ANALYSE DE DONNÉES

Résumé pratique du cours de Mme BENBRAHIM

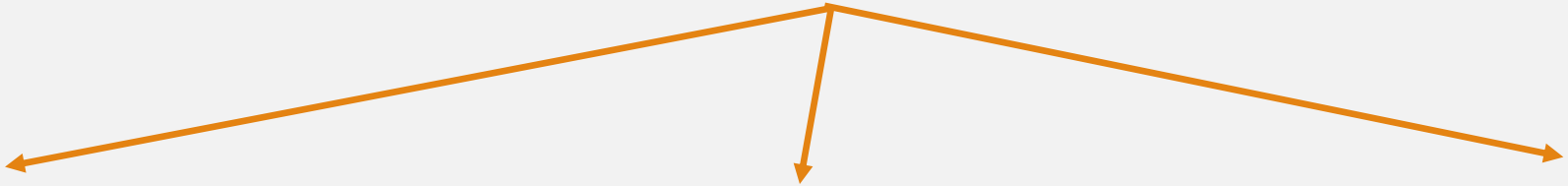
Réalisé par Meryam Daoudi

STRUCTURE DU COURS

1. Description des données
2. Régression linéaire Simple/Multiple
3. ACP

COMMENT DÉCRIRE MES DONNÉES ?

Je calcule les INDICATEURS STATIQUES



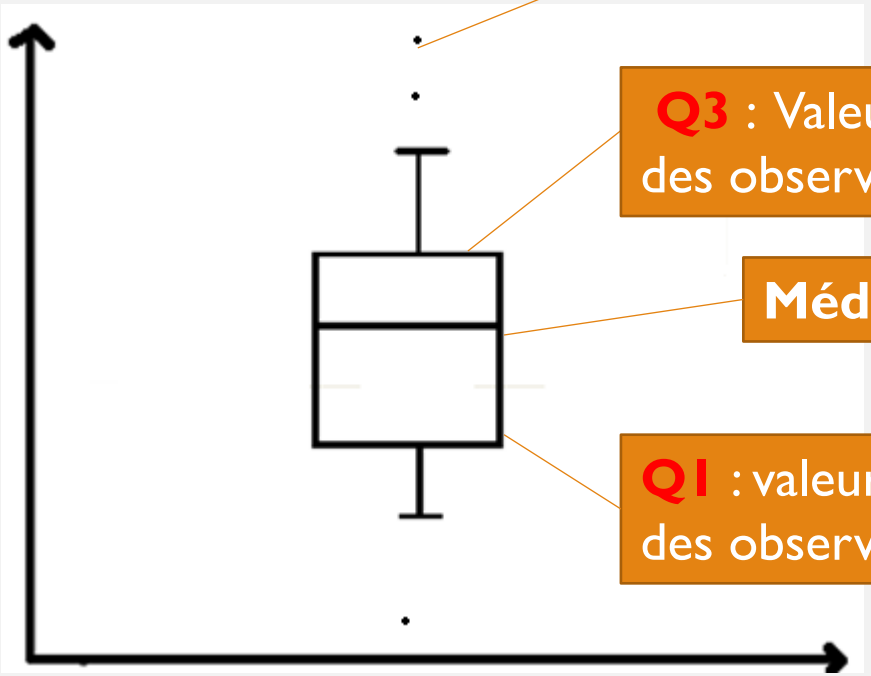
Caractéristique de tendance centrale :
- Moyenne
- Médiane
- mode

Dispersion
- Étendue (max-min)
- Écart type

Forme
- Quartile (Q1,Q2,Q3)
- Indice symétrique (skewness)
- Indice d'applatissement (kurtosis)

