15/07/2023

Examen de statistiques D.U. Data Analyst

CY Cergy Paris Université

Ibrahima Ly

Table des matières

Références :

- *D' haultfoeuille X Givord P 2014 La régression quantile en pratique Economie et statistique, 471 1 85 111*

- *Econométrie : manuel et exercices corrigés - 11ème édition Mars 2021 Auteur : Bourbonnais Régis*

- *Probabilités, analyse des données et Statistique de Gilbert Saporta*

I. Chapitre 1 : Introduction

1.1 Types de variables:

a. Variables Quantitatives: Elles reflètent une notion de grandeur ou de quantité. Il existe deux types de variables quantitatives :

- ✓ **Variable quantitative continue:** Ces variables peuvent avoir une infinité de valeurs possibles. Exemples:
  - Chiffre d'affaires des entreprises des TIC
  - Nombre de personnes utilisant internet au sein d'un pays

- ✓ **Variable quantitative discrète:** Ces variables ont un nombre limité de valeurs possibles. Exemples:
  - Nombre d'applications téléchargées sur le téléphone portable
  - Heures passées à regarder Netflix

b. Variables Qualitatives: Elles contiennent des informations relatives à une situation, une caractéristique. Ses valeurs sont des « catégories » ou « modalités ». Il existe deux types de variables qualitatives :

- ✓ **Variable qualitative ordonnée:** Il existe un ordre naturel entre chaque catégorie. Exemples:
  - Fréquence d'utilisation d'Instagram
  - Satisfaction liée à l'utilisation d'une enceinte connectée

- ✓ **Variable qualitative non ordonnée:** Il n'y a pas de classement naturel des modalités. Exemples:
  - Type de connexion internet utilisée à domicile
  - Marque du téléphone portable

1.2 Choix du type de graphique:

a. Graphique linéaire: Il est utilisé pour représenter l'évolution des variables quantitatives dans le temps.

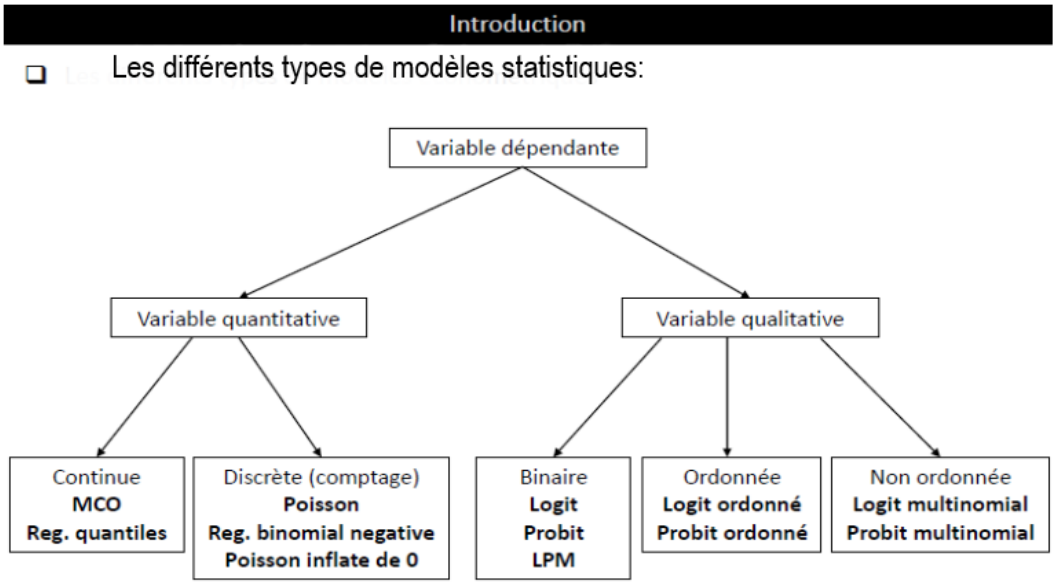
b. Graphique à barres: Il est utilisé pour représenter la fréquence des modalités des variables qualitatives, la valeur moyenne d'une variable quantitative en fonction d'une unité de temps ou des modalités d'une variable qualitative, et pour représenter la répartition d'une variable quantitative entre ses composantes.

c. Graphique circulaire (camembert): Il est utilisé pour représenter la répartition entre les modalités d'une variable qualitative.

d. Histogramme: Il est utilisé pour représenter la distribution d'une variable quantitative continue ou discrète.

e. Boite à moustaches: Il est utilisé pour représenter la distribution d'une variable quantitative continue ou discrète.

f. Nuage de points: Il est utilisé pour représenter la relation entre des variables quantitatives.



