Économétrie

Liliane Bonnal

Université de Poitiers

4 septembre 2016

Les chapitres : L3

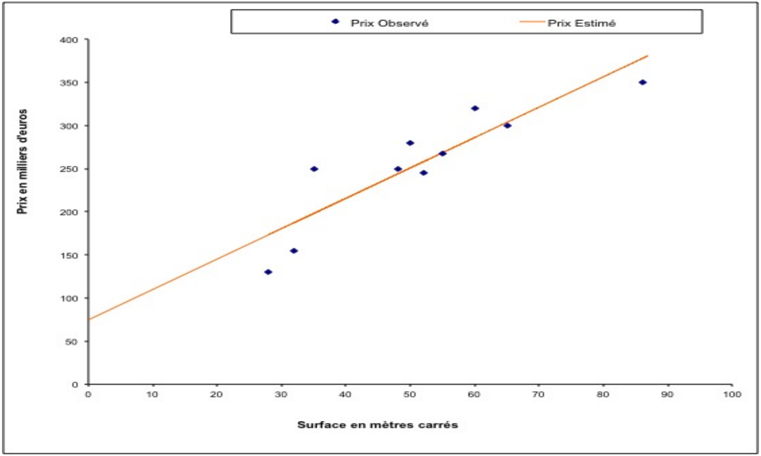
Chapitre 2 :

"le modèle de régression simple"

Plan

GPS

Relation entre le prix d'un logement et la surface du logement : "droite d'ajustement"



Commentaire du graphique

- Il existe un lien linéaire **positif** entre la surface des appartement et leur prix car le nuage de points est "réparti" autour d'une droite **croissante**.
- Faisons l'hypothèse que la relation entre le prix d'un logement et la surface du logement est linéaire.
  On peut alors écrire la relation :
  $$prix_i = \beta_0 + \beta_1 surface_i, \quad \forall\, i = 1, \cdots, n$$
  Avec n le nombre total de logements considérés dans l'échantillon.
- Bien que la droite résume bien le nuage de points, on peut quand même noter que les points ne sont pas tous alignés et l'écart entre la droite et les points est plus ou moins grand.

Nous allons appeler cet écart une "erreur". Par conséquent, cette équation est "vraie à une erreur de mesure près"

On va ajouter un terme d'erreur noté u_i à l'équation précédente.

On a donc

$$prix_i = \beta_0 + \beta_1 surface_i + u_i, \quad \forall\, i = 1, \cdots, n$$

GPS

GPS

Notions de base

Considérons, pour l'ensemble de la population, le modèle suivant :

$$Y = \beta_0 + \beta_1 X + U$$

Notions de base

Considérons, pour l'ensemble de la population, le modèle suivant :

$$Y = \beta_0 + \beta_1 X + U$$

Y : Variable dépendante
Variable à expliquer
Variable endogène
Variable du côté gauche.

Notions de base

Considérons, pour l'ensemble de la population, le modèle suivant :

$$Y = \beta_0 + \beta_1 X + U$$

Y : Variable dépendante
Variable à expliquer
Variable endogène
Variable du côté gauche.

X : Variable indépendante
Variable explicative, variable de contrôle
Variable exogène, covariable, régresseur
Variable du côté droit.

GPS

Hyp : La valeur moyenne de U est nulle $\Rightarrow \mathrm{E}(U) = 0$

Cette restriction n'est pas très forte car on peut toujours reparamétriser la constante pour normaliser $\mathrm{E}(U)$ à 0

Signification : U est une variable aléatoire telle que, en moyenne le modèle est bien spécifié.

Lien entre X et U : X n'a pas "d'influence" sur U (X et U ne sont pas corrélés)

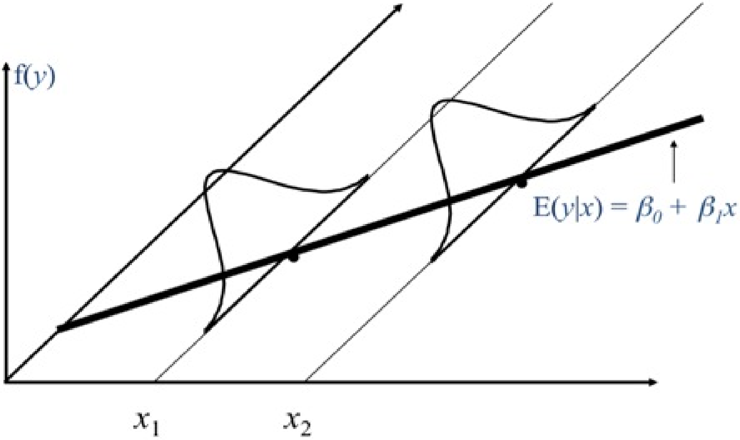
$$\mathrm{E}(U|X) = \mathrm{E}(U) = 0 \Rightarrow$$

$$\mathrm{E}(Y|X) = \beta_0 + \beta_1 X$$

Y est une VA dont la distribution est centrée sur $\mathrm{E}(Y|X)$, où l'espérance est une fonction linéaire de x, $\forall x$

Remarque : Pour simplifier nous allons supposer que X n'est pas aléatoire (exogène)

Espérance conditionnelle de Y, $\mathrm{E}(Y|X)$

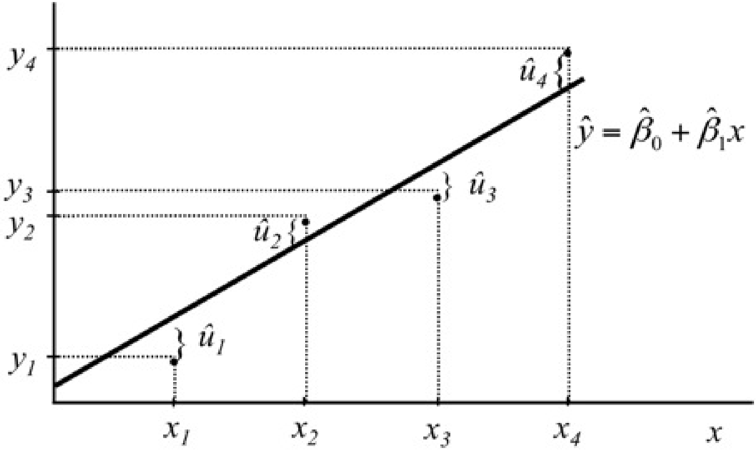


GPS

MCO ou OLS : Ordinary Least Squares

- Idée de base : Estimer les paramètres inconnus β_0 et β_1 associés à la population à partir d'un échantillon.
- Soit $\{(x_i, y_i); i = 1, \cdots, n\}$ un échantillon aléatoire de taille n.
- Pour chaque observation de cet échantillon on peut écrire la relation : $y_i = \beta_0 + \beta_1 x_i + u_i$
- Principe de la méthode : Déterminer les paramètres β tels que l'écart entre la valeur observée y_i et la valeur prédite (sur la droite d'ajustement ou de régression) soit la plus petite possible.
- Notons $\widehat{u_i}$ cet écart,
  MCO : Minimiser la somme des écarts au carré

Droite de régression, Nuage de points, Erreurs de mesure estimées



Estimation

- On va **estimer** les paramètres β_0 et β_1 qui minimisent

$$\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

  Cela revient à résoudre le programme mathématique suivant :

$$\underset{\beta_0,\beta_1}{\mathrm{Min}} \sum_{i=1}^{n} u_i^2 = \underset{\beta_0,\beta_1}{\mathrm{Min}}\ S(\beta_0, \beta_1)$$

  Mathématiquement :
  Condition de premier ordre (CIO) : on annule les dérivées (gradient)
  Condition de second ordre (C2O) : La matrice des dérivées secondes (matrice Hessienne) doit être semi-définie positive.

