



Chapitre 2

Régression linéaire multiple

2. I Le modèle linéaire général

Quand on construit un modèle économétrique, on suppose qu'il existe une relation entre les observations d'une variable y , dite à expliquer (dépendante), et k variables, dites explicatives (indépendantes), x_1, x_2, \dots, x_k , chacune étant affectée d'un coefficient β .

On admet que la relation n'est pas exacte de sorte qu'une réalisation de y est la somme de deux composantes :

$$y_i = \boxed{\beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}} + \boxed{u_i} \quad (2.I)$$

partie observée

partie inobservée

2. I Le modèle linéaire général

Le terme aléatoire u « capte les insuffisances » du modèle :

- le modèle n'est qu'une caricature/un résumé de la réalité
- les variables qui ne sont pas prises en compte dans le modèle
- les fluctuations liés à l'échantillonnage (si on change d'échantillon, on peut obtenir un résultat différent)

2. I Le modèle linéaire général

On ne connaît pas les k valeurs des coefficients β_1, \dots, β_k

L'économétrie permet la construction d'estimateurs de ces coefficients, notés $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$

Lecture des coefficients : $\frac{\delta y}{\delta x_j} = \beta_j$

$\hat{\beta}_j$: Variation estimée de y quand x_j augmente de 1 unité, toutes choses étant égales par ailleurs.

2.2 La démarche de la modélisation

La démarche de modélisation consiste toujours à :

- estimer les paramètres β_1, \dots, β_k en exploitant les données
- évaluer la précision de ces estimateurs (biais, efficacité)
- mesurer le pouvoir explicatif global du modèle
- évaluer l'influence des variables dans le modèle
- utiliser les résultats pour des prévisions/ préconisations/ aides à la décision

2.2.1 Le choix des estimateurs $\hat{\beta}$

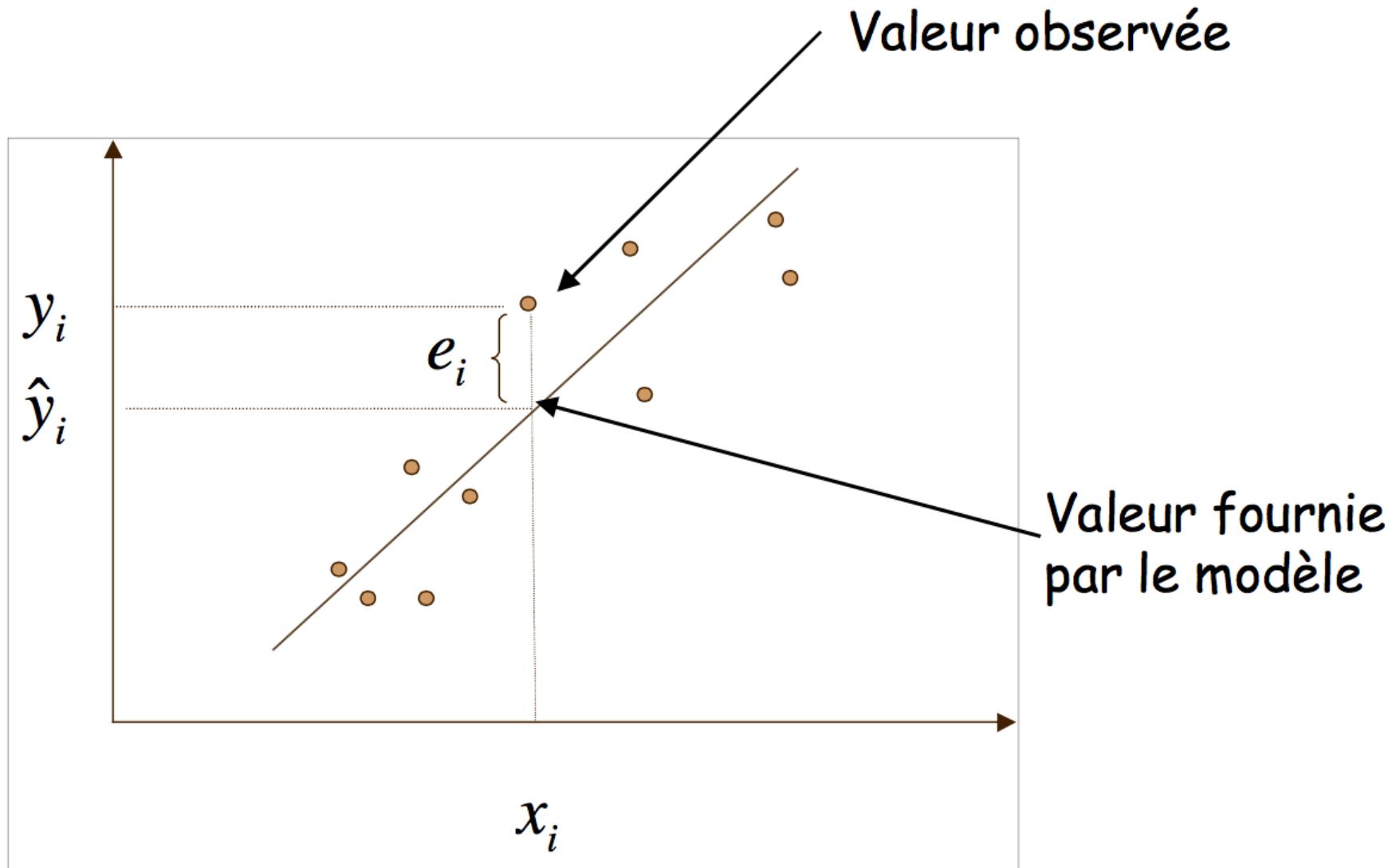
Idée générale du choix de la technique d'estimation : trouver le meilleur vecteur $\hat{\beta}$ qui génère les plus petites erreurs observées possibles.