1.3 Tester les associations bivariées entre variables : Plusieurs tests peuvent être utilisés selon le type de variables à comparer :

a. **Corrélation de Pearson (pwcorr) :** Il est utilisé pour tester deux variables quantitatives normales

b. **Corrélation de Spearman (spearman) :** Il est utilisé pour tester deux variables quantitatives non normales

c. **Test t (ttest) :** pour une variable binaire et une quantitative normale

d. **Test de Wilcoxon-Mann-Whitney (ranksum) :** pour une variable binaire et une quantitative non normale

e. **Test du khi-deux (chi2) ou test de Fisher (exact)** : après une tabulation croisée pour comparer deux variables binaires ou catégorielles avec une fréquence croisée inférieure à 5

f. **ANOVA (anova) :** pour une variable catégorielle (>2 modalités) et une quantitative normale

g. **Test de Kruskal-Wallis (kwallis) :** pour une variable catégorielle (>2 modalités) et une quantitative non normale.

Récapitulatif des tous les tests utilise			
Nom	**Statistique**	**Hypothèses**	**Régression auxiliaire**
Tests des significativité			
Student	$\lvert T_c \rvert = \frac{\hat{\beta}}{\sigma\hat{\beta}} \sim T_{lue}\ (1.96)$	H_0 : non-significatif H_1 : significatif	
Fisher	$F_{stat} = \frac{R^2/_{k-1}}{1-R^2/_{n-k}} \sim F_{1-p}(v_1, v_2)$	H_0 : non-significatif H_1 : significatif	
Tests d'autocorrélation			
Durbin-Watson	$DW = \frac{\Sigma(\varepsilon_t - \varepsilon_{t-1})^2}{\varepsilon_t^2}$	H_0 : ρ=0 : pas d' autocorrélation H_1 : ρ≠0 :autocorrélation	$\varepsilon_t = \rho\ \varepsilon_{t-1} + \eta_t$
Ljung-Box	$Q=n(n+2)\Sigma_1^k \frac{r_k^2}{n-k} \sim \chi^2(k)$	H_0 : ϕ =0 :pas d'autocorrélation H_1 : ϕ ≠0 :autocorrélation	$\varepsilon_t = \phi_1\ \varepsilon_{t-1} + \phi_2\ \varepsilon_{t-2} + \phi_3\ \varepsilon_{t-3} + \ldots\ldots\ldots + \phi_n\ \varepsilon_{t-k} + \eta_t$
Test d'hétéroscédasticité			
Glejser	$\lvert T_c \rvert = \frac{\hat{\beta}}{\sigma\hat{\beta}} \sim T_{lue}\ (1.96)$	H_0 : β =0 :homoscédasticité H_1 : β ≠0 :hétéroscédasticité	$\lvert e_t \rvert = \alpha + \beta * X_1 + \varepsilon_t$ $\lvert e_t \rvert = \alpha + \beta * X_2 + \varepsilon_t$
White	$W = nR^2 \sim \chi^2(k)$ K= no des variables	H_0 : α =0 : homoscédasticité H_1 : α ≠0 : hétéroscédasticité	$e_t^2 = \alpha_0$+tous les var + carre des tous var + croise + ε_t
ARCH	$LM = nR^2 \sim \chi^2(k)$ K= no des retards	H_0 : ϕ =0 : homoscédasticité H_1 : ϕ ≠0 : hétéroscédasticité	$\varepsilon_t^2 = c + \phi_1\varepsilon^2_{t-1} + \phi_2\varepsilon^2_{t-2} + \ldots\ldots + \phi_n\varepsilon^2_{t-n} + \eta_t$
Breusch-Pagan	$BP = nR^2 \sim \chi^2(k)$ K = no des variables	H_0 : γ =0 : homoscédasticité H_1 : γ ≠0 : hétéroscédasticité	$\sigma_\varepsilon^2 = \gamma_0 + \gamma_1 * X_1 + \gamma_2 * X_2 + \gamma_3 * X_3 + \ldots\ldots + \gamma_n * X_n + \upsilon_t$
Test de Normalité			
Jarque-Bera	$= N*[\frac{Skw^2}{6} + \frac{(K-3)^2}{24}] \sim \chi^2(2)$	H_0 : Normalité H_1 : Non-Normalité	