- On note $\widehat{\beta_0}$ et $\widehat{\beta_1}$ les solutions de ce programme. Ce sont des valeurs que l'on appellera les **estimations** des paramètres β_0, β_1.

Les estimateurs

La solution du problème de minimisation est donnée par :

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \widehat{X})(Y_i - \widehat{Y})}{\sum_{i=1}^{n}(X_i - \widehat{X})^2} = \frac{\mathrm{Cov}(X, Y)}{\mathrm{V}(X)}$$

$$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X}$$

$\widehat{\beta_0}$ et $\widehat{\beta_1}$ sont les **estimateurs** des MCO de β_0 et β_1 pour le modèle défini par $y_i = \beta_0 + \beta_1 x_i + u_i$

Les **estimations** sont les valeurs prises par $\widehat{\beta_0}$ et $\widehat{\beta_1}$, appelé aussi les coefficients (ou les paramètres) estimés. Elles dépendent donc des observations c'est-à-dire de l'échantillon considéré (cf. exercice 1, dossier 3 de TD)

Remarque : On estime d'abord $\widehat{\beta_1}$ car il est utilisé dans le calcul de $\widehat{\beta_0}$.

Caractéristiques des estimateurs

- $\widehat{\beta_1}$ est déterminé par la Cov(x,y), par conséquent le signe de $\widehat{\beta_1}$ dépend du signe de la covariance $\Rightarrow$
  - Si Cov$(x, y) > 0 \Rightarrow \widehat{\beta_1} > 0$
  - Si Cov$(x, y) < 0 \Rightarrow \widehat{\beta_1} < 0$
- $\widehat{\beta_1} \neq 0$ si V$(x) \neq 0$
- $\widehat{\beta_1}$ est une fonction linéaire des y
- Les estimateurs dépendent seulement des valeurs de x et de y, ils sont donc faciles à calculer.
- On dit que ces estimateurs sont **ponctuels** : la droite de régression passe par le point moyen$(\overline{x}, \overline{y})$.

Estimation et terminologie

- L'equation de la droite de régression notée encore la régression estimée, (obtenue par la régression de Y sur X) est définie par :

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i, \quad \forall \ i = 1, \cdots, n$$

Pour un x_i donné, $i = 1, \cdots, n, \widehat{y_i}$ est appelée :

  - la valeur ajustée,
  - la valeur prédite ou la prédiction,
  - la valeur estimée ou l'estimation,
  - la prévision.

- $\widehat{u_i}$ est appelé le **résidu** ou **l'erreur estimée**. Il est défini par

$$\widehat{u_i} = y_i - \widehat{y_i} = y_i - \widehat{\beta_0} - \widehat{\beta_1} x_i, \quad \forall \ i = 1, \cdots, n$$

Reprise de l'exemple, estimation

surface (x)	prix (en €, y)	xy	xx	prix estimé	erreur
28	20	560	784	26,63	-6,63
50	43	2150	2500	38,32	4,68
55	41	2255	3025	40,97	0,03
60	49	2940	3600	43,63	5,37
48	38	1824	2304	37,25	0,75
35	38	1330	1225	30,35	7,65
86	53	4558	7396	57,44	-4,44
65	46	2990	4225	46,28	-0,28
32	24	768	1024	28,75	-4,75
52	37	1924	2704	39,38	-2,38
somme					
511	389	21299	28787	389	2,8422E-14

xbar	51,1
ybar	38,9
num	1421,1
den	2674,9
cov(x,y)	142,11
var(x) (éch)	297,211111
var(x) (pop)	267,49

beta1	0,5312722
beta0	11,7519907
beta1 (ech)	0,47814498
beta1 (pop)	0,5312722

Reprise de l'exemple, estimation

CODE SAS : Création d'une base SAS

```
libname toto 'c :\travail\enseignement\poitiers\économétrieL3' ;
data toto.exemplecours ;
input surface prix ;
28 20
50 43
55 41
60 49
48 38
35 38
86 53
65 46
32 24
52 37
;
run ;
```

CODE SAS : Analyse exploratoire "minimale"

proc means data=toto.exemplecours n mean std var ;

run ;

Procédure MEANS

Variable	N	Moyenne	Ecart-type	Variance
surface	10	51.1000000	17.2398118	297.2111111
prix	10	38.9000000	10.3112668	106.3222222

CODE SAS : Régression linéaire

```
proc reg data=toto.exemplecours;
model prix = surface;
run;
quit;
```

Procédure REG
Modèle : MODEL1
Variable dépendante : prix

Nombre d'observations lues	10
Nombre d'observations utilisées	10

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	754.99092	754.99092	29.91	0.0006
Erreur	8	201.90908	25.23864		
Total sommes corrigées	9	956.90000			

Root MSE	5.02381	R carré	0.7890
Moyenne dépendante	38.90000	R car. ajust.	0.7626
Coeff Var	12.91467		

Valeurs estimées des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	11.75199	5.21168	2.25	0.0541
surface	1	0.53127	0.09714	5.47	0.0006

GPS

propriétés des MCO

- La somme des résidus est égale à 0 : $\sum_{i=1}^{n} \widehat{u}_i = 0$

  $\Rightarrow \overline{\widehat{u}} = 0$

- Etant donnée l'hypothèse faite sur les X, on a :
  $\sum_{i=1}^{n} x_i \widehat{u}_i = 0 \quad \Rightarrow \text{Cov}(X, \widehat{U}) = 0$

GPS

Mesure de l'ajustement

Pour chaque observation i de l'échantillon, une partie de y_i est expliquée grâce à x_i (variable observée) et une partie n'est pas expliquée, erreur u_i (non observée, aléatoire).

On sait que $\widehat{u_i} = y_i - \widehat{y_i} \Longleftrightarrow y_i = \widehat{y_i} + \widehat{u_i}$

Définissons les éléments suivants :

- $\sum_{i=1}^{n}(y_i - \overline{y})^2$: Somme Totale des Carrés (STC, SST, TSS)
- $\sum_{i=1}^{n}(\widehat{y_i} - \overline{\widehat{y}})^2 = \sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2$: Somme des Carrés Expliquée (SCE, SSE, ESS)
- $\sum_{i=1}^{n}(\widehat{u_i} - \overline{\widehat{u}})^2 = \sum_{i=1}^{n}(\widehat{u_i})^2$: Somme des Carrés des Résidus (SCR, SSR, RSS)

Mesure de la qualité de l'ajustement

On peut alors écrire la relation suivante :

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2 + \sum_{i=1}^{n}(\widehat{u_i})^2$$

$$\iff \text{STC} = \text{SCE} + \text{SCR}$$

Comment savoir si la droite d'ajustement reproduit correctement les données de l'échantillon ?
En calculant le **coefficient de détermination** noté **R carré** ou $\mathbf{R}^2$.

$$\mathrm{R}^2 = \frac{\text{SCE}}{\text{STC}} = 1 - \frac{\text{SCR}}{\text{STC}}$$

Propriétés du R^2

- $0 \leq \mathrm{R}^2 \leq 1$
- $\mathrm{R}^2 = 1 \Rightarrow$ ajustement parfait
- $\mathrm{R}^2 = 0 \Rightarrow$ absence de relation entre la variable dépendante Y et le régresseur X, $\Rightarrow \beta_1 = 0$.

