Chapitre 1 : Rappels et regression linéaire

Econométrie 1

Licence 3 Economie-Gestion E2AD-EGL, MIASHS-Economie

A. Fadhuile (adelaide.fadhuile@univ-grenoble-alpes.fr)
Univ Grenoble Alpes

Année 2023-2024

1 – Introduction

1.1 – Objectifs du chapitre

- Rappels de Statistiques Descriptives : moyennes, variances, covariances, corrélations, regression et droite d'ajustement.
- Positionnement économique de la méthode
- Applications grâce à l'utilisation d'un logiciel: "gretl"

1 – Introduction

1.1 – Objectifs du chapitre

- Ce que vous savez ?
  - Faire un ajustement linéaire
  - Résoudre un programme de maximisation/d'optimisation
  - Calculer : une espérance, une variance, un coefficient de corrélation, un intervalle de confiance.
- Ce que nous allons apprendre ?
  - Estimateur des Moindres Carrés Ordinaires (MCO)
  - Estimation des paramètres par les MCO.
  - Définition de la précision d'un estimateur
  - Définitions et interprétations des hypothèses
  - Définitions et interprétation des paramètres estimés
- Applications à deux exemples

Exemple 1 Une fonction de production $\Rightarrow$ production expliquée par la quantité de travail

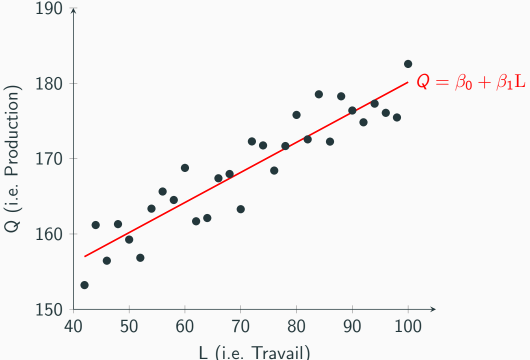
Exemple 2 Une fonction de gain $\Rightarrow$ niveau de salaire expliqué par le nombre d'années d'études

1 – Introduction

1.2 – Intuitions graphiques

1.2.1 – *Exemple 1*

Figure 1: Production (Q) en fonction du travail (L), N=30

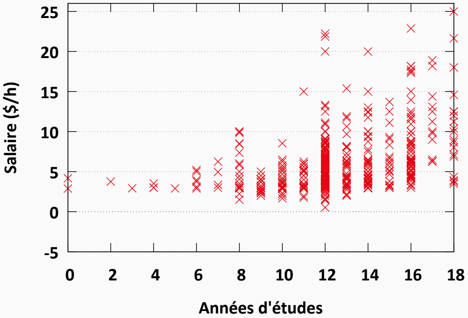


1 – Introduction

1.2 – Intuitions graphiques

1.2.2 – Exemple 2

Figure 2: Salaire en fonction du nombre d'années d'études, N=568

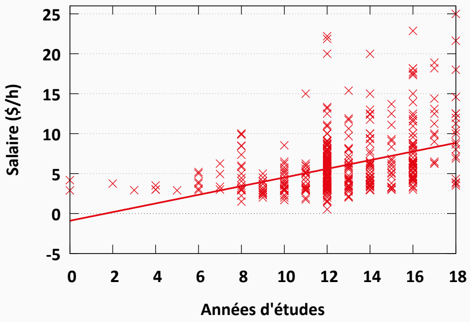


1 – Introduction

1.2 – Intuitions graphiques

1.2.2 – Exemple 2

Figure 3: Salaire en fonction du nombre d'années d'études, N=568



Plan du cours

2 – Rappels

2.1 – Les données

2.1.1 – La table des données

- Considérons une population comprenant n individus.
- Soit x_i la valeur de la variable x relative à l'individu i.
  - Elle peut prendre les valeurs $1, 2, ..., n$ auxquelles correspondent les valeurs de la variable

$$x_1, x_2, ..., x_i, ..., x_n$$

- La table de données sera du type:

Individu	y	x
1	y_1	x_1
2	y_2	x_2
$\vdots$	$\vdots$	$\vdots$
i	y_i	x_i
$\vdots$	$\vdots$	$\vdots$
n	y_n	x_n

Table 2.1: Exemple d'une table de données

2 – Rappels

2.2 – La nature de la base de données

- Trois types de données peuvent être mobilisées pour réaliser des analyses économétrique
  1. Des données individuelles en **coupe transversale**, elles seront indicées en i
     - Il y aura une observation par individu
     - Ex: des consommateurs, des entreprises, des pays...
  2. Des données temporelles en **série temporelles**, elles seront indicées en t
     - Ex : des années, des trimestres, des mois...
  3. Les 2 dimensions (1+2) , on parlera de **données de panel**, elles seront indicées en it (les 2 en même temps).
- **Méthodes économétriques devront être adaptées**

2 – Rappels

2.3 – La nature des variables

- Rappel (L1-Stat) : les variables peuvent être quantitative et/ou qualitatives
  - Quantitatives : sont **toujours numériques**...
    - **discrètes** : la variable prend ses valeurs dans un **ensemble dénombrable**. *Ex: nombre d'enfants*
    - **continues** : la variable prend ses valeurs dans un **ensemble continu** *Ex: salaire, PIB*
  - Qualitatives
    - **catégorielles**: aucun ordre naturel n'est possible *Ex: origine du bac, sexe*
    - **ordinales**: un ordre existe *Ex: niveau de diplôme, niveau de satisfaction*
  - Remarque: des codes numériques affectés à ces variables n'en font pas pour autant des variables quatitatives

    Ex: nomenclature des CSP avec des codes de professions de 1 à 8; les numéro des départements/des régions.
- **Attention :** L'interprétation changera selon la nature de chaque variable

2 – Rappels

2.4 – Notation symbolique d'une Sommes

- Σ (sigma) : symbolise la **Somme des valeurs** x_i de la variable x.
  - Par définition : $\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_i + \cdots + x_n$
  - *Ex: Si* $x_1 = 2$, $x_2 = 5$, *et* $x_3 = 6$, *alors:* $N = 3 \Rightarrow \sum_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 2 + 5 + 6 = 13$
- Quelques propriétés

$$\sum_{i=1}^{n}(x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n}(ax_i) = a\sum_{i=1}^{n} x_i, \text{ (où a est une constante)}$$

$$\sum_{i=1}^{n} x_i y_i \neq (\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i)$$

2 – Rappels

2.5 – Rappels statistiques descriptives

$$\overline{a} = \frac{\sum_{i=1}^{N} a_i}{N}$$

$$var(a) = \frac{1}{N}\sum_{i=1}^{N}(a_i - \overline{a})^2 = \frac{1}{N}\sum_{i=1}^{N} a_i^2 - \overline{a}^2 \Rightarrow \sigma_a = \sqrt{VAR(a)}$$

$$cov(a,b) = \frac{1}{N}\sum_{i=1}^{N}(a_i - \overline{a})(b_i - \overline{b}) = \frac{1}{N}\sum_{i=1}^{N} a_i b_i - \overline{a}\overline{b}$$

$$corr(a,b) = \rho_{ab} = r(a,b) = \frac{cov(a,b)}{\sqrt{var(a)var(b)}} = \frac{\sigma_{ab}}{\sigma_a \sigma_b}$$

Plan du cours

Rappels

Les modèles linéaires déterministes

- Exemple d'une fonction de production
- Forme fonctionnelle et interprétation
- Modèle niveau-niveau
- Modèle log-niveau
- Modèle niveau-log

Modèles linéaires aléatoires

Propriétés des MCO

3 – Les modèles linéaires déterministes

3.1 – Exemple d'une fonction de production

- Considérons une fonction de production de type *Cobb-douglas*

$$Q = AL^{\alpha}K^{\beta}$$

  - avec Q la production, K le capital et L le travail.
  - Le niveau de production est supposé endogène,
  - est déterminé par les niveaux de travail et de capital (variables exogènes)
- La fonction de production peut être linéarisée en appliquant le log

$$\log Q = \log A + \alpha \log L + \beta \log K$$

- En généralisant à N firmes i, chaque firme à la fonction de production:

$$\log Q_i = b_0 + b_1 \log L_i + b_2 \log K_i$$

$$\text{avec:} b_0 = \log A \text{ et } b_1 = \alpha \text{ et } b_2 = \beta$$

3 – Les modèles linéaires déterministes

3.1 – Exemple d'une fonction de production

3.1.1 – *Elasticité et Productivité du facteur travail*

Soit le modèle : $\log Q_i = b_0 + b_1 \log L_i + b_2 \log K_i$

- b_1: élasticité du facteur travail:

$$b_1 = \frac{\partial \log Q}{\partial \log L} = \frac{\partial Q}{\partial L} \times \frac{L}{Q}$$

- **Productivité marginale** ?

$$Pm_L = \frac{\partial Q}{\partial L}$$

- Comment passer de l'un à l'autre ?

$$b_1 = \frac{\partial \log Q}{\partial \log L} = \frac{\partial Q}{\partial L} \times \frac{L}{Q}$$

$$\Leftrightarrow Pm_L = \frac{\partial Q}{\partial L} = \frac{Q}{L} \times b_1$$

- Avec des données: **estimation** de l'élasticité et la productivité marginale du facteur travail pour une population donnée.

3 – Les modèles linéaires déterministes

3.1 – Exemple d'une fonction de production

3.1.2 – *Elasticité et Productivité du facteur capital ?*

Soit le modèle : $\log Q_i = b_0 + b_1 \log L_i + b_2 \log K_i$

- b_2 : élasticité du facteur capital:

$$b_2 =$$

- **Productivité marginale** ?

$$Pm_K =$$

-

- Comment passer de l'un à l'autre ?

$$b_2 =$$

$$\Leftrightarrow Pm_K =$$

3 – Les modèles linéaires déterministes

3.1 – Exemple d'une fonction de production

3.1.3 – b0 ?

$$\text{Soit le modèle : } \log Q_i = b_0 + b_1 \log L_i + b_2 \log K_i$$

- C'est la valeur de $\log Q_i$ si $\log L_i = 0$ et $\log K_i$
- Autrement dit, c'est le log du niveau de production commun à toutes les observations, mais qui n'est expliqué ni par le travail, ni par le capital.
- Comment revenir au niveau de production ?
    -

3 – Les modèles linéaires déterministes

3.2 – Forme fonctionnelle et interprétation

- 4 "grandes familles" de formes fonctionnelles
  - Modèle niveau-niveau
  - Modèle log-log
  - Modèle log-niveau
  - Modèle niveau-log
- Dans tous les cas, la forme fonctionnelle est linéaire en paramètres.

3 – Les modèles linéaires déterministes

3.3 – Modèle niveau-niveau

- Si $y = b_0 + b_1 x_1 + b_2 x_2$
  - b_1 est un **effet marginal** car:

$$b_1 = \frac{\partial y}{\partial x_1}$$

  - Une augmentation de x d'**UNE unité** augmente y de b_1 **UNITÉS**
  - On parle de spécification en **Niveau-Niveau**
- **Il est possible de passer de l'élasticité à l'effet marginal et inversement.**

3 – Les modèles linéaires déterministes

3.3 – Modèle niveau-niveau

3.3.1 – Modèle log-log

- Considérons le modèle économique suivant:

$$\log y = b_0 + b_1 \log x_1 + b_2 \log x_2$$

$$b_1 = \frac{\partial \log y}{\partial \log x_1} = \frac{\frac{\partial y}{y}}{\frac{\partial x_1}{x_1}} = \frac{\partial y}{\partial x_1} \times \frac{x_1}{y}$$

  - b_1 est une **élasticité**
  - Une augmentation de x d'**1 %** augmente y de b_1 **%**
  - On parle de spécification **log-log**

3 – Les modèles linéaires déterministes

3.4 – Modèle log-niveau

- Si $\log y = b_0 + b_1 x_1 + b_2 x_2$

$$b_1 = \frac{\partial \log y}{\partial x_1} = \frac{\frac{\partial y}{y}}{\partial x_1}$$

  - Alors : b_1 est une **semi-élasticité** car:
  - Une augmentation de x_1 d'**UNE UNITÉ** augmente y de $b_1 * 100$ **%**
  - On parle de spécification en **Log-Niveau**

3 – Les modèles linéaires déterministes

3.5 – Modèle niveau-log

- Si $y = b_0 + b_1 \log x_1 + b_2 x_2$
    - $$b_1 = \frac{\partial y}{\partial \log x_1} = \partial y \times \frac{x_1}{\partial x_1}$$
    - Une augmentation de x d'**UN %** augmente y de $b_1/100$ **UNITÉS**
    - On parle de spécification en **Niveau-log**

3 – Les modèles linéaires déterministes

3.5 – Modèle niveau-log

- Comment choisir la **forme fonctionnelle** adaptée ?
  - Adapter à la nature de chaque variable : quantitative, qualitative (discrete, ordonnée)
  - Analyser graphiquement de la relation entre les variables et les statistiques descriptives.
  - Analyser statistiques de la relation entre les variables : statistiques descriptives (e.g. corrélations, covariance)

