Sessions de révision – Session 1

Statistique descriptive

Certificat de chargé d'études statistiques

Martin Chevalier (INSEE)

Objectifs de la session

Rappeler l'ensemble des **outils de la statistique descriptive** uni- et bivariée : représentations graphiques, tableaux, mesures d'association.

Insister sur le **contexte d'utilisation** des différents outils.

Mettre en avant les **considérations pratiques** dans l'utilisation des différents outils et leurs **limites** respectives.

De l'importance de la nature des variables

Les outils à mobiliser sont déterminés par la nature des variables à analyser :

- **Variables quantitatives** : leurs modalités peuvent être précisément exprimées les unes en fonction des autres (cardinalité). On distingue les variables discrètes des variables continues.
- **Variables qualitatives** : leurs modalités ne peuvent pas être précisément exprimées les unes par rapport aux autres. On distingue les variables ordonnées des variables non-ordonnées.

Exemples : Âge (quantitative discrète), salaire (quantitative continue), temps de travail en tranches (qualitative ordonnée), position sur le marché du travail (qualitative non-ordonnée).

Plan de la session

Statistique univariée sur variable qualitative

Tri à plat

Le tri à plat permet d'afficher l'**effectif**, la **fréquence** (sous forme de pourcentages) ainsi que l'effectif et la fréquence **cumulés** dans l'ordre des modalités.

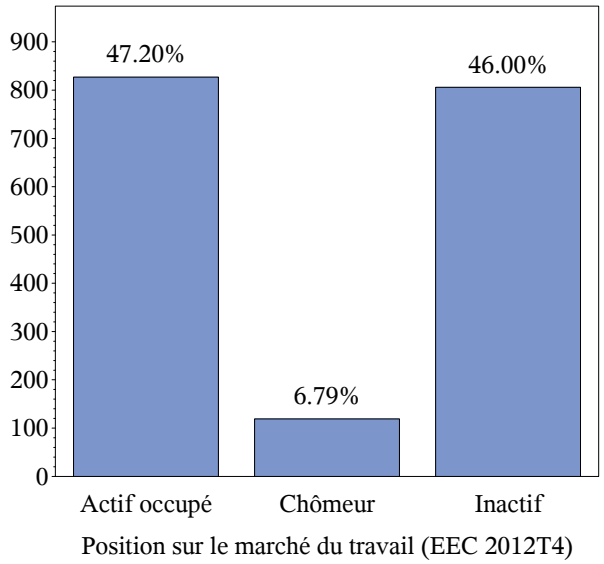
Position sur le marché du travail (EEC 2012T4)				
ACTEU	**Frequency**	**Percent**	**Cumulative Frequency**	**Cumulative Percent**
Actif occupé	827	47.20	827	47.20
Chômeur	119	6.79	946	54.00
Inactif	806	46.00	1752	100.00

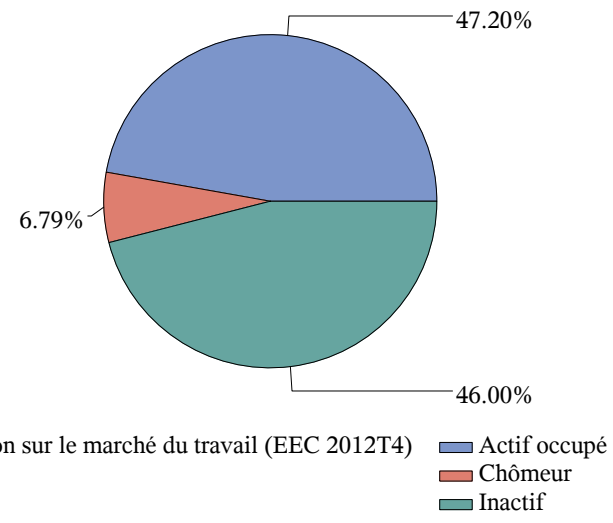
Remarque : Attention à bien utiliser une variable de pondération quand c'est nécessaire (données d'enquête).

Statistique univariée sur variable qualitative

Diagrammes

Ce tri à plat peut être représenté par plusieurs types de diagrammes, en particulier le **diagramme en bâtons** (gauche) et le **diagramme circulaire** (droite).

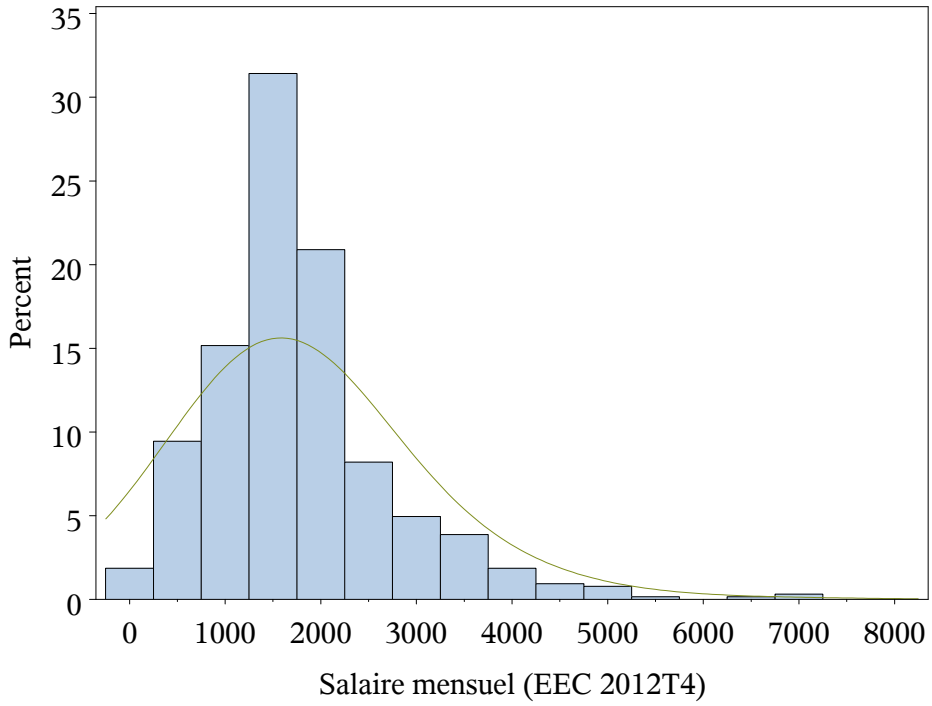




Statistique univariée sur variable quantitative

Histogramme

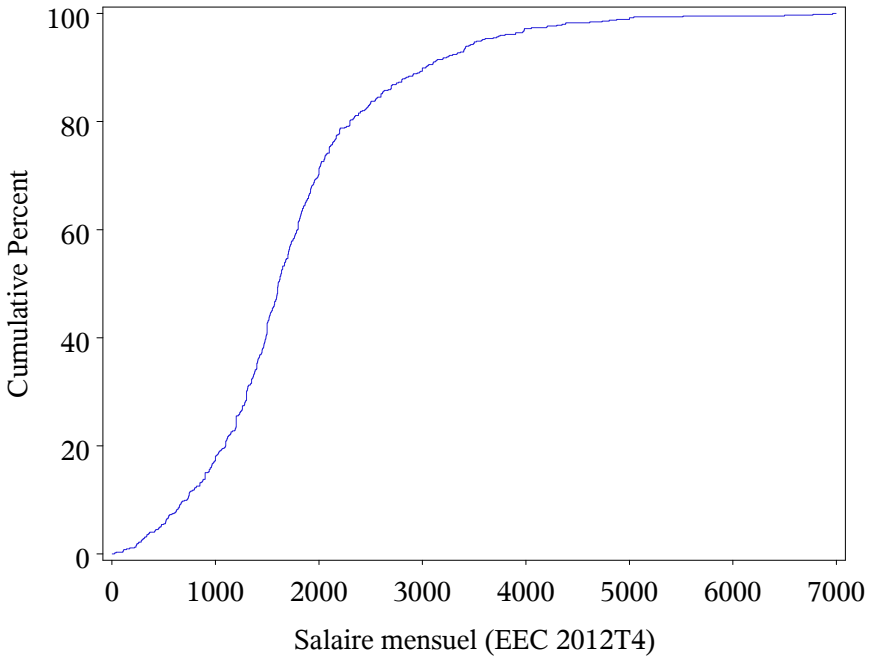
La distribution d'une variable quantitative peut être représentée par un **histogramme**. Son allure dépend de la **largeur des tranches** utilisées.



Statistique univariée sur variable quantitative

Courbe cumulative

La **courbe cumulative** (notée F(a)) représente, pour une valeur donnée *a*, la part des observations dont la valeur est inférieure à *a*.



Statistique univariée sur variable quantitative

Caractéristiques de valeur centrale

Moyenne (arithmétique) $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

→ Limite : très sensible aux valeurs extrêmes.
→ Variantes : moyenne tronquée, moyenne winsorisée.

Médiane Valeur prise par X qui sépare l'ensemble des individus en deux groupes de même taille.
→ Remarque : se lit directement sur la courbe cumulative ($F(Med) = 1/2$).
→ Avantage : très peu sensible aux valeurs extrêmes.

Mode Valeur prise le plus souvent par Y.
→ Remarque : se lit directement sur l'histogramme.
→ Avantage : très peu sensible aux valeurs extrêmes.
→ Limite : peu utile en pratique, en particulier quand la variable est continue.

Statistique univariée sur variable quantitative

Caractéristiques de position

Les quantiles (appelés aussi fractiles) sont construits **à partir de la courbe cumulative** : le quantile de niveau α est la plus petite valeur prise par Y telle que $F(y) \geq \alpha$.

Les **trois quartiles** Q1, Q2 et Q3 séparent la population en quatre sous-ensembles d'effectifs égaux.

De même, les **neuf déciles** D1,...,D9 séparent la population en dix sous-ensembles d'effectifs égaux.

On peut également définir des quintiles, des centiles, etc.

Remarque : La médiane est aussi le deuxième quartile et le cinquième décile.

Statistique univariée sur variable quantitative

Caractéristiques de dispersion

Écart inter-quartile Q3-Q1

Variance empirique $V(Y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$
$\rightarrow$ Inconvénient : d'ordre quadratique par rapport à Y.

Écart-type $s_Y = \sqrt{V(Y)}$
$\rightarrow$ Avantage : du même ordre de grandeur que Y.

Coefficient de variation $CV(Y) = \frac{s_Y}{\bar{y}}$
$\rightarrow$ Avantage : mesure de dispersion relative de Y.

Statistique univariée sur variable quantitative

Concentration : La courbe de Lorenz (1)

Quand la variable quantitative analysée est additive, il est possible de construire des indicateurs de concentration.

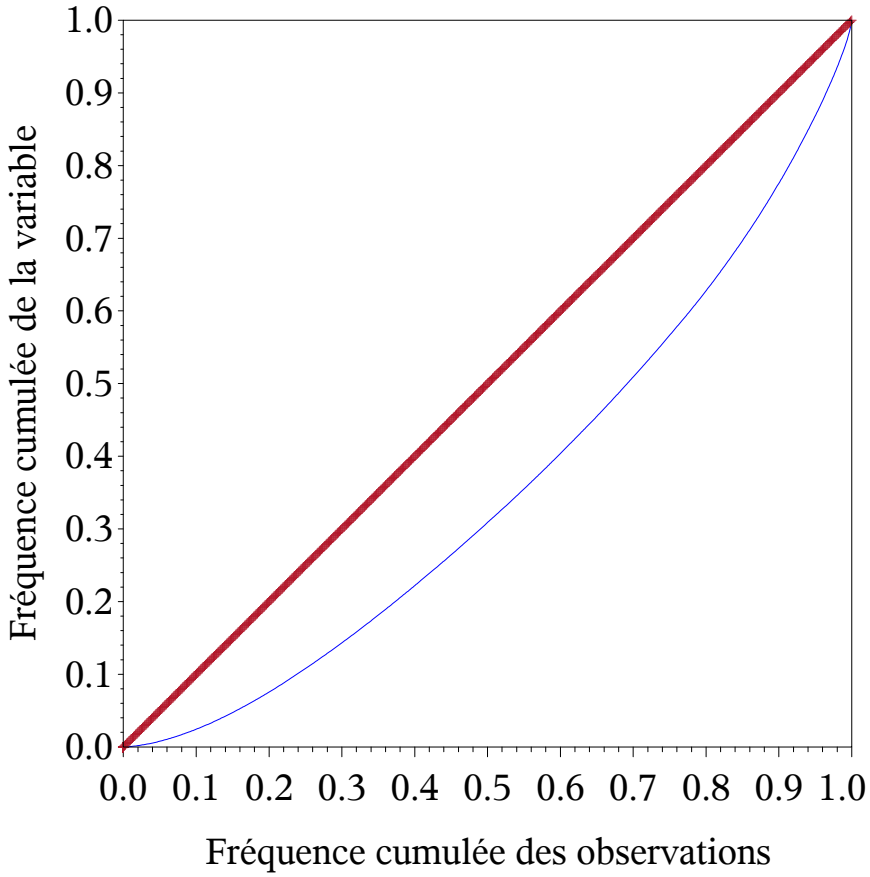
La **courbe de Lorenz** représente ainsi la **fréquence cumulée de la variable** en fonction de la **fréquence cumulée des observations**.

Son interprétation permet de mettre en évidence les **inégalités dans la répartition** de la variable analysée.

Plus la courbe de Lorenz est **éloignée de la première bissectrice**, plus la répartition de la variable analysée est **inégalitaire**.

Statistique univariée sur variable quantitative

Concentration : La courbe de Lorenz (2)



Statistique univariée sur variable quantitative

Concentration : L'indice de Gini

L'indice de Gini G est un **indice synthétique de concentration** d'une variable additive construit à partir de la courbe de Lorenz.

Il correspond à **deux fois l'aire entre la courbe de Gini et la première bissectrice** :

- $G = 0$: la courbe de Lorenz est confondue avec la première bissectrice (**égalité parfaite**) ;
- $G = 1$: la courbe de Lorenz est confondue avec les côtés du carré de côté 1 (**inégalité parfaite**).

Statistique bivariée sur variables qualitatives

Tableau de contingence

Le **tableau de contingence** correspond à la ventilation des observations selon les modalités des deux variables analysées.

Table of ACTEU by SEXE			
ACTEU(Position sur le marché du travail (EEC 2012T4))	SEXE(Sexe)		
	Homme	Femme	Total
Actif occupé	419	408	827
Chômeur	56	63	119
Inactif	360	446	806
Total	835	917	1752

Table of ACTEU by SEXE			
ACTEU(Position sur le marché du travail (EEC 2012T4))	SEXE(Sexe)		
	Homme	Femme	Total
Actif occupé	23.92	23.29	47.20
Chômeur	3.20	3.60	6.79
Inactif	20.55	25.46	46.00
Total	835 47.66	917 52.34	1752 100.00

Les **effectifs marginaux** du tableau coïncident avec le tri à plat sur chacune des variables prise séparément.

Statistique bivariée sur variables qualitatives

Pourcentage en ligne et en colonne

Le pourcentage en ligne (ou profil-ligne) correspond à la distribution de la variable en colonne **conditionnellement aux modalités** de la variable en ligne (gauche).

Table of ACTEU by SEXE			
ACTEU(Position sur le marché du travail (EEC 2012T4))	SEXE(Sexe)		
	Homme	Femme	Total
Actif occupé	50.67	49.33	
Chômeur	47.06	52.94	
Inactif	44.67	55.33	
Total	835	917	1752

Table of ACTEU by SEXE			
ACTEU(Position sur le marché du travail (EEC 2012T4))	SEXE(Sexe)		
	Homme	Femme	Total
Actif occupé	50.18	44.49	
Chômeur	6.71	6.87	
Inactif	43.11	48.64	
Total	835	917	1752

De même, le pourcentage en colonne correspond à la distribution de la variable en ligne conditionnellement aux modalités de la variable en colonne (droite).

Statistique bivariée sur variables qualitatives

Situation d'indépendance

On montre qu'en cas d'indépendance totale des deux variables, l'effectif de chaque cellule devrait être le **produit des effectifs marginaux rapporté au nombre total d'observations**.

Expected

Table of ACTEU by SEXE			
ACTEU(Position sur le marché du travail (EEC 2012T4))	SEXE(Sexe)		
	Homme	Femme	Total
Actif occupé	394.15	432.85	
Chômeur	56.715	62.285	
Inactif	384.14	421.86	
Total	835	917	1752

Statistique bivariée sur variables qualitatives

Statistique du χ^2

Intuition : Plus les écarts entre distribution théorique sous l'hypothèse d'indépendance et distribution effective sont importants, plus les variables sont liées statistiquement.

On définit ainsi la **statistique du χ^2** :

$$D^2 = \sum_{p=1}^{P} \sum_{q=1}^{Q} \frac{(n_{p,q}^{obs} - n_{p,q}^{th})^2}{n_{p,q}^{th}}$$

où les variables X et Y ont respectivement P et Q modalités.

Quand D^2 vaut 0, les deux variables sont indépendantes. **Plus D^2 est grand, plus on est fondé à penser que les deux variables sont liées** (*cf.* session 2 sur le test du χ_2).

Statistique bivariée sur variables qualitatives

Contribution au χ^2

Chacun des termes $\frac{(n_{p,q}^{obs} - n_{p,q}^{th})^2}{n_{p,q}^{th}}$ peut être vu comme la **contribution de la cellule** à la statistique du χ^2.

Cell Chi-Square	Table of ACTEU by SEXE		
	ACTEU(Position sur le marché du travail (EEC 2012T4))	SEXE(Sexe)	
		Homme	Femme	Total
	Actif occupé	1.5672	1.427	
	Chômeur	0.009	0.0082	
	Inactif	1.5168	1.3811	
	Total	835	917	1752

Une valeur élevée permet de détecter une **sur- ou sous-représentation importante** relative à la case.

Statistique bivariée sur variables qualitatives

V de Cramer

La comparaison de statistiques du χ^2 issues de tris croisés différents n'a en général **pas de sens**.

Pour mener à bien ce type de comparaison, on a recours à une statistique construite à partir du χ^2, **le V de Cramer** :

$$V = \left(\frac{D^2}{n \times \min(P-1, Q-1)} \right)$$

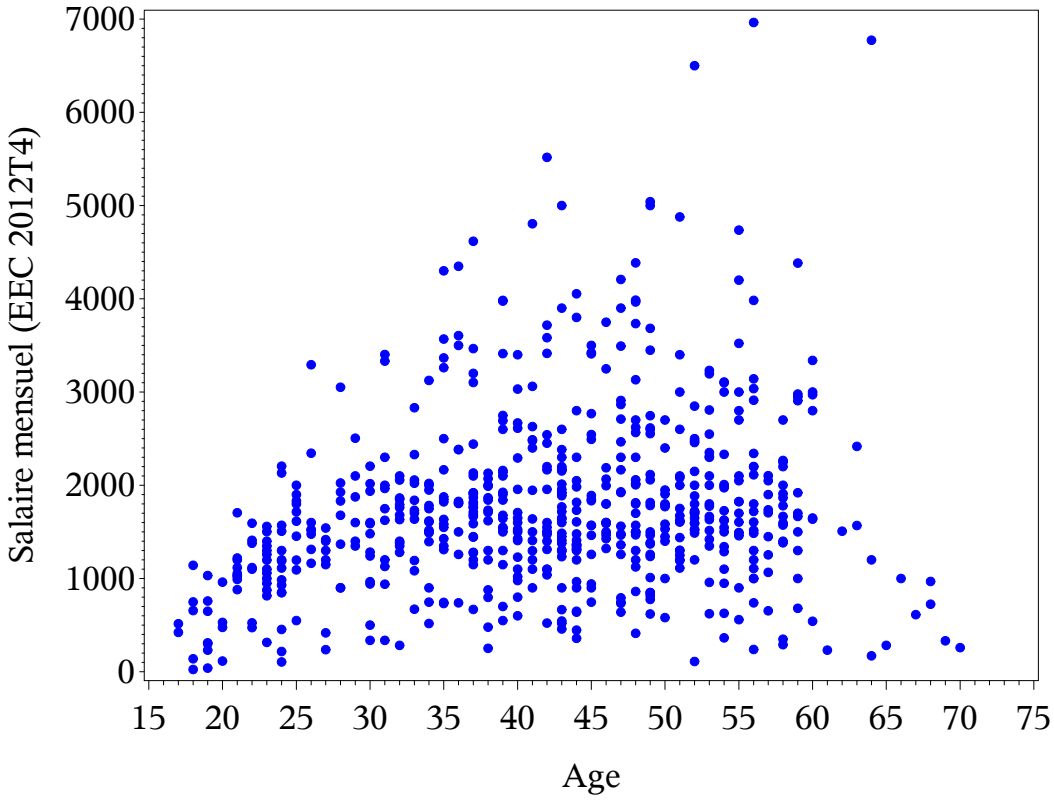
Plus le V de Cramer est proche de 1, plus l'association entre les variables est importante.