II. Chapitre 2 : Régression

2.1 Différence entre une régression linéaire simple et une régression linéaire multiple :

Une régression linéaire simple :

Examine la relation entre une seule variable explicative (indépendante) et une variable à expliquer (dépendante). L'équation de cette régression est :

$$Yi = \alpha + \beta 1 X 1 i + \varepsilon$$

Une régression linéaire multiple :

Quant à elle, examine la relation entre deux variables explicatives ou plus et une variable à expliquer. L'équation de cette régression est :

$$Yi = \alpha + \beta 1 X 1 i + \beta 2 X 2 i + \ldots + \beta n X n i + \varepsilon i$$

$$E(Yi|Xi) = \alpha + \beta 1 X 1 i + \beta 2 X 2 i + \ldots + \beta n X n i + \varepsilon i \ (2)$$

L'équation 2 est la Modélisation espérance conditionnelle de Y moyenne.

L'indice *"i"* représente l'unité d'observation, qui peut être un individu, un pays, une entreprise, etc.

La variable *"Yi"* représente la variable expliquée ou dépendante, c'est-à-dire la variable que l'on cherche à expliquer ou prédire.

Les variables *"X1i" à "Xni"* représentent les variables explicatives ou indépendantes, qui sont utilisées pour expliquer les variations de la variable dépendante.

La constante *"α"* est la valeur de "Yi" lorsque la valeur de toutes les variables explicatives est fixée à 0.

Les paramètres *β1 à βn* sont les paramètres à estimer. Ils mesurent l'effet d'une augmentation d'une unité de la variable explicative sur la valeur de Yi. Ces paramètres permettent de quantifier l'importance de chaque variable explicative dans l'explication de Yi.

Le terme d'erreur *"εi"* représente toutes les informations manquantes dans l'explication linéaire des valeurs de Yi. Il peut inclure des problèmes de spécifications, des variables non prises en compte ou d'autres facteurs non capturés par le modèle linéaire.

Pour interpréter les résultats sous forme d'élasticités de la regression simple ou multiple, la variable dépendante et/ou les variables indépendantes peuvent être exprimées en logarithme :

$$ln(Yi) = \alpha + \beta 1 ln(X1i) + \beta 2 ln(X2i) + \ldots + \beta n ln(Xni) + \varepsilon i$$

Dans cette équation :

Les coefficients β1 à βn représentent les élasticités. Ils mesurent l'effet d'une augmentation d'un pourcentage d'une variable explicative sur la variation en pourcentage de Yi.

Si les variables ne sont pas exprimées en logarithme, la forme de l'équation devient :

$$Yi = \alpha + \beta 1 X1i + \beta 2 X2i + \ldots + \beta n Xni + \varepsilon i$$

Dans cette équation :

- ✓ **Les coefficients β1 à βn** représentent des semi-élasticités,
- ✓ Ils mesurent l'effet d'une augmentation d'une unité de la variable explicative sur la variation de **Yi.**
- ✓

En résumé, lorsque les variables sont exprimées en logarithme, les coefficients représentent des élasticités qui mesurent les variations en pourcentage, tandis que lorsque les variables ne sont pas en logarithme, les coefficients représentent des semi-élasticités qui mesurent les variations absolues.

La régression standardisée est une autre méthode utilisée pour comparer l'effet relatif des différentes variables explicatives sur la variable expliquée. Dans cette méthode, la variable dépendante ainsi que les variables indépendantes sont standardisées de manière à avoir une moyenne de 0 et un écart type de 1.

L'équation de la régression standardisée s'écrit comme suit :

$$sYi = \alpha + \beta 1 sXi + \beta 2 sXi + \ldots + \beta n sXi + \varepsilon i$$

Dans cette équation :

Les coefficients β1 à βn représentent l'effet d'une augmentation d'un écart type dans la variable explicative sur la variation en termes d'écart type de **Yi.**

En d'autres termes, ils mesurent l'importance relative de chaque variable explicative sur la variable dépendante lorsque les variables ont été standardisées.

Dans le contexte de la régression standardisée, la variable explicative ayant le coefficient β le plus élevé (et significatif) est celle qui exerce l'influence la plus importante sur **Yi.** Cela signifie que cette variable a un impact relatif plus fort sur la variation de la variable dépendante par rapport aux autres variables explicatives.

Le modèle MCO (Moindres Carrés Ordinaires) repose sur un certain nombre d'hypothèses :

Linéarité : ***Yi*** est une fonction linéaire des variables X1 à Xn et du terme d'erreur εi.