La boîte à moustache est le résumé graphique de la distribution

COMMENT LIRE LA BOÎTE À MOUSTACHE



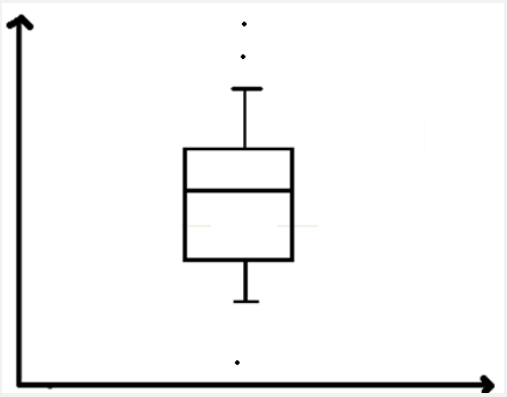
Outlier (valeur atypique)

Q3 : Valeur en dessous de laquelle se trouve 75% des observations

Médiane

Q1 : valeur en dessous de laquelle se trouve 25% des observations

COMMENT INTERPRÉTER LA BOÎTE À MOUSTACHE



- **La dispersion des données** se définit par la longueur de la boite à moustache
- **L'asymétrie (de la distribution)**correspond à la déviation de la ligne médiane (Q2) du centre de la boîte à moustaches par rapport à la longueur de la boîte,
- **L'asymétrie des moustaches** (moustache plus longue que l'autre)
- **Présence des point atypiques (Outliers)**

RÉGRESSION LINÉAIRE (SIMPLE/MULTIPLE)

LA RÉGRESSION LINÉAIRE ?

- C’est un modèle qui cherche à établir une relation linéaire entre 2 variables

- Exemple : Trouver une équation linéaire qui va prédire le prix d’une voiture en fonction du modèle et de l’année grâce à un échantillon que je possède au préalable,

Avant d'entamer tout ce travail

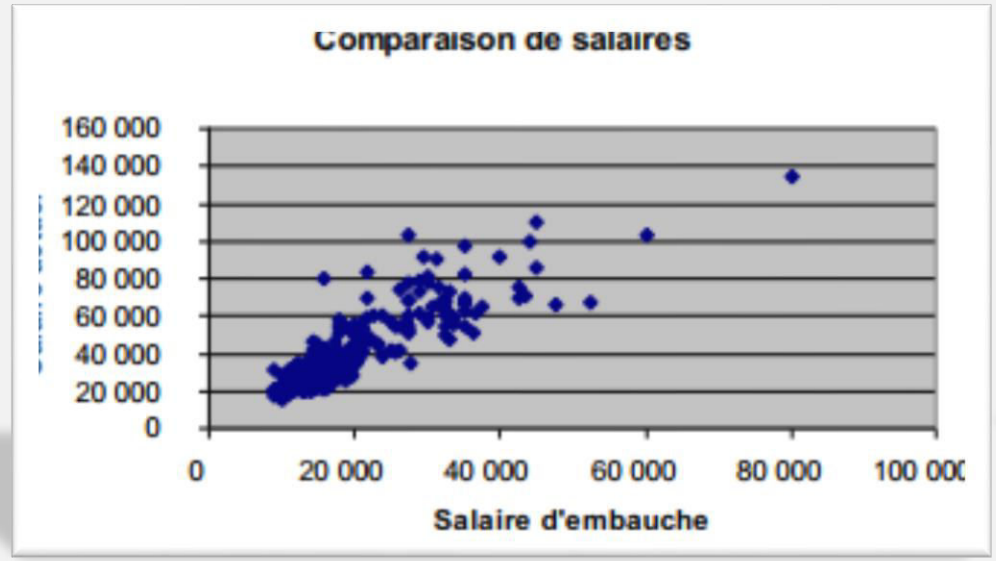
Y-a-il déjà une relation entre ma(mes) variable(s) exogènes et ma variable endogène ?!