Remarque : De nombreux indicateurs de liaison entre variables qualitatives existent, les plus connus étant le coefficient de contingence C et le coefficient ϕ.

Statistique bivariée sur variables quantitatives

Nuage de points

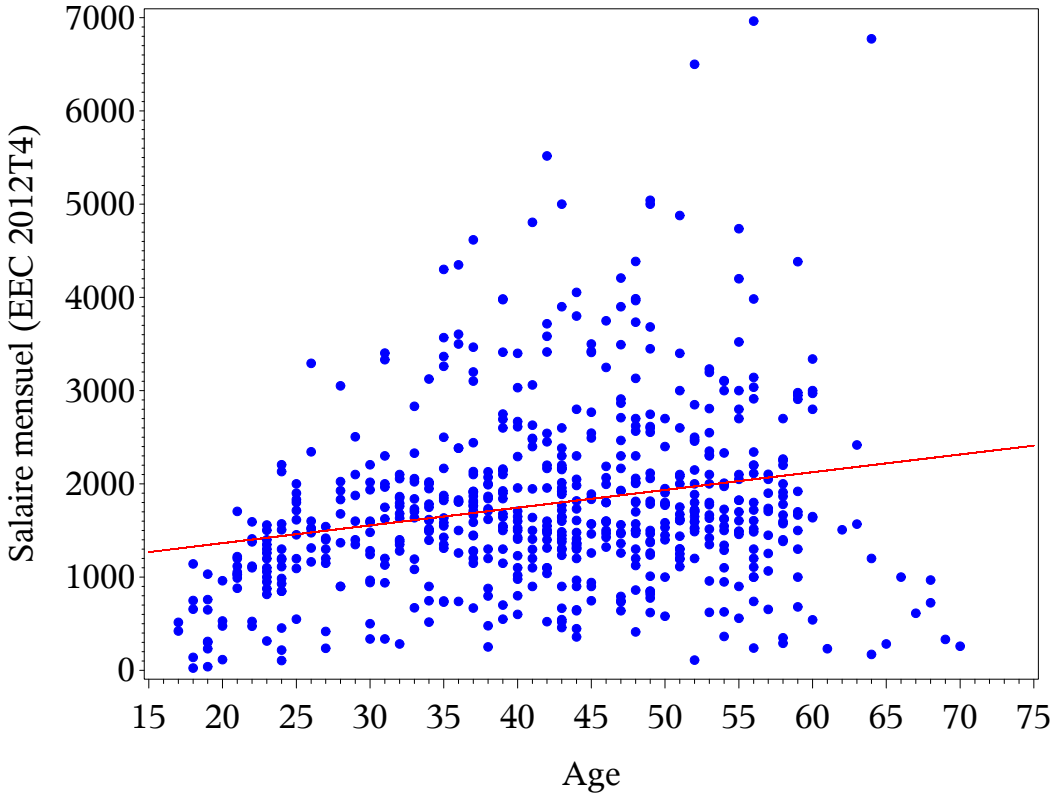
Le **nuage de points** permet de représenter simplement la relation entre deux variables quantitatives.



Statistique bivariée sur variables quantitatives

Droite de régression

Il est également possible de faire apparaître sur le nuage de points la **droite de la régression** linéaire correspondante.



Statistique bivariée sur variables quantitatives

Covariance et coefficient de corrélation de Pearson

Covariance La covariance empirique de deux variables Y et X est définie par $Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$

→ Interprétation :

- $Cov(X, Y) > 0$: Y et X varient dans le même sens ;
- $Cov(X, Y) < 0$: Y et X varient dans des sens contraires.

→ Inconvénient : la valeur de la covariance n'est pas bornée.

Coefficient de corrélation de Pearson

$$r_{X,Y} = \frac{Cov(X, Y)}{\sqrt{V(X)V(Y)}}$$

→ Avantage : $r_{X,Y}$ est compris entre -1 et 1.
→ Limite : comme la moyenne, $r_{X,Y}$ est très sensible aux valeurs extrêmes.

Statistique bivariée sur variables quantitatives

Coefficient de corrélation des rangs de Spearman

La sensibilité aux valeurs extrêmes du coefficient de corrélation de Pearson conduit à lui chercher des **alternatives**.

Le **coefficient de corrélation des rangs** de Spearman en est une :

$$r^S_{X,Y} = r_{R_X,R_Y} = \frac{Cov(R_X, R_Y)}{\sqrt{V(R_X)V(R_Y)}}$$

où R_X et R_Y sont les rangs des observations pour les variables X et Y respectivement.

Comme $r_{X,Y}$, $-1 \leq r^S_{X,Y} \leq 1$ mais $r^S_{X,Y}$ est **beaucoup moins sensible aux valeurs extrêmes**.

Statistique bivariée sur variables quantitatives

τ de Kendall (1)

Intuition : Pour chaque paire d'observations, examiner si la relation entre X et Y est dans le même sens ou de sens contraire.

La paire d'observations i et j...

- ... est dite **concordante** si $\{X_i < X_j \text{ et } Y_i < Y_j\}$ ou si $\{X_i > X_j \text{ et } Y_i > Y_j\}$.
- ... est dite **discordante** si $\{X_i > X_j \text{ et } Y_i < Y_j\}$ ou si $\{X_i < X_j \text{ et } Y_i > Y_j\}$.

Exemple : observations A, B et C

	X	Y
A	1	6
B	2	7
C	3	5

A-B : $X_A < X_B$ et $Y_A < Y_B$ donc **concordante**
A-C : $X_A < X_C$ mais $Y_A > Y_C$ donc **discordante**
B-C : $X_B < X_C$ mais $Y_B > Y_C$ donc **discordante**

Statistique bivariée sur variables quantitatives

τ de Kendall (2)

On note n_c et n_d le nombre de paires respectivement concordantes et discordantes parmi les $n(n-1)/2$ paires au total, et on définit le τ de Kendall par :

$$\tau_{X,Y} = \frac{n_c - n_d}{n(n-1)/2}$$

Principales propriétés :

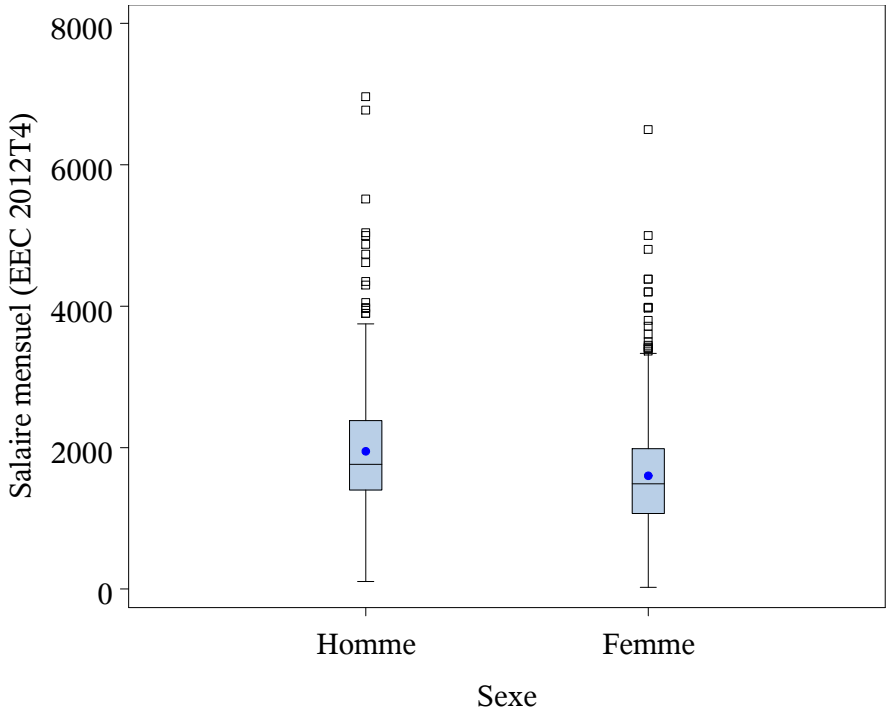
- $-1 \leq \tau_{X,Y} \leq 1$
- $\tau_{X,Y} = 1$: toutes les paires sont concordantes.
- $\tau_{X,Y} = -1$: toutes les paires sont discordantes.

Cet indicateur est insensible aux valeurs extrêmes et peut capter une relation non-linéaire.

Statistique bivariée sur variables quali et quanti

Boîtes à moustaches

Les « boîtes à moustaches » (ou boîtes de Tukey) permettent de **synthétiser côte-à-côte la distribution** d'une variable quantitative selon les modalités d'une variable qualitative.



Statistique bivariée sur variables quali et quanti

Décomposition de la variance

Plus généralement, il est possible de **décomposer la variance** d'une variable expliquée Y selon les K modalités d'une variable explicative X de la façon suivante :

$$V(Y) = \underbrace{\sum_{k=1}^{K} \frac{n_k}{n-1}(\bar{y}_k - \bar{y})^2}_{\text{Variance inter}} + \underbrace{\sum_{k=1}^{K} \frac{n_k - 1}{n-1} V_k(Y)}_{\text{Variance intra}}$$

où $\bar{y}_k$ et $V_k(Y)$ sont **respectivement la moyenne et la variance de Y pour les observations telles que $X = k$.**

Remarque : Cette décomposition de la variance est au fondement des méthodes d'analyse de la variance (*cf.* session 3) et de régression linéaire (*cf.* session 4) .

Statistique bivariée sur variables quali et quanti

Rapport de corrélation

Intuition : Plus la variance inter est grande devant la variance intra, meilleur est le pouvoir explicatif de X sur Y.

On définit ainsi le **rapport de corrélation** :

$$\eta^2_{Y|X} = \frac{\text{Variance inter}}{\text{Variance totale}}$$

Propriétés :

- $0 \leq \eta^2_{Y|X} \leq 1$ (car $0 \leq$ Variance inter $\leq$ Variance totale)
- plus $\eta^2_{Y|X}$ est proche de 1, plus la liaison entre X et Y est forte.

Remarque : Le rapport de corrélation est aussi le R^2 de la régression linéaire de Y sur X (*cf.* session 4).

Sessions de révision – Session 2

Statistique inférentielle

Certificat de chargé d'études statistiques

Martin Chevalier (INSEE)

Objectifs de la session

Revenir sur le principe de la statistique inférentielle : **décrire le comportement d'une statistique à partir de lois de probabilités**.

Insister sur la **construction** et l'**interprétation de tests statistiques** : erreurs de première ou de seconde espèce, statistique de test, valeur critique et p-valeur.

Appliquer les résultats de la statistique inférentielle à des **cas pratiques issus de la statistique descriptive** : test du χ^2, test de significativité du coefficient de corrélation de Pearson.

Plan de la session

Principe de la statistique inférentielle

Objectif : Aller au-delà de l'estimation ponctuelle

La statistique descriptive fournit un grand nombre d'outils permettant de synthétiser la relation entre plusieurs variables.

En tant que tels, ceux-ci ne permettent cependant pas déterminer si deux variables sont statistiquement liées.

Exemple : Un coefficient de corrélation de Pearson de 0,30 est-il suffisamment éloigné de 0 pour considérer que les deux variables sont significativement corrélées ?

Cela revient à mesurer à quel point la statistique d'intérêt S (moyenne, coefficient de corrélation, etc.) est estimée avec précision par les données dont on dispose.

Principe de la statistique inférentielle

Statistique inférentielle et variable aléatoire

En pratique, chaque observation de la variable à analyser est vue comme une **réalisation indépendante** d'une **suite de variables aléatoires de même loi**.

Une variable aléatoire est une **fonction qui associe à chaque événement** issu d'une expérience aléatoire une **valeur numérique**.

Si la loi suivie par les variables aléatoires est connue (*i.e.* tabulée), alors il est possible de **modéliser le comportement des statistiques construites à partir des données**.

En particulier, on peut à partir des données estimer les paramètres de la loi et les utiliser pour construire un **intervalle de confiance** et des **tests** sur les statistiques d'intérêt.

Principe de la statistique inférentielle

Exemple : Jet d'un dé parfaitement équilibré

L'ensemble des éventualités est : $\Omega = \{\boxdot, \boxdot, \boxdot, \boxdot, \boxdot, \boxdot\}$ et la probabilité de chacune est $1/6$. Dans ce contexte, on peut assez naturellement définir la variable aléatoire X :

$$X : \begin{array}{ccc} \Omega & \to & \{1,2,3,4,5,6\} \\ \omega & \mapsto & \begin{cases} 1 \text{ si } \boxdot, 2 \text{ si } \boxdot, 3 \text{ si } \boxdot, \\ 4 \text{ si } \boxdot, 5 \text{ si } \boxdot, 6 \text{ si } \boxdot \end{cases} \end{array}$$

On sait alors (par définition) que la variable aléatoire X suit une loi uniforme sur $\{1,2,3,4,5,6\}$.

Remarque : On aurait aussi bien pu définir la variable aléatoire $\tilde{X}$:

$$\tilde{X} : \begin{array}{ccc} \Omega & \to & \{0,1\} \\ \omega & \mapsto & \begin{cases} 0 \text{ si } \boxdot, \boxdot \text{ ou } \boxdot \\ 1 \text{ si } \boxdot, \boxdot \text{ ou } \boxdot \end{cases} \end{array}$$

Principe de la statistique inférentielle

Caractéristiques d'une variable aléatoire

Fonction de répartition : en notant $\mathbb{P}$ la loi de probabilité associée à la variable aléatoire X, on note F sa fonction de répartition définie par :

$$\begin{array}{rccc} & \mathbb{R} & \to & [0;1] \\ F & : & a & \mapsto & \mathbb{P}(X \leq a) \end{array}$$

Remarques : La **fonction cumulative** (*cf.* session 1) est la contrepartie empirique de la fonction de répartition. La **fonction quantile** est la réciproque de la fonction de répartition.

L'**espérance** et la **variance** sont définies par (cas d'une variable aléatoire discrète prenant K valeurs distinctes) :

$$E(X) = \sum_{k=1}^{K} k \times \mathbb{P}(X = k) \quad \text{et} \quad V(X) = \sum_{k=1}^{K} (k - E(X))^2 \times \mathbb{P}(X = k)$$

Principe de la statistique inférentielle

Quelques lois à connaître (1)

Loi normale (ou gaussienne) La loi normale d'espérance m et de variance σ^2 est notée $\mathcal{N}(m, \sigma^2)$. Sa fonction de répartition est :

$$F(a) = \int_{-\infty}^{a} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}} dx$$

Loi du χ^2 La loi du χ^2 à p degrés de liberté est notée χ^2_p. Si $X_1, ...X_p$ sont indépendantes et suivent une loi normale $\mathcal{N}(0, 1)$, alors $Z = \sum_{k=1}^{p} X_k^2 \hookrightarrow \chi^2_p$ et

$$E(Z) = p \quad \text{et} \quad V(Z) = 2p$$

Principe de la statistique inférentielle

Quelques lois à connaître (2)

Loi de Student La loi de Student à p degrés de liberté est notée $\mathcal{T}_p$. Si $Y \hookrightarrow \mathcal{N}(0,1)$ et $X \hookrightarrow \chi^2_p$ sont indépendantes, alors $Z = \dfrac{Y}{\sqrt{X/p}} \hookrightarrow \mathcal{T}_p$.

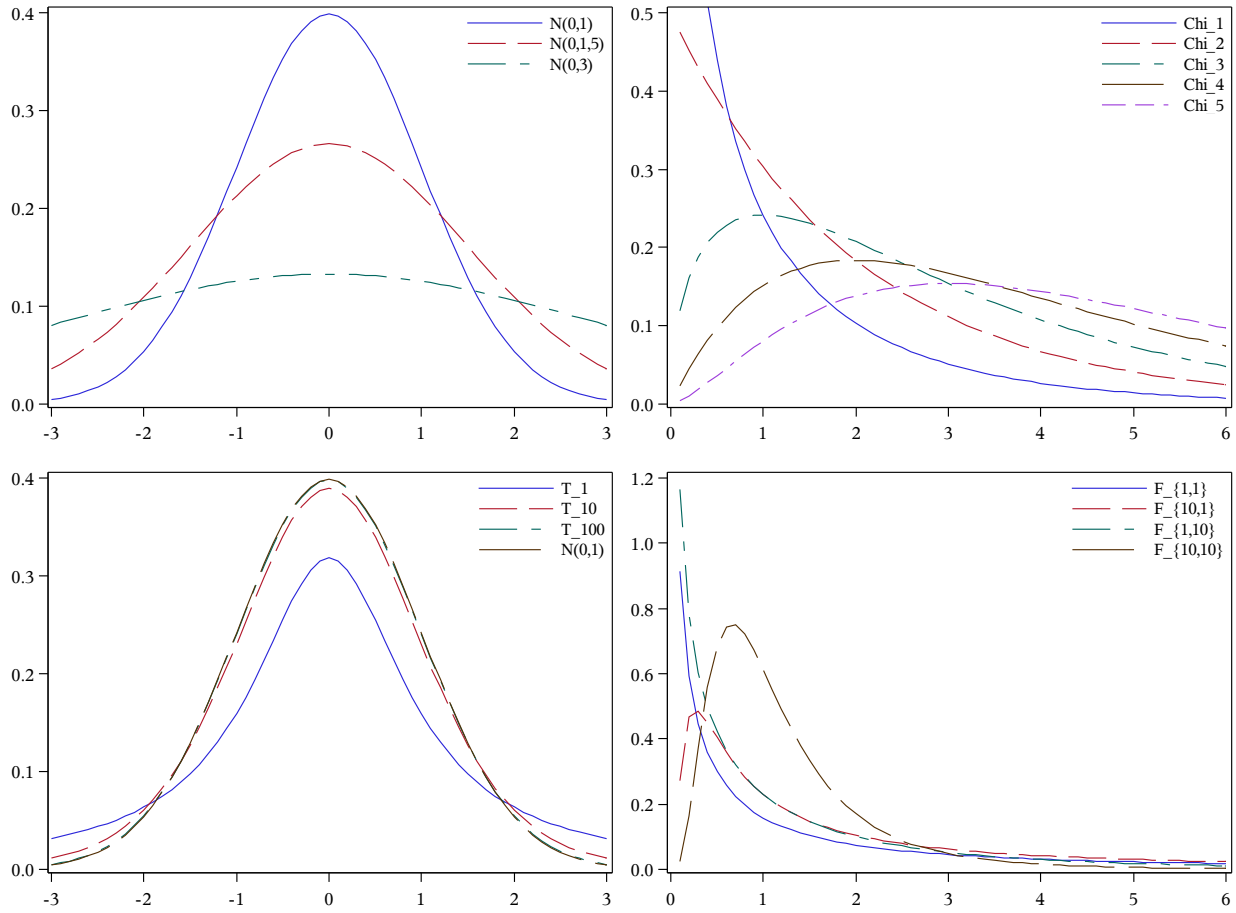
Remarque : Si $Z \hookrightarrow \mathcal{T}_p$ alors $Z \xrightarrow[p\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$

Loi de Fisher La loi de Fisher à q et p degrés de liberté est notée $F_{q,p}$. Si $Y \hookrightarrow \chi^2_q$ et $X \hookrightarrow \chi^2_p$ sont indépendantes, alors $Z = \dfrac{\sqrt{Y/q}}{\sqrt{X/p}} \hookrightarrow F_{q,p}$.

Remarque : $Z \hookrightarrow \mathcal{T}_p \Leftrightarrow Z^2 \hookrightarrow F_{1,p}$

Principe de la statistique inférentielle

Quelques lois à connaître (3)



Principe de la statistique inférentielle

Loi des variables aléatoires en statistique inférentielle

En statistique inférentielle, la loi de la variable aléatoire est en général **pré-déterminée** :

- soit elle est **connue** : expérience explicitement aléatoire (tirage au sort, etc.), phénomène physique, etc. ;
- soit elle est une **hypothèse** : **à distance finie** (moins de 30 observations), il est fréquent que l'on suppose que certaines quantités ont une distribution **gaussienne** ;
- soit elle est **suivie asymptotiquement** : on peut dans certains cas montrer que quand le nombre d'observations est grand (plus de 30), **la variable aléatoire coïncide avec une variable aléatoire suivant la loi**.