Les Xn sont déterminés sans erreurs : Les variables explicatives sont mesurées avec précision, sans erreurs de mesure.

E(εi|X1, Xn) = 0 : Le modèle est bien spécifié en moyenne, ce qui signifie qu'il ne manque pas de variables explicatives importantes pour expliquer la variation de Yi.

Var(εi|X1, Xn) = σ^2 : Le terme d'erreur a la même variance pour toutes les observations, ce qui est appelé l'homoscédasticité.

Cov(εi, εj) = 0 : Il n'y a pas d'autocorrélation des erreurs entre les différentes observations.

Absence de colinéarité entre les variables explicatives : Aucune variable explicative n'est une combinaison linéaire des autres variables. Les variables explicatives doivent être indépendantes les unes des autres.

Normalité des résidus : Les résidus ***εi*** doivent suivre une distribution normale. La normalité des résidus est nécessaire pour la validité des tests d'hypothèses et pour obtenir des intervalles de confiance précis.

Ces hypothèses sont importantes pour garantir la validité des résultats obtenus à partir du modèle MCO et pour interpréter correctement les coefficients estimés.

Que signifie la multicolinéarité ? A quoi sert le facteur d'inflation de la variance (VIF en anglais) dans le cadre de la lutte contre la multicolinéarité :

La multicolinéarité fait référence à une situation dans laquelle deux ou plusieurs variables explicatives dans un modèle de régression multiple sont fortement corrélées entre elles. Cela peut rendre difficile l'identification de l'effet indépendant de chaque variable explicative sur la variable à expliquer.
La multicolinéarité est un outil de test de diagnostic du modelé MCO.

Les diagnostics du modèle MCO permettent de mettre en place les outils de test pour :

- ✓ La détection des valeurs aberrantes et influentes.
- ✓ L'évaluation de la normalité des résidus.
- ✓ L'analyse de l'homoscédasticité des résidus.
- ✓ **La détection de la multicolinéarité entre les variables explicatives.**
- ✓ L'évaluation de la spécification du modèle.
- ✓ Ces diagnostics sont essentiels pour évaluer la validité et la fiabilité du modèle MCO, ainsi que pour identifier d'éventuels problèmes ou violations des hypothèses sous-jacentes.

Détection de la muticollinéarité :

La colinéarité implique que deux variables sont des combinaisons linéaires quasi parfaites l'une de l'autre (si combinaison parfaite impossible d'estimer le modèle de régression)

Lorsque plus de deux variables sont impliquées, on parle de multicollinéarité.

Le Termes de « collinéarité » et « multicollinéarité » souvent utilisés de manière interchangeable

À mesure que le degré de multicollinéarité augmente :

- ✓ Les estimations des coefficients du modèle de régression deviennent instables
- ✓ Les erreurs type des coefficients augmentent artificiellement
- ✓ Risque de considérer comme non significatif un coefficient qui l'est dans la réalité

Méthode permettent de tester la multicolinéarité dans un data Frame est le Facteur d'Inflation de la Variance (VIF) :

Le facteur d'inflation de la variance (VIF) : est un indicateur utilisé pour détecter la multicolinéarité. Un VIF élevé pour une variable explicative suggère que cette variable est fortement corrélée avec les autres variables explicatives.

VIF (variance inflated factor) : à utiliser après avoir fait la régression pour tester le niveau de collinéarité de chaque variable explicative et de la régression dans son ensemble

Règle générale : Un VIF supérieur à 5 ou 10 est source de préoccupation.

crime	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
pctmetro	7.828935	1.254699	6.24	0.000	5.304806	10.35306
poverty	17.68024	6.94093	2.55	0.014	3.716893	31.6436
single	132.4081	15.50322	8.54	0.000	101.2196	163.5965
_cons	-1666.436	147.852	-11.27	0.000	-1963.876	-1368.996

```
. vif
```

Variable	VIF	1/VIF
single	1.63	0.612873
poverty	1.53	0.654829
pctmetro	1.14	0.873510
Mean VIF	1.43	

Aucun VIF n'est > 10

➢ La multicollinéarité n'est pas un problème

Que signifie l'homoscédasticité des résidus ? Pourquoi est-ce important ?

L'homoscédasticité fait référence à une situation dans laquelle la variance des erreurs (ou des résidus) est constante sur tous les niveaux d'une variable indépendante. C'est une hypothèse importante en régression linéaire car si elle n'est pas respectée (si les données sont hétéroscédastiques), les estimations des erreurs standards peuvent être biaisées, ce qui peut à son tour conduire à des tests d'hypothèses incorrectes.