La technique des **moindres carrés ordinaires** (MCO, OLS) cherche la meilleure estimation des paramètres en

$$\text{Min} \sum_i e_i^2$$

où e est l'erreur observée (le résidu)

2.2.1 Le choix des estimateurs $\hat{\beta}$



2.2.1 Le choix des estimateurs $\hat{\beta}$

Une fois que les $\hat{\beta}$ seront trouvés, l'estimation, notée \hat{y} , est simplement :

$$\hat{y}_i = \hat{\beta}_1 \cdot x_{i1} + \hat{\beta}_2 \cdot x_{i2} + \dots + \hat{\beta}_k \cdot x_{ik} = \sum_{j=1}^k \hat{\beta}_j \cdot x_{ij}$$

e est l'écart existant entre la réalisation de y et l'estimateur \hat{y} :

$$e_i = y_i - \hat{y}_i$$

Les réalisations de la série e découlent directement du choix fait sur les estimateurs des coefficients

2.2.1 Le choix des estimateurs $\hat{\beta}$

Il est important de comprendre la différence existant entre le **terme d'erreur théorique** et l'**erreur empirique**

$$u_i = y_i - \sum_{j=1}^k \beta_j x_{ij}$$

$$e_i = y_i - \sum_{j=1}^k \hat{\beta}_j x_{ij}$$

L'erreur empirique (résidu) est la somme de deux composantes :

- erreur d'estimation de la partie expliquée de y (directement associée au choix des estimateurs)
- erreur théorique u (qui lui est totalement étranger à ce choix)

2.2.2 Principe de calcul

Par souci de concision, on adopte l'écriture matricielle

$$\begin{pmatrix} y_1 \\ \dots \\ y_i \\ \dots \\ y_T \end{pmatrix} = \begin{pmatrix} 1 & x_{12} & \dots & x_{1k} \\ 1 & & & \\ 1 & x_{i2} & x_{ij} & x_{ik} \\ 1 & & & \\ 1 & x_{T2} & & x_{Tk} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ \dots \\ u_i \\ \dots \\ u_T \end{pmatrix}$$

N.B. Noter la colonne représentant la constante

$$Y = X\beta + u$$

$$\hat{Y} = X\hat{\beta}$$

$$\hat{e} = (y_1 - \hat{y}_T) - (\hat{\beta}_1 - \hat{\beta}_2) \begin{pmatrix} \hat{x}_{12} & \hat{x}_{i2} & \hat{x}_{Tk} \\ x_{1j} & x_{ij} & x_{Tk} \\ x_{1k} & x_{ik} & x_{Tk} \end{pmatrix} e = Y - \hat{Y}$$

2.2.2 Principe de calcul

Le vecteur $\hat{\beta}$ cherché est donc la solution du problème $\text{Min } e'e$.

On doit résoudre $\frac{\delta e'e}{\delta \hat{\beta}} = 0$: il y a k équations dites « équations normales » à résoudre

On doit commencer par écrire la quantité $e'e$ en fonction de β :

$$e'e = Y'Y - 2Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}$$

e au sens de la transposée

2.2.2 Principe de calcul

Condition du premier ordre :

$$X'X\hat{\beta} = X'Y \text{ : système d'équations normales}$$

solution du problème : $\hat{\beta} = (X'X)^{-1}X'Y$

si XX' est inversible

$\Rightarrow X$ de rang k

2.2.2 Principe de calcul

Exemple numérique : Nombre d'abonnés Haut Débit

Nombre d'abonnements à Internet haut et très haut débit en France (en millions)											
T2 2013	T3 2013	T4 2013	T1 2014	T2 2014	T3 2014	T4 2014	T1 2015	T2 2015	T3 2015	T4 2015	T1 2016
24,4	24,7	24,9	25,2	25,4	25,7	26,0	26,2	26,3	26,6	26,9	27,1
source : ARCEP											

Un modèle simple expliquant l'évolution du nombre d'abonnés Haut Débit est de supposer qu'elle est une fonction croissante du temps. Dans ces conditions, une variable explicative, notée t , décrivant le passage de ce temps paraît adaptée. D'où le modèle retenu :