Plan du cours

4 – Modèles linéaires aléatoires

- En toute généralité, le modèle de régression linéaire s'écrit:

$$y_i \quad = \quad \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i \;\; \forall i = 1, \ldots, N$$

  - k est le nombre de variables explicatives du modèle
  - x_{ji} sont les variables explicatives
  - les β_k sont les paramètres du modèle
  - u_i est un terme d'erreur
- Suite à une variation de **chacun** des x_k d'1 unité, y varie de β_k : **effet marginal**.
- Le raisonnement s'effectue toutes choses égales par ailleurs $\Rightarrow$ i.e. **une seule variable varie d'une seule unité.** $\Leftrightarrow$ **CETERIS PARIBUS**.
- **Dans ce chapitre nous allons considérer une seule variable explicative :** x_{1i}

4 – Modèles linéaires aléatoires

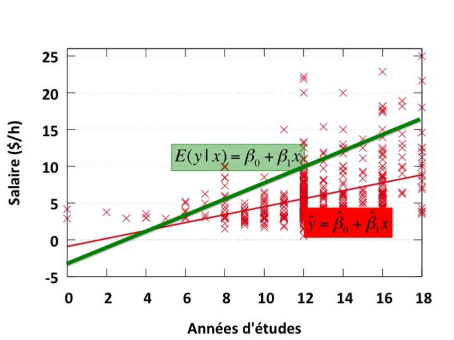
- Avec le modèle suivant :

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i \;\; \forall i = 1, \cdots, N$$

  et selon les livres, les termes suivants sont équivalents :
  - y_i est la variable dépendante/variable expliquée/variable de réponse/variable endogène
  - x_i est la variable explicative/la variable indépendante/covariable/variable de contrôle/variable exogène
  - β_0 et β_1 sont des paramètres
  - u_i est le terme d'erreurs
    - $\Rightarrow$ Traduit les perturbations qui affectent y, mais qui ne proviennent pas de x.
    - $\Rightarrow$ Ensemble des facteurs inobservées
- Le modèle de regression linéaire simple **cherche à expliquer y en fonction de x_1**
- Il s'agit de **l'équation à ESTIMER**.

4 – Modèles linéaires aléatoires

Figure 4: Droite de régression des MCO et fonction (inconnue) de régression de la population



4 – Modèles linéaires aléatoires

- Il est donc important d'incorporer un terme aléatoire dans le modèle : une **perturbation notée u**

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i$$

- Cette perturbation traduit principalement:
  - Les facteurs **inobservables**, qui illustrent les comportements des individus;
  - La forme fonctionnelle n'est pas forcément linéaire dans les paramètres;
  - Le fait que cette fonction peut varier en fonction du temps ou des individus observés;
  - L'omission de variables dans le modèle économique; Des erreurs de mesure sur les variables ... etc
- On peut donc avoir u_i à partir de l'équation:

$$y_i - \beta_0 - \beta_1 x_{1i} = u_i$$

4 – Modèles linéaires aléatoires

4.1 – Raisonnement

- Si les autres facteurs compris dans u sont maintenus constants alors

$$\Delta u = 0 \;\Rightarrow\; \Delta y = \beta_1 \Delta x \tag{1}$$

  - β_1 est le coefficient de la **pente** dans la relation entre y et x, et les autres facteurs dans u sont maintenus constants.
  - β_0 est la **constante** (ou ordonnée à l'origine).
- Remarque importante
  - β_1 mesure l'effet de x sur y en supposant que tous les autres facteurs sont fixes (y compris u)
  - Il va donc falloir poser des hypothèses sur x, u et leur lien !

4 – Modèles linéaires aléatoires

4.1 – Raisonnement

- L'espérance de l'erreur est nulle

$$E(u) = 0 \tag{2}$$

$\Leftrightarrow$ Les facteurs non observés ont une moyenne égale à zéro sur l'intégralité de la pop°

i.e. Ex2 *Les facteurs autres que le niveau d'étude, ont un effet moyen nul ds la pop°.*

- La valeur espérée de u (ou la moyenne de u) peut être décrite par la valeur de x pour une partie ds la pop°, quelque soit x:

$$E(u|x) = E(u) \tag{3}$$

  - la valeur moyenne des variables non observées est la même pour toutes les parties de la pop°
  - La moyenne commune à ces parties est égale à la moyenne de u sur l'ensemble de la pop°.
  - le terme d'erreur u n'est pas corrélé avec x dans la population.

i.e. Ex2 Supposons que u représente l'aptitude innée d'une personne $\Rightarrow$ non observable

- Cela veut donc dire que le niveau moyen de l'aptitude innée est identique quelque soit les nombre d'années d'études

4 – Modèles linéaires aléatoires

4.1 – Raisonnement

- En combinant les équations (2) et (3), on obtient :

$$\left.\begin{aligned} &E(u) = 0 \\ &E(u|x) = E(u) \end{aligned}\right\} \Rightarrow E(u|x) = E(u) = 0 \tag{4}$$

- L'espérance conditionnelle est égale à zéro (effet toutes choses égales par ailleurs).
- Cela permet d'écrire la **fonction de régression de la population**, notée $E(y|x)$:

$$E(y|x) = E(\beta_0 + \beta_1 x + u) = \beta_0 + \beta_1 x \tag{5}$$

- Cette équation donne la valeur moyenne de y pour différents niveaux de x

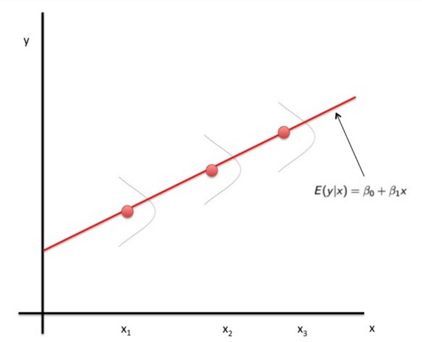
Ex2 C'est le gain de salaire que vous pouvez espérer avoir en moyenne après la licence.

- Attention, cela ne veut pas dire que chaque personne gagnera exactement E(y—x) ! certaines gagneront plus, et d'autres moins!

4 – Modèles linéaires aléatoires

4.1 – Raisonnement

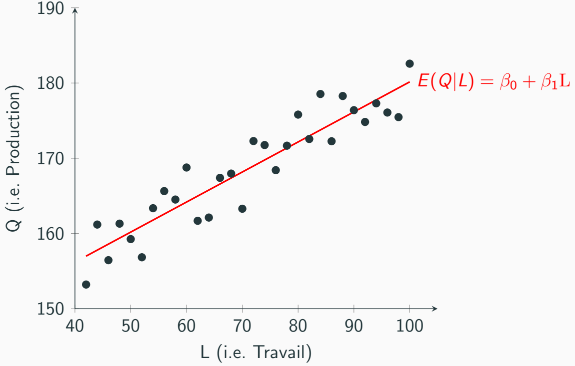
Figure 5: $E(y|x)$ en tant que fonction linéaire de x



4 – Modèles linéaires aléatoires

4.2 – Intuitions graphiques

Figure 6: Dispersion de la production (Q) et du travail (L), et fonction de regression de la population $E(Q|L) = \beta_0 + \beta_1 \mathrm{L}$



- Pour estimer les paramètres β_0 et β_1, nous avons besoin d'un échantillon issu de la population.

4 – Modèles linéaires aléatoires

4.3 – Dérivation de l'estimateur des MCO

- Utilisons les deux hypothèse précédentes ($E(u|x) = E(u) = 0$)
- Implication $\Rightarrow$ le terme d'erreur u n'est pas corrélé avec x dans la population.
- i.e. la valeur espérée de u est égale à zéro et la covariance entre x et u est aussi nulle.

$$E(u) = 0 \text{ et } cov(x, u) = E(x\, u) = 0 \tag{6}$$

- Cela implique que:

$$E(y - \beta_1 x - \beta_0) = 0 \text{ et } cov(x, u) = E(x(y - \beta_1 x - \beta_0)) = 0 \tag{7}$$

- Il y a donc 2 paramètres inconnus à estimer (β_0 et β_1) sous contrainte de minimiser la somme des erreurs au carré. Cela nécessite l'utilisation d'un échantillon pour trouver des **estimateurs fiables** de ces paramètres. Il seront notés : $\widehat{\beta}_0$ et $\widehat{\beta}_1$.

4 – Modèles linéaires aléatoires

4.3 – Dérivation de l'estimateur des MCO

- On va chercher β_0 et β_1 afin de minimiser la somme des carrés des écarts par rapport aux valeurs observées, ils minimisent la **Somme des Carrés des Résidus (SCR)**

$$\sum_{i=1}^{N} {u_i}^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{8}$$

- Les paramètres qui minimisent la **Somme des Carrés des Résidus (SCR)** sont: $\widehat{\beta}_0$ et $\widehat{\beta}_1$.
- Ces paramètres égalisent certains moments théoriques avec leurs contreparties empiriques :

$$n^{-1} \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0 \tag{9}$$

$$n^{-1} \sum_{i=1}^{n} x_i (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0 \tag{10}$$

4.3 – Dérivation de l'estimateur des MCO

- l'équation 9 devient

$$n^{-1}\sum_{i=1}^{n} y_i - \widehat{\beta}_0 - \widehat{\beta}_1 n^{-1}\sum_{i=1}^{n} x_i = 0 \Leftrightarrow \overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1\overline{x} \Leftrightarrow \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x} \tag{11}$$

- En réintroduisant l'équation 11 dans l'équation 10, on obtient:

$$n^{-1}\sum_{i=1}^{n} x_i(y_i - \overline{y} + \widehat{\beta}_1\overline{x} - \widehat{\beta}_1 x_i) = 0 \tag{12}$$

$$\Leftrightarrow n^{-1}\sum_{i=1}^{n} x_i(y_i - \overline{y}) = n^{-1}\sum_{i=1}^{n} x_i(-\widehat{\beta}_1\overline{x} + \widehat{\beta}_1 x_i) \tag{13}$$

$$\Leftrightarrow \sum_{i=1}^{n} x_i(y_i - \overline{y}) = \widehat{\beta}_1 \sum_{i=1}^{n} x_i(x_i - \overline{x}) \tag{14}$$

4 – Modèles linéaires aléatoires

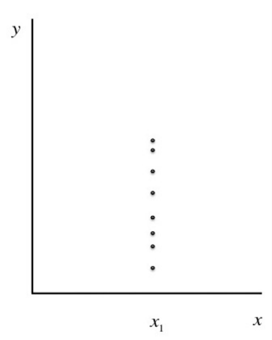
4.3 – Dérivation de l'estimateur des MCO

- Si x est constant pour toutes les observations, on a :

$$\sum_{i=1}^{n} x_i(x_i - \overline{x}) = 0$$

- On ne peut pas exploiter la différence entre les variables observées pour mesurer l'effet de x sur y.

Figure 7: Dispersion du niveau de production si la quantité de travail est toujours à x_1



4 – Modèles linéaires aléatoires

4.3 – Dérivation de l'estimateur des MCO

- L'équation 14 a une solution ssi $\sum_{i=1}^{n} x_i(x_i - \overline{x}) \neq 0$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \overline{y})}{\sum_{i=1}^{n} x_i(x_i - \overline{x})} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{cov(x, y)}{var(x)} \tag{15}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \tag{16}$$

- Interprétation
  - β_1 est la pente de la droite d'ajustement qui est égale à la covariance entre x et y divisée par la variance de x

4 – Modèles linéaires aléatoires

4.3 – Dérivation de l'estimateur des MCO

- $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont les **estimateurs** des Moindres Carrés Ordinaires (MCO) de β_0 et β_1.
- Pour tout $\widehat{\beta}_0$ et $\widehat{\beta}_1$, on peut calculer une valeur ajustée de y, notée $\widehat{y}$ telle que:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \tag{17}$$

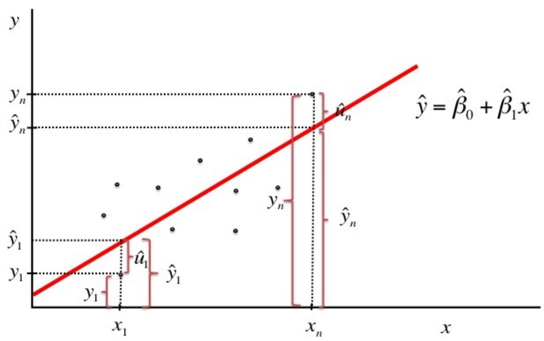
- S'il existe une valeur ajustée pour chaque observation, il existe aussi un **résidu** qui est égal à la différence entre la valeur ajustée et la valeur observée:

$$\widehat{u}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \tag{18}$$

4 – Modèles linéaires aléatoires

4.3 – Dérivation de l'estimateur des MCO

Figure 8: Valeurs ajustées et résidus



4.3 – Dérivation de l'estimateur des MCO

- Supposons que $\widehat{\beta}_0$ et $\widehat{\beta}_1$ doivent minimiser la somme des carrés des écarts par rapport aux valeurs observées, on dit qu'ils vont minimiser la **Somme des Carrés des Résidus (SCR)**
- Déterminés par:

$$\sum_{i=1}^{N} \widehat{u_i}^2 = \sum_{i=1}^{N} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \tag{19}$$

4 – Modèles linéaires aléatoires

4.3 – Dérivation de l'estimateur des MCO

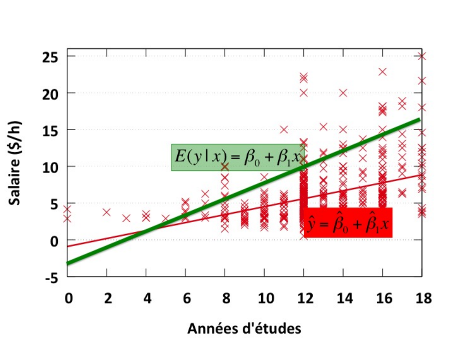
- Nous avons donc une droite de régression des MCO

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \tag{20}$$

4 – Modèles linéaires aléatoires

4.3 – Dérivation de l'estimateur des MCO

Figure 9: Droite de régression des MCO et fonction (inconnue) de régression de la population



4 – Modèles linéaires aléatoires

4.3 – Dérivation de l'estimateur des MCO

- Nous avons donc une droite de régression des MCO

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \tag{21}$$

- Mais, il s'agit de la version **estimée de la fonction de régression de l'échantillon**
- Les paramètres peuvent changer d'un échantillon à l'autre !
- Il est donc nécéssaire d'appréhender la notion de précision de l'estimateur par l'analyse de ses propriétés.