On parle de l'homoscédasticité lorsque le terme d'erreur a la même variance pour toutes les observations. Autrement dit, le terme d'erreur a la même variance pour toutes les observations.

$$Var(\varepsilon i|X1, Xn) = \sigma^2$$

L' homoscédasticité est l'une des principales hypothèses de la régression MCO qui vérifie l'homogénéité de la variance des résidus, c'est-à-dire, s'ils sont homoscédastiques.

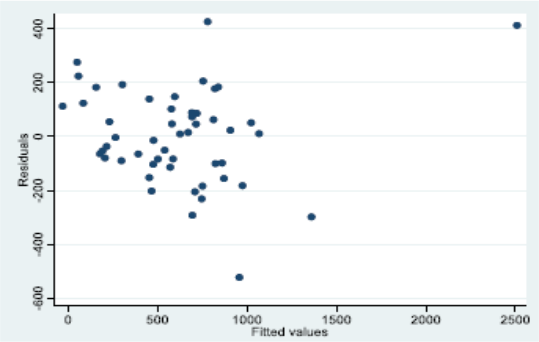
Si la variance des résidus est non constante, on dit que la variance des résidus est hétéroscédastique.

Il existe des méthodes graphiques et non graphiques pour détecter l'hétéroscédasticité.

Une méthode graphique couramment utilisée consiste à tracer les résidus en fonction des valeurs prédites

Si les résidus sont homoscédastiques il ne devrait pas y avoir de relation (tendance du point de vue graphique) entre résidus et les valeurs prédites.

Exemple de Test de Breusch-Pagan / Cook-Weisberg pour l'Hétéroscédasticité : Analyse de Résidus sous Stata :



Il semble bien y avoir une relation entre les résidus et les valeurs prédites

- Semble signaler un problème d'hétéroscedasticité

- Testons plus formellement la présence d'hétéroscedasticité à l'aide du test de Breusch-Pagan (commande *hettest*)

```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of crime

         chi2(1)      =    12.86
         Prob > chi2  =   0.0003
```

Rejet de H0

- La variance des résidus n'est pas homogène
- Problème d'hétéroscedasticité confirmé

Illustration de l'Hétéroscédasticité Confirmée via l'Analyse de Résidus avec Stata

Le graphique ci-dessus illustre une relation significative entre les résidus et les valeurs prédites ($p < 0.001$), ce qui suggère un problème d'hétéroscédasticité. Le test de Breusch-Pagan / Cook Weisberg effectué avec Stata confirme ce problème, rejetant l'hypothèse nulle d'une variance constante.

Les conséquences de l' hétéroscédasticité :

a. Absence de biais dans l'estimation des coefficients

✓ Malgré la présence d'hétéroscédasticité, les estimateurs des moindres carrés ordinaires (MCO) restent sans biais. Cela signifie que les estimations de coefficients restent fidèles à leurs vraies valeurs en moyenne.

b. Perte de l'efficacité des estimateurs

✓ Les estimations ne sont plus les « Best Linear Unbiased Estimators » (BLUE). En présence d'hétéroscédasticité, le MCO ne fournit pas l'estimation ayant la plus petite variance parmi tous les estimateurs sans biais. Cela signifie que nous pourrions avoir de meilleures estimations en utilisant d'autres méthodes.

c. Poids égal accordé à toutes les observations

✓ Le MCO donne le même poids à toutes les observations, ce qui peut être problématique en présence d'hétéroscédasticité. Les observations dont la variance des erreurs est plus grande fournissent en réalité moins d'informations.

d. Biais des erreurs types et des statistiques de test

✓ En présence d'hétéroscédasticité, les erreurs types sont biaisées, ce qui conduit à des statistiques de test biaisées et à des intervalles de confiance incorrects. Cela peut conduire à des conclusions incorrectes sur les variables significatives du modèle.

Comment Corriger l'Hétéroscédasticité?

✓ Parfois, l'hétéroscédasticité résulte d'une mauvaise spécification du modèle. Par exemple, cela peut être dû à des différences dans les effets des variables explicatives entre les sous-groupes, des effets non linéaires des variables explicatives, ou l'omission de variables explicatives importantes.

Utilisation des Erreurs Types Robustes:

✓ Pour gérer l'hétéroscédasticité, nous pouvons utiliser des erreurs types robustes avec l'option robust. Cette méthode permet de corriger les erreurs types en relâchant l'hypothèse de variance homogène.