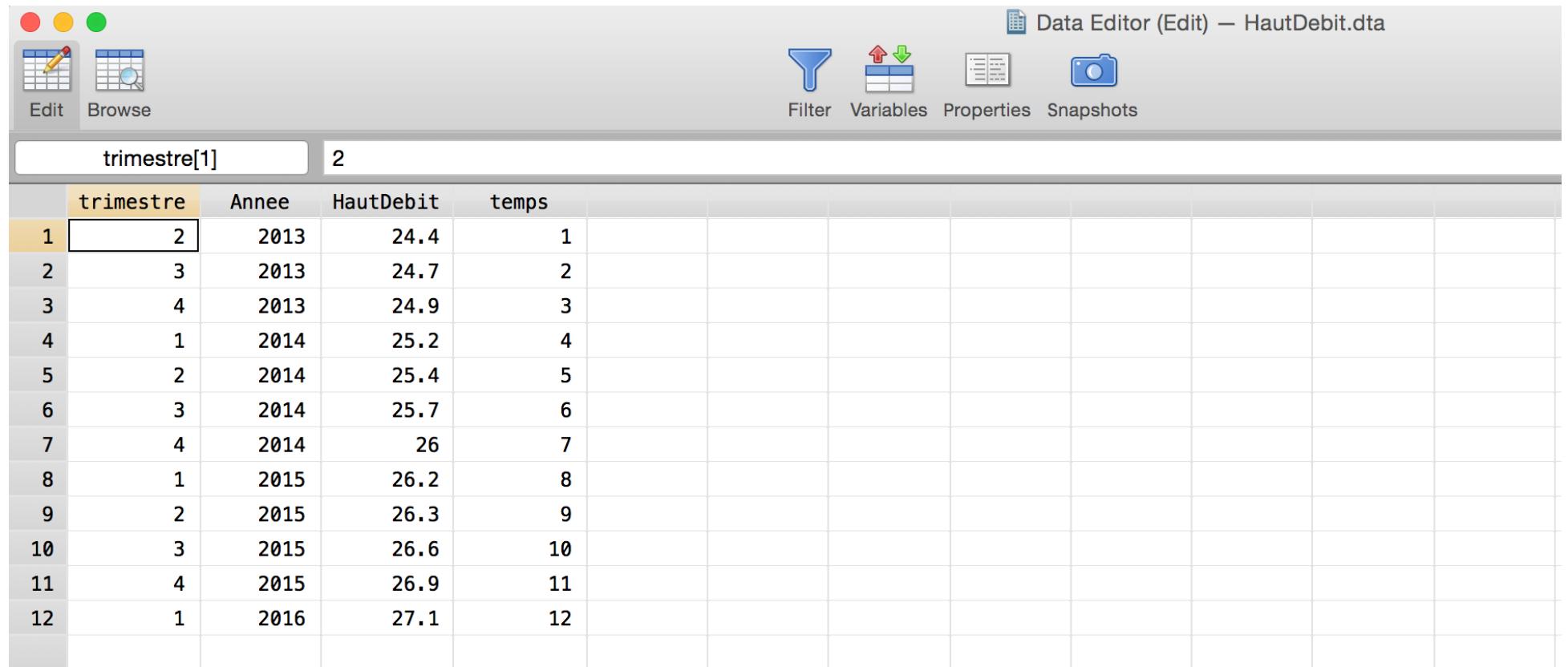
$$y_i = \beta_1 + \beta_2 t_i + u_i$$

Calculons les estimateurs MCO de β_1 et β_2

2.2.2 Principe de calcul

Exemple numérique : Nombre d'abonnés Haut Débit

Solution avec le logiciel Stata



The screenshot shows the Stata Data Editor interface. The title bar reads "Data Editor (Edit) — HautDebit.dta". The toolbar includes icons for Edit, Browse, Filter, Variables, Properties, and Snapshots. The main window displays a dataset with four columns: "trimestre", "Année", "HautDebit", and "temps". The first row is highlighted in yellow, and the value "2" is selected in the "trimestre" column. The data shows quarterly values from 2013 to 2016.

	trimestre	Année	HautDebit	temps
1	2	2013	24.4	1
2	3	2013	24.7	2
3	4	2013	24.9	3
4	1	2014	25.2	4
5	2	2014	25.4	5
6	3	2014	25.7	6
7	4	2014	26	7
8	1	2015	26.2	8
9	2	2015	26.3	9
10	3	2015	26.6	10
11	4	2015	26.9	11
12	1	2016	27.1	12

2.2.2 Principe de calcul

Exemple numérique : Nombre d'abonnés Haut Débit

Solution avec le logiciel Stata

```
. reg HautDebit temps
```

Source	SS	df	MS	Number of obs	=	12
Model	8.46881082	1	8.46881082	F(1, 10)	=	3040.23
Residual	.027855824	10	.002785582	Prob > F	=	0.0000
Total	8.49666664	11	.77242424	R-squared	=	0.9967
				Adj R-squared	=	0.9964
				Root MSE	=	.05278

HautDebit	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
temp	.2433566	.0044136	55.14	0.000	.2335226 .2531907
_cons	24.20152	.032483	745.05	0.000	24.12914 24.27389

β_2

β_1

Autre exemple : équation de salaire

Tableau 6 : équation de salaire pour les emplois occupés en 2000 : influence des formations structurée et informelle

Variables	Paramètre	P. value
Constante	1,900	0,000
Sexe (homme=1)	0,145	0,000
Probabilité de suivre une formation structurée	0,394	0,000
Probabilité de suivre une formation structurée*sexé	-0,055	-0,535
Probabilité de suivre une formation en cours d'emploi	0,264	-0,072
Probabilité de suivre une formation en cours d'emploi*sexé	0,041	-0,751
Probabilité d'obtenir une promotion	0,159	-0,004
Probabilité d'obtenir une promotion*sexé	0,016	-0,831
Marié	0,084	0,000
En couple	0,084	0,000
Enfants à charges	0,000	-0,998
Ancienneté avec l'employeur	0,001	0,000
Ancienneté au carré /100	0,000	-0,110
Expérience à temps plein	0,019	0,000
Expérience au carré/100	-0,032	0,000
Niveau de scolarité (Réf. = Pas de diplômes d'études secondaires)		
- Diplômes d'études secondaires	0,076	0,000
- Certificat	0,108	0,000
- Diplôme collégial	0,118	0,000
- Diplôme universitaire	0,259	0,000
Heures travaillées		
- Temps partiel	0,017	-0,324
- Horaire de travail flexible	0,014	-0,259
- Travail entre 6h et 18h	-0,013	-0,444
- Heures habituelles non rémunérées	0,008	0,000
Statut d'emploi		
- Emploi permanent	0,002	-0,926
- Emploi de supervision	0,042	-0,001
- Emploi syndiqué	0,049	-0,002
- Procédure d'évaluation du rendement	-0,018	-0,186
Etablissement à but non lucratif	-0,007	-0,760
Etablissement à propriété étrangère	0,064	-0,002

Profession (Réf. = Personnel technique/métiers)		
- Gestionnaires	0,187	0,000
- Professionnels	0,176	0,000
- Commercialisation ou ventes	-0,092	-0,003
- Personnel de bureau	-0,077	0,000
- Personnel non qualifié	-0,109	0,000
Branche d'activité (Réf. = Commerce de détail)		
- Exploitation des ressources naturelles	0,397	0,000
- Industries de la fabrication	0,245	0,000
- Constructions	0,359	0,000
- Transport, entreposage et commerce de gros	0,233	0,000
- Communications et autres services publics	0,270	0,000
- Finance et assurances	0,223	0,000
- Services immobiliers et de location	0,274	0,000
- Services aux entreprises	0,251	0,000
- Enseignement et services de soins de santé	0,213	0,000
- Information et industries culturelles	0,264	0,000
Taille de l'établissement (Réf. = 500 employés et plus)		
- Moins de 20 employés	-0,124	0,000
- Entre 20 et 99 employés	-0,147	0,000
- Entre 100 et 499 employés	-0,078	0,000
Région (Réf. = Ontario)		
- Colombie-Britannique	0,024	-0,282
- Alberta	-0,090	0,000
- Provinces des Prairies	-0,153	0,000
- Québec	-0,036	-0,174
- Provinces de l'Atlantique	-0,223	0,000
<i>R</i> ²	54,91%	
Nombre d'observations	18 870	

2.2.3 Les propriétés des estimateurs

L'économètre n'a pas pour unique objectif l'obtention des $\hat{\beta}$: il veut par exemple tester une théorie, faire des prévisions et donc tirer des conclusions sur les véritables valeurs de β .