Reprise de l'exemple, estimation

Procédure REG
Modèle : MODEL1
Variable dépendante : prix

Nombre d'observations lues	10
Nombre d'observations utilisées	10

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	754.99092	754.99092	29.91	0.0006
Erreur	8	201.90908	25.23864		
Total sommes corrigées	9	956.90000			

Root MSE	5.02381	R carré	0.7890
Moyenne dépendante	38.90000	R car. ajust.	0.7626
Coeff Var	12.91467		

Valeurs estimées des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	11.75199	5.21168	2.25	0.0541
surface	1	0.53127	0.09714	5.47	0.0006

GPS

GPS

Hypothèses complémentaires

Rappel :
Modèle pour la population : $Y = \beta_0 + \beta_1 X + U$

Généralement, on travaille avec un échantillon représentatif de taille n, caractérisé par des couples $\{(x_i, y_i); i = 1, \cdots, n\}$, pour lequel la relation précédente est définie par

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Hypothèses du modèle :

H1 : $\mathrm{E}(u_i) = \mathrm{E}(u_i | x_i) = 0$
H2 : Les x_i varient,
H3 : $\mathrm{Var}(u_i) = \sigma^2$
H4 : $\mathrm{Cov}(u_i, u_j) = 0, \quad \forall i, j; i \neq j$
H5 : $\mathrm{Cov}(u_i, x_i) = 0, \quad \forall i$
H6 : Les erreurs u_i sont iid et suivent une loi normale $\Rightarrow u_i \sim \mathrm{N}(0, \sigma^2), \forall i$

Conséquences de ces hypothèses :

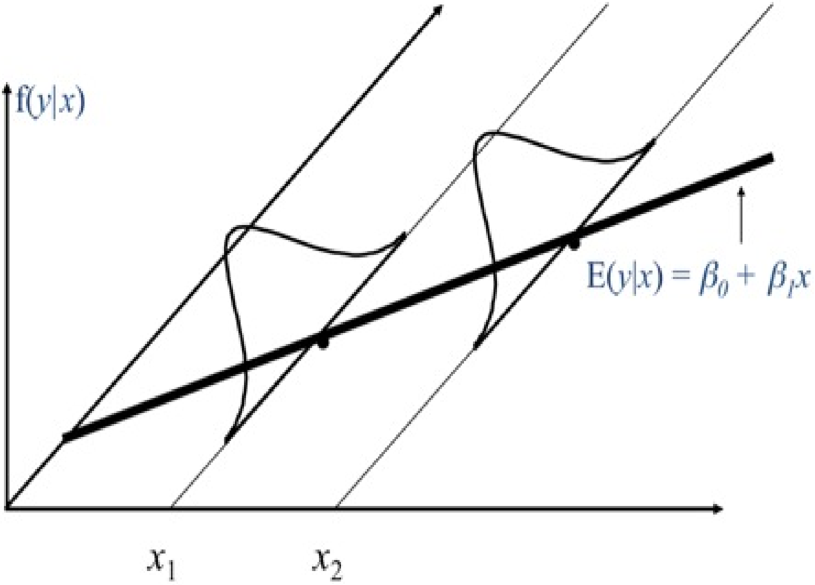
- les $y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2)$
- les estimateurs β sont aléatoires (car ils dépendent des y) et suivent aussi des distributions normales dont nous allons déterminer les moments.

Remarque :

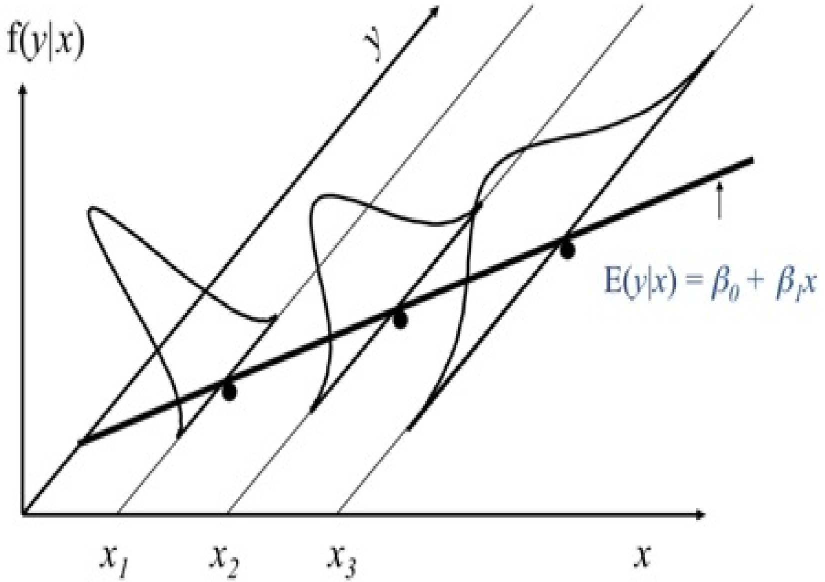
L'hypothèse H3 signifie que les erreurs ont une variance **homoscédastique** $\Rightarrow V(u_i) = \sigma^2, \forall i$.

Si H3 n'est pas vérifiée, les erreurs ont une variance **hétéroscédastique** $\Rightarrow V(u_i) = \sigma_i^2, \forall i$

Variance homoscédastique de y



Variance hétéroscédastique de y



GPS

1. Sans biais

Définition : Un estimateur est sans biais si son espérance est égale à sa vraie valeur (paramètre associé à la population) : $E(\widehat{\beta}) = \beta$

Sous les hypothèses H1 à H6, les estimateurs MCO sont sans biais ⇒

$$E(\widehat{\beta_0}) = \beta_0 \quad \text{et} \quad E(\widehat{\beta_1}) = \beta_1$$

Avant de montrer ces deux affirmations commençons par vérifier ces 3 égalités :

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \overline{x}) &= 0 \\
\sum_{i=1}^{n}(x_i - \overline{x})x_i &= \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x}) \\
\sum_{i=1}^{n}(x_i - \overline{x})y_i &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) \\
\sum_{i=1}^{n}(y_i - \overline{y})x_i &= \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})
\end{aligned}$$

estimateurs MCO : estimateurs sans biais

D'après les affirmations précédentes et la formule de l'estimateur MCO de β_1 on a :

$$\begin{aligned}
\widehat{\beta_1} &= \frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \\
&= \frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sum_{i=1}^{n}(x_i-\overline{x})y_i \\
&= \frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sum_{i=1}^{n}(x_i-\overline{x})(\beta_0+\beta_1 x_i+u_i) \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(x_i-\overline{x})u_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}
\end{aligned}$$

Calcul de $E(\beta_1)$ et $E(\beta_0)$

D'après les affirmations précédentes et la formule de l'estimateur MCO de β_1 on a :

$$\begin{aligned}
E\left(\widehat{\beta_1}\right) &= E\left(\beta_1 + \frac{\sum_{i=1}^{n}(x_i-\overline{x})u_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right) \\
&= \beta_1 + \frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sum_{i=1}^{n}(x_i-\overline{x})E(u_i) \\
&= \beta_1
\end{aligned}$$

$$\begin{aligned}
E\left(\widehat{\beta_0}\right) &= E\left(\overline{y}-\widehat{\beta_1}\overline{x}\right) \\
&= \beta_0
\end{aligned}$$

Variance des erreurs u_i et des y_i

- D'après l'hypothèse H3, la variance des erreurs est homoscédastique est égale à $V(u_i|x_i) = \sigma^2 \ \forall i$

  Mais, $V(u_i|x_i) = E(u_i^2|x_i) - [E(u_i|x_i)]^2$

  D'après l'hypothèse H1, $E(u_i|x_i) = 0$, on a donc :
  $V(u_i|x_i) = E(u_i^2|x_i) = \sigma^2$