Le principal exemple de la troisième configuration est le **théorème central limite**.

Principe de la statistique inférentielle

Théorème central limite (1)

On tire indépendamment $P \times n$ variables aléatoires $X_1^{(1)}, X_2^{(1)}, ..., X_{n-1}^{(P)}, X_n^{(P)}$, suivant une même loi quelconque d'espérance m et de variance σ^2 finies avec $\sigma^2 \neq 0$.

Pour chacun des P groupes de taille n, on calcule la moyenne

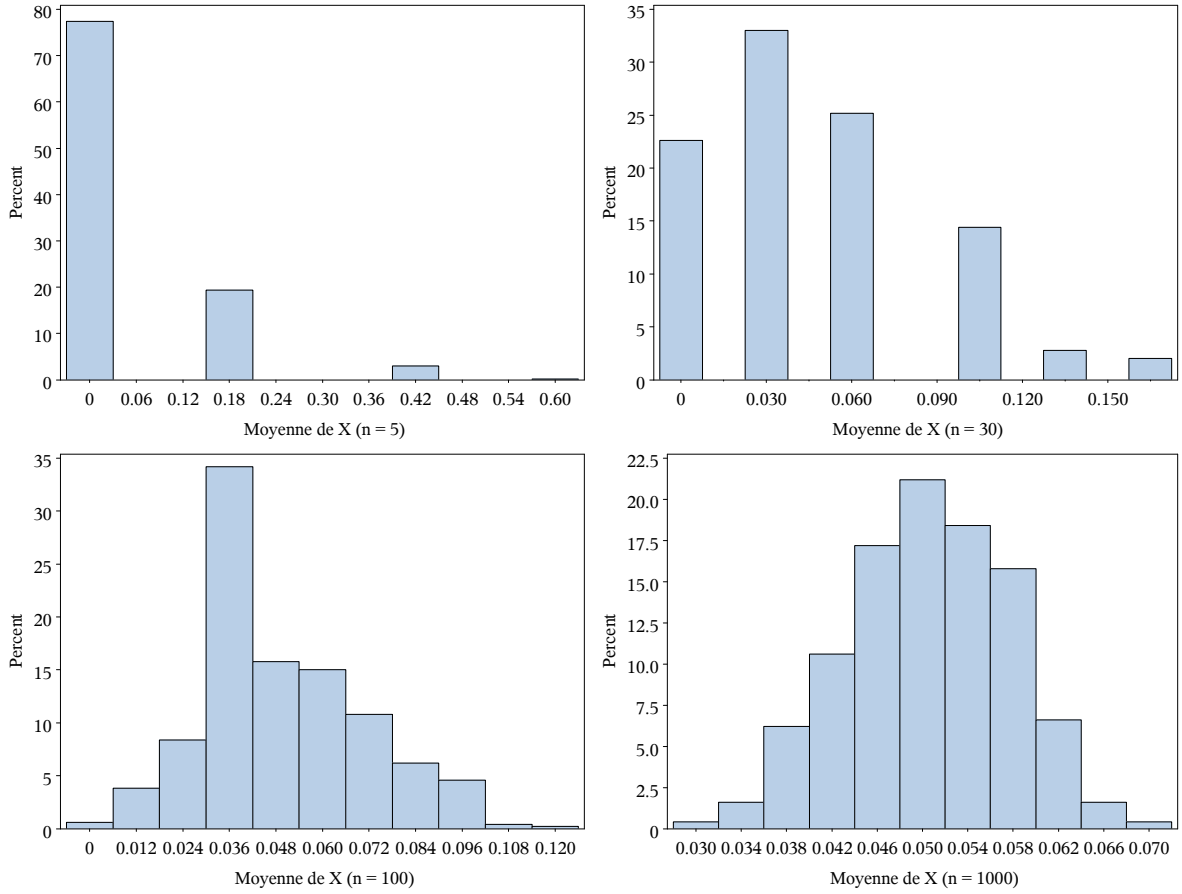
$$\bar{X}^{(p)} = \frac{1}{n} \sum_{i=1}^{n} X_i^{(p)}$$

Si on note $\bar{X}$ la variable aléatoire correspondant à la distribution des P moyennes ainsi calculées, le théorème central limite énonce que :

$$\bar{X} \xrightarrow[n \to +\infty]{\mathcal{L}} \mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$$

Principe de la statistique inférentielle

Théorème central limite (2)



Construire un intervalle de confiance

Cas général

Dès lors que la loi suivie par la statistique d'intérêt S est connue, il est possible de construire un intervalle de confiance au niveau de confiance $(1-\alpha)$ % choisi (en général 95 %).

On suppose ainsi : $S \hookrightarrow \mathcal{L}(\theta)$ ou $S \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{L}(\theta)$ où $\mathcal{L}$ est une loi de paramètre θ.

Par définition, on a alors : $\mathbb{P}\left(q_{\alpha/2}^{\mathcal{L}(\theta)} \leq S \leq q_{1-\alpha/2}^{\mathcal{L}(\theta)}\right) = 1-\alpha$ où $q_x^{\mathcal{L}(\theta)}$ est le quantile d'ordre x de la loi $\mathcal{L}(\theta)$.

On dit alors que l'intervalle $\left[q_{\alpha/2}^{\mathcal{L}(\theta)}; q_{1-\alpha/2}^{\mathcal{L}(\theta)}\right]$ est un intervalle de confiance bilatéral de niveau $(1-\alpha)$ % pour la statistique S.

Construire un intervalle de confiance

Intervalle de confiance d'une moyenne (1)

D'après le théorème central-limite, on sait que quelle que soit la distribution de X :

$$\bar{X} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}\left(m, \frac{\sigma^2}{n}\right)$$

Soit alors $Z = \dfrac{\bar{X} - m}{\sigma/\sqrt{n}}$ la variable aléatoire centrée réduite :

$$Z \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1)$$

En pratique σ^2 est estimée par la variance empirique

$$V(X) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Dans ces conditions $Z = \dfrac{\bar{X} - m}{\sqrt{V(X)/n}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{T}_{n-1}$

Construire un intervalle de confiance

Intervalle de confiance d'une moyenne (2)

En notant $q_x^{\mathcal{T}_{n-1}}$ le quantile d'ordre x de la loi de Student à $n-1$ degrés de liberté, on a alors :

$$\begin{aligned}
1-\alpha &= \mathbb{P}\left(q_{\alpha/2}^{\mathcal{T}_{n-1}} \leq Z \leq q_{1-\alpha/2}^{\mathcal{T}_{n-1}}\right) \\
&= \mathbb{P}\left(-q_{1-\alpha/2}^{\mathcal{T}_{n-1}} \leq Z \leq q_{1-\alpha/2}^{\mathcal{T}_{n-1}}\right) \\
&= \mathbb{P}\left(-q_{1-\alpha/2}^{\mathcal{T}_{n-1}} \leq \frac{\bar{X} - m}{\sqrt{V(X)/n}} \leq q_{1-\alpha/2}^{\mathcal{T}_{n-1}}\right) \\
&= \mathbb{P}\left(\bar{X} - q_{1-\alpha/2}^{\mathcal{T}_{n-1}}\sqrt{\frac{V(X)}{n}} \leq m \leq \bar{X} + q_{1-\alpha/2}^{\mathcal{T}_{n-1}}\sqrt{\frac{V(X)}{n}}\right)
\end{aligned}$$

On définit donc :

$$IC_{(1-\alpha)\ \%}(\bar{X}) = \left[\bar{X} - q_{1-\alpha/2}^{\mathcal{T}_{n-1}}\sqrt{\frac{V(X)}{n}}; \bar{X} + q_{1-\alpha/2}^{\mathcal{T}_{n-1}}\sqrt{\frac{V(X)}{n}}\right]$$

Construire un intervalle de confiance

Exemple : Salaire moyen dans l'EEC 2012T4

Sur les 647 observations de l'EEC 2012T4 utilisées, la variable de salaire a une moyenne $\bar{X}$ de 1 784 € et un écart-type empirique de 1 023 €.

On cherche à construire un intervalle de confiance à 95 % de la moyenne. On a donc $\alpha = 0,05$ et le quantile à $1 - \alpha/2 = 0,975$ d'une loi de Student à $647 - 1 = 646$ degrés de liberté vaut environ $1,96$.

$$IC_{95\ \%}(\bar{X}) = \left[1\ 784 - 1,96 \times \frac{1\ 023}{\sqrt{647}}; 1\ 784 + 1,96 \times \frac{1\ 023}{\sqrt{647}}\right]$$

$$IC_{95\ \%}(\bar{X}) = [1\ 705\ €; 1\ 863\ €]$$

Construire et interpréter un test statistique

Les tests statistiques : des outils pour une prise de décision binaire

Un intervalle de confiance permet d'enrichir considérablement l'analyse en fournissant une mesure de l'imprécision d'une estimation.

Cependant, dans certains cas c'est une **décision binaire** qui est recherchée : il s'agit de déterminer s'il est possible de rejeter l'hypothèse nulle H_0 au profit d'une hypothèse alternative H_1.

Si β est un paramètre aléatoire et c une constante, on peut par exemple poser le test :

$$H_0 : \beta = c \quad \text{contre} \quad H_1 : \beta \neq c$$

La théorie des tests permet de formaliser d'un point de vue statistique cette prise de décision.

Construire et interpréter un test statistique

Exemple : Test de dépistage (1)

Les **tests de dépistages** utilisées en santé publique peuvent être envisagés comme des tests statistiques.

En effet, ils ne sont pas totalement déterministes. Pour de multiples raisons, le procédé utilisé pour le dépistage peut produire des erreurs :

- soit il peut ne pas détecter un sujet infecté ;
- soit il peut détecter à tort un sujet sain comme infecté.

Le premier type d'erreur est dans ce cas clairement le plus grave : un sujet considéré à tort comme sain ne va pas pouvoir être soigné.

Construire et interpréter un test statistique

Les deux types d'erreur et choix de l'hypothèse H_0

Comme le test de dépistage, tout test peut produire deux types d'erreur, qui ne peuvent être minimisées simultanément.

RéalitéTest	H_0 acceptée	H_0 rejetée
H_0 vraie	*Niveau* 1 - α	Erreur de 1ère espèce α
H_0 fausse	Erreur de 2nd espèce β	*Puissance* 1 - β

Dans ce contexte, la stratégie dite de Neyman-Pearson conduit à distinguer les deux erreurs selon leur gravité :

- d'abord l'erreur de première espèce est contrôlée à un niveau α choisi ;
- puis l'erreur de seconde espèce est minimisée autant que possible.

Construire et interpréter un test statistique

Exemple : Test de dépistage (2)

L'erreur consistant à ne pas détecter des sujets infectés étant la plus grave, dans le cas des tests de dépistage le test statistique est posé de la façon suivante. On teste :

$$H_0 : \{\text{le sujet est infecté}\} \text{ contre } H_1 : \{\text{le sujet est sain}\}$$

L'erreur de première espèce est alors bien {le sujet est infecté mais pas détecté} et l'erreur de seconde espèce {le sujet est sain mais détecté comme infecté}.

On peut alors fixer un niveau aussi élevé que l'on souhaite (99,99 % par exemple) et chercher ensuite à minimiser autant que possible la part de sujets sains détectés comme infectés.

Les erreurs de première espèce couramment acceptées sont 5 % (0,05) et 1 % (0,01). 10 % (0,10) est parfois également acceptée.

Construire et interpréter un test statistique

Choix de la statistique de test

Les propriétés d'un test (en particulier sa puissance à un niveau donné) résultent du choix de la statistique de test t.

Remarque : On parle également de « variable de décision » ou de « statistique pivotale ».

Une statistique de test présente deux propriétés :

- « statistique » : elle doit pouvoir être calculée à partir des données dont on dispose ;
- « de test » : sous H_0

$$t \hookrightarrow \mathcal{L}(\theta) \quad \text{ou} \quad \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{L}(\theta)$$

avec $\mathcal{L}(\theta)$ une loi tabulée de paramètre θ.

Construire et interpréter un test statistique

Région critique et valeur critique

Intuition : En comparant la valeur prise par la statistique de test aux valeurs qu'elle prend en général quand H_0 est vérifiée, il est possible de juger si H_0 semble une hypothèse raisonnable.

On utilise les quantiles de $\mathcal{L}(\theta)$ pour déterminer une **région critique** W telle que

$$\mathbb{P}(t \in W) = \alpha$$

où α est le risque de première espèce souhaité.

En pratique, on peut dans la plupart des cas se ramener à une région critique de la forme $[q; +\infty[$ où q est la **valeur critique** : c'est la valeur d'un quantile de la loi $\mathcal{L}(\theta)$.

Construire et interpréter un test statistique

Détermination de la valeur critique

Pour un niveau de confiance $1 - \alpha$ donné, le quantile q de $\mathcal{L}(\theta)$ à utiliser pour déterminer la valeur critique dépend du caractère **bilatéral** ou **unilatéral** du test.

Test bilatéral $\quad H_0 : \beta = c \quad$ contre $\quad H_1 : \beta \neq c$

La valeur critique est le quantile à $1 - \alpha/2$ de $\mathcal{L}(\theta)$.

Exemple : Test de significativité d'un coefficient dans une régression (*cf.* sessions 4 et 5).

Test unilatéral $\quad H_0 : \beta = c \quad$ contre $\quad H_1 : \beta > c$

La valeur critique est le quantile à $1 - \alpha$ de $\mathcal{L}(\theta)$.

Exemple : Test de significativité globale dans une régression (*cf.* sessions 4 et 5).

Construire et interpréter un test statistique

Exemple : Statistique de test suivant une loi normale centrée réduite

$$H_0 : \beta = c \quad \text{contre} \quad \beta \neq c$$

On dispose d'une statistique de test t telle que sous H_0 $t \hookrightarrow \mathcal{N}(0,1)$ et on souhaite mener le test au seuil $\alpha = 0,05$. La valeur de t est 1,70.

Le test est un test bilatéral donc la valeur critique est le quantile à $1 - \alpha/2 = 97,5$ % d'une loi normale centrée réduite, c'est-à-dire 1,96 : t $<$ 1,96 donc on ne peut pas rejeter H_0 à 5 %.

Si maintenant on mène le test $H_0 : \beta = c \quad \text{contre} \quad \beta > c$ qui est unilatéral, alors la valeur critique est le quantile à $1 - \alpha =$ 95 % d'une loi normale centrée réduite, c'est-à-dire 1,65 : t $>$ 1,65 donc on peut dans ce cas rejeter H_0 à 5 %.

Construire et interpréter un test statistique

Lecture du test et p-valeur

Pour interpréter un test à partir de la valeur de la statistique de test, il est donc nécessaire de consulter la table des quantiles de la loi que suit la statistique de test sous H_0.

Également calculable à partir du test, la **p-valeur** permet de s'affranchir de cette contrainte.

Elle peut être vue comme le **seuil limite au-delà duquel il n'est plus possible de rejeter l'hypothèse nulle**.

Si ce seuil limite est inférieur à un des seuils couramment utilisés (0,10, 0,05, 0,01) alors on peut rejeter l'hypothèse nulle à ce seuil.

Exemple : Si la p-valeur d'un test vaut 0,025, alors on peut rejeter l'hypothèse nulle au seuil de 5 % mais pas au seuil de 1 %.

Application : Le test du χ^2

Objectif et hypothèses du test

Le test d'indépendance entre deux variables qualitatives est une des applications les plus courantes de la théorie des tests en statistique descriptive.

Ce test cherche à déterminer si deux variables qualitatives X et Y sont statistiquement liées.

Dans une approche prudente, on pose le test de la façon suivante :

$$H_0 : \{X \text{ et } Y \text{ sont indépendantes}\}$$
$$\text{contre}$$
$$H_1 : \{X \text{ et } Y \text{ ne sont pas indépendantes}\}$$

On contrôle ainsi le risque d'affirmer à tort que X et Y sont liées ($1^{\text{ère}}$ espèce) en minimisant autant que possible le risque de ne pas détecter des associations significatives (2^{nd} espèce).

Application : Le test du χ^2

Statistique de test

L'analyse du tableau de contingence a conduit à construire la statistique D^2 :

$$D^2 = \sum_{p=1}^{P} \sum_{q=1}^{Q} \frac{(n_{p,q}^{obs} - n_{p,q}^{th})^2}{n_{p,q}^{th}}$$

Cette quantité est calculable sur les données et on peut montrer que sous l'hypothèse H_0 d'indépendance elle suit une loi du χ^2 à $(P-1)(Q-1)$ degrés de liberté.

D^2 est donc une statistique de test permettant de mener à bien le test d'indépendance entre X et Y.

Application : Le test du χ^2

Exemple : Position sur le marché du travail et sexe dans l'EEC 2012T4

La valeur critique pour ce test unilatéral au seuil de 5 % est $q^{\chi^2_2}_{0,95} = 5,99$. $D^2 = 5,91 < 5,99$ donc on ne peut pas rejeter H_0 au seuil de 5 %.

En revanche $5,91 > q^{\chi^2_2}_{0,90} = 4,61$ donc on peut rejeter H_0 au seuil de 10 %.

Statistic	DF	Value	Prob
Chi-Square	2	5.9093	0.0521
Likelihood Ratio Chi-Square	2	5.9127	0.0520
Mantel-Haenszel Chi-Square	1	5.8896	0.0152
Phi Coefficient		0.0581	
Contingency Coefficient		0.0580	
Cramer's V		0.0581	

On aboutit directement à la même conclusion en interprétant la p-valeur, qui est supérieure à 0,05 mais inférieure à 0,10.

Application : Le test de significativité de $r_{X,Y}$

Objectif et hypothèses du test

Le test de significativité de $r_{X,Y}$ vise à déterminer si deux variables quantitatives sont statistiquement liées.

En pratique, cela revient à tester si le coefficient de corrélation de Pearson $r_{X,Y}$ est significativement différent de 0.

Dans une approche prudente, on pose le test de la façon suivante :

$$H_0 : r_{X,Y} = 0 \quad \text{contre} \quad H_1 : r_{X,Y} \neq 0$$

On contrôle ainsi le risque d'affirmer à tort que X et Y sont liées ($1^{\text{ère}}$ espèce) en minimisant autant que possible le risque de ne pas détecter des associations significatives (2^{nd} espèce).

Application : Le test de significativité de $r_{X,Y}$

Statistique de test

On peut montrer que sous l'hypothèse H_0 de nullité du coefficient de corrélation de Pearson $r_{X,Y}$:

$$t = r_{X,Y} \times \sqrt{\frac{n-2}{1-r_{X,Y}^2}} \hookrightarrow \mathcal{T}_{n-2}$$

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations		
	SALRED	**age**
SALRED Salaire mensuel (EEC 2012T4)	1.00000 647	0.21024 <.0001 647
age Age	0.21024 <.0001 647	1.00000 1752

En pratique dans SAS, on utilise la p-valeur du test pour accepter ou rejeter l'hypothèse pour un niveau de confiance $1 - \alpha$ donné.

Sessions de révision – Session 3

Analyse de variance

Certificat de chargé d'études statistiques

Martin Chevalier (INSEE)

Objectifs de la session

Articuler les différentes méthodes d'analyse de la variance présentées lors des différentes sessions : test d'égalité des moyennes (T-test), ANOVA, ANCOVA.

Souligner les liens entre analyse de la variance et statistique descriptive d'une part, analyse de la variance et méthodes de régression d'autre part.

Présenter des exemples de mise en œuvre pratique.

Plan de la session

Définition, contexte et hypothèses

Définition

L'analyse de la variance (ANOVA) d'une variable Y quantitative est la décomposition de sa variance empirique $V(Y)$ selon les modalités d'une ou plusieurs variables explicatives de nature qualitative.

Remarque : Cette décomposition est déjà exploitée en statistique descriptive pour calculer le rapport de corrélation $\eta_{Y|X}$.

Plus précisément, l'analyse de la variance désigne l'interprétation de tests statistiques construits à partir de cette décomposition.

On parle d'analyse de la covariance (ANCOVA) de Y dès lors qu'une des variables explicatives est de nature quantitative.

Définition, contexte et hypothèses

Formulation du test

L'objectif du test associé à l'analyse de la variance est de déterminer si les variables Y et X sont statistiquement liées avec un niveau de confiance raisonnable (95 % ou 99 %).