Variables explicatives

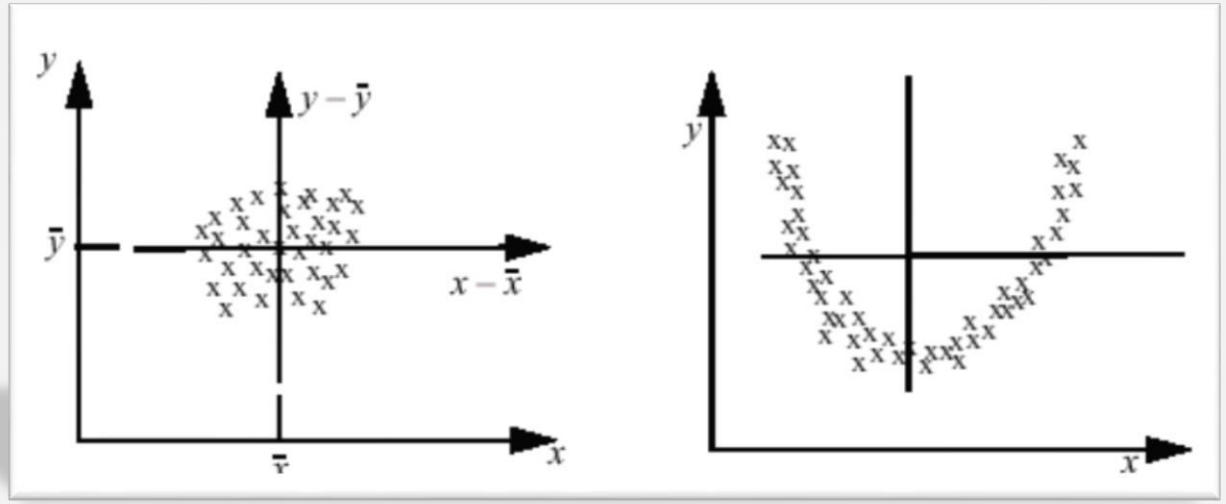
Variable que je cherche à prédire

Solution

- Calculer le coefficient de corrélation (ou notre cas le lire dans la matrice de corrélation)
- Observer graphiquement (Nuage des points)



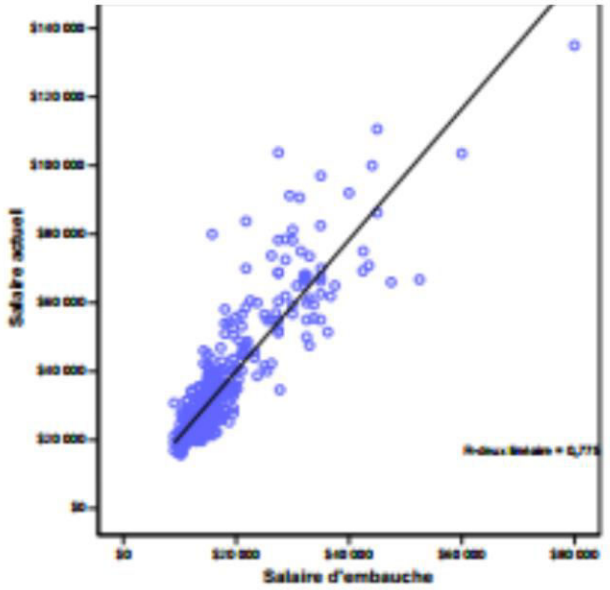
Il existe une liaison linéaire –> je peux tenter une régression linéaire



Pas de liaison linéaire

Méthode pas trop fiable dans les cas limites (On ne saura pas distinguer à l'œil)

Le coefficient de corrélation est plus adapté !



Matrice de corrélation

	SALDEB	*SALACT*
SALDEB	1	
SALACT	0,88	1

Coefficient de corrélation dans l'exemple précédent = 0,88 révèle une forte liaison linéaire entre le salaire actuel et le salaire de début

Plus la valeur absolue du coefficient de corrélation est proche de 1 , plus il y a une forte liaison linéaire

« Y-a-t-il une liaison linéaire entre la variable X et Y ? »