- σ, l'écart-type des erreurs est inconnu

- $V(y_i|x_i) = V(\beta_0 + \beta_1 x_i + u_i|x_i) = \sigma^2$

Calcul de $V(\beta_1)$ et $V(\beta_0)$

D'après les affirmations précédentes et la formule de l'estimateur MCO de β_1 on a :

$$
\begin{aligned}
V\left(\widehat{\beta_1}\right) &= V\left(\beta_1 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})u_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right) = \left(\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^2 V\left(\sum_{i=1}^{n}(x_i - \overline{x})u_i\right) \\
&= \left(\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^2 \sum_{i=1}^{n}(x_i - \overline{x})^2 V(u_i) \\
&= \left(\frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)^2 \sigma^2 \sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}
\end{aligned}
$$

Calcul de $V(\beta_1)$ et $V(\beta_0)$

$$V\left(\widehat{\beta_1}\right) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$V\left(\widehat{\beta_0}\right) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right)$$

Quelques remarques sur ces variances

- Plus σ^2 est grande, plus les variances des estimateurs β_0 et β_1 seront grandes
- Plus les x_i auront de la variabilité (variance des X grande), plus les variances des estimateurs seront faibles
- Plus la taille de l'échantillon n est grande, plus les variance des estimateurs seront faibles
- Plus les variances des estimateurs sont faibles, plus les estimateurs seront précis
- Problème : σ^2 est inconnu $\Rightarrow$ les variances des estimateurs ne sont pas connues
  Pourquoi a-t-on besoin de connaitre les variances des estimateurs ?
  Pour tester la fiabilité des estimations (tests, IC ...)
  $\Rightarrow$ il faut estimer la variance des erreurs σ^2.

Estimation des variances des erreurs et des paramètres

- Les erreurs u_i ne sont pas connues mais nous connaissons les $\widehat{u_i}$. Nous allons utilisées les erreurs estimées pour calculer l'estimateur de la variance $\widehat{\sigma}^2$.
- Un estimateur sans biais de σ^2 est donné par :

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{u_i}^2 = \frac{1}{n-2}SCR$$

- On a donc :

$$\widehat{\mathrm{V}}\left(\widehat{\beta_1}\right) = \frac{\widehat{\sigma}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$$

$$\widehat{\mathrm{V}}\left(\widehat{\beta_0}\right) = \widehat{\sigma}^2\left(\frac{1}{n}+\frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)$$

Reprise de l'exemple, estimation

Procédure REG
Modèle : MODEL1
Variable dépendante : prix

Nombre d'observations lues	10
Nombre d'observations utilisées	10

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	754.99092	754.99092	29.91	0.0006
Erreur	8	201.90908	25.23864		
Total sommes corrigées	9	956.90000			

Root MSE	5.02381	R carré	0.7890
Moyenne dépendante	38.90000	R car. ajust.	0.7626
Coeff Var	12.91467		

Valeurs estimées des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	11.75199	5.21168	2.25	0.0541
surface	1	0.53127	0.09714	5.47	0.0006

Propriétés des estimateurs : Résumé

Sous les hypothèses H1 à H6 :

- Les estimateurs MCO sont sans biais.
- Les variances des estimateurs MCO sont minimales (la précision des estimateurs MCO est maximale) $\Rightarrow$ Ce sont les meilleurs estimateurs linéaires.

  Théorème de Gauss-Markov :

  - L'EMCO est le meilleur estimateur linéaire sans biais : BLUE (Best Linear Unbiased Estimator)
  - l'EMCO est l'estimateur de Gauss-Markov
  - l'EMCO est un estimateur efficace (il a la plus petite variance).

GPS

Loi des estimateurs

- Les erreurs u_i suivent une loi normale.
  Etant donnée la relation linéaire entre les y_i et les u_i
  $\Rightarrow$ les y_i suivent aussi une loi normale.
- L'estimateur $\widehat{\beta_1}$ est une fonction linéaire des y_i
  $\Rightarrow$ $\widehat{\beta_1}$ suit une loi normale tel que : $\widehat{\beta_1} \sim N\left(E(\widehat{\beta_1}), V(\widehat{\beta_1})\right)$
- L'estimateur $\widehat{\beta_0}$ est une fonction linéaire de $\widehat{\beta_1}$ et de $\overline{y}$
  $\Rightarrow$ $\widehat{\beta_0}$ suit une loi normale tel que : $\widehat{\beta_0} \sim N(E\left(\widehat{\beta_0}\right), V(\widehat{\beta_0})\big)$
- Etant donné que l'estimateur $\widehat{\beta_0}$ dépend de $\widehat{\beta_1}$
  $\Rightarrow \mathrm{Cov}(\widehat{\beta_0}, \widehat{\beta_1}) \neq 0$
- $\widehat{\beta_0}$ et $\widehat{\beta_1}$ suivent une loi normale bivariée.

Loi de $\widehat{\beta}$

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} \sim N\left(E(\widehat{\beta}), V(\widehat{\beta})\right)$$

Avec :

$$\begin{aligned}
E(\widehat{\beta}) &= \begin{bmatrix} E(\widehat{\beta}_0) \\ E(\widehat{\beta}_1) \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \\
V(\widehat{\beta}) &= \begin{bmatrix} V(\widehat{\beta}_0) & Cov(\widehat{\beta}_0, \widehat{\beta}_1) \\ Cov(\widehat{\beta}_0, \widehat{\beta}_1) & V(\widehat{\beta}_1) \end{bmatrix} \\
&= \sigma^2 \begin{bmatrix} \left(\dfrac{1}{n} + \dfrac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right) & \dfrac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \dfrac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} & \dfrac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{bmatrix}
\end{aligned}$$

GPS

IC de β_1

On sait que $\widehat{\beta_1} \sim N(\beta_1, \mathrm{V}(\widehat{\beta_1}) = \sigma^2_{\widehat{\beta_1}})$

Posons $Z_1 = \dfrac{\widehat{\beta_1} - \beta_1}{\sigma_{\widehat{\beta_1}}} = \left(\widehat{\beta_1} - \beta_1\right) \times \dfrac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sigma}$

$Z_1 \sim N(0, 1)$ si σ connu, or, σ inconnu mais nous connaissons son estimateur $\widehat{\sigma}^2 = \dfrac{1}{n-2} \sum_{i=1}^{n} \widehat{u_i}^2$.

Posons $Z_2 = (n-2)\dfrac{\widehat{\sigma}^2}{\sigma^2} = \dfrac{1}{\sigma^2}\sum_{i=1}^{n} \widehat{u_i}^2 \sim \chi^2_{n-2}$

IC de β_1

D'aprés les propriétés sur les lois (cf. chap 1), on sait que :

$$t = \frac{Z_1}{\sqrt{\frac{Z_2}{n-2}}} \sim t_{n-2}$$

Si on remplace Z_1 et Z_2 on obtient :

$$t = (\widehat{\beta_1} - \beta_1) \times \frac{1}{\sigma} \times \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} \times \frac{\sigma\ \sqrt{n-2}}{\sqrt{\sum_{i=1}^{n}\widehat{u_i}^2}} = \frac{\widehat{\beta_1} - \beta_1}{\widehat{\sigma}} \sim t_{n-2}$$

L'intervalle de confiance est défini par :

$$P\left[-t_{\alpha/2} \leq t \leq t_{\alpha/2}\right] = 1 - \alpha.$$

où $t_{\alpha/2}$ est la valeur de t (appelée valeur critique) obtenue à partir de la distribution d'un t-stat (t-student) pour niveau d'erreur $\frac{\alpha}{2}$ et $(n-2)$ ddl. On a donc

$$P\left[-t_{\alpha/2} \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} \leq t_{\alpha/2}\right] = 1 - \alpha.$$

$$\Leftrightarrow Pr\left[\widehat{\beta}_1 - t_{\alpha/2}\widehat{\sigma}_{\widehat{\beta}_1} \leq \beta_1 \leq \widehat{\beta}_1 + t_{\alpha/2}\widehat{\sigma}_{\widehat{\beta}_1}\right] = 1 - \alpha$$

Les intervalles de confiance de β_1 et β_0 sont définis par :

$$\mathrm{IC}_{\beta_1} = \left[\widehat{\beta}_1 \mp \ t_{\alpha/2} \ \widehat{\sigma}_{\widehat{\beta}_1}\right] \text{ et } \mathrm{IC}_{\beta_0} = \left[\widehat{\beta}_0 \mp \ t_{\alpha/2} \ \widehat{\sigma}_{\widehat{\beta}_0}\right]$$

IC : Commentaires

- On a $(1-\alpha)\%$ de chances que la vraie valeur du paramètre soit incluse dans l'IC de ce paramètre.