4 – Modèles linéaires aléatoires

4.4 – Application – Exemple 1

4.4.1 – Cadre général

Considérons l'**équation à estimer** pour expliquer le salaire suivante :

$$salaire_i = \beta_0 + \beta_1 etud_i + \epsilon_i$$

Afin d'analyser cette relation, nous dispsons d'une base de données composée de 526 observations, avec

- $wage_i$: le salaire horaire du salariés i en \$/heure
- $educ_i$: le nombre d'année d'études de l'individu i.

L'**équation estimée** est :

$$\widehat{salaire}_i = \widehat{\beta}_0 + \widehat{\beta}_1 etud_i$$

Estimer cette équation revient à estimer $\widehat{\beta}_0$ et $\widehat{\beta}_1$ ici par les MCO

4 – Modèles linéaires aléatoires

4.4 – Application – Exemple 1

4.4.2 – Estimation

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} etud_i(salaire_i - \overline{salaire})}{\sum_{i=1}^{n} etud_i(etud_i - \overline{etud})} = \frac{cov(etud, salaire)}{var(etud)} = \frac{2179.20}{4015.43} = 0.541$$

avec $\overline{salaire} = 5.89$ et $\overline{etud} = 12.56$

$$\widehat{\beta}_0 = \overline{salaire} - \widehat{\beta}_1 \overline{etud} = 5.89 - 12.56 \times 0.541 = -0.905$$

L'équation estimée est :

$$\widehat{salaire}_i = -0.905 + 0.541 etudes_i$$

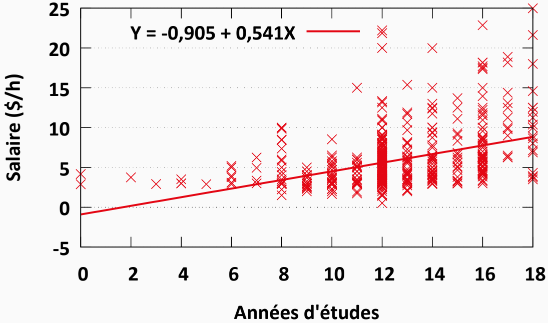
4 – Modèles linéaires aléatoires

4.4 – Application – Exemple 1

4.4.3 – *Interprétations*

$$\widehat{salaire_i} = -0.905 + 0.541 etudes_i$$

Figure 10: Salaire en fct du niveau d'études



- β_1 ?
  - $\beta_1 = \frac{\partial y}{\partial x} = \frac{\partial salaire}{\partial educ}$
  - La pente mesure le rendement d'une année d'étude supplémentaire sur le salaire.
  - $\widehat{\beta_1} = 0.541$: Effet marginal
  - Une année d'étude supplémentaire augmente le salaire de 0,541 unités.

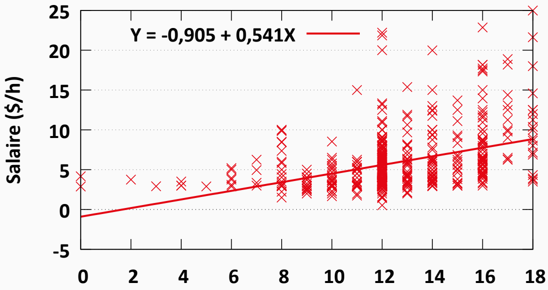
4 – Modèles linéaires aléatoires

4.4 – Application – Exemple 1

4.4.3 – Interprétations

$$\widehat{salaire}_i = -0.905 + 0.541 etudes_i$$

Figure 11: Salaire en fct du niveau d'études



- β_0 ?
  - $\widehat{\beta}_0 = -0.905$: constante
  - Si le nombre d'années d'études est nul

$$salaire_i = -0.905 + 0.541 \times 0 = -0.905$$

  - le salaire est de -0,9 si le nombre d'années d'études est nul.

4 – Modèles linéaires aléatoires

4.4 – Application – Exemple 1

4.4.4 – Valeurs ajustées

$$\widehat{salaire}_i = -0.905 + 0.541 etudes_i$$

Table 4.1: Salaire en fct du niveau d'études

ident	salaire	educ	$\widehat{salaire}$
1	3,1	11	0,905+0,541 *11=5,05
2	3,24	12	5,59
3	3	11	5,05
4	6	8	3,43
$\vdots$	$\vdots$	$\vdots$	$\vdots$
524	4,67	15	7,22
525	11,56	16	7,76
526	3,5	14	6,67

4 – Modèles linéaires aléatoires

4.4 – Application – Exemple 1

4.4.4 – Valeurs ajustées

$$\widehat{salaire}_i = -0.905 + 0.541 etudes_i$$

Table 4.2: Salaire en fct du niveau d'études

ident	salaire	educ	$\widehat{salaire}$
1	3,1	11	5,05
2	3,24	12	5,59
3	3	11	5,05
4	6	8	3,43
$\vdots$	$\vdots$	$\vdots$	$\vdots$
524	4,67	15	7,22
525	11,56	16	7,76

- $\widehat{salaire}_i - salaire_i \neq 0$
- Il y a bien un écart entre les **valeurs ajustées** et la valeurs observées de $salaire_i$ $C'estler\acute{e}sidudepr\acute{e}visiondumod\grave{e}le : \widehat{salaire}_i - salaire_i = \widehat{u}_i$

- selon "l'amplitude" de cette erreur, le modèle sera ou moins précis.

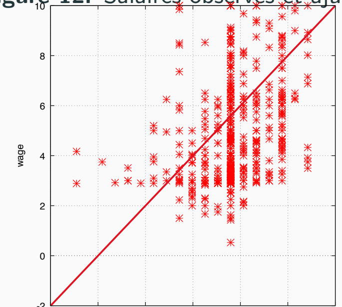
4 – Modèles linéaires aléatoires

4.4 – Application – Exemple 1

4.4.4 – Valeurs ajustées

$$\widehat{salaire}_i = -0.905 + 0.541 etudes_i$$

Figure 12: Salaires observés et ajustés



- $\widehat{salaire}_i - salaire_i \neq 0$
- Il existe une différence entre la **valeur ajustée** et la valeur observée de *$salaire_i$*
- C'est le résidu de prédit par le modèle:

$$\widehat{salaire}_i - salaire_i = \widehat{u}_i$$

- selon "l'amplitude" de cette erreur, le modèle sera ou moins précis.

UGA Université Grenoble Alpes

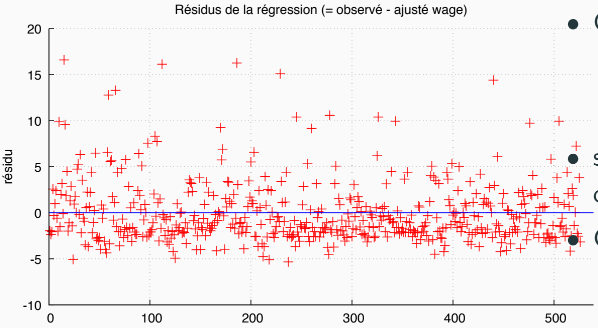
4 – Modèles linéaires aléatoires

4.4 – Application – Exemple 1

4.4.5 – *Résidus*

$$\widehat{salaire}_i = -0.905 + 0.541 etudes_i$$

Figure 13: Résidus par observations



- C'est le résidu de prévision du modèle:

$$\widehat{salaire}_i - salaire_i = \widehat{u}_i$$

- selon "l'amplitude" de cette erreur, le modèle sera ou moins précis
- Car on voudrait le "moins" d'écarts possibles"

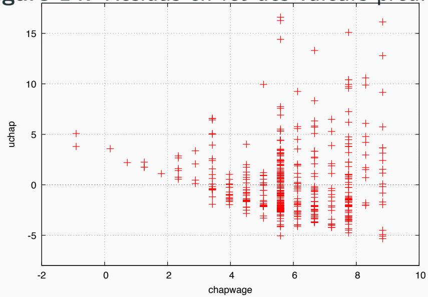
4 – Modèles linéaires aléatoires

4.4 – Application – Exemple 1

4.4.5 – Résidus

$$\widehat{salaire}_i = -0.905 + 0.541 etudes_i$$

Figure 14: Résidus en fct des valeurs prédites



- C'est le résidu de prévision du modèle:

$$\widehat{salaire}_i - salaire_i = \widehat{u}_i$$

- plus les $\widehat{salaire}_i$ sont importants, plus les $\widehat{u}_i$ sont dispersés.

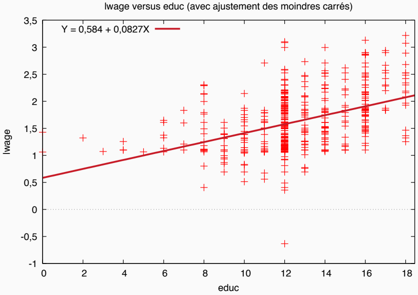
4 – Modèles linéaires aléatoires

4.4 – Application – Exemple 1

4.4.6 – Et si l'équation estimée était légèrement différente Considérons le modèle suivant:

$$\widehat{\log(salaire)}_i = \beta_0 + \beta_1 etudes_i$$

Figure 15: log Salaire en fct du niveau d'études



- Modèle Log-Niveau
- β_1 ?
  - $\beta_1 = \frac{\partial \log y}{\partial x} = \frac{\partial \log salaire}{\partial educ}$
  - Semi-élasticité
  - La pente mesure le rendement d'une année d'étude supplémentaire sur le **log** du salaire.
  - $\widehat{\beta_1} = 0.0827.$ Donc l'impact du nombre d'années d'études sur le niveau de salaire est de $0.0827 * 100\% = 8.27\%$
  - Cela veut dire que chaque année d'etude en plus augmente de salaire de 8.27%.

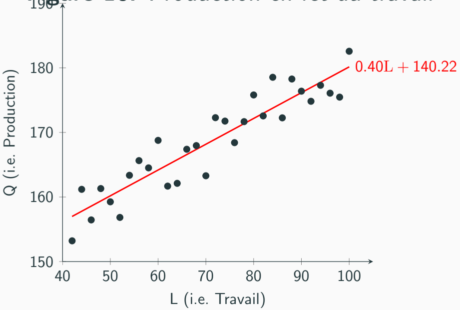
4 – Modèles linéaires aléatoires

4.5 – Application – Exemple 2

Considérons le modèle: $Q_i = \beta_0 + \beta_1 L_i + u_i$

Le modèle estimé s'écrit donc: $\widehat{Q}_i = \widehat{\beta}_0 + \widehat{\beta}_1 L_i \Leftrightarrow \widehat{Q}_i = 140.22 + 0.40 L_i$

Figure 16: Production en fct du travail



- $\widehat{\beta}_0 = 140.22$: constante
- $\widehat{\beta}_1 = 0.40$: Effet marginal
- Commenter

Plan du cours

5 – Propriétés des MCO

5.1 – Valeurs ajustées et résidus

- Supposons que $\widehat{\beta}_0$ et $\widehat{\beta}_1$ soient obtenus à partir d'un échantillon de données.
- On peut alors calculer $\widehat{y}_i$ et $\widehat{u}_i$
- Si $\widehat{u}_i > 0$, la droite des MCO "**sous-estime**" y_i
- Si $\widehat{u}_i < 0$, la droite des MCO "**sur-estime**" y_i
- Si $\widehat{u}_i = 0$, est le cas "**idéal**" car toutes les valeurs observées correspondent aux valeurs ajustées, elles se situent sur la droite des MCO.

5 – Propriétés des MCO

5.2 – Propriétés algébriques des statistiques issues de l'estimation par les MCO

5.2.1 – Somme des résidus nulle

- Si la sommes des résidus est nulle, alors la moyenne des résidus est également nulle.

$$\sum_{i=1}^{N} \widehat{u}_i = 0 \tag{22}$$

- Résultat qui vient naturellement, car $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont déterminés sous cette hypothèse.