En pratique, cela revient à tester l'égalité des moyennes $m_1, ..., m_K$ de Y au sein des groupes d'observations définis par les K modalités de la variable X.

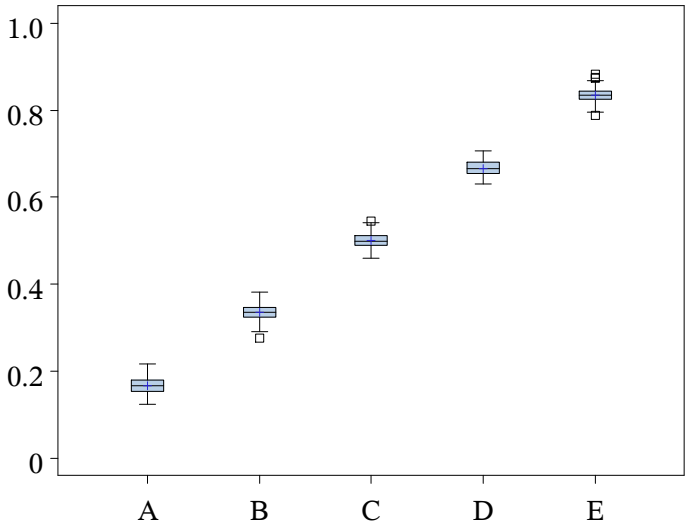
Le test associé à l'analyse de la variance de la variable Y se formule donc de la façon suivante :

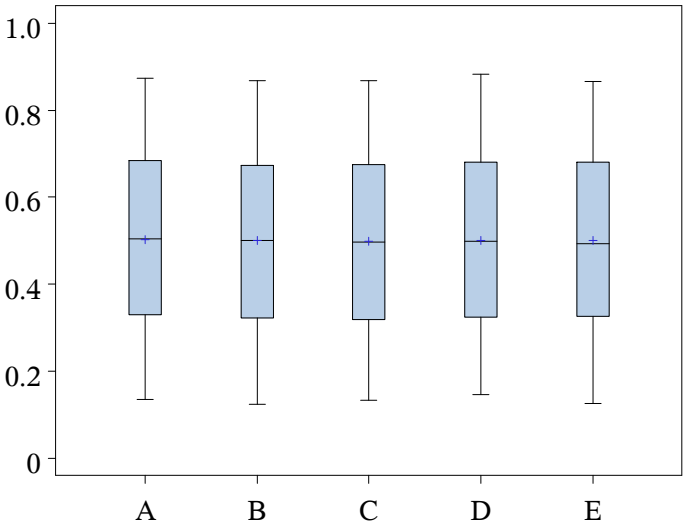
$$H_0 : m_1 = \ldots = m_K = m = \text{constante}$$
$$\text{contre}$$
$$H_1 : \exists k \in \{1, \ldots, K\} \quad m_k \neq m$$

Définition, contexte et hypothèses

Intuition du test

Si les moyennes observées de Y diffèrent fortement d'un groupe à l'autre (gauche), alors il est possible de conclure à une association significative entre Y et X.





Si les moyennes observées de Y sont très proches d'un groupe à l'autre (droite), alors il n'est pas possible de conclure à une association significative entre Y et X.

Définition, contexte et hypothèses

Analyse de la variance et T-test

Le test dit d'égalité des moyennes (ou T-test) cherche à déterminer si la moyenne de Y diffère significativement au sein de deux groupes.

Le test d'égalité des moyennes est donc un cas particulier d'analyse de variance, où la variable explicative X ne présente que deux modalités distinctes.

Les statistiques de test diffèrent mais sont liées, de même que les lois qu'elles suivent sous H_0 : on aboutit donc exactement au même résultat avec les deux méthodes.

Définition, contexte et hypothèses

Analyse de la variance et méthodes de régression

Toute analyse de la variance peut être vue comme un modèle de régression linéaire.

La différence majeure entre les deux approches réside dans l'interprétation que l'on souhaite effectuer :

- l'analyse de la variance permet de déterminer si une variable explicative **prise globalement** est statistiquement liée à la variable d'intérêt ;
- la régression linéaire permet de déterminer si **certaines modalités** de la variable explicative sont significativement associées à des niveaux différents de la variable expliquée.

L'analyse de la variance présente une dimension **plus descriptive** que les méthodes de régression, qui cherchent davantage à mettre en évidence des **relations de causalité**.

Définition, contexte et hypothèses

Hypothèses de l'ANOVA

L'analyse de la variance repose sur trois hypothèses principales :

- **Indépendance** Les observations constituant les différents groupes peuvent être considérées comme tirées indépendamment les unes des autres.
- **Normalité** Les observations correspondent à des variables aléatoires suivant des lois normales.
- **Homogénéité** La variance des lois suivies par les observations des différents groupes est la même d'un groupe à l'autre.

Remarque : L'hypothèse d'homogénéité correspond à l'hypothèse d'homoscédasticité dans le contexte des méthodes de régression.

Définition, contexte et hypothèses

Test des hypothèses de l'ANOVA

Test de l'hypothèse de normalité : Shapiro-Wilk (1965)

$$H_0 : \text{ la distribution est normale}$$
$$\text{contre}$$
$$H_1 : \text{ la distribution n'est pas normale}$$

On rejette l'hypothèse nulle quand la statistique de test W est éloignée de 1 (les valeurs critiques sont spécifiquement tabulées).

Test de l'hypothèse d'homogénéité : Bartlett (1937)

$$H_0 : \sigma_1^2 = \ldots = \sigma_K^2 = \sigma^2 = \text{constante}$$
$$\text{contre}$$
$$H_1 : \exists k \in \{1, \ldots, K\} \quad \sigma_k^2 \neq \sigma^2$$

La statistique de test B suit une loi χ^2_{K-1} : on rejette donc H_0 au seuil α dès lors que $B > q_{1-\alpha}^{\chi^2_{K-1}}$.

Définition, contexte et hypothèses

Exemple : Salaire dans l'EEC 2012T4

La variable de salaire de l'EEC au 2012T4 échoue à chacun des deux tests (croisement avec le sexe pour le test de Bartlett).

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.861441	**Pr < W**	<0.0001
Kolmogorov-Smirnov	**D**	0.131735	**Pr > D**	<0.0100
Cramer-von Mises	**W-Sq**	3.065624	**Pr > W-Sq**	<0.0050
Anderson-Darling	**A-Sq**	16.55847	**Pr > A-Sq**	<0.0050

Bartlett's Test for Homogeneity of SALRED Variance			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
SEXE	1	16.4052	<.0001

Définition, contexte et hypothèses

Exemple : Salaire simulé dans l'EEC 2012T4

Pour illustrer les méthodes d'ANOVA à l'aide de l'EEC 2012T4, une variable de salaire est resimulée.

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.997658	**Pr < W**	0.3202
Kolmogorov-Smirnov	**D**	0.022097	**Pr > D**	>0.1500
Cramer-von Mises	**W-Sq**	0.074374	**Pr > W-Sq**	0.2475
Anderson-Darling	**A-Sq**	0.465601	**Pr > A-Sq**	>0.2500

Bartlett's Test for Homogeneity of salred_simul Variance			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
SEXE	1	0.3691	0.5435

Variable explicative dichotomique : le T-test

Reformulation du test d'égalité des moyennes

La variable explicative X est dichotomique : elle partitionne les observations en deux groupes 1 et 2.

Dans ce cadre, il est possible de reformuler le test d'égalité des moyennes de façon plus simple :

$$H_0 : m_1 - m_2 = 0 \quad \text{contre} \quad H_1 : m_1 - m_2 \neq 0$$

Cette reformulation conduit à une statistique de test spécifique, dont le comportement sous H_0 est connu, que l'hypothèse d'homogénéité soit respectée ou non.

Remarque : On se place sous l'hypothèse que les deux échantillons sont effectivement indépendants (pas d'observations pairées).

Variable explicative dichotomique : le T-test

Cas 1 : Hypothèse d'homogénéité respectée

Sous les hypothèses d'indépendance, de normalité et d'homogénéité, on peut montrer que sous H_0 :

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \hookrightarrow \mathcal{T}_{n-2}$$

où S_p^2 est une somme pondérée de la variance empirique de Y dans les groupes 1 et 2 :

$$S_p^2 = \frac{1}{n-2}\left[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\right]$$

La statistique de test t est donc bien calculable : en comparant sa valeur à la valeur critique donnée par le quantile de $\mathcal{T}_{n-2}$ à $1 - \alpha/2$, il est possible de conclure.

Variable explicative dichotomique : le T-test

Cas 2 : Hypothèse d'homogénéité relâchée

Sous les hypothèses d'indépendance et de normalité, on peut montrer que sous H_0 :

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \hookrightarrow \mathcal{T}_{ddl}$$

avec

$$ddl = \frac{(n_1 - 1)(n_2 - 1)}{(n_1 - 1)C^2 + (n_1 - 1)(1 - C)^2}$$

et

$$C = \frac{S_1^2/n_1}{S_1^2/n_1 + S_2^2/n_2}$$

Remarque : Plusieurs méthodes existent pour approximer le nombre de degrés de liberté. La méthode de Satterthwaite (1946) implémentée par défaut dans SAS diffère de celle présentée ici.

Variable explicative dichotomique : le T-test

Exemple : Salaire simulé et sexe dans l'EEC 2012T4

SEXE	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	402	1903.9	621.6	31.0048	268.9	3651.4
2	401	1702.1	603.0	30.1145	-0.00227	3455.2
Diff (1-2)		201.8	612.4	43.2241		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	801	4.67	<.0001
Satterthwaite	Unequal	800.38	4.67	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	401	400	1.06	0.5435

Variable explicative dichotomique : le T-test

Exemple : Salaire effectif et sexe dans l'EEC 2012T4

SEXE	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	316	1975.0	1115.9	62.7721	106.0	10798.0
2	331	1601.5	889.8	48.9100	24.0000	6500.0
Diff (1-2)		373.5	1006.6	79.1670		

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	645	4.72	<.0001
Satterthwaite	Unequal	601.84	4.69	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	315	330	1.57	<.0001

Variable explicative polytomique

Décomposition de la somme des carrés totaux

La principale limite du test d'égalité des moyennes est qu'il n'est applicable que quand la variable X est dichotomique.

Pour généraliser ce test à $K > 2$, on repart de la formule de décomposition de la variance de Y :

$$\underbrace{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}_{V(Y)} = \underbrace{\sum_{k=1}^{K}\frac{n_k}{n-1}(\bar{y_k}-\bar{y})^2}_{\text{Variance inter}} + \underbrace{\sum_{k=1}^{K}\frac{n_k-1}{n-1}V_k(Y)}_{\text{Variance intra}}$$

En multipliant par $(n-1)$ et en remplaçant $V_k(Y)$ par sa valeur $\frac{1}{n_k-1}\sum_{i=1}^{n_k}(y_i-\bar{y_k})^2$, on obtient la formule de décomposition de la somme des carrés totaux :

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\substack{\text{Somme des carrés}\\ \text{totaux (SCT)}}} = \underbrace{\sum_{k=1}^{K}n_k(\bar{y_k}-\bar{y})^2}_{\substack{\text{Somme des carrés}\\ \text{expliqués (SCE)}}} + \underbrace{\sum_{k=1}^{K}\sum_{i=1}^{n_k}(y_i-\bar{y_k})^2}_{\substack{\text{Somme des carrés}\\ \text{résiduels (SCR)}}}$$

Variable explicative polytomique

Statistique de test

Sous les hypothèses d'indépendance, de normalité et d'homogénéité, on peut montrer que sous H_0 :

$$F = \frac{SCE/(K-1)}{SCR/(n-K)} \hookrightarrow F_{K-1,n-K}$$

Les quantités $SCE/(K-1)$ et $SCR/(n-K)$ sont appelées respectivement **carré moyen expliqué** et **carré moyen résiduel**.

Intuition : Plus la variance inter est importante, plus le **carré moyen expliqué est grand devant le carré moyen résiduel** et plus on va avoir tendance à rejeter l'hypothèse H_0 d'égalité des moyennes.

Variable explicative polytomique

Exemple : Salaire simulé et diplôme dans l'EEC 2012T4

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	102907602.4	20581520.5	88.29	<.0001
Error	641	149428278.9	233117.4		
Corrected Total	646	252335881.3			

R-Square	Coeff Var	Root MSE	salred_simul Mean
0.407820	26.98843	482.8224	1788.998

Source	DF	Anova SS	Mean Square	F Value	Pr > F
DDIPL	5	102907602.4	20581520.5	88.29	<.0001

Variable explicative polytomique

Exemple : Salaire simulé et sexe dans l'EEC 2012T4

En menant une analyse de la variance sur salaire simulé selon le sexe, on retrouve le même résultat qu'avec le T-test.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8176077.8	8176077.8	21.80	<.0001
Error	801	300427052.8	375065.0		
Corrected Total	802	308603130.5			

Source	DF	Anova SS	Mean Square	F Value	Pr > F
SEXE	1	8176077.769	8176077.769	21.80	<.0001

En particulier, on remarque que $F = t^2$: $21,80 = 4,67^2$ (aux arrondis près).

Variable explicative polytomique

Mise en œuvre du test de Tukey

Si l'association globale entre Y et X peut être jugée statistiquement significative, alors il est intéressant d'analyser les écarts entre les groupes pris deux-à-deux.

C'est le principe du test de Tukey : **pour chaque paire de modalités, la significativité de l'écart entre les moyennes de Y est testée**.

Remarque : Cette comparaison des moyennes est analogue à celle effectuée par le T-test, sans coïncider exactement pour autant : le test de Tukey tient en effet compte des autres modalités de la variable explicative intégrée dans l'ANOVA.

Variable explicative polytomique

Exemple : Salaire simulé et diplôme dans l'EEC 2012T4

Alpha	0.05
Error Degrees of Freedom	797
Error Mean Square	342750.7
Critical Value of Studentized Range	4.03993

Comparisons significant at the 0.05 level are indicated by *.**				
DDIPL Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
1 - 3	165.94	-37.00	368.87	
1 - 5	444.18	262.88	625.48	***
1 - 4	466.48	272.99	659.97	***
1 - 6	514.26	235.99	792.52	***
1 - 7	604.07	400.27	807.87	***
3 - 1	-165.94	-368.87	37.00	
3 - 5	278.24	89.21	467.28	***
3 - 4	300.55	99.78	501.31	***
3 - 6	348.32	64.96	631.68	***
3 - 7	438.13	227.42	648.85	***
5 - 1	-444.18	-625.48	-262.88	***
5 - 3	-278.24	-467.28	-89.21	***
5 - 4	22.30	-156.56	201.16	
5 - 6	70.08	-198.22	338.37	
5 - 7	159.89	-30.08	349.85	
4 - 1	-466.48	-659.97	-272.99	***
4 - 3	-300.55	-501.31	-99.78	***
4 - 5	-22.30	-201.16	156.56	
4 - 6	47.77	-228.91	324.46	
4 - 7	137.59	-64.05	339.22	

Variables explicatives multiples

De l'intérêt d'ajouter des variables explicatives

Une analyse de la variance de la variable Y qui n'intègre qu'une seule variable explicative qualitative court le risque d'être faussée par la présence d'**effets de structure**.

Exemple : Le temps partiel, majoritairement féminin, affecte fortement le salaire mensuel. Quand le temps partiel n'est pas intégré dans l'analyse de la variance du salaire selon le sexe, les écarts de salaire entre hommes et femmes sont surestimés.

Il s'agit d'une des manifestations du **biais de variable omise** abordé à propos des méthodes de régression (*cf* session 6).

Les variables explicatives supplémentaires jouent ainsi le rôle de **variables de contrôle** prémunissant contre ce type de biais.

Variables explicatives multiples

Variables croisées et variables quantitatives

Une analyse de variance multivariée peut en particulier inclure des **variables croisées** (ou **effets d'interaction**).

Exemple : Il est possible que l'association entre temps partiel et salaire varie entre homme et femme. Pour tenir compte de cette hypothèse dans le modèle, il faut intégrer la variable sexe $\times$ temps partiel dans l'ANOVA.

Il est également fréquent qu'une ou plusieurs variables quantitatives soient liées à la variable expliquée.

Exemple : L'âge, dans la mesure où il est lié à l'ancienneté et à l'expérience, exerce très probablement une influence sur le salaire.

Variables explicatives multiples

Construction de la statistique de test

Comme précédemment, la statistique de test s'appuie sur la décomposition de la somme des carrés totaux.

Néanmoins, cette décomposition fait intervenir **toutes les variables explicatives** (y compris quantitatives), ce qui la rend particulièrement **complexe à énoncer**.

Une approche alternative consiste à rapprocher ces cas particuliers d'analyse de la (co)variance des **tests d'hypothèses complexes** dans un modèle de régression.

La démarche précise de construction du test en présence de plusieurs variables explicatives sera donc revue à l'issue des sessions consacrées aux méthodes de régression (session 6).

Variables explicatives multiples

Exemple : Salaire simulé dans l'EEC 2012T4

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	117216159.6	13024017.7	53.96	<.0001
Error	793	191386970.9	241345.5		
Corrected Total	802	308603130.5			

R-Square	Coeff Var	Root MSE	salred_simul Mean
0.379828	27.24488	491.2693	1803.162

Source	DF	Type I SS	Mean Square	F Value	Pr > F
DDIPL	5	35430809.92	7086161.98	29.36	<.0001
SEXE	1	10219277.09	10219277.09	42.34	<.0001
age	1	27510716.82	27510716.82	113.99	<.0001
tpp	1	41833064.71	41833064.71	173.33	<.0001
SEXE*tpp	1	2222291.06	2222291.06	9.21	0.0025

Sessions de révision – Session 4

Régression linéaire

Certificat de chargé d'études statistiques

Martin Chevalier (INSEE)

Objectifs de la session

Contextualiser le modèle de régression linéaire et présenter ses principales caractéristiques.

Insister sur l'interprétation des coefficients et le test de leur significativité.

Illustrer la mise en œuvre des méthodes de régression avec un exemple suivi tout du long.

Plan de la session

Objectifs et démarche des méthodes de régression

Modéliser et expliquer : les deux objectifs des méthodes de régression

Les méthodes de régression ont deux objectifs :

1. objectif **prédictif** : être en mesure, à partir des seules variables explicatives, de reconstituer (« prédire ») au mieux la variable expliquée ;
2. objectif **explicatif** : déterminer si le lien entre la variable expliquée et une variable explicative est statistiquement significatif « toutes les autres variables du modèle étant contrôlées par ailleurs ».

Selon les contextes, l'importance de ces objectifs peut varier :

- en prévision, le pouvoir prédictif d'un modèle est crucial ;
- en analyse socio-économique, c'est la dimension explicative du modèle qui prime.

Objectifs et démarche des méthodes de régression

Démarche générale des méthodes de régression

La mise en œuvre des méthodes de régression se décompose en trois grandes étapes :

1. Construction d'un modèle pertinent : analyse uni- et bivariée de la variable expliquée, test de nombreux modèles et comparaison d'**indicateurs de qualité** ;

2. Analyse des relations entre variables : interprétation des coefficients et de leur **significativité** ;

3. (Si possible) Interprétation causale : identification des **sources de biais potentiel**, lien avec d'autres éléments d'analyse socio-économique ou financière.

Objectifs et démarche des méthodes de régression

La question de la causalité

« Corrélation n'est pas causalité » : le lien statistiquement significatif entre deux variables n'indique **pas nécessairement une relation de causalité** ; il conduit même parfois à des **conclusions fallacieuses** (*cf.* session 6, endogénéité).

En règle générale, **les modèles de régression seuls ne suffisent pas à établir de lien de causalité** : il faut étayer l'argumentaire sur d'autres éléments, notamment issus de l'analyse économique, financière, sociologique etc.

Dans certaines configurations néanmoins, le modèle économétrique peut permettre de dégager des relations de cause à effet (panel, cadre expérimental, etc.).

Objectifs et démarche des méthodes de régression

Nature de la variable expliquée et choix de modélisation

Le premier élément qui guide la construction du modèle est la nature de la variable expliquée.

Une modélisation est en effet adaptée à une nature de variable bien spécifique :

- Quantitative continue : régression linéaire ;
- Qualitative dichotomique : régression logistique ou probit dichotomique ;
- Qualitative polytomique : régression logistique ou probit ordonnée ou multinomiale ;
- Quantitative discrète de comptage : régression de Poisson.