- Les Estimateurs Huber/White, ou les estimateurs "sandwich" de la variance, sont des exemples de cette méthode.
- Il est important de noter que l'utilisation des erreurs types robustes ne modifie pas les coefficients estimés, mais fournit des p-values plus précises pour les tests d'hypothèses.

Hypothèses sous-jacentes à l'application du modèle linéaire : Il y a plusieurs hypothèses clés en régression linéaire, parmi lesquelles :

- ✓ **Linearité :** La relation entre les variables indépendantes et la variable dépendante est linéaire.
- ✓ **Indépendance :** Les observations sont indépendantes les unes des autres.
- ✓ **Homoscédasticité :** La variance des erreurs est constante à travers tous les niveaux des variables indépendantes.
- ✓ **Normalité :** Les erreurs sont normalement distribuées.

Utilité du qqplot :

Un qqplot (quantile-quantile plot) est un outil graphique pour évaluer si un ensemble de données suit une distribution particulière. C'est souvent préféré aux tests comme Shapiro ou Kolmogorov-Smirnov car il permet une évaluation visuelle qui peut être plus informative que les résultats d'un test d'hypothèse. Un qqplot permet également d'identifier des aspects spécifiques d'un écart par rapport à une distribution, comme les valeurs aberrantes ou la skewness.

Les régressions quantiles permettent de décrire plus précisément la distribution d'une variable d'intérêt conditionnelle à ses déterminants comparativement à la régression linéaire.

$$Médiane(Yi|Xi) = \alpha + \beta 1 X1i + \beta 2 X2i + \ldots + \beta nXni + \varepsilon i$$

$$Premier\_quartile(Yi|Xi) = \alpha + \beta 1 X1i + \beta 2 X2i + \ldots + \beta nXni + \varepsilon i$$

Plusieurs limites à la modélisation de la moyenne par régression MCO par rapport à la régression quantiles (qqplot):

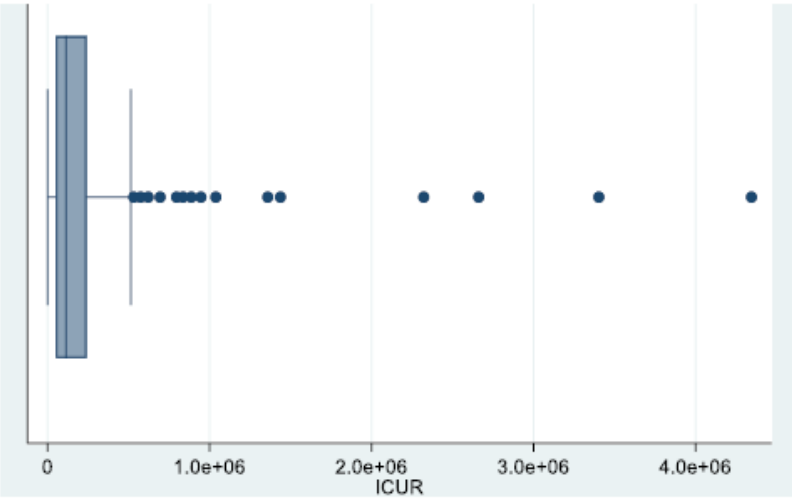
Information apportée par moyenne parfois trop partielle pour étudier certains phénomènes

Même si elle apporte une information essentielle, la modélisation de la moyenne reste limitée car elle n'apporte aucune information sur la distribution de la variable dépendante

Exemples:

L'étude du revenu moyen n'informe pas sur sa répartition ou inégale dans la population

L'étude du nombre moyen d'ordinateurs détenus par les foyers français ne renseigne pas sur la situation de « précarité numérique » de certains foyers



	Percentiles	Smallest		
1%	6111	662		
5%	17557	6111		
10%	26088	16526	Obs	123
25%	51447	16905	Sum of Wgt.	123
50%	112328		Mean	287821.3
		Largest	Std. Dev.	606220.2
75%	239145	2322263		
90%	620980	2661514	Variance	3.68e+11
95%	1038389	3403415	Skewness	4.509925
99%	3403415	4345650	Kurtosis	25.59775

L'analyse de ces données suggère que la variable "ICUR" présente une asymétrie positive avec des valeurs aberrantes potentiellement élevées. Cela est évident car la moyenne (287821.3) est significativement plus élevée que la médiane (112328), indiquant la présence de valeurs extrêmes qui déplacent la moyenne vers le haut. Ces valeurs aberrantes pourraient avoir un impact important sur l'analyse et la modélisation des données, et il pourrait être nécessaire d'utiliser des méthodes robustes ou de reconsidérer la spécification du modèle pour les gérer de manière appropriée.