L'économète a besoin de connaître la fiabilité/précision des estimateurs obtenus.

=> nécessité d'une discussion sur les propriétés des estimateurs : l'économète a besoin de savoir si l'estimateur des MCO est le meilleur estimateur possible (càd le plus précis possible).

2.2.3 Les propriétés des estimateurs

Critères :

En économétrie, les estimateurs que l'on cherche doivent avoir des caractéristiques particulières pour être considérés comme « optimaux » :

- non biaisés
- efficaces

2.2.3 Les propriétés des estimateurs

Absence de biais

Définition du biais : Soit $\hat{\beta}$ l'estimateur d'un vecteur β , estimé à partir d'un échantillon de taille T . On peut définir son biais par :

$$B(\hat{\beta}) = E(\hat{\beta} - \beta) = E(\hat{\beta}) - \beta$$

L'estimateur est sans biais si $B(\hat{\beta}) = 0 \Rightarrow E(\hat{\beta}) = \beta$

Il est asymptotiquement sans biais si $\lim_{T \rightarrow \infty} B(\hat{\beta}) = 0$

2.2.3 Les propriétés des estimateurs

Efficacité

Il est souhaitable pour la variance d'un estimateur d'être aussi petite que possible.

L'absence de biais ou la variance minimale ne sont pas seuls des critères suffisants pour une sélection parmi les estimateurs.

Il faut la combinaison des deux critères :

$$E(\hat{\beta}) = \beta$$

$$\Omega_{\hat{\beta}} = E(\hat{\beta} - E(\hat{\beta}))^2$$
 aussi faible que possible

Un tel estimateur est appelé *estimateur efficace*.

2.2.3 Les propriétés des estimateurs

Théorème de Gauss-Markov

Théorème : Si les hypothèses sous-jacentes du modèle de régression linéaire sont vérifiées, l'estimateur des moindres carrés ordinaires, dans la catégorie des estimateurs linéaires sans biais, a une variance minimale, c'est-à-dire qu'il est le BLUE (meilleur estimateur linéaire sans biais)

2.2.3 Les propriétés des estimateurs

Hypothèses

Hypothèses probabilistes (stochastiques)

- les X sont observés sans erreur (non aléatoires)
- en moyenne, le modèle est bien spécifié : $E(u_i) = 0$
- *homoscédasticité* : la variance de l'erreur est constante :

$$\forall i = 1, \dots, T; V(u_i) = \sigma^2$$

- *Absence d'autocorrélation des erreurs* : $E(u_i, u_j) = 0$
- L'erreur est indépendante des variables explicatives (*exogénéité*) :

$$Cov(u_i, X_i) = E(u_i X_i) = 0$$

2.2.3 Les propriétés des estimateurs

Hypothèses

Hypothèses structurelles

- *Absence de colinéarité parfaite* : il n'y a pas de relations linéaires parfaites parmi les variables explicatives. $(X'X)$ est régulière et son inverse existe.
- $(X'X)/T$ tend vers une matrice finie non singulière quand $T \rightarrow \infty$
- *Nombre d'observations (T) supérieur au nombre de paramètres du modèle (k)*

2.2.3 Les propriétés des estimateurs

Démonstration

L'estimateur MCO est sans biais

$$\begin{aligned}\hat{\beta} &= (X'X)^{-1}X'Y \\ \Rightarrow E(\hat{\beta}) &= E[(X'X)^{-1}X'Y] = E[(X'X)^{-1}X'(X\beta + u)] \\ &= E[(X'X)^{-1}X'X\beta + (X'X)^{-1}X'u)] \\ &= E(\beta) + (X'X)^{-1}X'E(u) \\ &= \beta\end{aligned}$$

2.2.3 Les propriétés des estimateurs

Démonstration

L'estimateur MCO est le BLUE

Il reste à démontrer que l'estimateur MCO est à variance minimale parmi la classe des estimateurs sans biais. D'où :

- i) Calculer la matrice de variance-covariance
- ii) La comparer avec celle des autres estimateurs

i) Calcul de la matrice de variance-covariance

$$\Omega_{\hat{\beta}} = E\{(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\}$$

Or $\hat{\beta} = (X'X)^{-1}X'Y = \beta + (X'X)^{-1}X'u$

$$\hat{\beta} - \beta = (X'X)^{-1}X'u \text{ et } (\hat{\beta} - \beta)' = u'X(X'X)^{-1}$$

$$(\hat{\beta} - \beta) \cdot (\hat{\beta} - \beta)' = (X'X)^{-1}X'u u'X(X'X)^{-1}$$

$$\Omega_{\hat{\beta}} = (X'X)^{-1}X'E(uu')X(X'X)^{-1}$$

$$\Omega_{\hat{\beta}} = (X'X)^{-1}X'\sigma^2 X(X'X)^{-1}$$

$$\Omega_{\hat{\beta}} = \sigma^2(X'X)^{-1}X'X(X'X)^{-1}$$

$$\Omega_{\hat{\beta}} = \sigma^2(X'X)^{-1}$$

ii) La comparer avec celle des autres estimateurs

- considérons la classe des estimateurs linéaires de Y

$$\hat{b} = MY$$

$$\hat{b} = [(X'X)^{-1}X' + N]Y$$

- on doit se restreindre à la classe des estimateurs linéaires sans biais

$$\hat{b} = [(X'X)^{-1}X' + N](X\beta + u)$$

$$\hat{b} = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u + NX\beta + Nu$$

$$\hat{b} = \beta + (X'X)^{-1}X'u + NX\beta + Nu$$

$$\Rightarrow E(\hat{b}) = \beta + NX\beta$$

On doit imposer la restriction $NX=0$ pour se trouver dans la classe des estimateurs sans biais