Objectifs et démarche des méthodes de régression

Relation entre les différentes modélisations

La régression linéaire correspond au modèle linéaire le plus simple, les **Moindres carrés ordinaires** (MCO ou OLS).

En pratique, sa mise en œuvre ne fait appel qu'à des éléments de calcul matriciel.

Les modélisations logistiques, probit ou de Poisson appartiennent à la **classe plus large des modèles linéaires généralisés**.

La mise en œuvre pratique de ces modèles implique la **maximisation d'une quantité**, la **vraisemblance**, à l'aide d'algorithmes adaptés (Newton-Raphson notamment).

Remarque : Le modèle de régression linéaire est un cas particulier de modèle linéaire généralisé.

Modèle de régression linéaire

Formulation

Le modèle linéaire multiple est un modèle de régression qui s'écrit sous la forme :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_K X_K + u$$

où $\beta_0, \ldots, \beta_K$ sont les paramètres du modèle et u un résidu.

Ce modèle est « linéaire » dans la mesure où **les paramètres n'interagissent avec les variables explicatives que par l'intermédiaire de relations linéaires** (sommes ou multiplications).

Exemple : Le modèle $Y = \beta_0 + X_1^{\beta 1} + u$ n'est pas linéaire.

Quand il n'y a qu'une seule variable explicative X_1, on parle de **régression linéaire simple** ; sinon on parle de **régression linéaire multiple**.

Modèle de régression linéaire

Exemple : Salaire dans l'EEC 2012T4

On cherche à déterminer certains déterminants du salaire à partir des données de l'EEC 2012T4.

Le salaire est une variable continue donc on peut sans difficulté lui appliquer le modèle de régression linéaire.

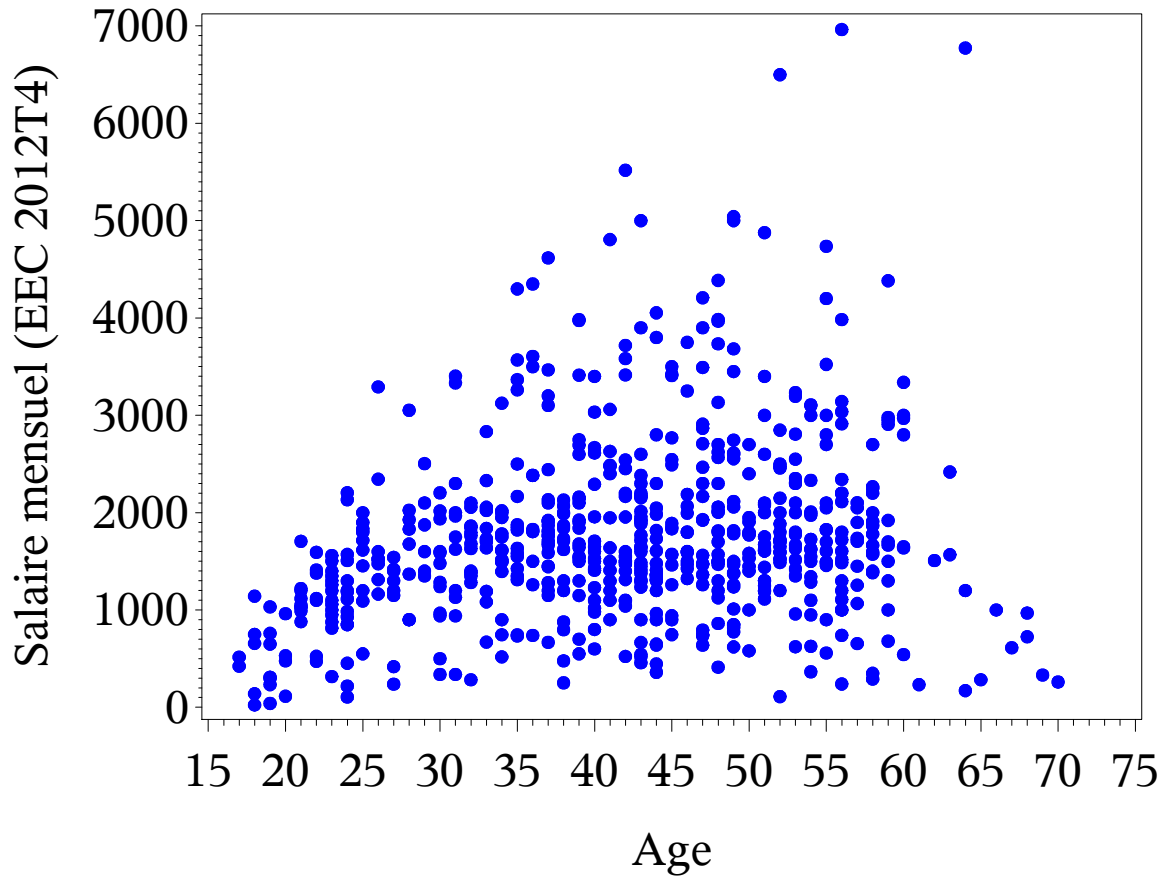
Les variables explicatives potentielles sont l'âge, le sexe, le niveau de diplôme, le fait de travailler ou non à temps partiel.

On peut donc commencer par construire le modèle de régression linéaire simple suivant :

$$salaire = \beta_0 + \beta_1 age + u$$

Modèle de régression linéaire

Exemple : Salaire dans l'EEC 2012T4



Modèle de régression linéaire

Estimation : Le cas de la régression linéaire simple

L'estimation du modèle est menée en cherchant les valeurs de $\beta_0, \ldots, \beta_K$ qui **maximise l'ajustement de la prédiction aux vraies valeurs** de Y.

Mathématiquement, on traduit cet objectif par la **minimisation de la somme des carrés des résidus** (SCR) : $\sum_{i=1}^{n} u_i^2$.

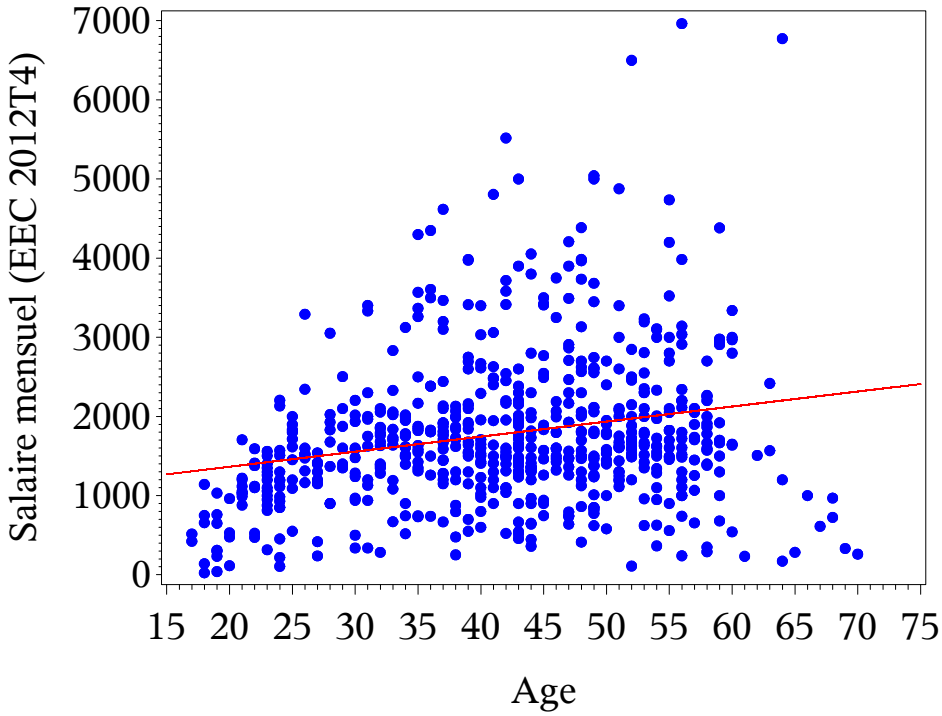
Dans le cas de la régression linéaire simple, les valeurs $\hat{\beta}_0$ et $\hat{\beta}_1$ qui minimisent cette quantité sont facilement calculables :

$$\hat{\beta}_1 = \frac{Cov(X,Y)}{V(X)} = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^{n}(y_i - \bar{Y})^2} \quad \text{et} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Modèle de régression linéaire

Exemple : Salaire dans l'EEC 2012T4

$$\hat{\beta}_1 = \frac{2\ 433}{128} = 19,01 \quad \text{et} \quad \hat{\beta}_0 = 1784 - 19,01 \times 42,05 = 984,62$$



Modèle de régression linéaire

Estimation : Le cas de la régression linéaire multiple

Les estimateurs $\hat{\beta}_0$, $\hat{\beta}_1$,..., $\hat{\beta}_K$ peuvent également être facilement obtenus dans le cas de la régression linéaire multiple.

Pour ce faire, il faut réécrire le modèle sous forme matricielle :

$$Y = \beta X + u$$

avec $\beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_K \end{pmatrix}$, $Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$ et $X = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{K,1} \\ \vdots & \vdots & \ldots & \vdots \\ 1 & x_{1,n} & \cdots & x_{K,n} \end{pmatrix}$

On peut alors montrer que la somme des carrés des résidus est minimisée par :

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

Modèle de régression linéaire

Principales hypothèses de la régression linéaire multiple (RLM)

RLM1 Linéarité dans les paramètres

RLM2 Echantillonnage aléatoire

RLM3 Absence de colinéarité parfaite
$\rightarrow$ Aucune des variables explicatives n'est constante, il n'existe pas de relation linéaire entre variables explicatives.

RLM4 Espérance conditionnelle des erreurs nulle

$$E(u|X_1, \ldots, X_K) = 0$$

$\rightarrow$ Sinon : endogénéité (*cf.* session 6)

RLM5 Homoscédasticité

$$V(u|X_1, \ldots, X_K) = \sigma^2 = \text{constante}$$

$\rightarrow$ Sinon : hétéroscédasticité (*cf.* session 6)

RLM6 Normalité de la distribution des erreurs.

Modèle de régression linéaire

Principales propriétés de la régression linéaire multiple

Sous RLM1-RLM4, l'estimateur de la régression linéaire multiple $\hat{\beta}$ est **sans biais** :

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

Sous RLM1-RLM5, l'estimateur de la régression linéaire multiple $\hat{\beta}$ est le **meilleur estimateur linéaire sans biais** (*Best linear unbiased estimator*, BLUE).

Autrement dit, quand ces hypothèses sont vérifiées on ne peut pas trouver d'estimateur linéaire sans biais qui soit plus précis que celui obtenu par la régression linéaire multiple.

Indicateurs de qualité du modèle

Pour juger de la qualité de l'ajustement des modèles de régression linéaire, deux types de statistiques sont disponibles :

- **le R^2 et le R^2 ajusté** : ils rendent compte de la part de la variance de Y expliquée par le modèle et mesurent son caractère prédictif ;
- **le test de significativité globale** : il teste la nullité simultanée de tous les coefficients (sauf la constante) et rend compte du caractère explicatif du modèle.

Dans tous les cas ces indicateurs sont construits à partir de la **somme des carrés des résidus** : plus elle est faible devant la variance de Y, meilleur est le modèle.

Indicateurs de qualité du modèle

Le R^2 et le R^2 ajusté

Le R^2 correspond au ratio de la variance de Y expliquée par le modèle (somme des carrés expliqués ou SCE) sur la variance totale de Y (somme des carrés totaux ou SCT) :

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

avec SCR la somme des carrés des résidus (ce que cherche à minimiser l'estimation du modèle).

Comme $SCT = SCE + SCR$, le R^2 est compris entre 0 et 1. Plus il est proche de 1, plus le modèle explique bien la variabilité de Y.

Par construction, le R^2 augmente dès qu'on ajoute une variable (même sans aucun lien avec Y). Pour corriger ce phénomène on définit le R^2 ajusté.

Remarque : Quand la régression n'intègre qu'une seule variable qualitative, le R^2 coïncide avec le rapport de corrélation $\eta_{Y|X}$.

Indicateurs de qualité du modèle

Le test de significativité globale

Le test de significativité globale est formulé de la façon suivante :

$$H_0 : \beta_1 = \ldots = \beta_K = 0 \quad \text{contre} \quad H_1 : \exists k,\ \beta_k \neq 0$$

On peut montrer que sous H_0 :

$$F = \frac{SCE/K}{SCR/(n-(K+1))} \hookrightarrow F_{K,n-(K+1)}$$

On peut donc facilement tester H_0 à un niveau de confiance usuel (90 %, 95 %, 99 %).

Remarque : Cette statistique de test est identique à celle utilisée en ANOVA, aux degrés de liberté près. La constante est implicitement comptée en ANOVA, mais pas ici ($K_{ANOVA} = K_{REG} + 1$).

Indicateurs de qualité du modèle

Exemple : Salaire dans l'EEC 2012T4

On estime le modèle comportant toutes les variables explicatives : âge, sexe, diplôme et temps partiel.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	8	267880662	33485083	53.06	<.0001
Error	633	399499191	631120		
Corrected Total	641	667379853			

Root MSE	794.43084	**R-Square**	0.4014
Dependent Mean	1790.93146	**Adj R-Sq**	0.3938
Coeff Var	44.35853		

Interprétation des coefficients et significativité

Interpréter les coefficients (1)

Au-delà de la modélisation, le plus souvent l'objectif d'une régression est d'interpréter la valeur des coefficients estimés.

L'interprétation du coefficient dépend de la nature de la variable explicative.

Constante Le coefficient $\hat{\beta}_0$ associé à la constante correspond à la valeur moyenne de la variable expliquée dans l'échantillon si l'ensemble des variables explicatives valaient 0.

Variable explicative quantitative Le coefficient $\hat{\beta}_k$ associé à la variable quantitative X_k signifie qu'en moyenne dans l'échantillon, une augmentation de 1 de la valeur de X_k est associée à une augmentation $\hat{\beta}_k$ de la valeur de Y.

Exemple : Dans l'exemple de régression linéaire simple, $\hat{\beta}_1 = 19$ € : à une année supplémentaire est associé en moyenne un salaire mensuel de 19 € supérieur.

Interprétation des coefficients et significativité

Intégrer des variables explicatives qualitatives

Dans la formulation mathématique du modèle, les variables explicatives $X_1, \ldots, X_K$ sont des variables numériques.

Pour intégrer des variables qualitatives, il convient de les dichotomiser.

Exemple : La variable `sexe` dont les valeurs sont `"1","2"` peut être dichotomisée en `sexe1` (1 pour les hommes, 0 pour les femmes) et `sexe2` (0 pour les hommes, 1 pour les femmes).

Par définition, la somme des variables indicatrices d'une variable explicative qualitative est 1 pour chaque observation.

Pour éviter une situation de colinéarité parfaite, une des variables indicatrices doit donc être exclue du modèle : c'est la **modalité de référence**.

Interprétation des coefficients et significativité

Interpréter les coefficients (2)

Variable explicative qualitative Le coefficient $\hat{\beta}_{k,j}$ de la modalité j d'une variable explicative qualitative x_k s'interprète relativement à la modalité de référence.

Dans ce contexte, $\hat{\beta}_{k,j}$ s'interprète de la façon suivante : en moyenne dans l'échantillon, **le fait d'avoir la modalité j plutôt que la modalité de référence** est associé à une augmentation $\hat{\beta}_{k,j}$ de la valeur de Y.

Exemple : On estime le modèle de régression linéaire simple :

$$salaire = \beta_0 + \beta_1 sexe2 + u$$

On obtient $\hat{\beta}_1 = -374$. Cela signifie qu'en moyenne dans l'échantillon, être une femme (plutôt qu'un homme) est associé à un salaire mensuel inférieur de 374 €.

Interprétation des coefficients et significativité

Régression linéaire multiple et théorème de Frisch-Vaugh

Dès lors que la régression linéaire intègre plus d'une variable explicative, l'interprétation des coefficients peut être effectuée « tous les autres paramètres du modèle étant maintenus constants par ailleurs ».

Autrement dit, chaque coefficient capte l'effet propre de sa variable explicative : les **effets de structure** associés aux autres variables du modèle sont pris en compte.

Ceci est la conséquence du **théorème de Frisch-Vaugh** : chaque coefficient de la régression linéaire multiple peut être obtenu par régression linéaire simple sur un résidu « purgé » de l'effet des autres variables du modèle.

Interprétation des coefficients et significativité

Exemple : Salaire dans l'EEC 2012T4

Root MSE	794.43084	**R-Square**	0.4014
Dependent Mean	1790.93146	**Adj R-Sq**	0.3938
Coeff Var	44.35853		

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	942.47113	134.96444	6.98	<.0001
age	Age	1	26.30629	2.93641	8.96	<.0001
sexe2	Femme	1	-259.45871	66.26836	-3.92	0.0001
ddipl1	Supérieur à BAC + 2	1	725.21454	102.60060	7.07	<.0001
ddipl3	BAC + 2	1	463.69871	108.14199	4.29	<.0001
ddipl5	CAP, BEP	1	-217.11880	96.31689	-2.25	0.0245
ddipl6	Brevet	1	-207.12477	150.42667	-1.38	0.1690
ddipl7	Aucun ou CEP	1	-544.13129	110.53885	-4.92	<.0001
tpp	Temps partiel	1	-914.69714	84.29250	-10.85	<.0001

Interprétation des coefficients et significativité

Test de significativité et t de Student

Une fois le modèle de régression posé, déterminer si l'association entre la variable Y et la variable X_k est statistiquement significative revient à tester la nullité de β_k.

On formule donc le test de significativité de β_k :

$$H_0 : \beta_k = 0 \quad \text{contre} \quad H_1 : \beta_k \neq 0$$

Sous les hypothèses RLM1-RLM6, on peut montrer que sous H_0 :

$$t_k = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)} \hookrightarrow \mathcal{T}_{n-(K+1)}$$

avec

$$se(\hat{\beta}_k) = \frac{SCR/(n-(K+1))}{(1-R^2_{-k})\sum_{i=1}^{n}(x_{k,i}-\bar{x}_k)}$$

l'**erreur standard** (*standard error*) de $\hat{\beta}_k$ et R^2_{-k} le R^2 de la régression sur toutes les variables sauf X_k.

Interprétation des coefficients et significativité

Interprétation des tests de significativité

Pour un niveau de confiance $1 - \alpha$ donné, on peut alors comparer la valeur de t_k au quantile à $1 - \alpha/2$ % de $\mathcal{T}_{n-(K+1)}$ et ainsi accepter ou rejeter H_0 au seuil α.

Comme toujours, une autre possibilité consiste à directement interpréter la p-valeur du test : si celle-ci est inférieure à α, alors on peut rejeter H_0 au seuil $1 - \alpha$.

Remarque : Il est également possible de construire des tests unilatéraux du type :

$$H_0 : \beta_k = 0 \quad \text{contre} \quad H_1 : \beta_k > 0$$

La statistique de test est alors inchangée. En revanche, dans ce contexte c'est le quantile à $1 - \alpha$ % de $\mathcal{T}_{n-(K+1)}$ qui donne la valeur critique du test (et non à $1 - \alpha/2$ %).

Sessions de révision – Session 5

Régression logistique

Certificat de chargé d'études statistiques

Martin Chevalier (INSEE)

Objectifs de la session

Justifier l'utilisation de modèles de régression logistique pour modéliser des données qualitatives.

Rappeler le cadre des modèles linéaires généralisés et le principe de l'estimation par maximum de vraisemblance.

Présenter les particularités des modèles de régression logistique : indicateurs de qualité, interprétation des résultats (*odds-ratio*).

Plan de la session

Données qualitatives et régression logistique

Inadaptation de la régression linéaire aux données qualitatives

Les modèles linéaires classiques sont adaptés à une variable expliquée de nature quantitative : ils reposent en effet sur la décomposition de la variance de la variable expliquée Y.

Mais quand la variable Y est qualitative, cette « variance » n'a pas beaucoup de sens.