III. Chapitre 3: Tests non paramétriques

1- Différence entre un test apparié et un test non-apparié :

Un test apparié est utilisé lorsque les observations sont recueillies en paires. Par exemple, si vous mesurez le poids de sujets avant et après un régime, les observations sont appariées car chaque mesure après le régime est associée à une mesure spécifique avant le régime.

Un test non apparié est utilisé lorsque les observations ne sont pas recueillies en paires. Par exemple, si vous mesurez le poids de deux groupes distincts de sujets (par exemple, un groupe qui suit un régime et un autre qui ne le suit pas), les observations ne sont pas appariées.

2- Contexte d'utilisation des tests non paramétriques :

Les tests non paramétriques sont utilisés lorsque les données ne respectent pas les hypothèses requises pour les tests paramétriques, comme la normalité. Ils sont également utiles pour les données catégorielles ou ordinales, ou lorsque les tailles des échantillons sont petites.

3-Exemples de tests non paramétriques :

- Test de Wilcoxon : utilisé pour comparer deux échantillons appariés.

- Test de Mann-Whitney : utilisé pour comparer deux échantillons indépendants.

- Test de Kruskal-Wallis : une extension du test de Mann-Whitney pour plus de deux échantillons indépendants.

- **Test du chi-carré de Pearson :** utilisé pour analyser les données de fréquence dans un tableau de contingence.

4- Origine de l'expression "test de rangs" :

L'expression "test de rangs" vient du fait que ces tests impliquent généralement le classement des données (l'attribution de rangs) avant de réaliser les comparaisons statistiques. Par exemple, dans le test de Wilcoxon ou le test de Mann-Whitney, les données de tous les groupes sont combinées et classées, puis les rangs sont utilisés pour réaliser le test.

IV. Chapitre 4 : ANOVA

Calcul d'une somme de carrés pour une variable donnée :

La somme des carrés est une mesure de la variabilité ou de la dispersion des données. Dans une ANOVA, nous calculons généralement deux types de sommes de carrés : la somme des carrés de traitement (SST) et la somme des carrés de l'erreur (SSE). La SST mesure la variabilité due aux différences entre les moyens des groupes, tandis que la SSE mesure la variabilité due aux différences à l'intérieur des groupes.

Ce tableau présente l'analyse de la variance pour un modèle de régression simple.

Source de variation	Somme des carrés	Degré de liberté	Carrés moyens
x	$SCE = \sum_t (\hat{y}_t - \overline{y})^2$	1	$SCE/1$
Résidu	$SCR = \sum_t e_t^2$	$n-2$	$SCR/(n-2)$
Total	$SCT = \sum_t (y_t - \overline{y})^2$	$n-1$	

Tableau d'Analyse de la variance pour une régression multiple

Source de variation	Somme des carrés	Degré de liberté	Carrés moyens
$x_1, x_2, \ldots, x_k$	$SCE = \sum_t (\hat{y}_t - \overline{y})^2$	k	SCE/k
Résidu	$SCR = \sum_t e_t^2$	$n-k-1$	$SCR/(n-k-1)$
Total	$SCT = \sum_t (y_t - \overline{y})^2$	$n-1$	

2- Calcul de la statistique F pour une variable donnée :

La statistique F est le rapport de la variance entre les groupes (mesurée par la SST) sur la variance à l'intérieur des groupes (mesurée par la SSE). Une valeur F élevée suggère que les moyennes des groupes sont significativement différentes.

$$F^* = \frac{\frac{SCE}{ddl_{SCE}}}{\frac{SCR}{ddl_{SCR}}} = \frac{\frac{\sum_t (\hat{y}_t - \overline{y})^2}{1}}{\frac{\sum_t e_t^2}{(n-2)}}$$

Ou encore :

$$F^* = \frac{\frac{SCE}{ddl_{SCE}}}{\frac{SCR}{ddl_{SCR}}} = \frac{\frac{\sum_t (\hat{y}_t - \overline{y})^2}{1}}{\frac{\sum_t e_t^2}{(n-2)}} = \frac{\frac{R^2}{1}}{\frac{(1-R^2)}{(n-2)}}$$