➡ Je vérifie la matrice de corrélation

MODÈLE DE RÉGRESSION

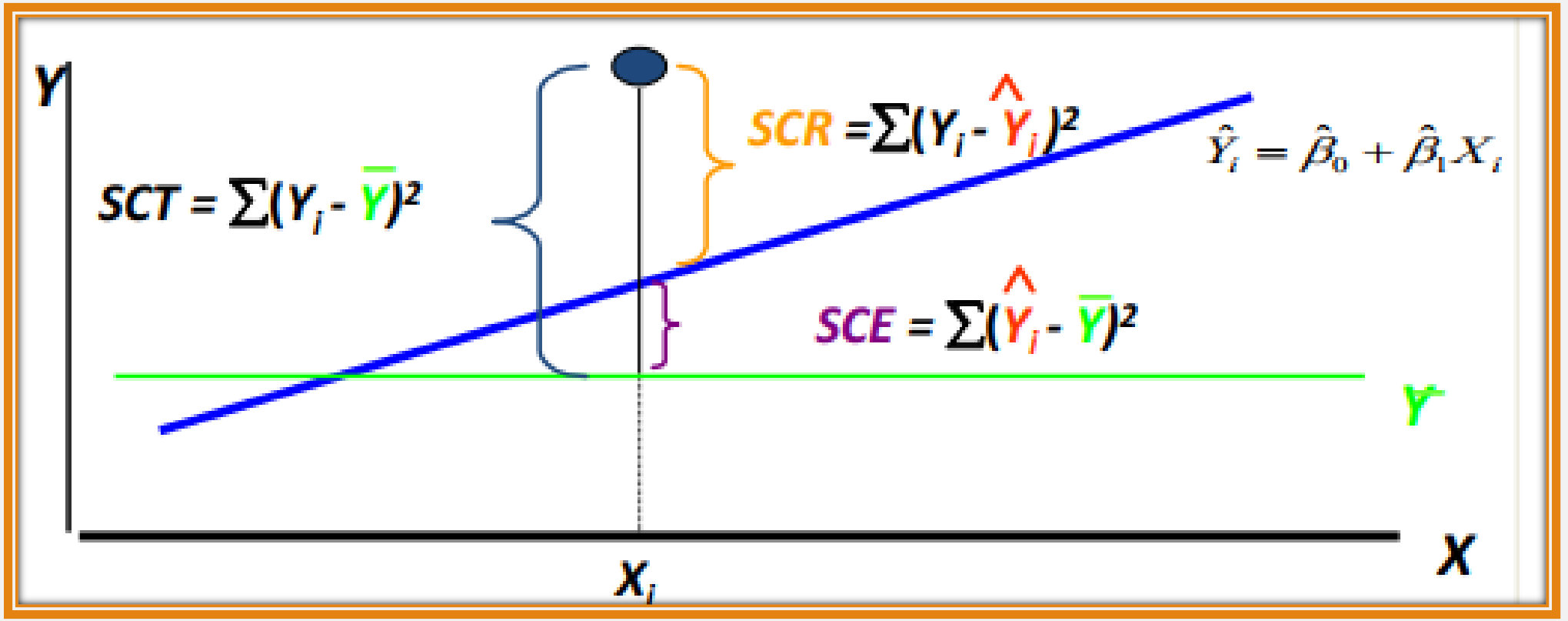
- Modéliser cette relation par une équation

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

- βi sont les coefficient que l'on va chercher
- Epsilon est l'erreur (ce que notre équation n' pas pris en compte)

NB : c'est pas à nous de calculer ces coefficients

QUALITÉ D'UN MODÈLE (I)



SCT : ∑ des infos disponibles dans les données de l'éch.

SCE : ∑ des infos que notre modèle a expliqué

SCR : ∑ des infos résiduels que notre modèle n'a pas expliqué

Ainsi

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

Coefficient de détermination :

(il indique la proportion des infos de l'échantillon que notre modèle a pu expliqué)

QUALITÉ D'UN MODÈLE (2)

R^2 (étant la proportion expliquée), donc bien évidemment :

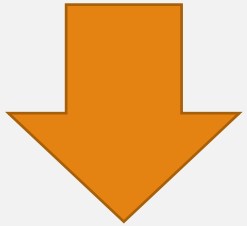
- $0<R^2<1$
- $R^2=1$ correspond au modèle idéal
- Plus notre R^2 tends vers 1 , plus notre modèle est bien

Récapitulatif des modèles[b]

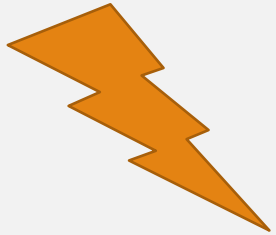
Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation	Durbin-Watson
1	,916[a]	,839	,806	8,347	1,696

Le *R2-Ajusté*, est davantage utilisé que le R^2 car il ne dépend
pas du nombre de variables

- Supposons que notre modèle est bien (R=0,87 par exemple), Est-ce que c'est suffisant ?