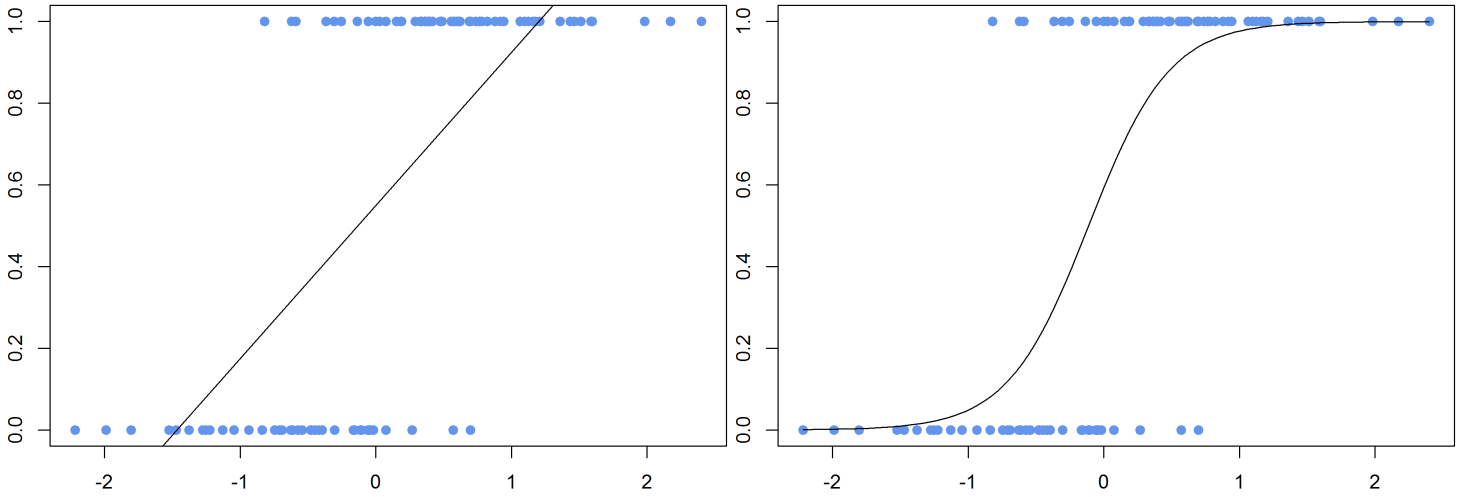
Quand Y est dichotomique, il est néanmoins possible de l'envisager comme une variable quantitative à deux modalités, 0 ou 1.

Même dans ce cas, il est souvent préférable d'adopter une modélisation adaptée à la nature de Y.

Données qualitatives et régression logistique

Exemple : Données simulées

Dans ce contexte, la modélisation linéaire (gauche) semble moins adaptée que la modélisation logistique (droite).



Données qualitatives et régression logistique

Facilité d'interprétation de la régression logistique

En pratique néanmoins, il n'est pas rare que les résultats qualitatifs obtenus par les deux méthodes soient proches : même signe et même significativité pour les coefficients.

L'intérêt de la régression logistique réside néanmoins également dans l'**interprétation de ses coefficients**.

Une transformation simple permet en effet de les exprimer en termes de « rapport de chances » (*odds-ratio*), formulation qui simplifie leur interprétation.

Exemple : En 2002, un enfant d'enseignants avait 14 fois plus de chances d'être bachelier qu'un enfant d'ouvriers non-qualifiés (Observatoire des inégalités, `http://www.inegalites.fr/spip.php?page=article&id_article=272`)

Le modèle linéaire généralisé et son estimation

Formulation

Le modèle utilisé en régression logistique appartient à la classes des modèles linéaires généralisés.

Ce sont des modèles qui s'écrivent sous la forme :

$$Y = f(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \beta_K X_K + u)$$

où f est une fonction prédéfinie. De façon plus synthétique (en notation matricielle) :

$$y = f(\beta X + u)$$

Remarque Ces modèles restent linéaires en leurs paramètres : $y = \beta_0 + X_1^{\beta_1} + u$ n'est pas un modèle linéaire généralisé.

Le modèle linéaire généralisé et son estimation

Stratégie d'estimation

Comme dans la régression linéaire classique, l'objectif est de trouver les paramètres $\hat{\beta}_0, \dots, \hat{\beta}_K$ qui maximisent l'ajustement du modèle aux données.

Néanmoins, contrairement au modèle linéaire classique, **il n'existe aucune formule** qui donne directement la valeur de $\hat{\beta}$.

On utilise donc des **algorithmes d'optimisation** pour maximiser une certaine **fonction objectif**, la **vraisemblance** du modèle (*likelihood* en anglais).

Le modèle linéaire généralisé et son estimation

Estimation (1) : Vraisemblance du modèle

De manière générale, la vraisemblance d'un modèle portant sur n observations de la variable expliquée Y étant donné le jeu de variables explicatives X s'écrit :

$$L_n(y_1, \ldots, y_n, X_1, \ldots, X_n) = \mathbb{P}_{y_1,\ldots,y_n}[y_1, \ldots, y_n | X_1, \ldots, X_n]$$

Cette quantité s'interprète comme la probabilité d'observer les valeurs $(y_1, \ldots, y_n)$ étant données les valeurs $(X_1, \ldots, X_n)$.

L'estimation d'un modèle de régression linéaire généralisé consiste à **trouver le vecteur $\hat{\beta}$ qui maximise cette probabilité**.

Le modèle linéaire généralisé et son estimation

Estimation (2) : Contribution des observations à la vraisemblance

Sous l'hypothèse que les observations sont indépendantes et identiquement distribuées, on peut réécrire L_n :

$$L_n = \mathbb{P}[y_1|X_1] \times \ldots \times \mathbb{P}[y_n|X_n] = \prod_{i=1}^{n} \mathbb{P}[y_i|X_i]$$

Chaque facteur de ce produit s'interprète comme une **contribution à la vraisemblance totale** de l'échantillon :

- si y_i est peu vraisemblable étant données les variables X_i, le terme $\mathbb{P}[y_i|X_i]$ est faible ;
- si y_i est vraisemblable étant données les variables X_i, le terme $\mathbb{P}[y_i|X_i]$ est élevé.

Le modèle linéaire généralisé et son estimation

Estimation (3) : La famille binomiale

Comme Y est dichotomique, on peut réécrire sa probabilité :

$$\mathbb{P}[y_i|X_i] = \mathbb{P}[y_i = 1|X_i]^{y_i} \times \mathbb{P}[y_i = 0|X_i]^{1-y_i}$$

- Si $y_i = 1$, on cherche $\hat{\beta}$ qui maximise la probabilité d'avoir $y_i = 1$ sachant X_i.
- Si $y_i = 0$, on cherche $\hat{\beta}$ qui maximise la probabilité d'avoir $y_i = 0$ sachant X_i.

On dit dans ce cas que le modèle linéaire généralisé appartient à la **famille binomiale** (car Y n'a que deux modalités).

Comme par ailleurs $\mathbb{P}[y_i = 0|X_i] = 1 - \mathbb{P}[y_i = 1|X_i]$, on peut réécrire L_n en fonction de $\mathbb{P}[y_i = 1|X_i]$ uniquement :

$$L_n = \prod_{i=1}^{n} \mathbb{P}[y_i = 1|X_i]^{y_i} \times (1 - \mathbb{P}[y_i = 1|X_i])^{1-y_i}$$

Le modèle linéaire généralisé et son estimation

Estimation (4) : La fonction de lien logistique

Pour estimer le modèle en cherchant $\hat{\beta}$ qui maximise L_n, il ne reste donc plus qu'à pouvoir calculer $\mathbb{P}[y_i = 1|X_i]$.

C'est là le rôle de la **fonction de lien** f. Dans le cas de la famille binomiale, $f(X_i\beta) = \mathbb{P}[y_i = 1|X_i]$

On parle de **régression logistique** quand la fonction de lien est la **fonction logit** :

$$f(x) = logit(x) = \frac{e^x}{1 + e^x}$$

Remarque : D'autres fonctions de lien sont utilisables dans le cas binomial, notamment la fonction de répartition de d'une loi normale centrée réduite (modèle probit, *cf. session 6*).

Le modèle linéaire généralisé et son estimation

Exemple : Données simulées

À partir de données simulées (variables Y et X_1), on construit le modèle de régression logistique simple suivant :

$$\mathbb{P}[y_i = 1|x_{1,i}] = logit(\beta_0 + \beta_1 x_{1,i})$$

Dans le courant de l'estimation du modèle, on cherche à comparer la vraisemblance obtenue pour deux jeux de paramètres (β_0^A, β_1^A) et (β_0^B, β_1^B) différents.

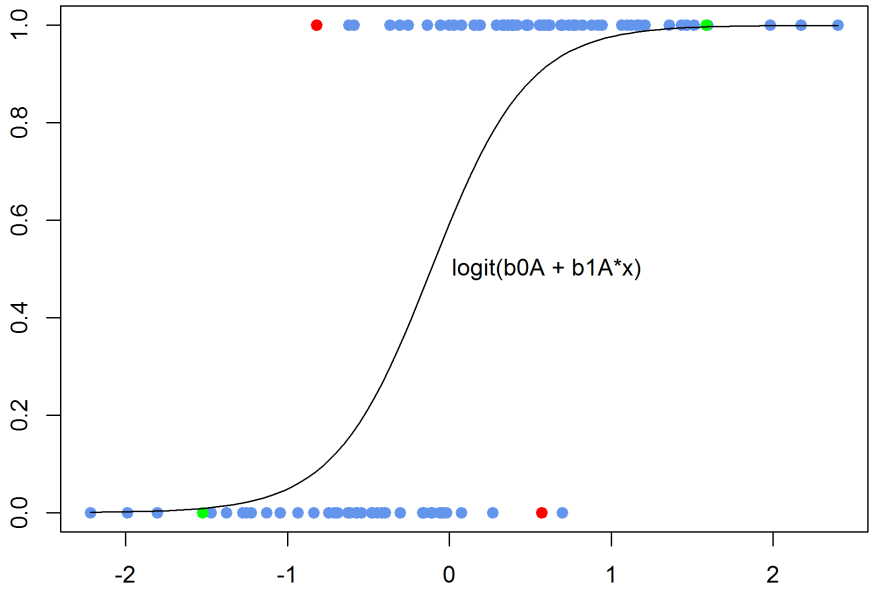
Sur les deux graphiques qui suivent sont représentées les fonctions :

$$y = logit(\beta_0^A + \beta_1^A x) \quad \text{et} \quad y = logit(\beta_0^B + \beta_1^B x)$$

Le modèle linéaire généralisé et son estimation

Exemple : Données simulées

$$\mathbb{P}[y_i|x_{1,i}] = logit(\beta_0^A + \beta_1^A x_{1,i})^{y_i} \times \left[1 - logit(\beta_0^A + \beta_1^A x_{1,i})\right]^{1-y_i}$$

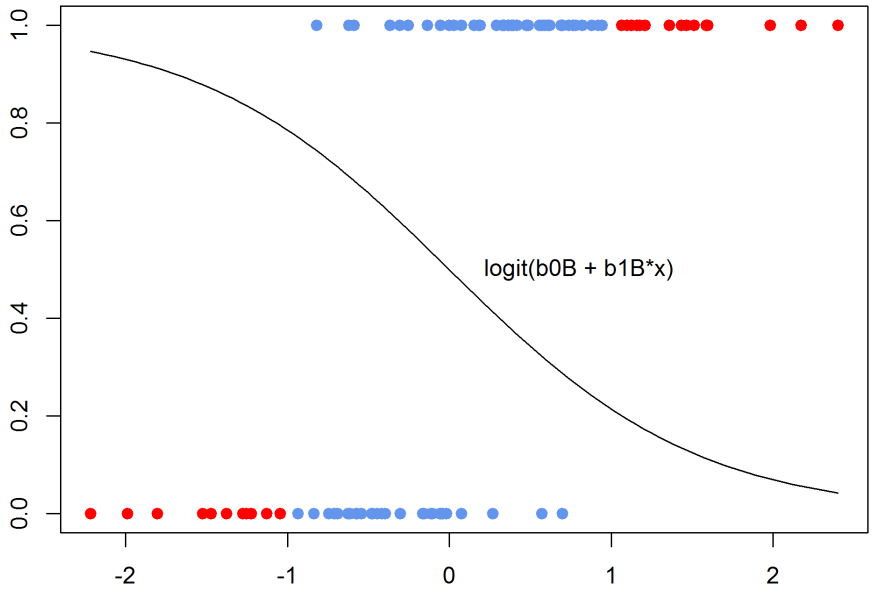


La position de la courbe semble indiquer que la valeur de $\mathbb{P}[y_i|x_{1,i}]$ est souvent élevée : **cette modélisation est vraisemblable**.

Le modèle linéaire généralisé et son estimation

Exemple : Données simulées

$$\mathbb{P}\left[y_i|x_{1,i}\right] = logit(\beta_0^B + \beta_1^B x_{1,i})^{y_i} \times \left[1 - logit(\beta_0^B + \beta_1^B x_{1,i})\right]^{1-y_i}$$



La position de la courbe semble indiquer que la valeur de $\mathbb{P}\left[y_i|x_{1,i}\right]$ est souvent faible : **cette modélisation est peu vraisemblable**.

Le modèle linéaire généralisé et son estimation

Bilan sur l'estimation d'un modèle logistique

Pour synthétiser, un modèle logistique est estimé en cherchant $\hat{\beta}$ qui maximise la quantité :

$$L_n = \prod_{i=1}^{n} \left[f(X_i\beta)^{y_i} \times [1 - f(X_i\beta)]^{1-y_i}\right]$$

En pratique, c'est le **logarithme de cette quantité** (la **log-vraisemblance**, *log-likelihood*) qui est maximisé :

$$\ell_n = \sum_{i=1}^{n} [y_i ln(f(X_i\beta)) + (1 - y_i) ln(1 - f(X_i\beta))]$$

Les algorithmes utilisés sont en généralisé l'**algorithme du score de Fisher ou l'algorithme de Newton-Raphson**.

Le modèle linéaire généralisé et son estimation

Remarque : Régression linéaire et modèle linéaire général

La régression linéaire classique peut être vue comme un cas particulier de modèle linéaire général appartenant à la **famille gaussienne** et avec la **fonction identité comme fonction de lien**.

Model Information		
Data Set	WORK.E	
Distribution	Normal	
Link Function	Identity	
Dependent Variable	SALRED	Salaire mensuel (EEC 2012T4)

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept	1	942.4711	134.0151	679.8064	1205.136	49.46	<.0001
age	1	26.3063	2.9158	20.5915	32.0211	81.40	<.0001
sexe2	1	-259.459	65.8022	-388.429	-130.489	15.55	<.0001
ddipl1	1	725.2145	101.8789	525.5356	924.8935	50.67	<.0001
ddipl3	1	463.6987	107.3813	253.2352	674.1622	18.65	<.0001
ddipl5	1	-217.119	95.6394	-404.569	-29.6690	5.15	0.0232
ddipl6	1	207.125	149.3686	-409.882	85.6322	1.92	0.1655

Indicateurs de qualité du modèle

Comme dans les méthodes de régression linéaire classiques, il est indispensable de disposer d'outils pour juger de l'adéquation du modèle aux données.

La plupart des statistiques d'ajustement sont construites à partir de la log-vraisemblance : AIC et SC.

Le test de significativité globale par le ratio de vraisemblance (*likelihood ratio test* en anglais) est utilisé pour juger du caractère explicatif d'un modèle.

Le pourcentage de concordance ou la courbe ROC peuvent être utilisés pour évaluer le caractère prédictif d'un modèle.

Indicateurs de qualité du modèle

Statistiques construites à partir de la vraisemblance

Pour comparer deux modèles portant sur la même variable expliquée, on peut comparer les valeurs de leur vraisemblance : **on privilégie le modèle présentant la plus grande vraisemblance**.

Cependant, quand un modèle comporte davantage de variables explicatives, son pouvoir prédictif augmente mécaniquement (comme pour le R^2).

On peut alors utiliser des indicateurs qui pénalisent la vraisemblance par le nombre de variables :

- *Akaike information criterion* : $AIC = -2\ell_n + 2(K+1)$
- Critère de Schwartz (aussi appelé *Bayesian information criterion*) : $SC = -2\ell_n + ln(n)(K+1)$

Indicateurs de qualité du modèle

Test de significativité globale par le ratio de vraisemblance

Pour évaluer le pouvoir explicatif d'un modèle, on peut comparer sa vraisemblance à celle du modèle ne comportant que la constante.

Il est possible de formaliser cette comparaison dans le cadre du test du **ratio de vraisemblance**.

On peut en effet montrer que sous l'hypothèse H_0 d'égalité des deux vraisemblances,

$$LR = -2ln\left(\frac{L^0}{L_n}\right) = 2(\ell_n - \ell^0) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_K$$

avec ℓ^0 la log-vraisemblance du modèle ne comportant que la constante.

Indicateurs de qualité du modèle

Exemple : Chômage parmi les actifs dans l'EEC 2012T4

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	717.763	649.131
SC	722.615	687.949
-2 Log L	715.763	633.131

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	82.6318	7	<.0001
Score	80.9885	7	<.0001
Wald	71.6570	7	<.0001

Indicateurs de qualité du modèle

Pourcentage de concordance

Le modèle de régression permet d'obtenir, pour chaque individu de l'échantillon, une probabilité prédite p_i sur la base des variables explicatives.

On peut alors classer chaque paire d'observations selon trois catégories :

- concordante : $y_1 = 0,\ y_2 = 1$ et $p_1 < p_2$ ou $y_1 = 1,\ y_2 = 0$ et $p_1 > p_2$
- discordante : $y_1 = 0,\ y_2 = 1$ et $p_1 > p_2$ ou $y_1 = 1,\ y_2 = 0$ et $p_1 < p_2$
- *ex-aequo* : $y_1 = y_2$ ou $p_1 = p_2$.

On peut alors calculer un **pourcentage de paires concordantes** rapporté au nombre de paires non *ex-aequo*.

Indicateurs de qualité du modèle

Performance de la classification et courbe ROC

De façon plus générale, on peut évaluer la performance d'un modèle dichotomique à partir de la courbe ROC (*Receiver operating characteristics*).

Bien souvent, l'objectif d'un modèle est d'aboutir à une classification binaire : le radar détecte-t-il un avion ennemi ? le message est-il un *spam* ?

En sortie du modèle, on a pour chaque individu une probabilité prédite. Toute la question est de savoir où classer l'individu à partir de sa probabilité p_i.

Où placer la probabilité seuil p^* entre les cas à classer comme positifs ($p_i > p^*$) et les cas à classer comme négatifs ($p_i < p^*$) ?

Indicateurs de qualité du modèle

Performance de la classification et courbe ROC

Le meilleur modèle est celui qui n'a **aucun faux négatif et aucun faux positif**. Mais on affaire à un arbitrage :

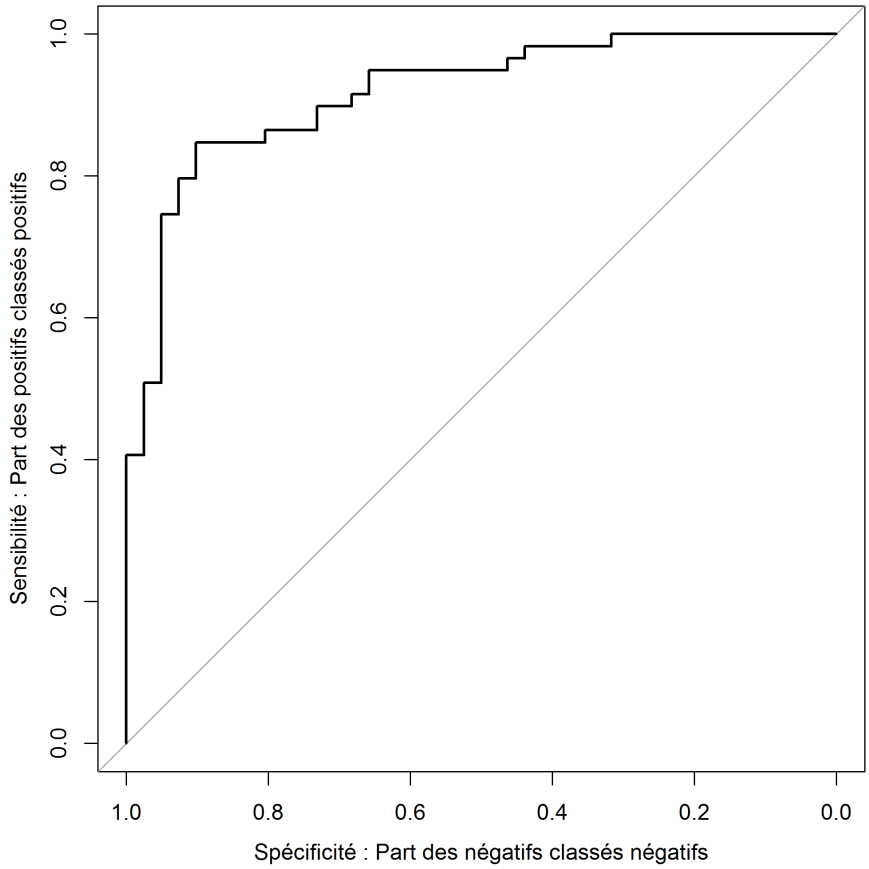
1. Si le seuil est trop haut, certains individus positifs risquent d'être classés comme négatifs (faux négatifs).
2. Si le seuil est trop bas, certains individus négatifs risquent d'être classés comme positifs (faux positifs).