- on doit évaluer la matrice de variance-covariance de ces estimateurs :

$$\Omega_{\hat{b}} = E[(\hat{b} - \beta)(\hat{b} - \beta)']$$

Or $\hat{b} = \beta + [(X'X)^{-1}X' + N]u$

$$\Rightarrow \hat{b} - \beta = Pu$$

$$= E[Puu'P'] = PE(uu')P'$$

$$= \sigma^2 PP'$$

$$\Rightarrow \Omega_{\hat{b}} = \Omega_{\hat{\beta}} + \sigma^2.NN'$$

L'estimateur MCO est sans biais à variance minimale

2.2.4 Le pouvoir explicatif global du modèle

Tableau d'analyse de variance

Equation d'analyse de la variance -

Décomposition de la variance

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

SCT → SCE → SCR
 Variabilité totale Variabilité expliquée par le modèle Variabilité non-expliquée (variabilité résiduelle)

Source de variation	Sommes des carrés	Degré de liberté	Carrés moyens
Modèle	SCE	k-1	SCE/(k-1)
Résiduel	SCR	T-k	SCR/(T-k)
Total	SCT	T-1	

. reg HautDebit temps

Source	SS	df	MS
Model	8.46881082	1	8.46881082
Residual	.027855824	10	.002785582
Total	8.49666664	11	.77242424

Tableau d'analyse de variance

2.2.4 Le pouvoir explicatif global du modèle

Coefficient de détermination

Définition : Le coefficient de détermination, ou de corrélation multiple, noté R^2 , est utilisé pour mesurer la qualité de l'ajustement réalisé au moyen des estimateurs. R^2 est défini comme la part de la variance expliquée dans la variance totale :

$$R^2 = \frac{V(\hat{y})}{V(y)} = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

R^2 est compris entre 0 et 1

2.2.4 Le pouvoir explicatif global du modèle

Coefficient de détermination ajusté (corrigé)

Le R^2 augmente mécaniquement avec le nombre de variables explicatives introduites - même si les variables additionnelles ne sont pas pertinentes.

Conséquence : On ne peut pas comparer des modèles de complexité différente (càd avec un nombre de variables explicatives différent) à partir du R^2

Il faut utiliser le R^2 ajusté, noté \bar{R}^2 , qui est un R^2 corrigé par les degrés de liberté

$$\bar{R}^2 = 1 - \frac{SCR/(T - k)}{SCT/(T - 1)}$$

2.2.4 Le pouvoir explicatif global du modèle

Exemple

```
. reg HautDebit temps
```

Source	SS	df	MS	Number of obs	=	12
Model	8.46881082	1	8.46881082	F(1, 10)	=	3040.23
Residual	.027855824	10	.002785582	Prob > F	=	0.0000
				R-squared	=	0.9967
Total	8.49666664	11	.77242424	Adj R-squared	=	0.9964
				Root MSE	=	.05278

HautDebit	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
temp	.2433566	.0044136	55.14	0.000	.2335226
_cons	24.20152	.032483	745.05	0.000	24.12914

2.2.5 Evaluer l'influence des variables dans le modèle : les tests d'hypothèses

On peut évaluer l'influence des variables dans le modèle en se plaçant à différents niveaux :

- globalement (toutes les $k-1$ variables)
- individuellement (chaque variable)
- un bloc de variables (q variables , $q \leq k - 1$)

Par exemple, pour savoir si une variable explicative a un réel impact sur la variable à expliquer, il faut tester si son coefficient est différent de 0 ou pas. Regarder la seule valeur ne suffit pas car on travaille avec des approximations. On a besoin de tests d'hypothèses.

Pour répondre à ce type de questions, il faut préciser la loi des paramètres.

2.2.5 Evaluer l'influence des variables dans le modèle : les tests d'hypothèses

Distribution de $\hat{\beta}$

On doit faire l'hypothèse additionnelle : $u_i \rightarrow N(0, \sigma^2)$

Cette hypothèse de normalité permet d'effectuer des tests sur les paramètres : la loi de u étant précisée, on peut déduire celle des estimateurs de β et de σ^2 .

On va ainsi obtenu un résultat central, le théorème de Cochrane, à la base de tous les tests effectués à partir des MCO.

2.2.5 Evaluer l'influence des variables dans le modèle : les tests d'hypothèses

Distribution de $\hat{\beta}$

Proposition : Sous les hypothèses standard du modèle de régression linéaire et l'hypothèse de normalité des termes d'erreurs :

1. L'estimateur $\hat{\beta}_{MCO} \rightarrow N(\beta, \sigma^2(X'X)^{-1})$
2. L'estimateur s^2 de σ^2 , convenablement normalisé, est distribué selon une loi Chi-deux

$$(T - k) \frac{s^2}{\sigma^2} \sim \chi^2(T - k)$$

3. $\hat{\beta}_{MCO}$ et s^2 sont indépendants.