5 – Propriétés des MCO

5.2 – Propriétés algébriques des statistiques issues de l'estimation par les MCO

5.2.2 – Covariance

- La covariance des variables explicatives et les résidus des MCO est nulle.
- Ce résultat découle de l'équation 10 que l'on peut réecrire de la façon suivante:

$$\sum_{i=1}^{N} x_i \widehat{u}_i = 0 \tag{23}$$

- Comme la moyenne des résidus est nulle, la partie gauche est proportionnelle à la covariance entre x_i et $\widehat{u}_i$

5 – Propriétés des MCO

5.2 – Propriétés algébriques des statistiques issues de l'estimation par les MCO

5.2.3 – Le point moyen est toujours sur la droite

- Par construction, **le point moyen est toujours sur la droite de regression**, donc:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i \tag{24}$$

$$\Leftrightarrow \overline{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x} \tag{25}$$

$$\text{et } \overline{y} = \overline{\widehat{y}} \tag{26}$$

5 – Propriétés des MCO

5.3 – Analyse de la variance

5.3.1 – *Equation d'analyse de la variance (ANOVA)*

$$\text{Variance de } y = \text{Variance de } \hat{y} + \text{Variance de } \hat{u}$$

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

- **SCT : Somme des carrés totaux**

  $\Rightarrow$ mesure de la variation totale entre les y_i, i.e. degré de dispersion des y_i dans l'échantillon
- **SCE : Somme des carrés expliqués**

  $\Rightarrow$ mesure de la dispersion au sein des $\widehat{y_i}$
- **SCR : Somme des carrés des résidus**

  $\Rightarrow$ mesure de la dispersion au sein des $\widehat{u_i}$

5 – Propriétés des MCO

5.3 – Analyse de la variance

5.3.2 – *Qualité de l'ajustement*

- On en déduit le **coefficient de détermination**, noté R^2

$$R^2 = \frac{\sum_{i=1}^{N} (\widehat{y}_i - \overline{y})^2}{\sum_{i=1}^{N} (y_i - \overline{y})^2} = \frac{SCE}{SCT}$$

- Ratio entre la variation expliquée par la variance totale
- Par construction il est **TOUJOURS** compris entre 0 et 1
- Autre écriture

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

5 – Propriétés des MCO

5.3 – Analyse de la variance

5.3.3 – *Interprétations*

- Par construction il est **TOUJOURS** compris entre 0 et 1
  - $R^2 = 1$: Si tous les points correspondant aux données se trouvent sur la droite d'ajustement.
  - $R^2 = 0$: Les variations entre les $\widehat{y_i}$ ne capturent quasiment rien de la variation observée entre les y_i.
  - Remarque: un faible R^2 n'implique pas forcément que la régression des MCO ne sert à rien! mais que d'autres "problèmes" peuvent expliquer ce résultat.
  - Il peut de fait aussi s'exprimer en pourcentage : $100 \times R^2$ % c'est le pourcentage de la variation de la variation de y présente dans l'échantillon qui est expliquée par x (compris entre 0 et 100%).

5 – Propriétés des MCO

5.3 – Analyse de la variance

5.3.4 – *Application 1 suite*

$$\text{Variance de } wage = \text{Variance de } \widehat{wage} + \text{Variance de } \widehat{u}$$

$$\underbrace{\sum_{i=1}^{N}(wage_i - \overline{wage})^2}_{SCT} = \underbrace{\sum_{i=1}^{N}(\widehat{wage_i} - \overline{wage})^2}_{SCE} + \underbrace{\sum_{i=1}^{N}(\widehat{u_i} - \overline{\widehat{u}})^2}_{SCR}$$

- **SCT : Somme des carrés totaux** $=7160.41$
- **SCE : Somme des carrés expliqués** $=1179.73$
- **SCR : Somme des carrés des résidus** $= 5980.68$

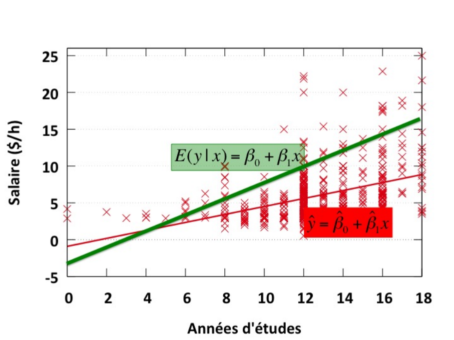
$\Rightarrow R^2 = \frac{SCE}{SCT} = \frac{1179.73}{5980.68} = 0.164$

- Interprétation :

Plan du cours

6 – Espérance et Variance

Figure 17: Droite de régression des MCO et fonction (inconnue) de régression de la population



⇒ Les droites rouges et vertes devraient être confondues!

On abordera la notion de **biais**.

6 – Espérance et Variance

Nous allons monter que sous 5 hypothèses "fondamentales", les estimateurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$ des MCO sont **sans biais ET à variance minimale**.

Ces hypothèses sont :

H_1: $E[u_i] = 0, \forall\ t$

H_2: H_2: $V(u_i) = E[(u_i - E(u_i))^2] = E[u_i^2] = \sigma_u^2 \forall\ i$

H_3: la variable explicative x_i est non aléatoire

H_4: le modèle est correctement spécifié

H_5: la variable explicative n'est pas constante pour toutes les observations

6 – Espérance et Variance

6.1 – Absence de biais de l'estimateur des MCO

- Qu'est ce qu'un estimateur sans biais ?
  - L'estimateur $\widehat{\beta}$ de β est sans biais si

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

- Cela veut dire qu'en moyenne le paramètre estimé est égal à la vraie valeur du paramètre.

6 – Espérance et Variance

6.1 – Absence de biais de l'estimateur des MCO

- Nous allons donc calculer : $E(\widehat{\beta}_0)$ et $E(\widehat{\beta}_1)$
- Et vérifier sous quel(les) hypothèse(s) on retrouve:

$$E(\widehat{\beta}_0) = \beta_0 \text{ et } E(\widehat{\beta}_1) = \beta_1$$

- Avec

$$\textbf{[H5]} \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^n x_i(y_i - \overline{y})}{\sum_{i=1}^n x_i(x_i - \overline{x})} \tag{27}$$

$$= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$$= \frac{\sum_{i=1}^n (x_i - \overline{x})y_i}{\sum_{i=1}^n x_i(x_i - \overline{x})}$$

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x} \tag{28}$$

6 – Espérance et Variance

6.1 – Absence de biais de l'estimateur des MCO

6.1.1 – $E(\widehat{\beta}_1)$

- Avec la spécification suivante: $y_i = \beta_0 + \beta_1 x_i + u_i$ **[H4]**
- Au numérateur:

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

$$= \beta_0 \sum_{i=1}^{n}(x_i - \overline{x}) + \beta_1 \sum_{i=1}^{n} x_i(x_i - \overline{x}) + \sum_{i=1}^{n} u_i(x_i - \overline{x})$$

- comme : $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$ et $\sum_{i=1}^{n} x_i(x_i - \overline{x}) = \sum_{i=1}^{n}(x_i - \overline{x})^2$, on obtient:

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n} u_i(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

6 – Espérance et Variance

6.1 – Absence de biais de l'estimateur des MCO

6.1.1 – $E(\widehat{\beta}_1)$

- Le calcul de l'espérance va nécessité d'utiliser d'hypothèses !!!

$$\begin{aligned}
E[\widehat{\beta}_1] &= E\left[\beta_1 + \frac{\sum_{i=1}^n u_i(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] \\
&= \beta_1 + E\left[\frac{\sum_{i=1}^n u_i(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}\right] \textbf{[H3 : Matrice des x non aléatoire]} \\
&= \beta_1 + \frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2} \sum_{i=1}^n [(x_i - \overline{x})E(u_i)] \textbf{[H1: [E[u]=0, et H3: E(x u)]]} \\
E[\widehat{\beta}_1] &= \beta_1
\end{aligned}$$

$\Rightarrow$ L'estimateur $\widehat{\beta}_1$ des MCO de β_1 est sans biais.

6 – Espérance et Variance

6.1 – Absence de biais de l'estimateur des MCO

6.1.2 – $E(\widehat{\beta}_0)$

- On sait que : $\overline{y} = \beta_0 + \beta_1 \overline{x} + \overline{u}$
- On vient de démontrer que $\widehat{\beta}_1$ était sans biais, donc $E(\widehat{\beta}_1) = \beta_1$
- Tant que x **n'est pas aléatoire**, on obtient:

$$
\begin{aligned}
\widehat{\beta}_0 &= \overline{y} - \widehat{\beta}_1 \overline{x} = \beta_0 + \beta_1 \overline{x} + \overline{u} - \widehat{\beta}_1 \overline{x} \\
&= \beta_0 + \overline{x}[\beta_1 - \widehat{\beta}_1] + \overline{u} \\
E[\widehat{\beta}_0] &= E\left[\beta_0 + \overline{x}[\beta_1 - \widehat{\beta}_1] + \overline{u}\right] \\
&= \beta_0 + \overline{x} E[\beta_1 - \widehat{\beta}_1] + E\left[\overline{u}\right] \\
&= \beta_0 + \overline{x}\beta_1 - \overline{x} \underbrace{E[\widehat{\beta}_1]}_{E[\widehat{\beta}_1] = \beta_1} + \underbrace{E\left[\overline{u}\right]}_{E[\overline{u}] = 0} = \beta_0
\end{aligned}
$$

6 – Espérance et Variance

6.1 – Absence de biais de l'estimateur des MCO

6.1.3 – Bilan

- L'estimateur des MCO est sans biais, sous les hypothèses suivantes:

  H_1: $E[u_i] = 0, \forall\ i$

  H_3: la variable explicative x_i est non aléatoire

  H_4: le modèle est correctement spécifié

  H_5: la variable explicative n'est pas constante pour toutes les observations

- Cela veut dire que : $E(\widehat{\beta}_0) = \beta_0$ et $E(\widehat{\beta}_1) = \beta_1$

6 – Espérance et Variance

6.1 – Absence de biais de l'estimateur des MCO

6.1.3 – Bilan

- Cette absence de biais implique que
  - Les coefficients estimés peuvent être plus forts ou plus faibles selon les échantillons
  - Mais en moyenne, ils seront égaux aux valeurs qui caractérisent la vraie relation entre y et x dans la population.
  - "En moyenne" : si on répète plusieurs fois l'estimation sur différents échantillons alors en moyenne on obtiendrait le même résultat.

6 – Espérance et Variance

6.2 – Calcul de la variance

- Nous avons vu que la distribution d'échantillonnage de $\widehat{\beta_1}$ est centrée sur β_1 ($\widehat{\beta_1}$ est sans biais).
- Nous allons chercher dans quelle mesure $\widehat{\beta_1}$ sera éloigné de β_1 en moyenne.
- La mesure de dispersion la plus fréquemment retenue pour une distribution est sa variance, ou son écart-type.
- Sous les hypothèses précédentes, on peut calculer les variances : $V(\widehat{\beta_0})$ et $V(\widehat{\beta_1})$

6 – Espérance et Variance

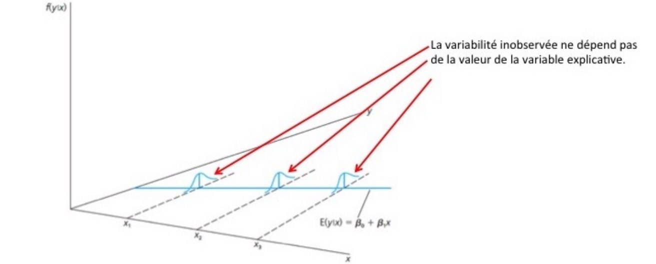
6.2 – Calcul de la variance

- Hypothèse complémentaire: la variance de l'erreur u, conditionnelle à x, est constante :

$$V(u|x) = \sigma_u^2$$

$\Rightarrow$ il s'agit de l'hypothèse d'**homoscédasticité** ou de "**variance constante**"

Figure 18: Modèle de régression simple sous l'hypothèse d'homoscédasticité

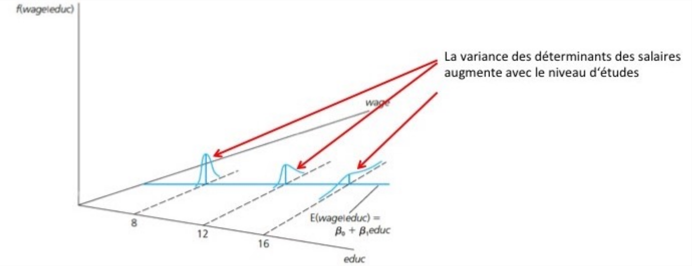


6 – Espérance et Variance

6.2 – Calcul de la variance

- Si cette hypothèse n'est pas respectée: on parle **hétéroscédasticité**
  - $\Rightarrow$ La variabilité inobservée dépend de la valeur de la variable explicative, ici elle augmente si x augmente

Figure 19: Modèle de régression simple sous l'hypothèse d'hétéroscédasticité



6 – Espérance et Variance

6.2 – Calcul de la variance

- Hypothèses :