Afin de limiter le risque de faux négatifs on est amené à tolérer un certain nombre de faux positifs, ou inversement.

La courbe ROC permet de représenter cet arbitrage.

Indicateurs de qualité du modèle

Exemple : Données simulées



Indicateurs de qualité du modèle

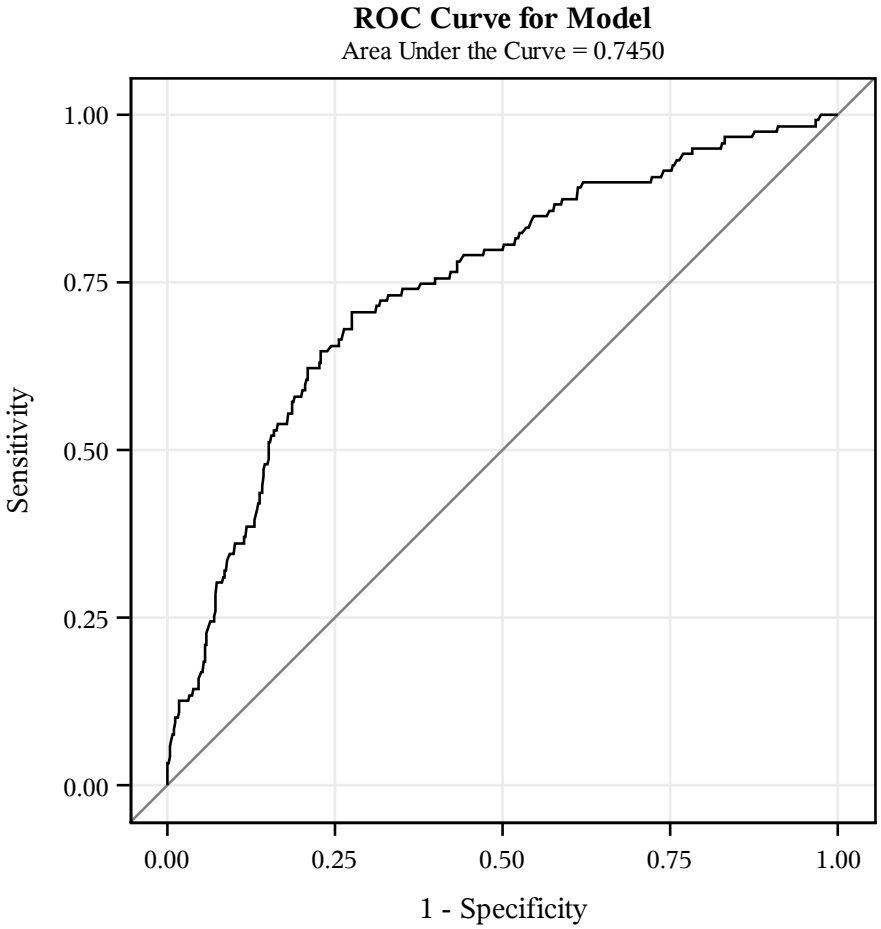
Performance de la classification et courbe ROC

Construction de la courbe ROC

1. Estimer le modèle et classer les observations par probabilités prédites croissantes ;
2. Pour chaque observation :
   - calculer la part des positifs classés positifs (**sensibilité**) si l'observation en question est le seuil entre positif et négatif ;
   - calculer la part des négatifs classés négatifs (**spécificité**) si l'observation en question est le seuil entre positif et négatif ;
3. La courbe ROC d'un test est obtenue en **représentant sa sensibilité en fonction de sa spécificité**.

Indicateurs de qualité du modèle

Exemple : Chômage parmi les actifs dans l'EEC 2012T4



Indicateurs de qualité du modèle

Exemple : Chômage parmi les actifs dans l'EEC 2012T4

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	74.4	**Somers' D**	0.490
Percent Discordant	25.4	**Gamma**	0.491
Percent Tied	0.2	**Tau-a**	0.108
Pairs	98413	**c**	0.745

Interprétation des coefficients et significativité

Lecture des coefficients

La **valeur** des coefficients en elle-même n'a **pas d'interprétation claire** : au sein d'un même modèle, on peut en revanche comparer le **signe** et l'**amplitude** des coefficients.

Les coefficients des variables explicatives s'interprètent différemment selon leur nature :

- pour les variables explicatives quantitatives, on peut interpréter directement le coefficient ;
- pour les variables explicatives qualitatives, l'interprétation des différentes indicatrices est relative à une modalité de référence.

Attention : SAS choisit la modalité de la variable expliquée à modéliser et dichotomise les variables explicatives de façon peu intuitive.

Interprétation des coefficients et significativité

Exemple : Chômage parmi les actifs dans l'EEC 2012T4

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	0.4493	0.3705	1.4706	0.2252
age	1	-0.0724	0.00940	59.3178	<.0001
sexe2	1	0.2719	0.2091	1.6911	0.1935
ddipl1	1	-0.4014	0.3811	1.1091	0.2923
ddipl3	1	-0.3383	0.3738	0.8190	0.3655
ddipl5	1	0.2527	0.3119	0.6563	0.4179
ddipl6	1	0.9879	0.3990	6.1307	0.0133
ddipl7	1	0.8976	0.3291	7.4370	0.0064

Interprétation des coefficients et significativité

Les *odds-ratio* et leur interprétation (1)

Mathématiquement, les *odds-ratio* d'un modèle de régression logistique correspondent à l'exponentielle de la valeur des coefficients :

$$OR_k = e^{\beta_k} = exp(\beta_k)$$

Les *odds-ratio* constituent ainsi une **présentation des résultats alternative aux coefficients**, qui est **plus aisément interprétable**, notamment quand la variable explicative est qualitative.

Remarque : Cette propriété est liée à l'utilisation de la fonction de lien logit. Elle n'est pas valable dans les modèles probit (*cf.* session 6)

Interprétation des coefficients et significativité

Les *odds-ratio* et leur interprétation (2)

Définition L'*odds-ratio* associé à une modalité d'une variable qualitative correspond au rapport de chances pour les individus présentant cette modalité d'être dans la situation modélisée, par rapport aux individus présentant la modalité de référence.

Exemple Soit le modèle :

$$chomage = \beta_0 + \beta_1 jeune + u$$

Si $OR_1 = e^{\beta_1} = 2$, cela signifie qu'**une personne jeune a deux fois plus de chances d'être au chômage qu'une personne plus âgée**.

Interprétation des coefficients et significativité

Exemple : Chômage parmi les actifs dans l'EEC 2012T4

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age	0.930	0.913	0.947
sexe2	1.312	0.871	1.977
ddipl1	0.669	0.317	1.413
ddipl3	0.713	0.343	1.483
ddipl5	1.288	0.699	2.373
ddipl6	2.686	1.229	5.870
ddipl7	2.454	1.287	4.677

Interprétation des coefficients et significativité

Significativité des coefficients : test de Wald

Comme dans les modèles linéaires classiques, le test de significativité du coefficient β_k associé à la variable X_k est formulé :

$$H_0 : \beta_k = 0 \quad \text{contre} \quad H_1 : \beta_k \neq 0$$

On peut alors montrer que sous H_0 :

$$z = \left(\frac{\hat{\beta}_k}{ase(\hat{\beta}_k)} \right)^2 \xrightarrow[n \to +\infty]{\mathcal{L}} \chi^2_1$$

avec $ase(\hat{\beta}_k)$ l'erreur-standard asymptotique de $\hat{\beta}_k$ (produit par l'estimation par maximum de vraisemblance).

Pour un niveau de confiance $1 - \alpha$ donné, la valeur critique du test est donc le quantile à $1 - \alpha$ % d'un χ^2_1. Si $z > q^{\chi^2_1}_{1-\alpha}$ alors on peut rejeter H_0 au seuil α.

Sessions de révision – Session 6

Compléments

Certificat de chargé d'études statistiques

Martin Chevalier (INSEE)

Objectifs de la session

Approfondir certains aspects abordés lors des sessions consacrées aux méthodes de régression.

Aborder des problématiques communes aux différentes méthodes de régression et à l'analyse de variance.

Revenir sur la mise en œuvre pratique des méthodes de régression.

Plan de la session

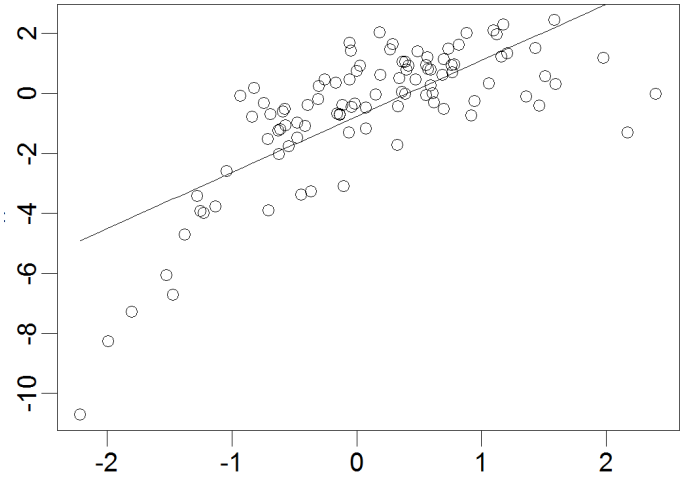
Choix des variables explicatives

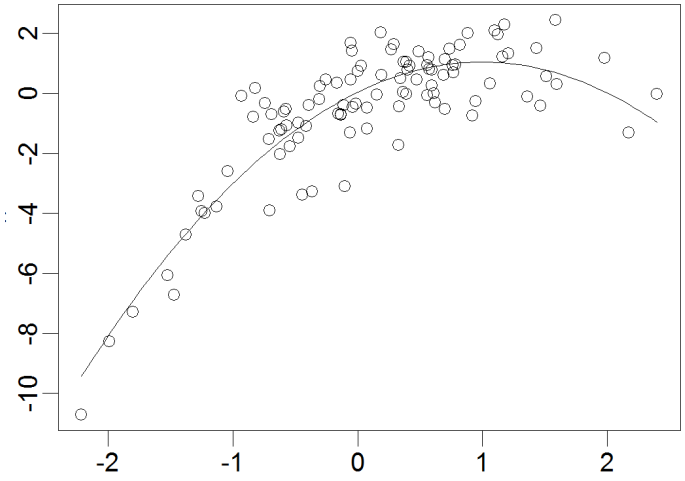
Plan de la partie

Choix des variables explicatives

Intégrer une variable quantitative et son carré (1)

Quelle que soit la modélisation (même en régression linéaire), il est possible et souvent pertinent d'intégrer certaines variables quantitatives accompagnées de leur carré.





Exemple : Le modèle

$$salaire = \beta_0 + \beta_1 age + \beta_2 age^2 + u$$

reste un modèle linéaire.

Choix des variables explicatives

Intégrer une variable quantitative et son carré (2)

Quand une variable est introduite avec son carré, il faut **interpréter conjointement les deux coefficients de cette variable**. Si

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + u$$

alors

$$\frac{\delta y}{\delta X_1} = \beta_1 + 2\beta_2 X_1$$

Cela signifie que le lien entre la variable X_1 et la variable y varie en fonction de la valeur de X_1 elle-même.

Exemple Le gain moyen de salaire associé à une année d'expérience supplémentaire peut être plus important en début qu'en fin de carrière.

Choix des variables explicatives

Choisir la modalité de référence d'une variable qualitative (1)

Quand une variable explicative qualitative comporte plus de deux modalités, le choix de la modalité de référence n'est pas neutre.

Il influence en effet la valeur des coefficients estimés mais aussi la significativité des tests.

De manière générale, quand les variables sont plus ou moins ordonnées (ex : niveau de diplôme) on choisit comme modalité de référence une valeur plutôt centrale et regroupant suffisamment d'observations.

Les hypothèses que l'on souhaite tester peuvent également conduire à privilégier une modalité de référence particulière.

Choix des variables explicatives

Choisir la modalité de référence d'une variable qualitative (2)

Quand on modifie la modalité de référence pour la variable diplôme :

- les coefficients des autres variables ne sont pas affectés ;
- les coefficients et les tests sur le diplôme sont affectés.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1182.18019	378.69623	-3.12	0.0019
age	Age	1	132.92657	19.30754	6.88	<.0001
age2	Age au carré	1	-1.27980	0.23422	-5.46	<.0001
sexe2	Femme	1	-486.44700	67.27258	-7.23	<.0001
ddipl1	Supérieur à BAC + 2	1	744.90024	109.56585	6.80	<.0001
ddipl3	BAC + 2	1	472.93725	115.53969	4.09	<.0001
ddipl5	CAP, BEP	1	-213.06679	102.44886	-2.08	0.0379
ddipl6	Brevet	1	-132.95576	160.31922	-0.83	0.4072
ddipl7	Aucun ou CEP	1	-593.15038	117.36148	-5.05	<.0001

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1775.33057	386.45648	-4.59	<.0001
age	Age	1	132.92657	19.30754	6.88	<.0001
age2	Age au carré	1	-1.27980	0.23422	-5.46	<.0001
sexe2	Femme	1	-486.44700	67.27258	-7.23	<.0001
ddipl1	Supérieur à BAC + 2	1	1338.05063	117.57476	11.38	<.0001
ddipl3	BAC + 2	1	1066.08763	124.34786	8.57	<.0001
ddipl4	BAC	1	593.15038	117.36148	5.05	<.0001
ddipl5	CAP, BEP	1	380.08359	107.57682	3.53	0.0004
ddipl6	Brevet	1	460.19462	163.40500	2.82	0.0050

Choix des variables explicatives

Intégrer des variables croisées (1)

Il est fréquent que l'on souhaite mesurer le lien entre Y et l'**interaction entre plusieurs variables explicatives**.

Exemple : Dans un modèle de salaire, on peut souhaiter tester le lien avec le sexe et l'âge, mais aussi un lien spécifique avec l'âge différencié selon le sexe.

Pour ce faire, on construit des variables croisées :

- **quali** $\times$ **quanti** : on autorise une variation du coefficient de la variable quantitative selon les modalités de la variable qualitative.
- **quali** $\times$ **quali** : on cherche à mesurer l'impact des croisements des deux variables sur Y.

Choix des variables explicatives

Intégrer des variables croisées (2)

$$salaire = \beta_0 + \beta_1 femme + \beta_2 age \times homme + \beta_3 age \times femme + u$$

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	986.66574	202.24657	4.88	<.0001
sexe2	Femme	1	-60.38107	298.29456	-0.20	0.8397
age_sexe1	Age x Homme	1	23.89774	4.70457	5.08	<.0001
age_sexe2	Age x Femme	1	15.80571	4.97500	3.18	0.0016

Choix des variables explicatives

Intégrer des variables croisées (3)

$$salaire = \beta_0 + \beta_1 homme \times tpp1 + \beta_2 femme \times tpp0 + \beta_3 femme \times tpp1 + u$$

Parameter Estimates						
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	Intercept	1	2095.75087	53.86442	38.91	<.0001
sexe1_tpp1	Homme x Temps partiel	1	-1484.75087	198.38800	-7.48	<.0001
sexe2_tpp0	Femme x Pas de temps partiel	1	-203.06391	80.91365	-2.51	0.0123
sexe2_tpp1	Femme x Temps partiel	1	-1148.40087	106.23725	-10.81	<.0001

Choix des variables explicatives

Utiliser des méthodes de sélection de modèle

Le choix des variables explicatives à intégrer dans un modèle peut dans une certaine mesure être **automatisé**, surtout quand le modèle a une visée prédictive.

Une des méthodes pour ce faire est la **régression pas-à-pas** :

- **pas-à-pas ascendant** : ajout progressif des variables hors-modèle si leur influence sur la variable expliquée est jugée significative ;
- **pas-à-pas descendant** : suppression progressive des variables non-significatives ;
- **meilleur sous-ensemble** : test de toutes les combinaisons de variables explicatives possibles et choix sur la base du test du score de vraisemblance.

Tests d'hypothèses complexes

Limites des tests déjà présentés

Les tests déjà présentés sont de deux types :

- les **tests globaux** : tests utilisés en analyse de variance à un facteur, tests de significativité globale des paramètres (statistique F ou ratio de vraisemblance) ;
- les **tests de significativité** : test de Student (régression linéaire) ou test de Wald (modèle linéaire généralisé).

Dans certains cas cependant, il est nécessaire de tester une hypothèse portant sur **plusieurs paramètres d'un modèle mais pas sur tous** :

- influence d'un facteur dans une analyse de variance à plusieurs facteurs ou test de significativité jointe de toutes les modalités d'une variable qualitative ;
- tests spécifiques quand une variable est introduite avec son carré ou en présence d'interactions.

Tests d'hypothèses complexes

Formulation d'hypothèses complexes

À des fins d'illustration, on se place dans le modèle :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

et on cherche à mener deux tests complexes :

$$H_0 : \beta_2 = 0 \text{ et } \beta_3 = 0 \quad \text{contre} \quad H_1 : \beta_2 \neq 0 \text{ ou } \beta_3 \neq 0$$

et d'autre part :

$$H_0 : \beta_2 = \beta_3 \quad \text{contre} \quad H_1 : \beta_2 \neq \beta_3$$

Remarque : Le premier cas correspond au test de significativité globale d'une variable qualitative en régression multiple et le second au test à mener en présence d'interactions.

Tests d'hypothèses complexes

Modèle contraint et modèle non-contraint (1)

La construction d'une statistique de test générale dans ces deux configurations s'appuie sur les notions de **modèle contraint** et de **modèle non-contraint**.

Le **modèle non-contraint** est le **modèle complet**, celui qui comporte le plus de paramètres distincts. Ici, pour les deux tests il s'agit de :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

Le **modèle contraint** est le modèle obtenu **quand l'hypothèse nulle est respectée** : la valeur des paramètres est alors « contrainte » à la valeur que l'on souhaite tester.

Tests d'hypothèses complexes

Modèle contraint et modèle non-contraint (2)

Pour le test

$$H_0 : \beta_2 = 0 \text{ et } \beta_3 = 0 \quad \text{contre} \quad H_1 : \beta_2 \neq 0 \text{ ou } \beta_3 \neq 0$$

le modèle contraint est donc :

$$Y = \beta_0 + \beta_1 X_1 + u$$

Pour le test

$$H_0 : \beta_2 = \beta_3 \quad \text{contre} \quad H_1 : \beta_2 \neq \beta_3$$

le modèle contraint est donc :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 (X_2 + X_3) + u$$

Tests d'hypothèses complexes

Intuition du test

Dans la mesure où il comporte plus de paramètres distincts, le modèle non-contraint conduit mécaniquement à un meilleur ajustement aux données :

- somme des carrés des résidus (SCR_{nc}) plus faible et donc R^2_{nc} plus élevé en régression linéaire ;
- log-vraisemblance ℓ_{nc} plus élevée dans le modèle linéaire généralisé.

Intuition : Si l'ajustement du modèle aux données est bien meilleur dans le modèle non-contraint que dans le modèle contraint ($R^2_{nc} >> R^2_c$ ou $\ell_{nc} >> \ell_c$), alors on aura tendance à rejeter la contrainte, c'est-à-dire H_0.

Tests d'hypothèses complexes

Statistique de test en régression linéaire

Pour un modèle comportant K variables explicatives (+ la constante) et si le test impose q restrictions, alors on peut montrer que sous H_0 :

$$F = \frac{(SCR_c - SCR_{nc})/q}{SCR_{nc}/(n-(K+1))} = \frac{(R^2_{nc} - R^2_c)/q}{(1-R^2_{nc})/(n-(K+1))} \hookrightarrow F_{q,n-(K+1)}$$

Dans les deux tests présentés $K = 3$. Dans le premier $q = 2$ et dans le second $q = 1$.