  Si $100(1-\alpha) = 95\%$ (5% d'erreur) cela signifie que dans 95% des cas, l'IC de β_1 par exemple contiendra la vraie valeur de β_1.

- La largeur de l'IC d'un paramètre est proportionnelle à l'écart-type de l'estimateur associé à ce paramètre :
  Plus l'écart-type de l'estimateur est grand, plus la probabilité (la certitude) d'estimer la vraie valeur du paramètre est faible

  L'écart-type d'un estimateur est une mesure de la précision de l'estimateur.

IC de σ^2

On vient de voir que $(n-2)\frac{\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$.

Posons $C^2 = (n-2)\frac{\widehat{\sigma}^2}{\sigma^2}$.

Ecrivons la probabilité suivante : $\mathrm{P}\left[\chi^2_{1-\frac{\alpha}{2}} \leq C^2 \leq \chi^2_{\alpha/2}\right] = 1-\alpha$.

$\chi^2_{1-\frac{\alpha}{2}}$ et $\chi^2_{\alpha/2}$ sont les deux valeurs critiques du χ^2 (valeurs théoriques) obtenus dans la table à $(n-2)$ ddl (Attention, la loi du χ^2 n'est pas symétrique).

Par substitution, on a :

$$\mathrm{P}\left[(n-2)\frac{\widehat{\sigma}^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq (n-2)\frac{\widehat{\sigma}^2}{\chi^2_{1-\frac{\alpha}{2}}}\right] = 1-\alpha$$

Intervalle de confiance pour σ^2 à 100 $(1-\alpha)$%.

Reprise de l'exemple, calcul de l'IC sous SAS

CODE SAS : Régression linéaire avec IC des paramètres

```
proc reg data=toto.exemplecours ;
model prix = surface / clb ;
run ;
quit ;
```

Option : clb : Confidence Limit of Beta

Par défaut SAS calcule les IC des paramètres pour un $\alpha = 0.05$. Si l'on veut modifier la commande il faut rajouter après clb l'instruction **alpha=0.01** pour faire un IC à 99% par exemple.

Reprise de l'exemple, estimation et IC

Procédure REG
Modèle : MODEL1
Variable dépendante : prix

Nombre d'observations lues	10
Nombre d'observations utilisées	10

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	754.99092	754.99092	29.91	0.0006
Erreur	8	201.90908	25.23864		
Total sommes corrigées	9	956.90000			

Root MSE	5.02381	R carré	0.7890
Moyenne dépendante	38.90000	R car. ajust.	0.7626
Coeff Var	12.91467		

Valeurs estimées des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à 95 %	
Intercept	1	11.75199	5.21168	2.25	0.0541	-0.26615	23.77014
surface	1	0.53127	0.09714	5.47	0.0006	0.30728	0.75527

GPS

Tests sur les valeurs des paramètres

Idée des tests : Mesurer la compatibilité entre les résultats des estimations et des hypothèses sur les vraies valeurs des paramètres.

Ex. $H_0 : \beta_1 = 0,5$ (hypothèse simple, bilatéral)

$H_1 : \beta_1 \neq 0,5$ (hypothèse composite)

Deux méthodes de tests :

- Par l'intervalle de confiance
- Test de signification

Approche par IC

1. Calculer l'intervalle de confiance du paramètre
2. Poser le test

   Ex. $H_0 : \beta_1 = 0,5$ (hypothèse simple, bilatéral)

   $H_1 : \beta_1 \neq 0,5$ (hypothèse composite)

3. Conclusion

   - Si la valeur du paramètre fixée sous H0 $\in$ IC, on ne rejette pas H0
   - Si la valeur du paramètre fixée sous H0 $\notin$ IC, on rejette H0

Pour l'exemple,

$\mathrm{IC}_{\beta_1} = [0,30728; 0,75527]$

0,5 $\in \mathrm{IC}_{\beta_1} \Rightarrow$ On ne rejette pas H0.

Approche par un test de signification

Définition : Un test de signification est un procédé par lequel les résultats d'un échantillon sont utilisés pour vérifier si une hypothèse nulle est vraie ou fausse.

On sait que :

$$t = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} = (\widehat{\beta}_1 - \beta_1) \times \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\widehat{\sigma}} \sim t_{(n-2)}$$

1. Poser le test

   H0 : $\beta_1 = \beta_{10}$ (valeur de β_1 sous H0)

   H1 : $\beta_1 \neq \beta_{10}$ (test bilatéral)

2. Calculer la valeur empirique du test sous H0

$$t_{emp} = \frac{\widehat{\beta}_1 - \beta_{10}}{\widehat{\sigma}_{\widehat{\beta}_1}}$$

3. Chercher la valeur théorique de la statistique de Student dans la table
4. Conclure

Exemple de test

Considérons le test défini précédemment pour β_1

Solution 1 : Le calcul

1. valeur empirique du test : $t_{emp} = \dfrac{0,53127 - 0,5}{0,09714} = \dfrac{0,03127}{0,09714} = 0,3219$
2. Valeur théorique : $t_8 = 2,306$
3. Conclusion $t_{emp} = 0,3219 < t_{th} = 2,306 \Rightarrow$ on ne rejette pas H0.
   Même conclusion qu'avec l'intervalle de confiance.

Solution 2 : SAS

CODE SAS : Régression linéaire et test

```
proc reg data=toto.exemplecours ;
model prix = surface ;
test surface=0.5 ;
run ;
quit ;
```

Reprise de l'exemple, Estimation et Test

Procédure REG
Modèle : MODEL1
Variable dépendante : prix

Nombre d'observations lues	10
Nombre d'observations utilisées	10

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	754.99092	754.99092	29.91	0.0006
Erreur	8	201.90908	25.23864		
Total sommes corrigées	9	956.90000			

Root MSE	5.02381	R carré	0.7890
Moyenne dépendante	38.90000	R car. ajust.	0.7626
Coeff Var	12.91467		

Valeurs estimées des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à 95 %	
Intercept	1	11.75199	5.21168	2.25	0.0541	-0.26615	23.77014
surface	1	0.53127	0.09714	5.47	0.0006	0.30728	0.75527

Procédure REG
Modèle : MODEL1

Résultats du test 1 pour la variable dépendante prix

Source	DDL	Moyenne quadratique	Valeur F	Pr > F
Numérateur	1	2.61592	0.10	0.7557
Dénominateur	8	25.23864		

Exemple $H0 : \beta_1 = 1; H1 : \beta_1 \neq 1$

Test Bilatéral

1. valeur empirique de la statistique :
$$t_{emp} = \frac{0.53127 - 1}{0.09714} = -4.8253 \sim t_8 \text{ ddl } = 2.3060$$
2. Conclusion du test
   - D'après la table $|t_{emp} > t_8| \Leftrightarrow 4.8253 > 2.3060$ ou encore $-4.8253 < -2.3060 \Rightarrow$ On rejette H0
   - $1 \notin IC_{\beta_1} = [0.30728; 0.75527]$
   - avec SAS :