  H_1: $E[u_i] = 0, \forall\ t$

  H_2: $V(u_i) = E[(u_i - E(u_i))^2] = E[u_i^2] = \sigma_u^2 \forall\ i$

  H_3: la variable explicative x_i est non aléatoire

  H_4: le modèle est correctement spécifié

  H_5: la variable explicative n'est pas constante pour toutes les observations

- Sous ces hypothèses:

$$V(\widehat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \qquad \Rightarrow \widehat{\sigma}_{\widehat{\beta}_1} = \sqrt{V(\widehat{\beta}_1)}$$

$$V(\widehat{\beta}_0) = \frac{\sigma_u^2}{n} \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \qquad \Rightarrow \widehat{\sigma}_{\widehat{\beta}_0} = \sqrt{V(\widehat{\beta}_0)}$$

6 – Espérance et Variance

6.2 – Calcul de la variance

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n u_i(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

$$V(\widehat{\beta}_1) = E[\widehat{\beta}_1 - E[\widehat{\beta}_1]]^2 = E\left[\frac{\sum_{i=1}^n u_i(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]^2$$

$$= \left[\frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]^2 \left[\sum_{i=1}^n (x_i - \overline{x})^2 \underbrace{V(u_i)}_{\text{Si } V(u_i) = \sigma_u^2 \text{ sous H2}}\right]$$

$$= \left[\frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]^2 \left[\sum_{i=1}^n (x_i - \overline{x})^2 \sigma_u^2\right] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$$

6 – Espérance et Variance

6.2 – Calcul de la variance

$$\begin{aligned}
\widehat{\beta}_0 &= \beta_0 + \overline{x}[\beta_1 - \widehat{\beta}_1] + \overline{u} \\
V(\widehat{\beta}_0) &= E[\widehat{\beta}_0 - E(\widehat{\beta}_0)]^2 = E[\widehat{\beta}_0 - \beta_0]^2 \\
&= E[\overline{x}[\beta_1 - \widehat{\beta}_1] + \overline{u}]^2 \\
&= \overline{x}^2 V(\widehat{\beta}_1) + V[\overline{u}] \\
&= \frac{\sigma_u^2 \overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} + \frac{\sigma_u^2}{n} \\
&= \sigma_u^2 \frac{n\overline{x}^2 + \sum_{i=1}^n (x_i - \overline{x})^2}{n \sum_{i=1}^n (x_i - \overline{x})^2} \\
&= \sigma_u^2 \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2}
\end{aligned}$$

6 – Espérance et Variance

6.3 – Estimateur de Gauss-Markov

Soit le modèle suivant : $y_i = \beta_0 + \beta_1 x_i + u_i \; \forall i = 1, ..., N$

Définition: Estimateur de Gauss Markov
Sous les hypothèses H_1 à H_5, $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont les meilleurs estimateurs linéaires sans biais de β_0 et β_1.

$\rightarrow$ aucun autre **estimateur linéaire sans biais** n'aura une variance inférieure.

$\Rightarrow$ les estimateurs $\widehat{\beta}_0$ et $\widehat{\beta}_1$ sont :

- sans biais **ET** à variance minimale

6 – Espérance et Variance

6.4 – Variance de l'erreur

- Les formules suivantes

$$V(\widehat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \qquad \Rightarrow \widehat{\sigma}_{\widehat{\beta}_1} = \sqrt{V(\widehat{\beta}_1)}$$

$$V(\widehat{\beta}_0) = \frac{\sigma_u^2}{n} \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \qquad \Rightarrow \widehat{\sigma}_{\widehat{\beta}_0} = \sqrt{V(\widehat{\beta}_0)}$$

- Permettent d'identifier les facteurs qui influencent $V(\widehat{\beta}_1)$, et $V(\widehat{\beta}_0)$.
- Mais ces formules contiennent des inconnues. Nous allons évaluer σ_u^2 en utilisant les données.

6 – Espérance et Variance

6.4 – Variance de l'erreur

6.4.1 – Erreurs et résidus

- En partant du modèle:

[éq° A] $y_i = \beta_0 + \beta_1 x_i + u_i \;\; \forall i = 1, ..., N,$

u_i est l'**erreur** relative à l'observation i.

  - Modèle écrit à partir de la population en fonction d'observations échantillonnées aléatoirement

- On sait aussi que :

[éq° B] $y_i = \widehat{y}_i + \widehat{u}_i \Leftrightarrow$ Donc $y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{u}_i$

$\Rightarrow$ L'**erreur** apparaît dans l'équation relative à la population (avec les paramètres β_k **(éq° A)**);

$\Rightarrow$ Les **résidus** apparaissent dans l'équation estimée (avec les paramètres $\widehat{\beta}_k$ **(éq° B)**);

- **Les erreurs ne peuvent jamais être observées alors que les résidus sont calculés à partir d'une base de données.**

6 – Espérance et Variance

6.4 – Variance de l'erreur

6.4.1 – Erreurs et résidus

- En utilisant les éq° A et B, on peut exprimer les résultas des résidus en fonction des erreurs:

$$\widehat{u}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i = (\beta_0 + \beta_1 x_i + u_i) - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$$

ou

$$\widehat{u}_i = u_i - (\widehat{\beta}_0 - \beta_0) - (\widehat{\beta}_1 - \beta_1) x_i$$

- On constate bien que $\widehat{u}_i \neq u_i \Rightarrow$ C'est la différence attendue entre ces 2 termes qui est égale à 0. Comme pour $(\widehat{\beta}_0 - \beta_0)$ et $(\widehat{\beta}_1 - \beta_1)$

6 – Espérance et Variance

6.4 – Variance de l'erreur

6.4.1 – Erreurs et résidus

- La variance de u ne dépend pas de x (i.e. elle est égale à la variance non conditionnelle):

$$V(\widehat{u}_i|x_i) = \sigma_u^2 = Var(u_i)$$

- La variance de l'erreur peut-être calculée à partir de la variance des résidus de l'échantillon, mais cet estimateur est biaisé,
    - si on remplace les erreurs par les résidus, on obtient: $\frac{\sum_{i=1}^{n} \widehat{u}_i^{\,2}}{n}$
    - Estimateur ne tient pas compte des CPO que l'estimateur des MCO doit respecter:

$$\sum_{i=1}^{n} \widehat{u}_i = 0 \text{ et} \sum_{i=1}^{n} x_i \widehat{u}_i = 0$$

    - autrement dit, si on connait la valeur des $n-2$ résidus dans notre échantillon, il faut que 2 contraintes soient respectées.

6 – Espérance et Variance

6.4 – Variance de l'erreur

6.4.1 – Erreurs et résidus

- Un estimateur sans biais de σ_u^2 est donné par:

$$\widehat{\sigma}_u^2 = \frac{\sum_{i=1}^{n} \widehat{u_i}^2}{N - (k+1)}$$

Ici k=1 (regression simple, 1 seule varaible explicative)

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

- Il incorpore l'ajustement relatif aux degrés de liberté.
- Calcul de l'espérance: $E[\sum_{i=1}^{n} \widehat{u_i}^2]$?

- Sous les hypothèses H_1 à H_5, la moyenne de l'éq° B :

$$\widehat{u}_i = u_i - (\widehat{\beta}_0 - \beta_0) - (\widehat{\beta}_1 - \beta_1)x_i \tag{29}$$

est équivalente à (sous l'hypothèse que la moyenne des résidus = 0)

$$0 = \overline{u} - (\widehat{\beta}_0 - \beta_0) - (\widehat{\beta}_1 - \beta_1)\overline{x} \tag{30}$$

- en effectuant la soustraction entre les eq° 29 et 30, on obtient:

$$\begin{aligned}
\widehat{u}_i =& (u_i - \overline{u}) - (\widehat{\beta}_1 - \beta_1)(x_i - \overline{x}) \\
\widehat{u}_i^2 =& (u_i - \overline{u})^2 + (\widehat{\beta}_1 - \beta_1)^2(x_i - \overline{x})^2 - 2(u_i - \overline{u})(\widehat{\beta}_1 - \beta_1)(x_i - \overline{x}) \\
\sum_{i=1}^{n} \widehat{u}_i^2 =& \sum_{i=1}^{n}(u_i - \overline{u})^2 + (\widehat{\beta}_1 - \beta_1)^2 \sum_{i=1}^{n}(x_i - \overline{x})^2 - 2(\widehat{\beta}_1 - \beta_1) \sum_{i=1}^{n}(u_i - \overline{u})(x_i - \overline{x})
\end{aligned}$$

6 – Espérance et Variance

6.4 – Variance de l'erreur

6.4.1 – *Erreurs et résidus*

$$E\left[\sum_{i=1}^{n}\widehat{u}_i^2\right] = E\left[\sum_{i=1}^{n}(u_i-\overline{u})^2 - (\widehat{\beta}_1-\beta_1)^2\sum_{i=1}^{n}(x_i-\overline{x})^2 - 2(\widehat{\beta}_1-\beta_1)\sum_{i=1}^{n}(u_i-\overline{u})(x_i-\overline{x})\right]$$

- 1er block $E\left[\sum_{i=1}^{n}(u_i-\overline{u})^2\right] = (n-1)\sigma_u^2$
- 2m block $E\left[(\widehat{\beta}_1-\beta_1)^2\sum_{i=1}^{n}(x_i-\overline{x})^2\right] = \sigma_u^2$ car $E\left[(\widehat{\beta}_1-\beta_1)^2\right] = V(\widehat{\beta}_1) = \frac{\sigma_u^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}$
- 3m block $E\left[2(\widehat{\beta}_1-\beta_1)\sum_{i=1}^{n}(u_i-\overline{u})(x_i-\overline{x})\right] = E[2(\widehat{\beta}_1-\beta_1)^2\sum_{i=1}^{n}(x_i-\overline{x})^2] = 2\sigma_u^2$
- Donc

$$E[\sum_{i=1}^{n}\widehat{u}_i^2] = (n-1)\sigma_u^2 + \sigma_u^2 - 2\sigma_u^2 = (n-2)\sigma_u^2 \Leftrightarrow E[SCR/(n-2)] = \sigma_u^2$$

6 – Espérance et Variance

6.5 – Application

- L'application des MCO pour estimer l'impact du niveau d'etude sur le salaire:

$$\widehat{salaire}_i = -0.905 + 0.541 etudes_i$$

- Nous allons donc estimer $\widehat{\sigma}_u^2$, $V(\widehat{\beta}_0)$ et $V(\widehat{\beta}_1)$.

$$\widehat{\sigma}_u^2 = \frac{1}{n-2}\sum_{i=1}^{n}\widehat{u}_i^2 = \frac{SCR}{n-2} = \frac{5980,68}{526-2} = 11,41$$

$$V(\widehat{\beta}_1) = \frac{\widehat{\sigma}_u^2}{\sum_{i=1}^{n}(etudes_i - \overline{etudes})^2} = \frac{11,41}{4025,43} \Rightarrow \widehat{\sigma}_{\widehat{\beta}_1} = 0,0532$$

$$V(\widehat{\beta}_0) = \frac{\widehat{\sigma}_u^2}{n}\frac{\sum_{i=1}^{n} etudes_i^2}{\sum_{i=1}^{n}(etudes_i - \overline{etudes})^2} = \frac{11,41}{526}\frac{87040}{4025,43} \Rightarrow \widehat{\sigma}_{\widehat{\beta}_0} = 0,6849$$

6 – Espérance et Variance

6.5 – Application

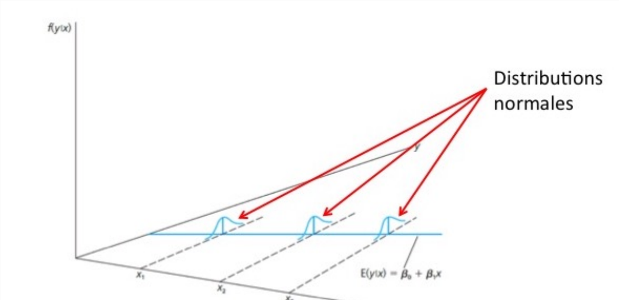
- Nous connaissons donc : l'espérance et la variance de l'estimateur des MCO, cela permet donc de connaître la précision de l'estimateur.
- Nous avons nous concentrer sur la distribution d'échantillonnage des $\widehat{\beta_k}$ pour réaliser de l'inférence statistique
- Nous allons donc:
    - Poser des hypothèses sur les paramètres de la population
    - Construire des intervalles de confiance

Plan du cours

7 – Inférence

- Sous l'hypothèse complémentaire de normalité des termes d'erreur, on peut écrire : $u_i \sim Normal(0, \sigma_u^2)$ indépendamment de $x_{i1}, x_{i2}, \cdots, x_{ik}$
- La forme et la distribution e la variance ne dépendent pas des variables explicatives.
- Cela implique que: $y|x \sim Normal(\beta_0 + \beta_1 x_1, \sigma_u^2)$

Figure 20: Distribution normale homoscédastique avec une seule variable explicative



7 – Inférence

- Le terme d'erreur est la “somme” de plusieurs facteurs inobservables
- La somme de ces facteurs suit une loi normale
- Quels problèmes ?
  - hétérogénéité des facteurs ?
  - Comment connaître la distribution de la somme des facteurs?
- La normalité des termes d'erreur est une question empirique
- Les termes d'erreur doivent avoir une distribution proche de la distribution normale
- Avec de grands échantillons, l'hypothèse de normalité n'est plus nécéssaire