Remarque : Quand le modèle contraint est le modèle ne comportant que la constante, alors $SCR_c = SCT$, $SCR_{nc} = SCR$ et $q = K$. On retrouve ainsi exactement le test de significativité globale (*cf.* session 4) :

$$F = \frac{(SCT - SCR)/K}{SCR/(n-(K+1))} = \frac{SCE/K}{SCR/(n-(K+1))} \hookrightarrow F_{K,n-(K+1)}$$

Tests d'hypothèses complexes

Statistique de test dans le modèle linéaire généralisé

Dans le modèle linéaire généralisé, ce test peut être posé comme un test du ratio de vraisemblance. On peut en effet montrer que sous H_0 :

$$LR = -2ln\left(\frac{L_c}{L_{nc}}\right) = 2(\ell_{nc} - \ell_c) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_q$$

où ℓ_{nc} est la log-vraisemblance du modèle non-contraint et ℓ_c la log-vraisemblance du modèle contraint.

Remarque : Quand le modèle contraint est le modèle ne comportant que la constante, alors $\ell_c = \ell^0$, $\ell_{nc} = \ell_n$ et $q = K$. On retrouve ainsi exactement le test de significativité globale (*cf.* session 5) :

$$LR = -2ln\left(\frac{L^0}{L_n}\right) = 2(\ell_n - \ell^0) \xrightarrow[n\to+\infty]{\mathcal{L}} \chi^2_K$$

Tests d'hypothèses complexes

Exemple : Salaire dans l'EEC 2012T4

Dans un modèle de salaire intégrant l'âge, le sexe et le diplôme, on teste la significativité globale du diplôme :

$$H_0 : \beta_{ddipl1} = \ldots = \beta_{ddipl7} = 0$$
contre
$$H_1 : \beta_{ddipl1} \neq 0 \quad \text{ou} \ldots \text{ou} \quad \beta_{ddipl7} \neq 0$$

Test 1 Results for Dependent Variable SALRED				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	5	27884722	37.05	<.0001
Denominator	639	752561		

Tests d'hypothèses complexes

Exemple : Salaire dans l'EEC 2012T4

On se place dans le modèle :

$$salaire = \beta_0 + \beta_1 femme + \beta_2 age \times homme + \beta_3 age \times femme + u$$

et on teste la significativité de l'écart de salaire associé à l'âge entre hommes et femmes :

$$H_0 : \beta_2 = \beta_3 \quad \text{contre} \quad H_1 : \beta_2 \neq \beta_3$$

Test 1 Results for Dependent Variable SALRED				
Source	**DF**	**Mean Square**	**F Value**	**Pr > F**
Numerator	1	1344466	1.40	0.2377
Denominator	643	962622		

Choix de la spécification du modèle

Plan de la partie

Modèles log-linéaires

Modèle probit

Modèles logit et probit polytomiques

Modèles de Poisson

Choix de la spécification du modèle

Modèles log-linéaires (1)

On désigne par modèles log-linéaires les modèles de la forme :

$$ln(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_K X_K + u$$

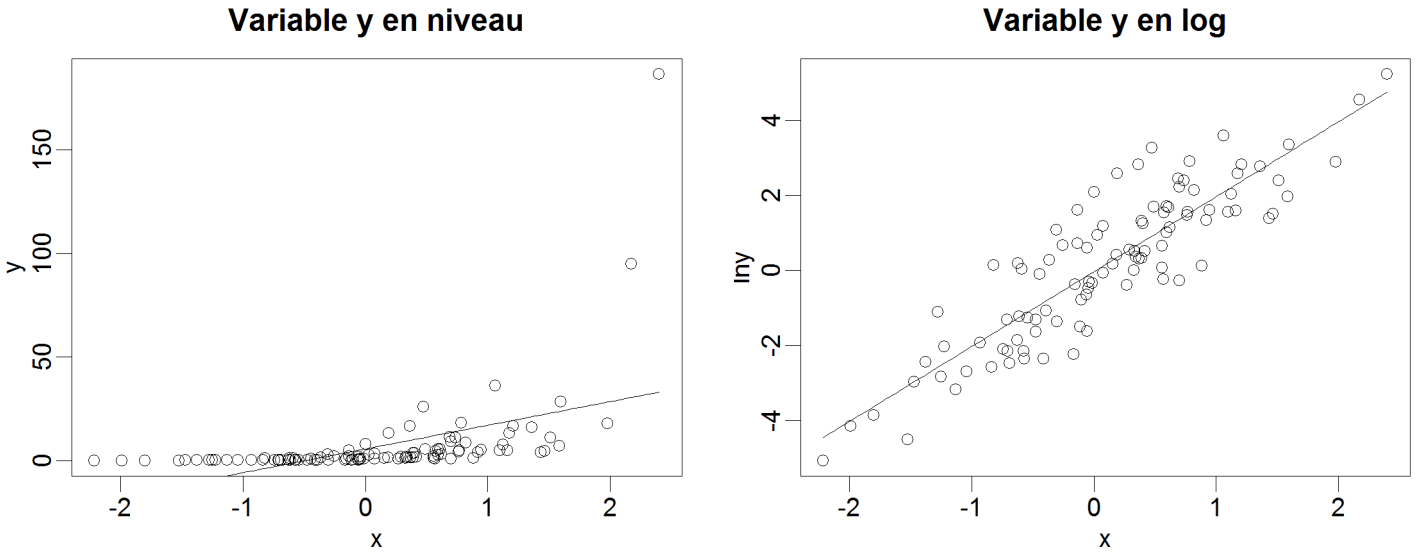
Il s'agit de modèles de régression linéaire dans lesquels la variable expliquée est remplacée par son logarithme.

Ces modèles sont utiles quand la variable d'intérêt ne prend que des valeurs strictement positives et présente une dispersion qui n'est pas constante.

Exemple : La modélisation du salaire est l'exemple le plus fréquent d'utilisation de modèles log-linéaires.

Choix de la spécification du modèle

Exemple : Données simulées



Choix de la spécification du modèle

Modèles log-linéaires (2)

L'interprétation des coefficients dans un modèle log-linaire diffère de celle des modèles linéaires classiques.

Le coefficient β_k associé à la variable X_k s'interprète ainsi comme l'**augmentation moyenne en pourcentages de la variable Y associée à une augmentation de X_k de 1**.

Remarque : Il est également possible d'introduire certaines variables explicatives en logarithme :

$$ln(Y) = \beta_0 + \beta_1 ln(X_1) + \beta_2 X_2 + u$$

Le coefficient $\hat{\beta}_1$ s'interprète alors comme l'**élasticité** de la variable Y à la variable X_1 : il correspond à l'augmentation en pourcentages de Y associée à une augmentation de 1 % de X_1.

Choix de la spécification du modèle

Modèle probit (1)

Le modèle probit est un modèle de régression pour données qualitatives dichotomiques :

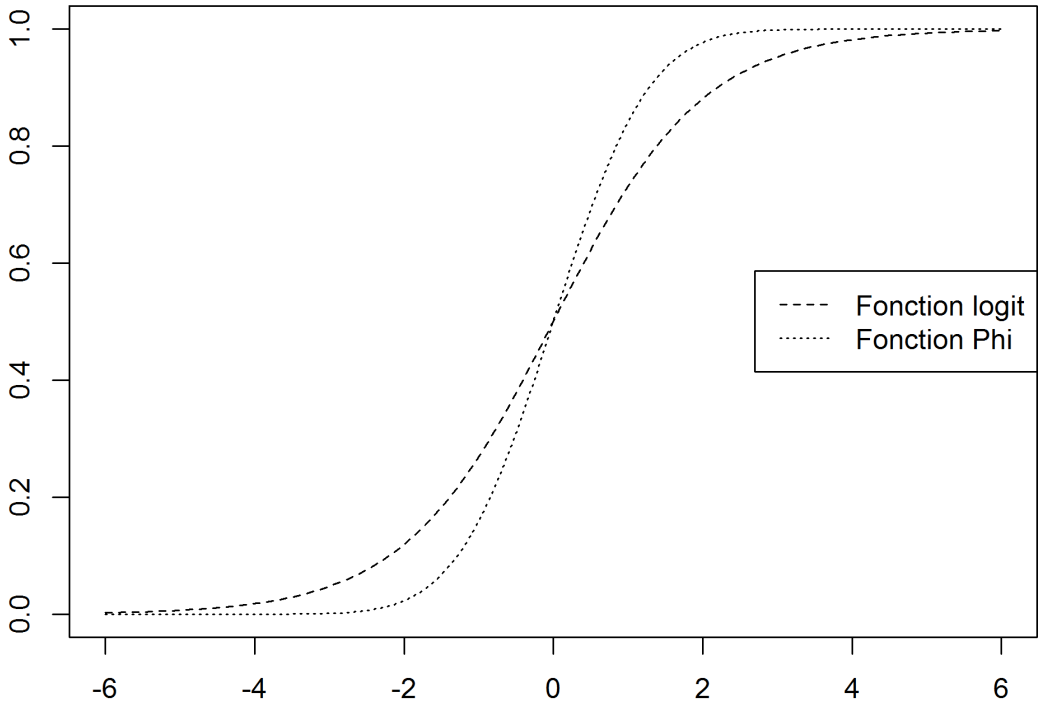
- comme le modèle logit, il appartient à la famille des modèles linéaires généralisés binomiaux ;
- à la différence du modèle logit, sa fonction de lien est la fonction de répartition de la loi normale centrée réduite :

$$f(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

Cette fonction a des propriétés proches de celles de la fonction logistique : qualitativement, les résultats obtenus avec une régression probit sont proches de ceux obtenus avec une régression logistique.

Choix de la spécification du modèle

Modèle probit (2)



Attention : En revanche, seuls les coefficients de la régression logistique peuvent donner lieu à une interprétation en termes d'*odds-ratio* !

Choix de la spécification du modèle

Modèles logit et probit polytomiques

En tant que tels, les modèles logit et probit ont été développés pour des variables expliquées qualitatives dichotomiques.

Cependant, ils peuvent être adaptés au cas des variables expliquées qualitatives polytomiques.

On distingue alors en général selon que la variable expliquée est ordonnée ou non-ordonnée.

Choix de la spécification du modèle

Modèles polytomiques ordonnés (ou ordinaux)

Le cadre des modèles logit et probit peut être **étendu au cas polytomique ordonné**, au prix d'une complexité supplémentaire.

Il est ainsi possible d'estimer des modèles logit ou probit ordonnés qui présentent les mêmes propriétés que les modèles dichotomiques, mais avec **deux particularités** :

- ils présentent plusieurs constantes (une par niveau de la variable expliquée - 1) ;
- leurs coefficients (ou *odds-ratio* dans le cas du modèle logit) s'interprètent en termes de probabilité moyenne de « passer à l'échelon supérieur » de la variable expliquée.

Choix de la spécification du modèle

Modèles polytomiques non-ordonnés

Quand les différentes modalités de la variable expliquée ne sont pas ordonnées, il est possible de mener une régression dite **multinomiale**.

Concrètement, cela revient à **choisir une modalité de la variable expliquée comme référence**, et à réaliser toutes les régressions dichotomiques des autres modalités contre celle-ci.

Cette méthodologie pose deux types de difficultés :

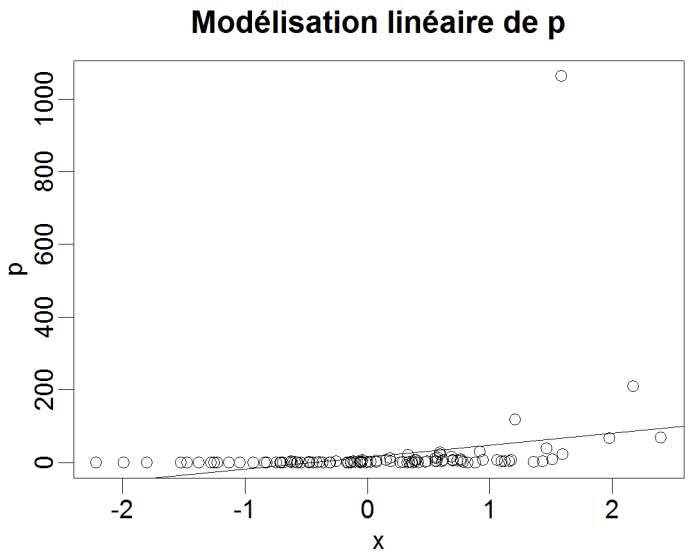
- la modalité de référence de la variable expliquée n'est pas toujours évidente ;
- le très grand nombre de coefficients à présenter et à interpréter peut perdre le lecteur.

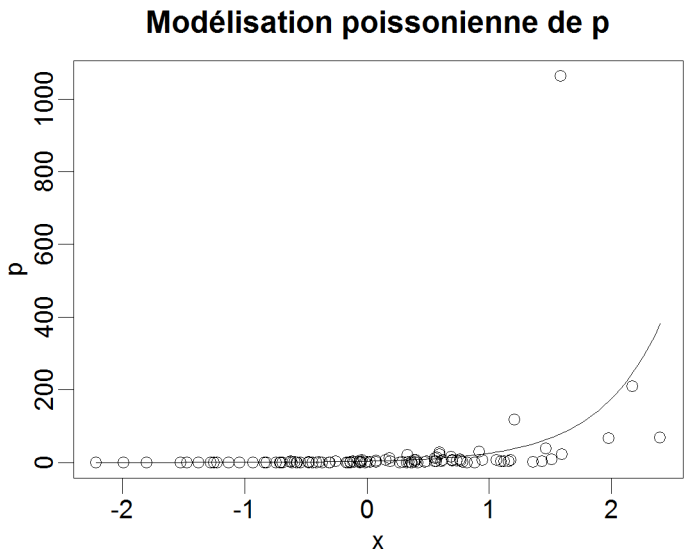
Choix de la spécification du modèle

Modèles de Poisson

Le modèle de Poisson correspond à une famille spécifique de modèles linéaires généralisés adaptée pour modéliser des données de comptage.

Exemple : Nombre d'accidents du travail dans une entreprise.





Retour sur les hypothèses de la modélisation

Hétéroscédasticité et erreurs standards robustes

Le modèle linéaire classique comme le modèle linéaire généralisé reposent sur l'hypothèse d'**homoscédasticité** :

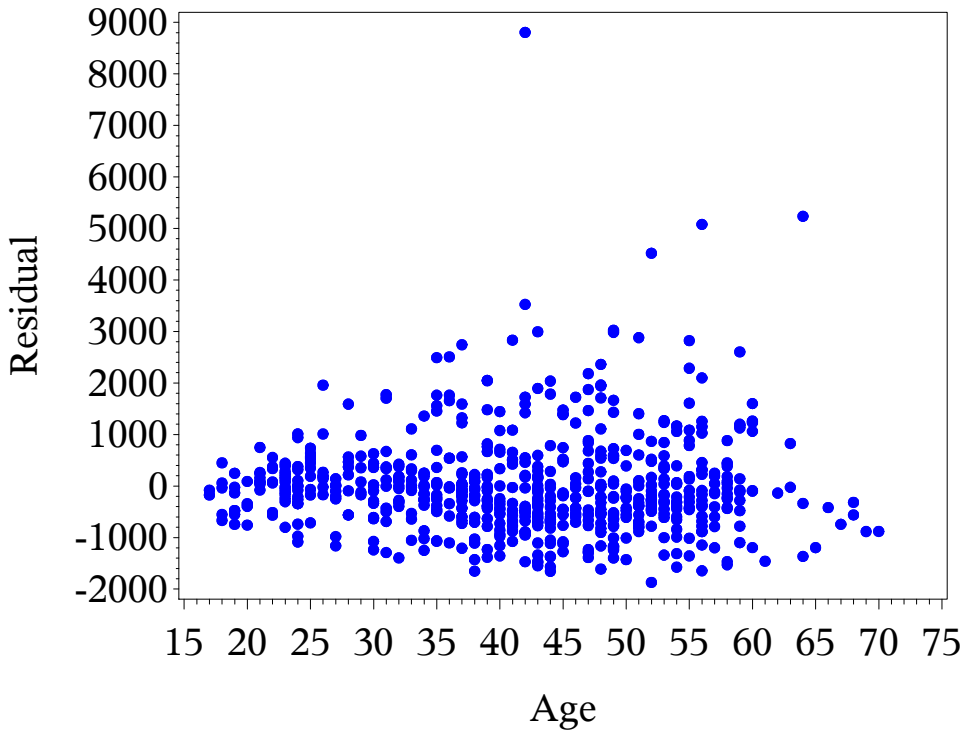
$$V(u|X_1, \ldots, X_K) = \sigma^2 = \text{constante}$$

Les erreurs standards étant calculées sous cette hypothèse d'homoscédasticité, quand elle n'est pas respectée **les tests sur les paramètres peuvent être faussés**.

Retour sur les hypothèses de la modélisation

Exemple : Salaire et âge dans l'EEC 2012T4

$$salaire = \beta_0 + \beta_1 age + +\beta_2 age^2 + u$$



Retour sur les hypothèses de la modélisation

Erreurs standards robustes

Il est néanmoins possible de prendre en compte l'hétéroscédasticité en utilisant une estimation dite « robuste » des erreurs standards.

Le principe de la méthode proposée par White (1980) est de faire intervenir la véritable distribution des résidus dans l'estimation de la matrice de variance-covariance du modèle.

On aboutit ainsi à une estimation dite « sandwich » de la matrice de variance-covariance, qui conduit à des erreurs standards en général (mais pas toujours) plus larges que les erreurs standards classiques.

Retour sur les hypothèses de la modélisation

Exemple : Salaire et âge dans l'EEC 2012T4

$$salaire = \beta_0 + \beta_1 age + +\beta_2 age^2 + u$$

Parameter Estimates									
							Heteroscedasticity Consistent		
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	-1512.77197	432.55151	-3.50	0.0005	367.44413	-4.12	<.0001
age	Age	1	151.70540	21.87161	6.94	<.0001	21.07460	7.20	<.0001
age2	Age au carré	1	-1.62583	0.26475	-6.14	<.0001	0.27456	-5.92	<.0001

Retour sur les hypothèses de la modélisation

Endogénéité et causalité

Une des hypothèses les plus importantes des modèles de régression est l'hypothèse d'espérance conditionnelle des erreurs nulle :

$$E(u|X_1, \ldots, X_K) = 0$$

Si cette hypothèse n'est pas respectée, les estimateurs obtenus ne sont plus sans biais : on parle alors d'**endogénéité**.

Les deux principales sources d'endogénéité que l'on rencontre en pratique sont le **biais de variable omise** et la **causalité inverse**. L'endogénéité peut également provenir d'erreurs de mesure sur la variable expliquée ou la variable d'intérêt.

Retour sur les hypothèses de la modélisation

Biais de variable omise

Quand une variable importante n'est pas intégrée dans le modèle, tous les coefficients sont susceptibles d'être biaisés.

Ce phénomène se comprend assez bien quand on observe l'évolution des coefficients lors de l'introduction de nouvelles variables (modèles imbriqués).

Exemple Introduire le temps partiel modifie la relation entre sexe et salaire : temps partiel et salaire d'une part et temps partiel et sexe d'autre part sont liés.

Le biais de variable omise est le signe des **effets de structure** qui relient les différentes variables explicatives.

Retour sur les hypothèses de la modélisation

Causalité inverse (ou simultanéité)

Dans de nombreuses situations, les relations entre variable expliquée et variable d'intérêt ne sont pas univoques : la causalité peut aller dans les deux sens.

Dans cette situation on parle de **simultanéité** ou de **causalité inverse**.

Exemple Relation entre mauvais état de santé et activité sur le marché du travail :

- être en mauvaise santé affecte la participation au marché du travail (recherche d'emploi plus difficile, incapacité, etc.) ;
- mais l'absence de participation au marché du travail peut avoir un impact très négatif sur la santé (conditions de vie, isolement, etc.).

Retour sur les hypothèses de la modélisation

Quelques pistes pour traiter l'endogénéité

Données plus complètes ou en panel Au stade de la conception et en fonction des objectifs de modélisation, certaines variables peuvent être ajoutées voire être observées à plusieurs moments du temps.

Logique expérimentale Une autre solution consiste à se rapprocher d'un protocole de type expérimental :

- expérimentation contrôlée : domaine médical, évaluation des politiques publiques ;
- expériences naturelles, discontinuités : ruptures spatiales, évolutions dans le temps.

Utiliser des variables instrumentales L'utilisation de certaines variables peut permettre de « filtrer » l'endogénéité de la variable explicative et donc de neutraliser le biais.