     CODE SAS : Test sur un paramètre

```
proc reg data=toto.ExempleCoursChap3 ;
model prix = surface nbpieces / i clb ;
test surface = 2 ;
run ;
quit ;
```

Reprise de l'exemple, H0 : $\beta_1 = 1$

Procédure REG
Modèle : MODEL1
Variable dépendante : prix

Nombre d'observations lues	10
Nombre d'observations utilisées	10

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	754.99092	754.99092	29.91	0.0006
Erreur	8	201.90908	25.23864		
Total sommes corrigées	9	956.90000			

Root MSE	5.02381	R carré	0.7890
Moyenne dépendante	38.90000	R car. ajust.	0.7626
Coeff Var	12.91467		

Valeurs estimées des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à 95 %	
Intercept	1	11.75199	5.21168	2.25	0.0541	-0.26615	23.77014
surface	1	0.53127	0.09714	5.47	0.0006	0.30728	0.75527

Procédure REG
Modèle : MODEL1

Résultats du test 1 pour la variable dépendante prix

Source	DDL	Moyenne quadratique	Valeur F	Pr > F
Numérateur	1	587.69092	23.29	0.0013
Dénominateur	8	25.23864		

Exemple $H0 : \beta_1 = -1; H1 : \beta_1 \neq -1$

Test Bilatéral

1. valeur empirique de la statistique :
   $$t_{emp} = \frac{0.53127 + 1}{0.09714} = 15.7635 \sim t_8 \text{ ddl } = 2.3060$$
2. Conclusion du test
   - D'après la table $|t_{emp} > t_8| \Leftrightarrow 15.7635 > 2.3060 \Rightarrow$ On rejette H0
   - $-1 \notin IC_{\beta_1} = [0.30728; 0.75527]$
   - avec SAS :

     CODE SAS : Test sur un paramètre

```
proc reg data=toto.ExempleCoursChap3 ;
model prix = surface nbpieces / i clb ;
test surface = 2 ;
run ;
quit ;
```

Reprise de l'exemple, H0 : $\beta_1 = -1$

Procédure REG
Modèle : MODEL1
Variable dépendante : prix

Nombre d'observations lues 10
Nombre d'observations utilisées 10

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	754.99092	754.99092	29.91	0.0006
Erreur	8	201.90908	25.23864		
Total sommes corrigées	9	956.90000			

Root MSE	5.02381	R carré	0.7890
Moyenne dépendante	38.90000	R car. ajust.	0.7626
Coeff Var	12.91467		

Valeurs estimées des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à 95 %	
Intercept	1	11.75199	5.21168	2.25	0.0541	-0.26615	23.77014
surface	1	0.53127	0.09714	5.47	0.0006	0.30728	0.75527

Procédure REG
Modèle : MODEL1

Résultats du test 1 pour la variable dépendante prix

Source	DDL	Moyenne quadratique	Valeur F	Pr > F
Numérateur	1	6272.09092	248.51	<.0001
Dénominateur	8	25.23864		

Test de significativité d'un paramètre

Avant de commenter les résultats d'une estimation, il faut **toujours** tester la **significativité** des paramètres.

1. Poser, pour chaque paramètre de la régression (β_0 et β_1) le test suivant :

   H0 : $\beta_j = 0$ pour $j = 0, 1$
   H1 : $\beta_j \neq 0$.

2. Calculer les valeurs empiriques des t pour les deux paramètres

$$t_{\beta_j} = \frac{\widehat{\beta}_j - 0}{\widehat{\sigma}_{\widehat{\beta}_1}}$$

3. On compare la valeur empirique t_{β_j} à la valeur théorique $t_{\alpha/2}$

Pour l'exemple : t_{β_0}=2,25 et t_{β_1}=7,47.
Conclusion des tests ?
On ne rejette pas H0 pour β_0 et on rejette H0 pour β_1

Commentaires

Important : Pour une régression, on ne commente que les paramètres qui sont significatifs (pour le α que l'on s'est fixé).

Interprétation de $\widehat{\beta_1}$

- Le coefficients est positif (et significatif) par conséquent on a une relation linéaire croissante entre la surface d'un appartement et le prix d'un appartement
  Si la surface de l'appartement $\nearrow$ ($\searrow$) $\Rightarrow$ le prix de l'appartement $\nearrow$ ($\searrow$)
- $\widehat{\beta_1} = \dfrac{\mathrm{d}y_i}{\mathrm{d}x_i}$: si la surface du logement augmente de 1 m^2, le prix du logement augmente en moyenne de 0,53127 (soit 531 €, (C_m) ; CM au M^2 : 671,3 €)

Interprétation de $\widehat{\beta_0}$ (significatif à seulement 10%)
Ici le paramètre constant n'est pas interprétable : prix d'un logement pour une surface égale à 0 !

Test Unilatéral

On sait que :

$$t = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} = (\widehat{\beta}_1 - \beta_1) \times \frac{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\widehat{\sigma}} \sim t_{(n-2)}$$

1. Poser le test

   H0 : $\beta_1 = \beta_{10}$ (valeur de β_1 sous H0)
   H1 : $\beta_1 > \beta_{10}$ ou $\beta_1 < \beta_{10}$ (test unilatéral)

2. Calculer la valeur empirique du test sous H0

$$t_{emp} = \frac{\widehat{\beta}_1 - \beta_{10}}{\widehat{\sigma}_{\widehat{\beta}_1}}$$

3. Chercher la valeur théorique de la statistique de Student dans la table

   Lorsque l'on lit une table il faut vérifier si la table est bilatérale ou unilatérale.

4. Conclure

Exemple $H0 : \beta_1 = 1; H1 : \beta_1 < 1$

Test Unilatéral

1. valeur empirique de la statistique :
   $$t_{emp} = \frac{0.53127 - 1}{0.09714} = -4.8253 \sim t_8 \text{ ddl } = -1.8595 \text{ (borne inférieure)}$$
2. Conclusion du test
   - D'après la table $|t_{emp}| > |t_8| \Leftrightarrow 4.8253 > 1.8595$ ou encore $-4.8253 < -2.3060 \Rightarrow$ On rejette H0
   - avec SAS :

     CODE SAS : Test sur un paramètre

     ```
     proc reg data=toto.ExempleCoursChap3 ;
     model prix = surface / clb ;
     test surface = 1 ;
     run ;
     quit ;
     ```

     SAS fait toujours des tests bilatéraux. Il est parfois difficile de conclure à un test unilatéral.

Reprise de l'exemple, H0 : $\beta_1 = 1; H1 : \beta_1 < 1$

Procédure REG
Modèle : MODEL1
Variable dépendante : prix

Nombre d'observations lues	10
Nombre d'observations utilisées	10

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	754.99092	754.99092	29.91	0.0006
Erreur	8	201.90908	25.23864		
Total sommes corrigées	9	956.90000			

Root MSE	5.02381	R carré	0.7890
Moyenne dépendante	38.90000	R car. ajust.	0.7626
Coeff Var	12.91467		

Valeurs estimées des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à 95 %	
Intercept	1	11.75199	5.21168	2.25	0.0541	-0.26615	23.77014
surface	1	0.53127	0.09714	5.47	0.0006	0.30728	0.75527

Procédure REG
Modèle : MODEL1

Résultats du test 1 pour la variable dépendante prix

Source	DDL	Moyenne quadratique	Valeur F	Pr > F
Numérateur	1	587.69092	23.29	0.0013
Dénominateur	8	25.23864		

Exemple $H0 : \beta_1 = 1; H1 : \beta_1 > 1$

Test Unilatéral

1. valeur empirique de la statistique :
   $$t_{emp} = \frac{0.53127 - 1}{0.09714} = -4.8253 \sim t_8 \text{ ddl } = 1.8595 \text{ (borne supérieure)}$$
2. Conclusion du test
   - D'après la table $t_{emp} < t_8 \Leftrightarrow -4.8253 < 1.8595 \Rightarrow$ On ne rejette pas H0
   - avec SAS :

     SAS fait toujours des tests bilatéraux. Il est parfois difficile de conclure à un test unilatéral.
     Les résultats de SAS ne nous permettent pas de conclure ce test.