7 – Inférence

7.1 – Test bilatéral

7.1.1 – Princippe

- Sous H_1 à H_5 auxquelles on ajoute la normalité, l'estimateur $\widehat{\beta_k}$ de β_k suit une loi normale

$$\widehat{\beta_k} \sim Normal(\beta_k, Var(\widehat{\beta}_k))$$

- L'estimateur standardisé suit donc une loi normale centrée réduite

$$\frac{\widehat{\beta_k} - \beta_k}{SD(\widehat{\beta}_k)} \sim Normal(0, 1)$$

- L'utilisation de l'écart-type estimé du paramètre permet d'écrire:

$$\frac{\widehat{\beta_k} - \beta_k}{\widehat{\sigma}_{\widehat{\beta}_k}} \sim t_{1-\alpha/2}(n-(k+1))\text{où k= de variables explicatives, ici: k=1}$$

7 – Inférence

7.1 – Test bilatéral

7.1.1 – *Princippe*

- Considérons le modèle suivant: $y_i = \beta_0 + \beta_k x_{ki} + u_i$
- On peut alors tester des hypothèses :
  - Pour savoir si le paramètre de la population est égal à 0 i.e. qu'après avoir contrôlé par la autres effets, x_k n'a pas d'effet sur y
- Testons, par exemple: H_0: $\beta_k = 0$, sous l'hypothèse H_0:

$$t_{\widehat{\beta}_k} = \frac{\widehat{\beta}_k - \overbrace{0}^{\beta_k = 0,\ \textbf{sous}\ H_0}}{\underbrace{\widehat{\sigma}_{\widehat{\beta}_k}}_{\textbf{dépend de l'écart-type estimé du paramètre}}} \sim t_{1-\alpha/2}(n-(k+1))$$

7 – Inférence

7.1 – Test bilatéral

7.1.1 – Princippe

- Considérons le modèle suivant: $y_i = \beta_0 + \beta_k x_{ki} + u_i$
- On peut alors tester des hypothèses :
  - Pour savoir si le paramètre de la population est égal à 0 i.e. qu'après avoir contrôlé par la autres effets, x_k n'a pas d'effet sur y
- Testons, par exemple: H_0: $\beta_k = 0$, sous l'hypothèse H_0:

$$t_{\widehat{\beta_k}} = \frac{\widehat{\beta_k} - 0}{\widehat{\sigma}_{\widehat{\beta_k}}} \sim t_{1-\alpha/2}(n-(k+1))$$

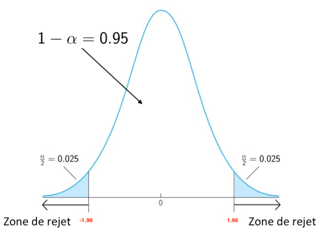
- L'objectif est de définir une règle de décision de sorte à rejeter l'hypothèse H_0 seulement avec une faible probabilité $\Rightarrow$ **avec un faible niveau de significativité**, par exemple 5% (i.e. $\alpha = 5\%$).

7 – Inférence

7.1 – Test bilatéral

7.1.1 – Princippe

- Hypothèse: $H_0 : \beta_k = 0$ vs $H_1 : \beta_k \neq 0$
- On rejette l'hypothèse H_0 pour préférer l'hypothèse alternative si et seulement si la valeur absolue de la statistique de test est dans la zone bleue



- Cela veut dire que si H_0 est vraie, elle est rejetée dans 5% des cas.

7 – Inférence

7.1 – Test bilatéral

7.1.2 – Réalisation d'un test de student

- Afin d'analyser l'effet du genre sur le niveau de salaire, on considère le modèle suivant:

$$wage_i = \beta_0 + \beta_1 male_i$$

- $\widehat{\beta_1}$ est la **différence de salaire** entre les hommes et les femmes.
- A partir de l'observation de 3000 individus, l'application des MCO donne les résultats suivants:

$$\widehat{wage}_i = \quad 5{,}098 \quad + \quad 1{,}403 \quad male_i$$

- $\widehat{\beta_0}$ est le salaire estimé des femmes, puisque si $male_i = 0$: $\widehat{wage_i^f} = 5,098$
  $\Rightarrow$ Le salaire **horaire des femmes** prédit par le modèle est de 5,098\$/h.
- le salaire des hommes est prédit avec $male_i = 1$: $\widehat{wage_i^h} = 6,504$
  $\Rightarrow$ Le salaire **horaire des hommes** prédit par le modèle est de 6,504\$/h.

7 – Inférence

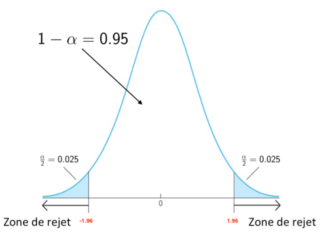
7.1 – Test bilatéral

7.1.2 – *Réalisation d'un test de student*

- L'équation estimée est :

$$\widehat{wage}_i = \underset{(0{,}090)}{5{,}098} + \underset{(0{,}1317)}{1{,}403}\; male_i$$

- Les valeurs sous les coefficients estimés sont les écarts-types.
- On va tester **au seuil** $\alpha\%$
  - H_0 : $\beta_1 = 0$, les femmes ont en moyenne le même taux de salaire horaire
  - H_1 : $\beta_1 \neq 0$, les femmes ont en moyenne un salaire différent de celui des hommes



7 – Inférence

7.1 – Test bilatéral

7.1.2 – *Réalisation d'un test de student*

- **Hypothèse testée, au seuil α%**
  - $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$
- **Statistique de test : Statistique de Student** :

$$t_{\widehat{\beta}_1} = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} \sim t_{1-\alpha/2}(n - (k+1))$$

- **Règle de decision**
  - si $|t_{\widehat{\beta}_1}| > t_{1-\alpha/2}(n-(k+1))$, on rejette H_0
  - Pour un seuil $\alpha = 5\%$, la statistique théorique doit être lue sur la Table de student

Table Student

7 – Inférence

7.1 – Test bilatéral

7.1.2 – Réalisation d'un test de student

- $H_0 : \beta_1 = 0$ et $H_1 : \beta_1 \neq 0$
- Sous H_0, la statistique calculée est donnée par

$$t_{\widehat{\beta}_1} = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} = \frac{1,402 - 0}{0,1317} = 10,64$$

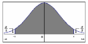
- Pour un seuil $\alpha = 5\%$, la statistique théorique est ▸ Lecture sur Table Student

7 – Inférence

7.1 – Test bilatéral

7.1.2 – *Réalisation d'un test de student*

Loi de Student



α	1	0,8	0,6	0,4	0,2	0,1	0,05	0,02	0,01	0,002	0,001
$1-\alpha$	0	0,2	0,4	0,6	0,8	0,9	0,95	0,98	0,99	0,998	0,999
ν = ddl											
1	0,0000	0,3249	0,7265	1,3764	3,0777	6,3137	12,706	31,821	63,656	318,29	636,58
2	0,0000	0,2887	0,6172	1,0607	1,8856	2,9200	4,3027	6,9645	9,9250	22,328	31,600
3	0,0000	0,2767	0,5844	0,9785	1,6377	2,3534	3,1824	4,5407	5,8408	10,214	12,924
4	0,0000	0,2707	0,5686	0,9410	1,5332	2,1318	2,7765	3,7469	4,6041	7,1729	8,6101
5	0,0000	0,2672	0,5594	0,9195	1,4759	2,0150	2,5706	3,3649	4,0321	5,8935	6,8685
6	0,0000	0,2648	0,5534	0,9057	1,4398	1,9432	2,4469	3,1427	3,7074	5,2075	5,9587
7	0,0000	0,2632	0,5491	0,8960	1,4149	1,8946	2,3646	2,9979	3,4995	4,7853	5,4081
8	0,0000	0,2619	0,5459	0,8889	1,3968	1,8595	2,3060	2,8965	3,3554	4,5008	5,0414
9	0,0000	0,2610	0,5435	0,8834	1,3830	1,8331	2,2622	2,8214	3,2498	4,2969	4,7809
10	0,0000	0,2602	0,5415	0,8791	1,3722	1,8125	2,2281	2,7638	3,1693	4,1437	4,5868
11	0,0000	0,2596	0,5399	0,8755	1,3634	1,7959	2,2010	2,7181	3,1058	4,0248	4,4369
12	0,0000	0,2590	0,5386	0,8726	1,3562	1,7823	2,1788	2,6810	3,0545	3,9296	4,3178
13	0,0000	0,2586	0,5375	0,8702	1,3502	1,7709	2,1604	2,6503	3,0123	3,8520	4,2209
14	0,0000	0,2582	0,5366	0,8681	1,3450	1,7613	2,1448	2,6245	2,9768	3,7874	4,1403
15	0,0000	0,2579	0,5357	0,8662	1,3406	1,7531	2,1315	2,6025	2,9467	3,7329	4,0728
16	0,0000	0,2576	0,5350	0,8647	1,3368	1,7459	2,1199	2,5835	2,9208	3,6861	4,0149
17	0,0000	0,2573	0,5344	0,8633	1,3334	1,7396	2,1098	2,5669	2,8982	3,6458	3,9651
18	0,0000	0,2571	0,5338	0,8620	1,3304	1,7341	2,1009	2,5524	2,8784	3,6105	3,9217
19	0,0000	0,2569	0,5333	0,8610	1,3277	1,7291	2,0930	2,5395	2,8609	3,5793	3,8833
20	0,0000	0,2567	0,5329	0,8600	1,3253	1,7247	2,0860	2,5280	2,8453	3,5518	3,8496
21	0,0000	0,2566	0,5325	0,8591	1,3232	1,7207	2,0796	2,5176	2,8314	3,5271	3,8193
22	0,0000	0,2564	0,5321	0,8583	1,3212	1,7171	2,0739	2,5083	2,8188	3,5050	3,7922
23	0,0000	0,2563	0,5317	0,8575	1,3195	1,7139	2,0687	2,4999	2,8073	3,4850	3,7676
24	0,0000	0,2562	0,5314	0,8569	1,3178	1,7109	2,0639	2,4922	2,7970	3,4668	3,7454
25	0,0000	0,2561	0,5312	0,8562	1,3163	1,7081	2,0595	2,4851	2,7874	3,4502	3,7251
26	0,0000	0,2560	0,5309	0,8557	1,3150	1,7056	2,0555	2,4786	2,7787	3,4350	3,7067
27	0,0000	0,2559	0,5306	0,8551	1,3137	1,7033	2,0518	2,4727	2,7707	3,4210	3,6895
28	0,0000	0,2558	0,5304	0,8546	1,3125	1,7011	2,0484	2,4671	2,7633	3,4082	3,6739
29	0,0000	0,2557	0,5302	0,8542	1,3114	1,6991	2,0452	2,4620	2,7564	3,3963	3,6595
30	0,0000	0,2556	0,5300	0,8538	1,3104	1,6973	2,0423	2,4573	2,7500	3,3852	3,6460
40	0,0000	0,2550	0,5286	0,8507	1,3031	1,6839	2,0211	2,4233	2,7045	3,3069	3,5510
50	0,0000	0,2547	0,5278	0,8489	1,2987	1,6759	2,0086	2,4033	2,6778	3,2614	3,4960
60	0,0000	0,2545	0,5272	0,8477	1,2958	1,6706	2,0003	2,3901	2,6603	3,2317	3,4602
70	0,0000	0,2543	0,5268	0,8468	1,2938	1,6669	1,9944	2,3808	2,6479	3,2108	3,4350
80	0,0000	0,2542	0,5265	0,8461	1,2922	1,6641	1,9901	2,3739	2,6387	3,1952	3,4164
90	0,0000	0,2541	0,5263	0,8456	1,2910	1,6620	1,9867	2,3685	2,6316	3,1832	3,4019
100	0,0000	0,2540	0,5261	0,8452	1,2901	1,6602	1,9840	2,3642	2,6259	3,1738	3,3905
200	0,0000	0,2537	0,5252	0,8434	1,2858	1,6525	1,9719	2,3451	2,6006	3,1315	3,3398
∞	0,0000	0,2533	0,5244	0,8416	1,2816	1,6449	1,9600	2,3263	2,5758	3,0903	3,2906