2.2.5 Evaluer l'influence des variables dans le modèle : les tests d'hypothèses

Résultat central pour les tests

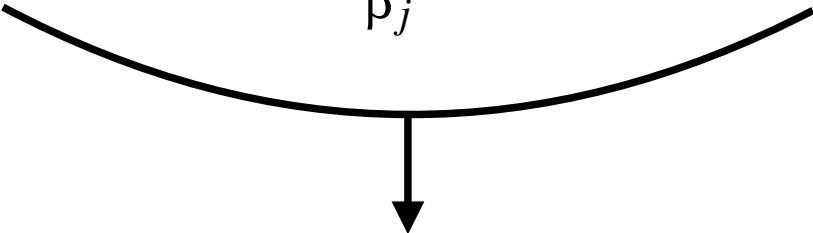
Résultat central : Sous les hypothèses standard du modèle de régression linéaire et l'hypothèse de normalité des termes d'erreurs, pour une composante j donnée du paramètre β , on a :

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t(T - k)$$

2.2.5 Evaluer l'influence des variables dans le modèle : les tests d'hypothèses

Résultat central pour les tests

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t(T - k)$$



On peut le mettre en oeuvre dans différents schémas :

- Test de significativité : $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$
- Test de conformité à un standard : $H_0 : \beta_j = c$ contre $H_1 : \beta_j \neq c$
- Intervalle de confiance au niveau ($1 - \alpha$)

Variance de l'erreur

© Théo Jalabert

$$\Omega_{\hat{\beta}} = \sigma^2 (X'X)^{-1}$$



$$\begin{aligned}\hat{\Omega}_{\hat{\beta}} &= \hat{\sigma}^2 (X'X)^{-1} \\ &= s^2 (X'X)^{-1}\end{aligned}$$

On peut démontrer que :

$$s^2 = \frac{e'e}{T - k} \text{ est tel que } E(s^2) = \sigma^2$$

Développons le résidu

$$\begin{aligned}e &= Y - \hat{Y} \\ &= (X\beta + u) - X\hat{\beta} \\ &= (X\beta + u) - X[\beta + (X'X)^{-1}X'u] \\ &= [I - X(X'X)^{-1}X']u\end{aligned}$$

.....

Appelée matrice M ,
symétrique et idempotente, de
taille (T, T)

$$e'e = u'M'Mu = u'Mu$$

On montre alors que :

$$E[e'e] = \sigma^2 \text{tr}(M)$$

$$= \sigma^2(T - k)$$

.....

Degrés de liberté

2.2.5 Evaluer l'influence des variables dans le modèle : les tests d'hypothèses

Inférence statistique sur les coefficients

Le test de significativité d'une variable :

La première étape pour effectuer des tests statistiques d'hypothèses est d'établir une hypothèse nulle et une hypothèse alternative.

Dans les tests de significativité, l'hypothèse nulle est que contrairement à la suggestion de la théorie économique, il n'y a pas de relations entre la variable à expliquer et chacune des variables explicatives. En d'autres termes, le modèle économique est faux.

Le test de significativité d'une variable :

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0$$

Statistique de test :

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}}$$

Règle de décision : rejet de H_0 si $|t| > t_{\alpha/2, T-k}$



valeur critique lue dans la table de Student pour un risque $\alpha/2$ et $(T - k)$ degrés de liberté

Interprétation :

Lorsque l'on rejette l'hypothèse H_0 , on dit que le coefficient est *statistiquement significatif*. La variable explicative associée a bien un impact significatif sur la variable à expliquer.

Le test de significativité d'une variable :

exemple : performances scolaires

$$api00_i = \beta_0 + \beta_1 enroll + \beta_2 mobility_i + \beta_3 ell_i + \beta_4 avg_{ed} + \beta_5 acsk3_i + \varepsilon_i$$

avec $api00_i$ une mesure des performances scolaires de l'étudiant i (plus cette variable est élevée, plus l'élève est brillant), $enroll$ le nombre moyen d'élèves dans ses classes de primaire, $mobility$ est une dichotomique qui vaut 1 si l'élève ne fréquentait pas cette école l'année précédente et 0 sinon, ell vaut 1 si la langue principale parlée à la maison n'est pas le français et 0 sinon, et avg_{ed} le niveau d'éducation moyen des parents et $acsk3$ vaut 1 si l'élève a débuté l'école l'année de ses 3 ans et 0 sinon.

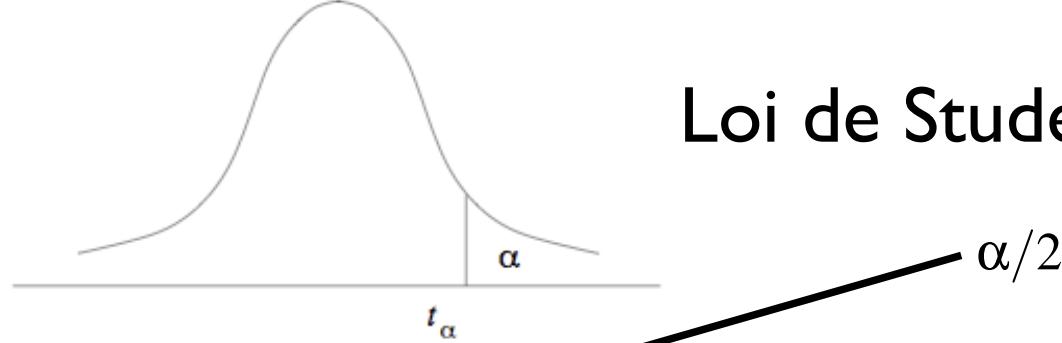
Le test de significativité d'une variable :

exemple : performances scolaires

$$(T - k)$$

Source	SS	df	MS	Number of obs	=	379
Model	5742392	5	1148478.4	F(5, 373)	=	???????
Residual	1937067.75	373	5193.21113	Prob > F	=	0.0000
Total	7679459.75	378	20316.0311	R-squared	=	???????
				Adj R-squared	=	???????
				Root MSE	=	72.064
<hr/>						
api00	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
enroll	-.029585	.0180449	-1.64	0.102	-.0650675	.0058974
mobility	-2.327182	.5240995	-4.44	0.000	???????????	???????????
ell	-2.431755	.2302196	-10.56	0.000	-2.884446	-1.979064
avg_ed	84.95083	7.076926	12.00	0.000	71.03515	98.8665
acs_k3	13.33766	2.81908	4.73	0.000	7.794376	18.88094
_cons	298.5516	59.51386	5.02	0.000	181.5269	415.5763