Reprise de l'exemple, H0 : $\beta_1 = 1; H1 : \beta_1 > 1$

Procédure REG
Modèle : MODEL1
Variable dépendante : prix

Nombre d'observations lues	10
Nombre d'observations utilisées	10

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	754.99092	754.99092	29.91	0.0006
Erreur	8	201.90908	25.23864		
Total sommes corrigées	9	956.90000			

Root MSE	5.02381	R carré	0.7890
Moyenne dépendante	38.90000	R car. ajust.	0.7626
Coeff Var	12.91467		

Valeurs estimées des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à 95 %	
Intercept	1	11.75199	5.21168	2.25	0.0541	-0.26615	23.77014
surface	1	0.53127	0.09714	5.47	0.0006	0.30728	0.75527

Procédure REG
Modèle : MODEL1

Résultats du test 1 pour la variable dépendante prix

Source	DDL	Moyenne quadratique	Valeur F	Pr > F
Numérateur	1	587.69092	23.29	0.0013
Dénominateur	8	25.23864		

Attention aux tables

Avant de lire la valeur théorique dans une table il faut vérifier si la table utilisée est bilatérale ou unilatérale.

Pour cela il faut regarder le graphique en début de table :

- Si deux zones sont hachurées c'est une table bilatérale ;
- Si seulement une zone est hachurée c'est une table unilatérale.

Comment lire une valeur théorique pour un test bilatéral dans une table unilatérale ?

Pour un test bilatéral nous avons quelque chose du type :
$P\left[-t_{\alpha/2} \leq t \leq t_{\alpha/2}\right] = 1 - \alpha$ (par exemple $\alpha = 0.05 = 5\%$)

Cette probabilité peut encore s'écrire :

$$\begin{array}{rrcl}
 & P(t \leq t_{\alpha/2}) - P(t \leq -t_{\alpha/2}) & = & 1 - \alpha \\
\Leftrightarrow & P(t \leq t_{\alpha/2}) - \left[1 - P(t \leq t_{\alpha/2})\right] & = & 1 - \alpha \\
\Leftrightarrow & 2P(t \leq t_{\alpha/2}) - 1 & = & 1 - \alpha \\
\Leftrightarrow & 2P(t \leq t_{\alpha/2}) & = & 2 - \alpha \\
\Leftrightarrow & P(t \leq t_{\alpha/2}) & = & 1 - \dfrac{\alpha}{2}
\end{array}$$

Conclusion :

- Pour trouver la valeur théorique d'un test bilatéral avec une table unilatérale, il faut prendre la valeur théorique associée à $\alpha/2$ dans la table.
  Exemple : si $\alpha = 0.05$ il faut prendre le t associé à $\alpha/2 = 0.025$
- Pour trouver la valeur théorique d'un test unilatéral avec une table bilatérale, il faut prendre la valeur théorique associée à $2 \times \alpha$ dans la table.
  Exemple : si $\alpha = 0.05$ il faut prendre le t associé à $2 \times \alpha = 0.10$

Tests : Résumé

Type d'hypothèse sur β_j	H_0	H_1	Règle de décision H0 rejetée si
Bilatérale	$\beta_j = \beta_{j0}$	$\beta_j \neq \beta_{j0}$	$\lvert t\rvert > t_{\alpha/2}(ddl)$
Droite	$\beta_j \leq \beta_{j0}$	$\beta_j > \beta_{j0}$	$t > t_{\alpha}(ddl)$
Gauche	$\beta_j \geq \beta_{j0}$	$\beta_j < \beta_{j0}$	$t < -t_{\alpha}(ddl)$

GPS

Nous avons montré que :

$$\begin{array}{ccccc} \sum_{i=1}^{n}(y_i - \bar{y})^2 & = & \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 & + & \sum_{i=1}^{n}\hat{u}_i^2 \\ \text{STC} & = & \text{SCE} & + & \text{SCR} \end{array}$$

Associons à ces sommes des ddl.

STC : $n-1$ ddl (on perd 1 ddl en calculant la moyenne).

SCR : $n-2$ ddl (2 : nombre de paramètres à estimer dans le modèle)
De façon générale $n-k$ ddl (où k est le nombre de paramètres estimés par le modèle)

SCE : 1 ddl (nombre de variables explicatives dans le modèle)
De façon générale $k-1$ ddl (où k est le nombre de paramètres estimés par le modèle)

Remarques

- La somme des ddl de SCR et de SCE = ddl SCT :
  $(n-2)+(1)=(n-1)$
  De manière générale : $(n-k)+(k-1)=(n-1)$ où k est le nombre de paramètres estimés par le modèle
- Ces sommes dont en fait des sommes de carré de VA suivant une loi normale $\Rightarrow$ On peut définir la statistique :

$$F=\frac{\text{SCE}/\text{ddl}_{SCE}}{\text{SCR}/\text{ddl}_{SCR}}\sim F\left(\text{ddl}_{SCE},\text{ddl}_{SCR}\right)$$

Avec 2 paramètres à estimer

- $\Rightarrow F\sim F(1,n-2)$; D'après la table de Fisher, $F_{th}(1,8)=5,318$
- Cela revient à réaliser un test de significativité du paramètre β_1

A quoi sert le R^2 ?

- Ces sommes permettent de calculer le $R^2 = \dfrac{\text{SCE}}{\text{STC}} = 1 - \dfrac{\text{SCR}}{\text{STC}}$.

  R^2 mesure le pourcentage de la variance (dispersion) de la variable à expliquer Y, expliquée par le modèle, c'est-à-dire par les variables explicatives (ici la variable explicative X).

  Dans l'exemple, 78,9 % de la variance du prix d'un logement est expliqué par la surface du logement.

  Plus le R^2 est élevé, plus la modélisation linéaire retenue est adaptée pour expliquer Y.

Stat du F, R^2

Procédure REG
Modèle : MODEL1
Variable dépendante : prix

Nombre d'observations lues	10
Nombre d'observations utilisées	10

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	754.99092	754.99092	29.91	0.0006
Erreur	8	201.90908	25.23864		
Total sommes corrigées	9	956.90000			

Root MSE	5.02381	R carré	0.7890
Moyenne dépendante	38.90000	R car. ajust.	0.7626
Coeff Var	12.91467		

Valeurs estimées des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|	Intervalle de confiance à 95 %	
Intercept	1	11.75199	5.21168	2.25	0.0541	-0.26615	23.77014
surface	1	0.53127	0.09714	5.47	0.0006	0.30728	0.75527

GPS

Moyennes

Terminologie

Prévisions : lorsque les données sont temporaires.
Ex : prévoir pour l'année suivante, le mois suivant ou la semaine suivante

Prédictions : lorsque les sonnées sont individuelles.
Objectif : Calculer la valeur estimée pour une valeur moyenne de X donnée, par exemple $\bar{x} = x_0$

On sait que $\mathrm{E}(u) = 0$ et $\mathrm{E}(\widehat{u}) = 0$.
Valeur estimée de y_0 pour une valeur moyenne $x_0 \Rightarrow$
$\widehat{y_0} = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$