7 – Inférence

7.1 – Test bilatéral

7.1.2 – *Réalisation d'un test de student*

α	1	0,8	0,6	0,4	0,2	0,1	0,05	0,02	0,01	0,002	0,001
$1-\alpha$	0	0,2	0,4	0,6	0,8	0,9	0,95	0,98	0,99	0,998	0,999
ν = ddl											
1	0,0000	0,3249	0,7265	1,3764	3,0777	6,3137	12,706	31,821	63,656	318,29	636,58
2	0,0000	0,2887	0,6172	1,0607	1,8856	2,9200	4,3027	6,9645	9,9250	22,328	31,600
3	0,0000	0,2767	0,5844	0,9785	1,6377	2,3534	3,1824	4,5407	5,8408	10,214	12,924
4	0,0000	0,2707	0,5686	0,9410	1,5332	2,1318	2,7765	3,7469	4,6041	7,1729	8,6101
5	0,0000	0,2672	0,5594	0,9195	1,4759	2,0150	2,5706	3,3649	4,0321	5,8935	6,8685
6	0,0000	0,2648	0,5534	0,9057	1,4398	1,9432	2,4469	3,1427	3,7074	5,2075	5,9587
7	0,0000	0,2632	0,5491	0,8960	1,4149	1,8946	2,3646	2,9979	3,4995	4,7853	5,4081
8	0,0000	0,2619	0,5459	0,8889	1,3968	1,8595	2,3060	2,8965	3,3554	4,5008	5,0414
9	0,0000	0,2610	0,5435	0,8834	1,3830	1,8331	2,2622	2,8214	3,2498	4,2969	4,7809
10	0,0000	0,2602	0,5415	0,8791	1,3722	1,8125	2,2281	2,7638	3,1693	4,1437	4,5868
11	0,0000	0,2596	0,5399	0,8755	1,3634	1,7959	2,2010	2,7181	3,1058	4,0248	4,4369
12	0,0000	0,2590	0,5386	0,8726	1,3562	1,7823	2,1788	2,6810	3,0545	3,9296	4,3178
13	0,0000	0,2586	0,5375	0,8702	1,3502	1,7709	2,1604	2,6503	3,0123	3,8520	4,2209
14	0,0000	0,2582	0,5366	0,8681	1,3450	1,7613	2,1448	2,6245	2,9768	3,7874	4,1403
15	0,0000	0,2579	0,5357	0,8662	1,3406	1,7531	2,1315	2,6025	2,9467	3,7329	4,0728
16	0,0000	0,2576	0,5350	0,8647	1,3368	1,7459	2,1199	2,5835	2,9208	3,6861	4,0149
17	0,0000	0,2573	0,5344	0,8633	1,3334	1,7396	2,1098	2,5669	2,8982	3,6458	3,9651
18	0,0000	0,2571	0,5338	0,8620	1,3304	1,7341	2,1009	2,5524	2,8784	3,6105	3,9217
19	0,0000	0,2569	0,5333	0,8610	1,3277	1,7291	2,0930	2,5395	2,8609	3,5793	3,8833
20	0,0000	0,2567	0,5329	0,8600	1,3253	1,7247	2,0860	2,5280	2,8453	3,5518	3,8496
21	0,0000	0,2566	0,5325	0,8591	1,3232	1,7207	2,0796	2,5176	2,8314	3,5271	3,8193
22	0,0000	0,2564	0,5321	0,8583	1,3212	1,7171	2,0739	2,5083	2,8188	3,5050	3,7922
23	0,0000	0,2563	0,5317	0,8575	1,3195	1,7139	2,0687	2,4999	2,8073	3,4850	3,7676
24	0,0000	0,2562	0,5314	0,8569	1,3178	1,7109	2,0639	2,4922	2,7970	3,4668	3,7454
25	0,0000	0,2561	0,5312	0,8562	1,3163	1,7081	2,0595	2,4851	2,7874	3,4502	3,7251
26	0,0000	0,2560	0,5309	0,8557	1,3150	1,7056	2,0555	2,4786	2,7787	3,4350	3,7067
27	0,0000	0,2559	0,5306	0,8551	1,3137	1,7033	2,0518	2,4727	2,7707	3,4210	3,6895
28	0,0000	0,2558	0,5304	0,8546	1,3125	1,7011	2,0484	2,4671	2,7633	3,4082	3,6739
29	0,0000	0,2557	0,5302	0,8542	1,3114	1,6991	2,0452	2,4620	2,7564	3,3963	3,6595
30	0,0000	0,2556	0,5300	0,8538	1,3104	1,6973	2,0423	2,4573	2,7500	3,3852	3,6460
40	0,0000	0,2550	0,5286	0,8507	1,3031	1,6839	2,0211	2,4233	2,7045	3,3069	3,5510
50	0,0000	0,2547	0,5278	0,8489	1,2987	1,6759	2,0086	2,4033	2,6778	3,2614	3,4960
60	0,0000	0,2545	0,5272	0,8477	1,2958	1,6706	2,0003	2,3901	2,6603	3,2317	3,4602
70	0,0000	0,2543	0,5268	0,8468	1,2938	1,6669	1,9944	2,3808	2,6479	3,2108	3,4350
80	0,0000	0,2542	0,5265	0,8461	1,2922	1,6641	1,9901	2,3739	2,6387	3,1952	3,4164
90	0,0000	0,2541	0,5263	0,8456	1,2910	1,6620	1,9867	2,3685	2,6316	3,1832	3,4019
100	0,0000	0,2540	0,5261	0,8452	1,2901	1,6602	1,9840	2,3642	2,6259	3,1738	3,3905
200	0,0000	0,2537	0,5252	0,8434	1,2858	1,6525	1,9719	2,3451	2,6006	3,1315	3,3398
∞	0,0000	0,2533	0,5244	0,8416	1,2816	1,6449	1,9600	2,3263	2,5758	3,0903	3,2906

7 – Inférence

7.1 – Test bilatéral

7.1.2 – *Réalisation d'un test de student*

200	0,0000	0,2537
∞	0,0000	0,2533

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

7 – Inférence

7.1 – Test bilatéral

7.1.2 – Réalisation d'un test de student

- **Hypothèse testée, au seuil** $\alpha\%$: $H_0 : \beta_1 = 0$ et $H_1 : \beta_1 \neq 0$
- Sous H_0, la statistique calculée est donnée par

$$t_{\widehat{\beta}_1} = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} = \frac{1,402 - 0}{0,1317} = 10,64 \text{ et } n = 3000, \text{donc } n - (k+1) = 3000 - 2 = 2998$$

- Pour un seuil $\alpha = 5\%$, la statistique théorique est Lecture Table Student $t_{1-\alpha/2}(2998) = 1,96$
- Donc: $|t_{1-\alpha/2}(2998)| < t_{\hat{\beta}_1} \Rightarrow$ On rejette H_0 au seuil de $\alpha = 5\%$
- On dira donc que le paramètre estimé est **significativement différent de 0** au seuil de $\alpha = 5\%$.
- On rejette l'hypothèse selon laquelle les hommes et les femmes ont significativement le même salaire.
- Le genre a un impact significativement différent de zéro pour expliquer le niveau de salaire dans notre échantillon.

7 – Inférence

7.2 – Intervalle de confiance

7.2.1 – *Définition*

- On peut alors construire l'intervalle de confiance pour un paramètre de la population β_k.
- Il donne l'ensemble des valeurs possibles du paramètre de la population.
- On utilise le fait que $\frac{\widehat{\beta}_k - \beta_k}{\widehat{\sigma}_{\widehat{\beta}_k}}$ suit une distribution de Student à $n-(k+1)$ degrés de liberté
- L'intervalle de confiance à $1-\alpha$ % est donc donné par:

$$[\widehat{\beta}_k - t_{1-\alpha/2}(n-(k+1))\,\widehat{\sigma}_{\widehat{\beta}_k} ; \widehat{\beta}_k + t_{1-\alpha/2}(n-(k+1))\,\widehat{\sigma}_{\widehat{\beta}_k}]$$

- De fait, si nous testons $\beta_k = a$ et que a n'est pas dans l'intervalle de confiance, alors on pourra dire que l'on rejette l'hypothèse $\beta_k = a$.

7 – Inférence

7.2 – Intervalle de confiance

7.2.2 – *Application*

- On peut alors construire l'intervalle de confiance a 95% pour le paramètre β_1

$$
\begin{aligned}
IC_{\beta_1,95\%} &= [\widehat{\beta}_1 - t_{1-\alpha/2}(n-(k+1))\,\widehat{\sigma}_{\widehat{\beta}_1}\,;\,\widehat{\beta}_1 + t_{1-\alpha/2}(n-(k+1))\,\widehat{\sigma}_{\widehat{\beta}_1}] \\
&= [1,402 - 1,96 \times 0,1317\,;\,1,402 + 1,96 \times 0,1317] \\
&= [1,144; 1,661]
\end{aligned}
$$

- L'intervalle de confiance pour le paramètre inconnu β_1 est $[1,144; 1,661]$. Cela signifie que dans la population, le différentiel de salaire entre les hommes et les femmes (β_1) devrait se trouver dans 95% des cas compris entre 1,144$ et 1,661$.

  $\Rightarrow$ l'intervalle de confiance du différentiel de salaire entre les hommes et les femmes.

  - 0 n'est pas dans l'intervalle, cela confirme le résultat du test précédent.

$$\widehat{salaire}_i = -0,905 + 0,541 etudes_i$$

N=596, $\widehat{\sigma}_{\widehat{\beta}_1} = 0,0532$ et $\widehat{\sigma}_{\widehat{\beta}_0} = 0,6849$

1. Interpréter l'équation estimée.
2. Le niveau d'étude est-il significatif pour expliquer le salaire, effectuer le test pour un de significativité de $\alpha = 5\%$.
   - 2.1 Hypothèse ?
   - 2.2 Statistique de test
   - 2.3 Règle de décision
   - 2.4 Application numérique
3. Quel est l'intervalle de confiance à 95% pour le paramètre β_1 ?

7 – Inférence

7.2 – Intervalle de confiance

7.2.3 – A vous de jouer?

- **Hypothèse testée, au seuil** $\alpha\%$
  - H_0 : $\beta_1 = 0$, le nombre d'années d'etudes **n'influence pas** le niveau de salaire
- Sous H_0, la statistique calculée est donnée par

$$t_{\widehat{\beta}_1} = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} = \frac{0,541 - 0}{0,0532} = 10,17 \quad \text{et } n = 596, \text{donc } n - (k+1) = 594$$

- Pour un seuil $\alpha = 5\%$, la statistique théorique est $t_{1-\alpha/2}(594) = 1,96$
- Donc: $|t_{1-\alpha/2}(594)| < t_{\hat{\beta}_1} \Rightarrow$ On rejette H_0 au seuil de significativité de 5%
- On dira donc que le paramètre estimé est **significativement différent de 0** au seuil de $\alpha = 5\%$.
- On rejette l'hypothèse selon laquelle le niveau d'étude n'a pas d'impact sur le salaire.

7.2.3 – A vous de jouer?

- On peut alors construire l'intervalle de confiance a 95% pour le paramètre β_1

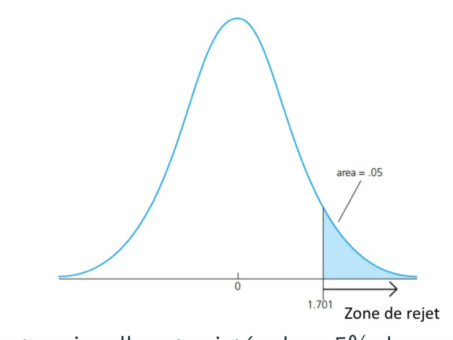
$$
\begin{aligned}
IC_{\beta_1,95\%} &= [\widehat{\beta}_1 - t_{1-\alpha/2}(n-(k+1))\,\widehat{\sigma}_{\widehat{\beta}_1}\,;\,\widehat{\beta}_1 + t_{1-\alpha/2}(n-(k+1))\,\widehat{\sigma}_{\widehat{\beta}_1}] \\
&= [0,541 - 1,96 \times 0,0532\,;\,0,541 + 1,96 \times 0,0532] \\
&= [0,438;0,646]
\end{aligned}
$$

- L'intervalle de confiance pour le paramètre inconnu β_1 est $[0,438;0,646]$. Cela signifie que dans la population, le rendement d'une année d'étude supplémentaire devrait se trouver, dans 95%, entre 0,438$ et 0,646$.
  - 0 n'est pas dans l'intervalle, cela confirme le résultat du test précédent.

7 – Inférence

7.3 – Test unilatéral

- Hypothèse: $H_0 : \beta_k \leq 0$ vs $H_1 : \beta_k > 0$
- On rejette l'hypothèse H_0 pour préférer l'hypothèse alternative si et seulement
    - la statistique de test est dans la zone bleue.
    - la statistique calculée est supérieure à $t_{1-\alpha}(n-(k+1))$

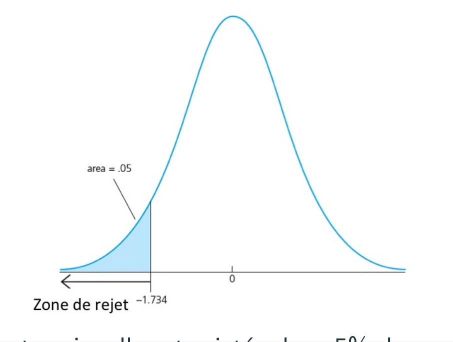


- Cela veut dire que si H_0 est vraie, elle est rejetée dans 5% des cas.

7 – Inférence

7.3 – Test unilatéral

- Hypothèse: $H_0 : \beta_k \geq 0$ vs $H_1 : \beta_k < 0$
- On rejette l'hypothèse H_0 pour préférer l'hypothèse alternative si et seulement
  - la statistique de test est dans la zone bleue.
  - la statistique calculée est inférieure à $-t_{1-\alpha(n-(k+1))}$



- Cela veut dire que si H_0 est vraie, elle est rejetée dans 5% des cas.