Loi de Student


 $(T - k)$

d.d.l.	α	0.10	0.05	0.025	0.01	0.005
1		3.078	6.314	12.706	31.821	63.657
2		1.886	2.920	4.303	6.965	9.925
3		1.638	2.353	3.182	4.541	5.841
4		1.533	2.132	2.776	3.747	4.604
5		1.476	2.015	2.571	3.365	4.032
6		1.440	1.943	2.447	3.143	3.707
7		1.415	1.895	2.365	2.998	3.499
8		1.397	1.860	2.306	2.896	3.355
20		1.325	1.725	2.086	2.528	2.845
21		1.323	1.721	2.080	2.518	2.831
22		1.321	1.717	2.074	2.508	2.819
23		1.319	1.714	2.069	2.500	2.807
24		1.318	1.711	2.064	2.492	2.797
25		1.316	1.708	2.060	2.485	2.787
26		1.315	1.706	2.056	2.479	2.779
27		1.314	1.703	2.052	2.473	2.771
28		1.313	1.701	2.048	2.467	2.763
29		1.311	1.699	2.045	2.462	2.756
∞		1.282	1.645	1.960	2.326	2.576

Le test de significativité d'une variable :

Autres possibilités pour tester la significativité :

- *utilisation des p-value*

p-value : seuil de risque tel que la valeur testée est exactement à la limite entre la zone de rejet et la zone d'acceptation; probabilité de rejeter à tort l'hypothèse nulle.

Si p-value < seuil de risque (5%), on rejette H0.

- *utilisation des intervalles de confiance*

Définition : Un intervalle de confiance pour le paramètre β_j au niveau α est un intervalle $[a, \bar{a}]$ tel que $P(\beta_j \in [a, \bar{a}]) = 1 - \alpha$

Le test de significativité d'une variable :

- utilisation des intervalles de confiance

Sous les hypothèses usuelles et de normalité, l'intervalle de confiance au niveau α est :

$$[\hat{\beta}_j - \hat{\sigma}_{\hat{\beta}_j} \cdot t_{\alpha/2, T-k}; \hat{\beta}_j + \hat{\sigma}_{\hat{\beta}_j} \cdot t_{\alpha/2, T-k}]$$

Si 0 n'appartient pas à l'intervalle de confiance, la variable explicative associée au coefficient est significative au seuil de risque α .

En statistique, la fiabilité d'un estimateur ponctuel est mesurée par son écart-type. Ainsi, au lieu d'avoir confiance dans un seul estimateur ponctuel, on peut construire un intervalle autour de cet estimateur, de sorte que cet intervalle ait, par exemple, 95% de chances d'inclure la valeur effective du paramètre.

Le test de significativité d'une variable :

- utilisation des intervalles de confiance

Sous les hypothèses usuelles et de normalité, l'intervalle de confiance au niveau α est :

$$[\hat{\beta}_j - \hat{\sigma}_{\hat{\beta}_j} \cdot t_{\alpha/2, T-k}; \hat{\beta}_j + \hat{\sigma}_{\hat{\beta}_j} \cdot t_{\alpha/2, T-k}]$$

Si 0 n'appartient pas à l'intervalle de confiance, la variable explicative associée au coefficient est significative au seuil de risque α .

En statistique, la fiabilité d'un estimateur ponctuel est mesurée par son écart-type. Ainsi, au lieu d'avoir confiance dans un seul estimateur ponctuel, on peut construire un intervalle autour de cet estimateur, de sorte que cet intervalle ait, par exemple, 95% de chances d'inclure la valeur effective du paramètre.

exemple : performances scolaires

Source	SS	df	MS	Number of obs	=	379
Model	5742392	5	1148478.4	F(5, 373)	=	???????
Residual	1937067.75	373	5193.21113	Prob > F	=	0.0000
Total	7679459.75	378	20316.0311	R-squared	=	???????
				Adj R-squared	=	???????
				Root MSE	=	72.064

api00	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
enroll	-.029585	.0180449	-1.64	0.102	-.0650675 .0058974
mobility	-2.327182	.5240995	-4.44	0.000	?????????? ??????????
ell	-2.431755	.2302196	-10.56	0.000	-2.884446 -1.979064
avg_ed	84.95083	7.076926	12.00	0.000	71.03515 98.8665
acs_k3	13.33766	2.81908	4.73	0.000	7.704376 18.88094
_cons	298.5516	59.51386	5.02	0.000	181.5269 415.5763

$$t_{\alpha/2, T-k} = 1,96$$

IC :

$[-2,327182 - 0,5240995 * 1,96 ; -2,327182 + 0,5240995 * 1,96]$
 $[-3,354417 ; -1,299947]$

Le test de conformité :

$$H_0 : \beta_j = c \text{ contre } H1 : \beta_j \neq c$$

Statistique de test : $t = \frac{\hat{\beta}_j - c}{\hat{\sigma}_{\hat{\beta}_j}}$



Règle de décision : rejet de H_0 si $|t| > t_{\alpha/2, T-k}$

Exemple : Fonction de consommation - revenu

La théorie keynésienne de la consommation est-elle vérifiée ? Selon Keynes, “en moyenne et la plupart du temps les hommes tendent à accroître leur consommation à mesure que leur revenu croît, mais non d'une quantité aussi grande que l'accroissement du revenu”.