NON, ce qu'on a prouvé jusqu'à mtn c'est que le modèle représente bien les données que j'ai déjà.



On dit chercher si mon modèle est SIGNIFIQUATIVEMENT GLOBAL (il est bon également pour les nouvelles valeurs)

ÉVALUATION GLOBALE DE LA RÉGRESSION

$$F = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}}$$

On effectue un test d'hypothèse H_0 : Notre modèle n'est pas significativement global

2 méthodes pour rejeter l'hypothèse

Comparer F avec la loi de Fisher correspondant

→ F > $F_{1-\alpha}(1, n-2)$ $\mathcal{F}(1, n-2)$

→ Modèle significativement global !

Comparer la p-value avec le risque (alpha)

→ p-value < alpha

→ Modèle significativement global !

la probabilité critique (p-value) : *probabilité que la loi de Fisher dépasse la* statistique calculée F

Tableau d'ANOVA (parfois F est appelé D)

ANOVA[b]

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	12673,101	7	1810,443	25,985	,000[a]
	Résidu	2438,527	35	69,672		
	Total	15111,628	42			

La p-value est appelé dans les tableaux Sig (significance)

Généralement, c'est la Sig qu'on compare avec le risque qu'elle nous donne

Il faut aussi s'assurer des Coefficient de notre équation (βi)

Dans un premier temps, on obtient une équation, mais il faut chercher si le coefficient vaut vraiment la valeur qu'on a trouvé ou sinon c'est 0

Coefficients[a]

Modèle	Coefficients non standardisés		Coefficients standardisés	t	Sig.
	A	Erreur standard	Bêta		
constante	13,951	8,309		1,679	,102
Var 1	6,575	1,455	,424	4,520	,000
Var2	9,216	1,959	,389	4,705	,000

MÊME BOULOT !

Notre Hypothèse Ho : c'est le coeff d'une variable=0
-> Pour la rejeter (et donc garder le terme) :
Il faut que la Sig < alpha (risque)

IL Y'A 2 grands Courant concernant la cte (certains disent qu'il ne faut pas effectuer le test, et d'autre qui disent qu'il faut le faire),
Mme Benbrahim dit qu'elle préferer le faire !