7 – Inférence

7.4 – Significativité GLOBALE du modèle

- **Hypothese testée** : $H_0 : \beta_i = 0$ contre $H_1 : \beta_i \neq 0$ avec $i \in \{1.2...., k\}$
- **Statistique de test**

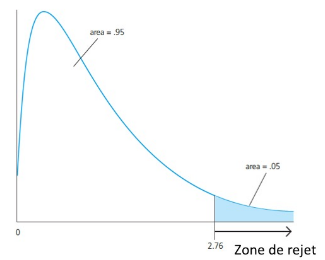
$$\begin{aligned} F &= \frac{SCE/k}{SCR/\left[n-(k+1)\right]} \\ &= \frac{R^2/k}{(1-R^2)/\left[n-(k+1)\right]} \sim F_{v_1,v_2} \end{aligned}$$

  où F est la statistique de Fisher calculée et F_{v_1,v_2} est la valeur critique dans la table usuelle statistique de Fisher à $v_1 = k$ et $v_2 = n-(k+1)$.
- **Régle de décision**
  - si $F > F_{v_1,v_2}$, on rejette $H_0 \Rightarrow$ le modèle est globalement significatif

7 – Inférence

7.4 – Significativité GLOBALE du modèle

- On rejette H_0 si zone bleue



7 – Inférence

7.4 – Significativité GLOBALE du modèle

7.4.1 – Application

$\widehat{salaire}_i = -0,905 + 0,541 etudes_i$

N=596, $\widehat{\sigma}_{\widehat{\beta}_1} = 0,0532$ et $\widehat{\sigma}_{\widehat{\beta}_0} = 0,6849$

- **Hypothese testée** : $H_0 : \beta_i = 0$ contre $H_1 : \beta_i \neq 0$ avec $i \in \{1.2...., k\}$
- $v_1 = 1$ et $v_2 = 524 \Rightarrow F_{1,\infty} = 3.84$
- Ici, $F = 103,362$

$\Rightarrow F_{1,\infty} < F$

- **Conclusion** On rejette l'hypothèse H_0, le modèle est globalement significatif.
- Interprétation en français ?

7 – Inférence

7.5 – Comment lire ces différents résultats sur un logiciel ?

```
Modèle 1: MCO, utilisant les observations 1-526
Variable dépendante: wage

             coefficient   éc. type    t de Student   p. critique
  ---------------------------------------------------------------
  const      -0,904852    0,684968      -1,321        0,1871
  educ        0,541359    0,0532480     10,17         2,78e-22   ***

Moyenne var. dép.      5,896103   Éc. type var. dép.   3,693086
Somme carrés résidus   5980,682   Éc. type régression  3,378390
R2                     0,164758   R2 ajusté            0,163164
F(1, 524)              103,3627   P. critique (F)      2,78e-22
Log de vraisemblance -1385,712   Critère d'Akaike     2775,423
Critère de Schwarz     2783,954   Hannan-Quinn         2778,764
```

Plan du cours

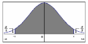
8 – Conclusion

8.1 – A Apprendre

- Définitions importantes
  - estimateur des MCO
  - Les 5 hypothèses
  - définition du biais/Calcul de la variance
- Quelles sont les notions importantes que vous avez identifié ?
  - Estimateur de Gauss-Markov : biais ? variance ?
- Notions importantes
  - Spécification / modèles
  - Commentaires d'estimation au niveau économétrique ET économique

Table Student

Loi de Student



α	1	0,8	0,6	0,4	0,2	0,1	0,05	0,02	0,01	0,002	0,001
$1-\alpha$	0	0,2	0,4	0,6	0,8	0,9	0,95	0,98	0,99	0,998	0,999
ν = ddl											
1	0,0000	0,3249	0,7265	1,3764	3,0777	6,3137	12,706	31,821	63,656	318,29	636,58
2	0,0000	0,2887	0,6172	1,0607	1,8856	2,9200	4,3027	6,9645	9,9250	22,328	31,600
3	0,0000	0,2767	0,5844	0,9785	1,6377	2,3534	3,1824	4,5407	5,8408	10,214	12,924
4	0,0000	0,2707	0,5686	0,9410	1,5332	2,1318	2,7765	3,7469	4,6041	7,1729	8,6101
5	0,0000	0,2672	0,5594	0,9195	1,4759	2,0150	2,5706	3,3649	4,0321	5,8935	6,8685
6	0,0000	0,2648	0,5534	0,9057	1,4398	1,9432	2,4469	3,1427	3,7074	5,2075	5,9587
7	0,0000	0,2632	0,5491	0,8960	1,4149	1,8946	2,3646	2,9979	3,4995	4,7853	5,4081
8	0,0000	0,2619	0,5459	0,8889	1,3968	1,8595	2,3060	2,8965	3,3554	4,5008	5,0414
9	0,0000	0,2610	0,5435	0,8834	1,3830	1,8331	2,2622	2,8214	3,2498	4,2969	4,7809
10	0,0000	0,2602	0,5415	0,8791	1,3722	1,8125	2,2281	2,7638	3,1693	4,1437	4,5868
11	0,0000	0,2596	0,5399	0,8755	1,3634	1,7959	2,2010	2,7181	3,1058	4,0248	4,4369
12	0,0000	0,2590	0,5386	0,8726	1,3562	1,7823	2,1788	2,6810	3,0545	3,9296	4,3178
13	0,0000	0,2586	0,5375	0,8702	1,3502	1,7709	2,1604	2,6503	3,0123	3,8520	4,2209
14	0,0000	0,2582	0,5366	0,8681	1,3450	1,7613	2,1448	2,6245	2,9768	3,7874	4,1403
15	0,0000	0,2579	0,5357	0,8662	1,3406	1,7531	2,1315	2,6025	2,9467	3,7329	4,0728
16	0,0000	0,2576	0,5350	0,8647	1,3368	1,7459	2,1199	2,5835	2,9208	3,6861	4,0149
17	0,0000	0,2573	0,5344	0,8633	1,3334	1,7396	2,1098	2,5669	2,8982	3,6458	3,9651
18	0,0000	0,2571	0,5338	0,8620	1,3304	1,7341	2,1009	2,5524	2,8784	3,6105	3,9217
19	0,0000	0,2569	0,5333	0,8610	1,3277	1,7291	2,0930	2,5395	2,8609	3,5793	3,8833
20	0,0000	0,2567	0,5329	0,8600	1,3253	1,7247	2,0860	2,5280	2,8453	3,5518	3,8496
21	0,0000	0,2566	0,5325	0,8591	1,3232	1,7207	2,0796	2,5176	2,8314	3,5271	3,8193
22	0,0000	0,2564	0,5321	0,8583	1,3212	1,7171	2,0739	2,5083	2,8188	3,5050	3,7922
23	0,0000	0,2563	0,5317	0,8575	1,3195	1,7139	2,0687	2,4999	2,8073	3,4850	3,7676
24	0,0000	0,2562	0,5314	0,8569	1,3178	1,7109	2,0639	2,4922	2,7970	3,4668	3,7454
25	0,0000	0,2561	0,5312	0,8562	1,3163	1,7081	2,0595	2,4851	2,7874	3,4502	3,7251
26	0,0000	0,2560	0,5309	0,8557	1,3150	1,7056	2,0555	2,4786	2,7787	3,4350	3,7067
27	0,0000	0,2559	0,5306	0,8551	1,3137	1,7033	2,0518	2,4727	2,7707	3,4210	3,6895
28	0,0000	0,2558	0,5304	0,8546	1,3125	1,7011	2,0484	2,4671	2,7633	3,4082	3,6739
29	0,0000	0,2557	0,5302	0,8542	1,3114	1,6991	2,0452	2,4620	2,7564	3,3963	3,6595
30	0,0000	0,2556	0,5300	0,8538	1,3104	1,6973	2,0423	2,4573	2,7500	3,3852	3,6460
40	0,0000	0,2550	0,5286	0,8507	1,3031	1,6839	2,0211	2,4233	2,7045	3,3069	3,5510
50	0,0000	0,2547	0,5278	0,8489	1,2987	1,6759	2,0086	2,4033	2,6778	3,2614	3,4960
60	0,0000	0,2545	0,5272	0,8477	1,2958	1,6706	2,0003	2,3901	2,6603	3,2317	3,4602
70	0,0000	0,2543	0,5268	0,8468	1,2938	1,6669	1,9944	2,3808	2,6479	3,2108	3,4350
80	0,0000	0,2542	0,5265	0,8461	1,2922	1,6641	1,9901	2,3739	2,6387	3,1952	3,4164
90	0,0000	0,2541	0,5263	0,8456	1,2910	1,6620	1,9867	2,3685	2,6316	3,1832	3,4019
100	0,0000	0,2540	0,5261	0,8452	1,2901	1,6602	1,9840	2,3642	2,6259	3,1738	3,3905
200	0,0000	0,2537	0,5252	0,8434	1,2858	1,6525	1,9719	2,3451	2,6006	3,1315	3,3398
∞	0,0000	0,2533	0,5244	0,8416	1,2816	1,6449	1,9600	2,3263	2,5758	3,0903	3,2906

Table de Fisher

Table de la loi de Fisher-Snedecor

(Valeurs de F ayant la probabilité P d'être dépassées)

	$\nu_1 = 1$		$\nu_1 = 2$		$\nu_1 = 3$		$\nu_1 = 4$		$\nu_1 = 5$	
ν_2	P = 0,05	P = 0,01	P = 0,05	P = 0,01	P = 0,05	P = 0,01	P = 0,05	P = 0,01	P = 0,05	P = 0,01
1	161,4	4052.00	199,5	4999.00	213,7	3403.00	224,6	5625.00	230,2	5764.00
2	18,51	98,49	19,00	99,00	19,16	99,17	19,25	99,25	19,30	99,30
3	10,13	34,12	9,55	30,81	9,28	29,46	9,12	28,71	9,01	28,24
4	7,71	21,20	6,94	18,00	6,59	16,69	6,39	13,98	6,26	13,32
5	6,61	16,26	5,79	13,27	5,41	12,06	5,19	11,39	5,03	10,97
6	3,99	13,74	3,14	10,91	4,76	9,78	4,53	9,13	4,39	8,75
7	3,39	12,23	4,74	9,35	4,33	8,43	4,12	7,85	3,97	7,45
8	3,32	11,26	4,46	8,63	4,07	7,39	3,84	7,01	3,69	6,63
9	5,12	10,56	4,26	8,02	3,86	6,99	3,63	6,42	3,48	6,06
10	4,96	10,04	4,10	7,56	3,71	6,33	3,48	5,99	3,33	5,64
11	4,84	9,65	3,98	7,20	3,59	6,22	3,36	5,67	3,20	5,32
12	4,75	9,33	3,88	6,93	3,49	5,93	3,26	5,41	3,11	5,06
13	4,67	9,07	3,80	6,70	3,41	5,74	3,18	5,20	3,02	4,86
14	4,60	8,86	3,74	6,31	3,34	5,56	3,11	5,03	2,96	4,69
15	4,34	8,68	3,68	6,36	3,29	5,42	3,06	4,89	2,90	4,56
16	4,49	8,53	3,63	6,23	3,24	5,29	3,01	4,77	2,85	4,44
17	4,45	8,40	3,59	6,11	3,20	5,18	2,96	4,67	2,81	4,34
18	4,41	8,28	3,53	6,01	3,16	5,09	2,93	4,58	2,77	4,25
19	4,38	8,18	3,52	5,93	3,13	5,01	2,90	4,50	2,74	4,17
20	4,35	8,10	3,49	5,85	3,10	4,94	2,87	4,43	2,71	4,10
21	4,32	8,02	3,47	5,78	3,07	4,87	2,84	4,37	2,68	4,04
22	4,30	7,94	3,44	5,72	3,05	4,82	2,82	4,31	2,66	3,99
23	4,28	7,88	3,42	5,66	3,03	4,76	2,80	4,26	2,64	3,94
24	4,26	7,82	3,40	5,61	3,01	4,72	2,78	4,22	2,62	3,90
25	4,24	7,77	3,38	5,37	2,99	4,68	2,76	4,18	2,60	3,86
26	4,22	7,72	3,37	5,33	2,98	4,64	2,74	4,14	2,39	3,82
27	4,21	7,68	3,33	5,49	2,96	4,60	2,73	4,11	2,37	3,78
28	4,20	7,64	3,34	5,43	2,95	4,57	2,71	4,07	2,56	3,75
29	4,18	7,60	3,33	5,42	2,93	4,34	2,70	4,04	2,34	3,73
30	4,17	7,56	3,32	5,39	2,92	4,31	2,69	4,02	2,53	3,70
40	4,08	7,31	3,23	5,18	2,84	4,31	2,61	3,83	2,43	3,31
60	4,00	7,08	3,15	4,98	2,76	4,13	2,32	3,65	2,37	3,34
120	3,92	6,85	3,07	4,79	2,68	3,93	2,43	3,48	2,29	3,17
∞	3,84	6,64	2,99	4,60	2,60	3,78	2,37	3,32	2,21	3,02