Les EMCO sont BLUE par conséquent, $\widehat{y_0}$est un estimateur BLUE dont les moments sont définis par :

$$
\begin{aligned}
E(\widehat{y_0}) &= E(\widehat{\beta_0} + \widehat{\beta_1} x_0) = \beta_0 + \beta_1 x_0 \\
V(\widehat{y_0}) &= V(\widehat{\beta_0} + \widehat{\beta_1} x_0) \\
&= V(\widehat{\beta_0}) + x_0^2 V(\widehat{\beta_1}) + 2\ Cov(\widehat{\beta_0}, \widehat{\beta_1} x_0) \\
&= \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} + \sigma^2 \times \frac{x_0^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&- 2 \times x_0 \times \frac{\sigma^2 \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&= \sigma^2 \left(\frac{n x_0^2 + \sum_{i=1}^{n} x_i^2 - 2 n x_0 \bar{x}}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \right)
\end{aligned}
$$

On peut simplifier cette variance en ajoutant et retranchant $n\widehat{x}^2$:

$$
\begin{aligned}
V(\widehat{y}_0) &= \sigma^2 \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 + n\bar{x}^2 + nx_0^2 - 2nx_0\bar{x}}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sigma^2 \left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(x_0 - \bar{x})^2\right]}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]
\end{aligned}
$$

$\Rightarrow \quad \widehat{y}_0 \sim \mathrm{N}\left(\mathrm{E}(\widehat{y}_0), \mathrm{V}(\widehat{y}_0)\right)$

Sachant que σ^2 est inconnu et estimé par $\widehat{\sigma}^2$ on a :

$$t_{y_0} = \frac{\widehat{y_0} - \mathrm{E}(\widehat{y_0})}{\widehat{\sigma}_{\widehat{y_0}}} \sim t_(n-2).$$

IC pour la prévision

$$\mathrm{P}\left[\widehat{\beta}_0 + \widehat{\beta}_1 x_0 - t_{\alpha/2}\widehat{\sigma}_{\widehat{y_0}} \leq y_0 \leq \widehat{\beta}_0 + \widehat{\beta}_1 x_0 + t_{\alpha/2}\widehat{\sigma}_{\widehat{y_0}}\right] = 1 - \alpha$$

$$\Rightarrow IC_{y_0} = \left[(\widehat{\beta}_0 + \widehat{\beta}_1 x_0) \mp t_{\alpha/2}\widehat{\sigma}_{\widehat{y_0}}\right]$$

Individuelles

La prévision individuelle est définie par $\widehat{y}_{i0} = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i0}$

De façon individuelle, on peut faire une erreur de mesure lors de la prévision $\Rightarrow$
$\widehat{u}_{i0} = y_{i0} - \widehat{y}_{i0} \neq 0$
$\widehat{u}_{i0} = y_{i0} - \widehat{y}_{i0} = (\beta_0 - \widehat{\beta}_0) + (\beta_1 - \widehat{\beta}_1) x_{i0} + u_{i0}$

Par conséquent les moments ce cet estimateur sont définis par :

$$\begin{aligned}
E(\widehat{u}_{i0}) &= 0 \\
V(\widehat{u}_{i0}) &= V(\widehat{\beta}_0) + V(\widehat{\beta}_1)\ x_{i0}^2 + 2\ x_{i0}\ \mathrm{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) + 2\mathrm{Cov}(\widehat{\beta}_0, u_{i0}) \\
&\quad +2\ x_{i0}\ \mathrm{Cov}(\widehat{\beta}_1, u_{i0}) + V(u_{i0}) \\
&= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{i0} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]
\end{aligned}$$

On peut donc déduire la statistique de test

$$t = \frac{(\widehat{y}_{i0} - y_{i0}) - \mathrm{E}(\widehat{u}_{i0})}{\sqrt{\mathrm{V}(y_{i0} - \widehat{y}_{i0})}} \sim t_{(n-2)}$$

CODE SAS : Régression linéaire avec prédictions et IC des prédictions

```
proc reg data=toto.exemplecours;
model prix = surface / clm cli;
run;
quit;
```

Remarques importantes :

Par défaut SAS calcule les IC des prévisions pour un $\alpha = 0.05$.

L'option **clm** permet de calculer les prévisions au point moyen. les écart-types sont calculés pour ces prévisions.

L'option **cli** permet de calculer les prévisions indivuelles. les écart-types ne sont pas calculés mais il est possible de les retrouver avec les IC.

Si l'on veut modifier le seuil d'erreur, il faut rajouter après clm l'instruction *alpha=0.01* pour faire un IC à 99% par exemple.

Reprise de l'exemple, Prédictions et IC

Procedure REG
Modèle : MODEL1
Variable dépendante : prix

Statistiques de sortie

Obs.	Variable dépendante	Valeur prédite	Prédiction de la moy. Erreur type	Moyenne de l'IC à 95 %		Prédiction de l'IC à 95 %		Résidus
1	20.0000	26.6276	2.7493	20.2877	32.9675	13.4214	39.8339	-6.6276
2	43.0000	38.3156	1.5923	34.6439	41.9874	26.1627	50.4685	4.6844
3	41.0000	40.9720	1.6332	37.2058	44.7382	28.7902	53.1537	0.0280
4	49.0000	43.6283	1.8087	39.4576	47.7991	31.3155	55.9411	5.3717
5	38.0000	37.2531	1.6170	33.5244	40.9818	25.0829	49.4232	0.7469
6	38.0000	30.3465	2.2293	25.2058	35.4872	17.6723	43.0208	7.6535
7	53.0000	57.4414	3.7438	48.8081	66.0747	42.9934	71.8894	-4.4414
8	46.0000	46.2847	2.0849	41.4769	51.0925	33.7417	58.8276	-0.2847
9	24.0000	28.7527	2.4425	23.1202	34.3852	15.8711	41.6343	-4.7527
10	37.0000	39.3781	1.5911	35.7091	43.0472	27.2261	51.5302	-2.3781

Somme des résidus	0
Somme des résidus du carré	201.90908
Somme des carrés des résidus prédits (PRESS)	391.84703

GPS

Relation entre le salaire horaire et le nombre d'années d'études

Considérons un échantillon de 526 individus actifs pour lesquels on observe le salaire horaire (salh) et le nombre d'années d'études (educ).
Soit la relation estimée par MCO suivante :

$$\begin{array}{cccccc} \widehat{\text{salh}}_i & = & -0.90 & + & 0.54 \ \text{educ}_i \\ & & (0.84) & & (0.18) \end{array}$$

les écart-types des paramètres sont donnés entre parenthèses.

1. Caractériser y_i et x_i.
2. Interpréter les coefficients estimés.
3. Quel est le salaire estimé d'une personne ayant 8 années d'études ? De combien augmenterait le salaire de cette personne si elle avait 4 ans d'études supplémentaires ? Commenter.

Nouvelle relation entre le salaire horaire et le nombre d'années d'études

Considérons une nouvelle relation entre le salaire horaire et le nombre d'année d'études estimée par MCO :

$$\begin{array}{ccccc} \widehat{\ln(\text{salh})}_i & = & \underset{(0.21)}{0.584} & + & \underset{(0.03)}{0.083}\ \text{educ}_i \end{array}$$

1. Caractériser y_i et x_i.
2. Interpréter les coefficients estimés.

Relation entre le salaire hebdomadaire (salw) et le nombre d'heures travaillées (nbh)

Soit la relation estimée par MCO :

$$\widehat{\ln(\text{salw})}_i = \underset{(1.21)}{4.822} + \underset{(0.10)}{0.257}\ \text{nbh}_i$$

1. Caractériser y_i et x_i.
2. Interpréter les coefficients estimés.

Fin