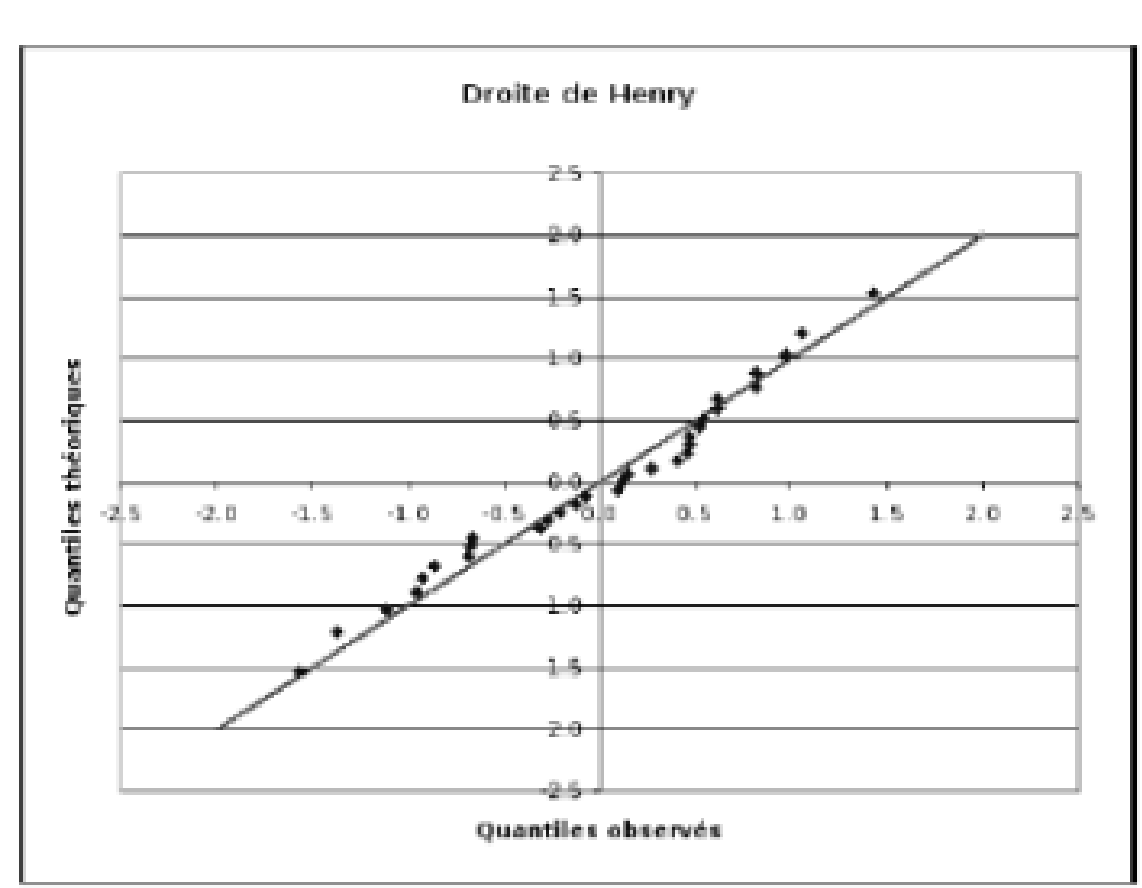
Les variables dont le coeff =0 alors ne figureront pas dans notre équation finale

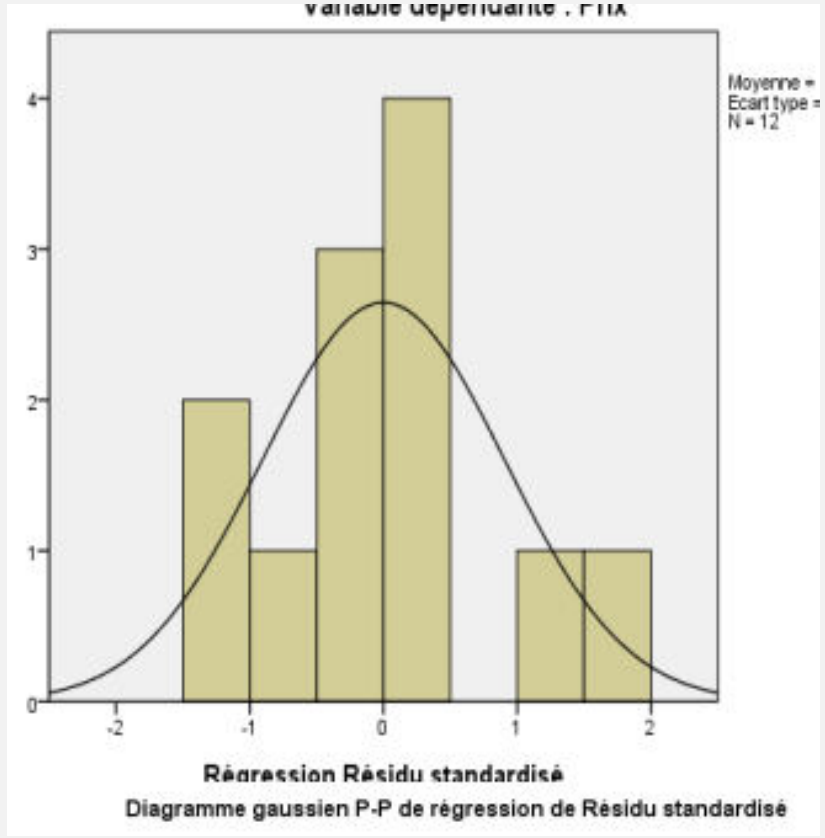
HYPOTHÈSES À VÉRIFIER

- Normalité des résidus
- Normalité de Y
- Homoscédasticité
- Non-auto corrélation