3- Passage du F à la p-value :

La p-value est la probabilité d'obtenir une valeur de F aussi extrême (ou plus) si l'hypothèse nulle est vraie (c'est-à-dire, si les moyennes des groupes sont en réalité égales). Elle est obtenue en référençant la valeur F à la distribution F, qui est une distribution théorique qui donne les valeurs attendues de F sous l'hypothèse nulle. Un graphique de la distribution F avec la région critique correspondant à la valeur F observée peut être utilisé pour illustrer comment la p-value est obtenue.

$$F^* = \frac{\sum_t (\hat{y}_t - \bar{y})^2 / k}{\sum_t e_t^2 / (n-k-1)} = \frac{R^2/k}{(1-R^2)/(n-k-1)}$$

4- Interaction entre deux facteurs dans une ANOVA :

Une interaction entre deux facteurs dans une ANOVA se produit lorsque l'effet d'un facteur sur la variable dépendante dépend du niveau de l'autre facteur. Par exemple, supposons que vous meniez une expérience pour tester l'effet de deux médicaments (facteur A et B) sur la pression artérielle. Si l'effet de la prise du médicament A dépend du fait que le sujet ait ou non pris le médicament B, alors il y a une interaction entre les deux facteurs.

V. Chapitre 5 : Régression logistique

Contexte d'utilisation des régressions de type "loi de Poisson" :

La régression de Poisson est généralement utilisée lorsque la variable dépendante est un nombre d'événements qui se produisent dans un intervalle de temps ou d'espace fixe et que ces événements sont indépendants.

La modélisation des données de comptage fait appel à différents modèles de régression en fonction de la distribution de la variable dépendante :

Modèle de poisson :

La moyenne de la variable dépendante est proche de sa variance équidispersion

- ✓ **Modèle négatif binomial :** surdispersion de la variable dépendante (variance moyenne)
- ✓ **Modèle de Poisson à inflation de 0 :** nombre excessif de 0 pour la variable dépendante

2- Utilisation du terme "Poisson" :

Le terme "Poisson" provient du nom du mathématicien français Siméon Denis Poisson, qui a introduit la distribution de Poisson en 1837. Cette distribution est largement utilisée pour modéliser le nombre d'événements qui se produisent dans un intervalle fixe de temps ou d'espace lorsque ces événements sont relativement rares et indépendants.

$$\beta = \ln(E(Y_i \mid X_i + 1)) - \ln(E(Y_i \mid X_i))$$

$$\leftrightarrow \beta = \ln\left(\frac{E(Y_i|X_i+1)}{E(Y_i|X_i)}\right)$$

3- Utilisation du terme "binomiale" :

Le terme "binomiale" est utilisé pour décrire un type de régression logistique lorsque la variable dépendante est une variable dichotomique (c'est-à-dire, elle prend deux valeurs possibles, souvent codées comme 0 et 1). Cela provient du fait que la distribution binomiale est utilisée pour modéliser le nombre de "succès" dans un nombre fixe d'essais indépendants.

$$\beta = \ln(E(Y_i \mid X_i + 1)) - \ln(E(Y_i \mid X_i))$$

$$\leftrightarrow \beta = \ln\left(\frac{E(Y_i|X_i+1)}{E(Y_i|X_i)}\right)$$

$$\leftrightarrow \boxed{e^{\beta} = \frac{E(Y_i|X_i+1)}{E(Y_i|X_i)}}$$

4- Odd-ratio et différence avec le risque relatif :

L'odd-ratio est le rapport des cotes d'un événement dans deux groupes. En épidémiologie, l'odd-ratio est souvent utilisé pour quantifier à quel point l'exposition à un certain facteur est associée à une maladie. Le risque relatif, quant à lui, est le rapport des probabilités d'un événement dans deux groupes. Les deux mesures sont similaires, mais l'odd-ratio peut être utilisé même lorsque l'événement est rare, tandis que le risque relatif est plus intuitif (il dit simplement combien de fois l'événement est plus probable dans un groupe que dans l'autre).

5- Conditions pour lesquelles les risques relatifs et OR sont approximativement les mêmes :

Lorsque l'événement d'intérêt est rare (c'est-à-dire, sa probabilité est faible), l'odd-ratio et le risque relatif seront très similaires. Cela est dû au fait que lorsque la probabilité d'un événement est faible, les "chances" de l'événement (c'est-à-dire, la probabilité de l'événement divisée par la probabilité de non-événement) seront très similaires à la probabilité elle-même.