{u}_i^2)/(n-2)$	$n-2$
Totale (SCT)	$\sum y_i^2$	$(\sum y_i^2)/(n-1)$	$n-1$

1 : sum of squares (SS)
2 : mean sum of squares (MSS) = SS / df
3 : degrees of freedom (df)
* : variation due à la variable X

$$F = \frac{SCE/1}{SCR/(n-2)} = \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum \hat{u}_i^2/(n-2)} = \frac{\hat{\beta}_2^2 \sum x_i^2}{\hat{\sigma}^2} \quad (1)$$

Si u_i suivent une loi normale, sous l'hyp nulle que $\beta_2 = 0$, F suit une loi de Fisher à 1 degré de liberté au numérateur et $(n-2)$ degrés de liberté au dénominateur.

A quoi cette variable F peut-elle servir ?

$$E(\hat{\beta}_2^2 \sum x_i^2) = \sigma^2 + \beta_2^2 \sum x_i^2 \quad (2)$$

$$E\left(\frac{\sum \hat{u}_i^2}{n-2}\right) = E(\hat{\sigma}^2) = \sigma^2 \quad (3)$$

> Si $\beta_2 = 0$, (2) & (3) fournissent une estimation identique de σ^2 ⟶ X n'a pas d'influence linéaire sur Y

> Si $\beta_2 \neq 0$, (2) & (3) sont ≠ ⟶ une partie de la variat° de Y sera attribuable à X.

$\Rightarrow$ la variable F permet de tester l'hyp nulle que $\beta_2 = 0$.

Exemple : Revenus et dep. de cons. des ménages.

Table 5.4. ANOVA pour exemple rev. – dép. cons.

Source de variation	Somme des carrés	df	Somme des carrés moyens
Due à la régression (SCE)	8552,73	1	8552,73
Due aux résidus (SCR)	337,27	8	42,159
Totale (SCT)	8890,00	9	

$$\Rightarrow \quad F = \frac{8552,73}{42,159} = 202,87$$

$$p\text{-value} = 0,0000001 \qquad \left(\alpha = \frac{1}{10.000.000}\right)$$

On rejette H_0 (on rejette l'hyp nulle que le revenu n'a pas d'influence significative sur les dép. de cons. des ménages).

Pour un modèle de rég. à 2 variables, test de Fisher est équivalent à un test en t. Pour régressions multiples, test de Fisher offre bcp d'ô possibilités.