Exemple : Fonction de consommation - revenu

. reg conso pib

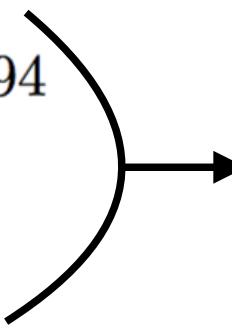
Source	SS	df	MS	Number of obs	= 22
Model	989.633	1	989.633	F(1, 20)	= 678.11
Residual	29.1888968	20	1.45940484	Prob > F	= 0.0000
Total	1018.8211	21	48.5152903	R-squared	= 0.9714
				Adj R-squared	= 0.9699
				Root MSE	= 1.2881

conso	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pib	.7443224	.0285833	26.04	0.000	.6846987 .883946
_cons	-72.04406	10.6845	-6.74	0.000	-94.33153 -49.75658

$$H_0 : \beta_{pib} = 1 \text{ vs. } H_1 : \beta_{pib} \neq 1$$

$$t = \frac{0,7443224 - 1}{0,0285833} = -8,9449994$$

$$t_{\alpha/2,T-k} = 2.086$$



on rejette H_0 :
 « confirmation » de
 l'hypothèse de la théorie de
 Keynes

Test sur un bloc de variables : Test de « c » contraintes linéaires sur les coefficients

$$H_0 : R.\beta = r \text{ vs. } H_1 : R.\beta \neq r$$

r : vecteur colonne (c, l) contenant les 2nd s membres des contraintes (au nombre de c)

R : matrice de dimension (c, k) et de rang q , contenant les coefficients des éléments de β dans chacune des contraintes.

La contrainte de rang imposée sur R exclut la présence de contraintes redondantes

test de significativité

$$\begin{cases} R = \begin{pmatrix} 0 & 1 & \cdots & 0 \end{pmatrix} \\ r = \begin{pmatrix} 0 \end{pmatrix} \end{cases}$$

Cas particuliers
test d'égalité de 2 coefficients

$$\begin{cases} R = (0 \ 1 \ -1 \ 0 \dots 0) \\ r = \begin{pmatrix} 0 \end{pmatrix} \end{cases}$$

test de significativité globale

$$R = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & & & & \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}; r = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Test sur un bloc de variables : Test de « c » contraintes linéaires sur les coefficients

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

Sous $\downarrow H_0$

$$R\hat{\beta} - r \sim N(0, \sigma^2 R(X'X)^{-1} R')$$

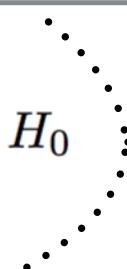
$$(R\hat{\beta} - r)' [\sigma^2 R(X'X)^{-1} R']^{-1} (R\hat{\beta} - r) \sim \chi^2(c)$$

Statistique de test :

$$F = \frac{(R\hat{\beta} - r)' [R(X'X)^{-1} R']^{-1} (R\hat{\beta} - r)/c}{s^2} \sim F(c, T - k)$$

↑
⋮
⋮

Sous H_0



$$\frac{(R\hat{\beta} - r)' [\sigma^2 R(X'X)^{-1} R']^{-1} (R\hat{\beta} - r)/c}{\frac{(T-k)s^2}{\sigma^2(T-k)}} \sim F(c, T - k)$$

$$(T - k) \frac{s^2}{\sigma^2} \sim \chi^2(T - k)$$

Règle de décision :

Si $F > F_\alpha$: on rejette H_0

Exemple : test de significativité globale

. reg conso pib

Source	SS	df	MS	Number of obs =	22
Model	989.633	1	989.633	F(1, 20) =	678.11
Residual	29.1888968	20	1.45940484	Prob > F =	0.0000
Total	1018.8211	21	48.5152983	R-squared =	0.9714
				Adj R-squared =	0.9699
				Root MSE =	1.2081

conso	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pib	.7443224	.0285833	26.04	0.000	.6846987 .883946
_cons	-72.04406	10.6845	-6.74	0.000	-94.33153 -49.75658

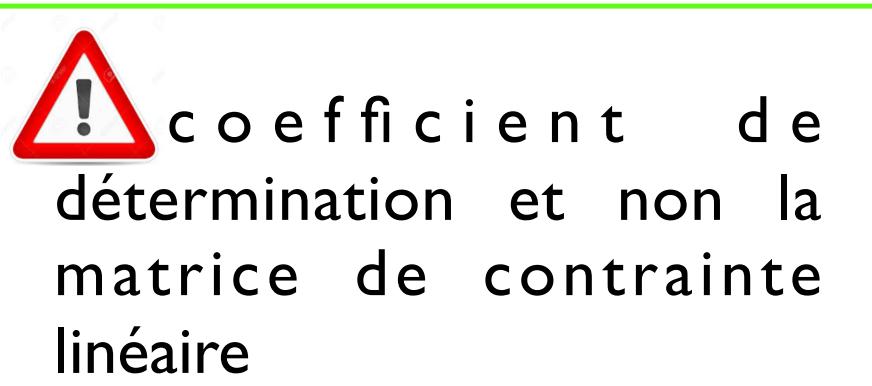
On rejette H_0 : tous les coefficients (hors constante) ne sont pas nuls. Il y a au moins une variable explicative significative. Le modèle a un certain pouvoir explicatif.

Exemple : test de significativité globale : $H_0 : \beta_2 = \dots = \beta_k = 0$

Dans le cas particulier du test de significativité globale, on peut montrer que la statistique de test de Fisher peut aussi se calculer de la manière suivante :

Sous H_0

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



exemple : performances scolaires

Source	SS	df	MS	Number of obs	=	379
Model	5742392	5	1148478.4	F(5, 373)	=	???????
Residual	1937067.75	373	5193.21113	Prob > F	=	0.0000
Total	7679459.75	378	20316.0311	R-squared	=	???????
				Adj R-squared	=	???????
				Root MSE	=	72.064

api00	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
enroll	-.029585	.0180449	-1.64	0.102	-.0650675 .0058974
mobility	-2.327182	.5240995	-4.44	0.000	?????????? ????????
ell	-2.431755	.2302196	-10.56	0.000	-2.884446 -1.979064
avg_ed	84.95083	7.076926	12.00	0.000	71.03515 98.8665
acs_k3	13.33766	2.81908	4.73	0.000	7.794376 18.88094
_cons	298.5516	59.51386	5.02	0.000	181.5269 415.5763

2.2.6 Utilisation des résultats comme aide à la décision

© Théo Jalabert



Il faut absolument mettre en perspectives les résultats lus à partir des logiciels.

- que peut-on en tirer de manière opérationnelle ?
(politique publique, stratégies commerciales, etc.)
- prédictions : prendre des profils « types » et faire des prédictions à partir de notre modèle