Attention ! Il faut les mentionner tous (soit la figure correspondante à ce test figure ou pas) !!!

NORMALITÉ DES RÉSIDUS

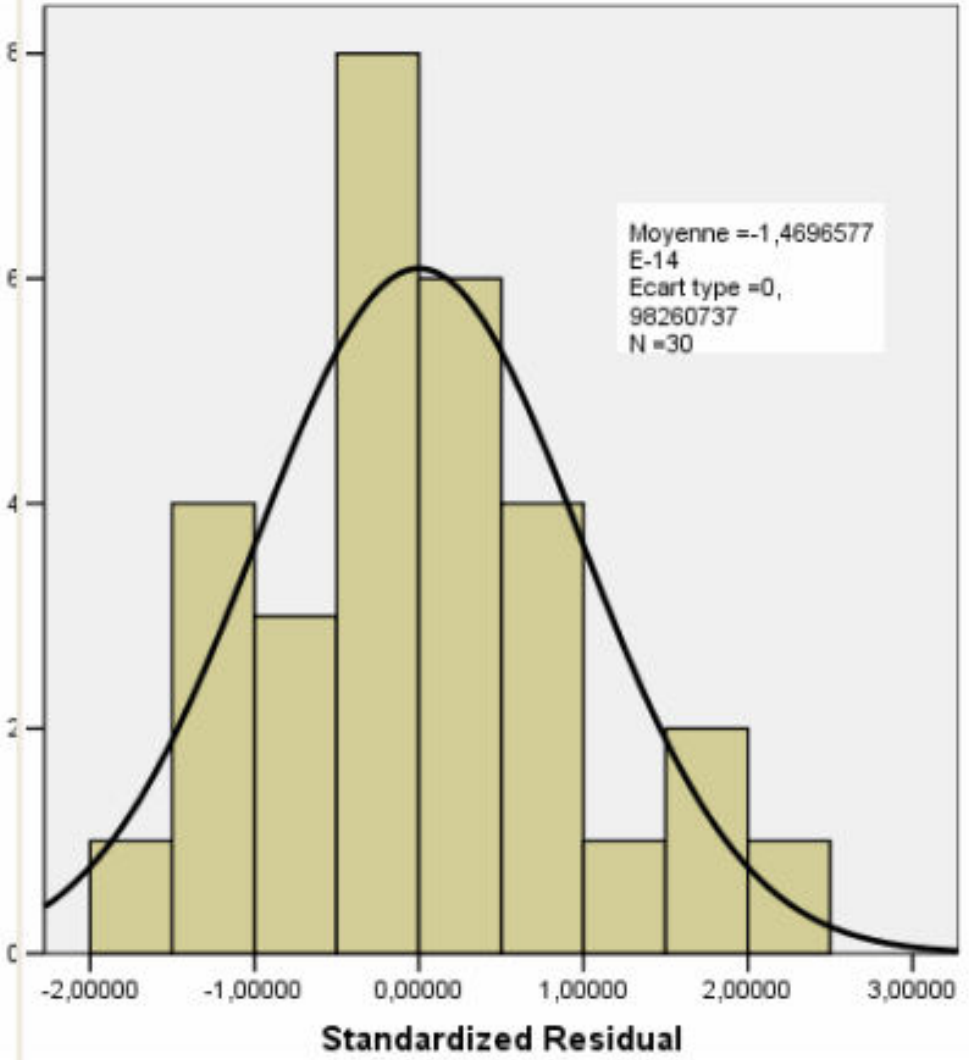


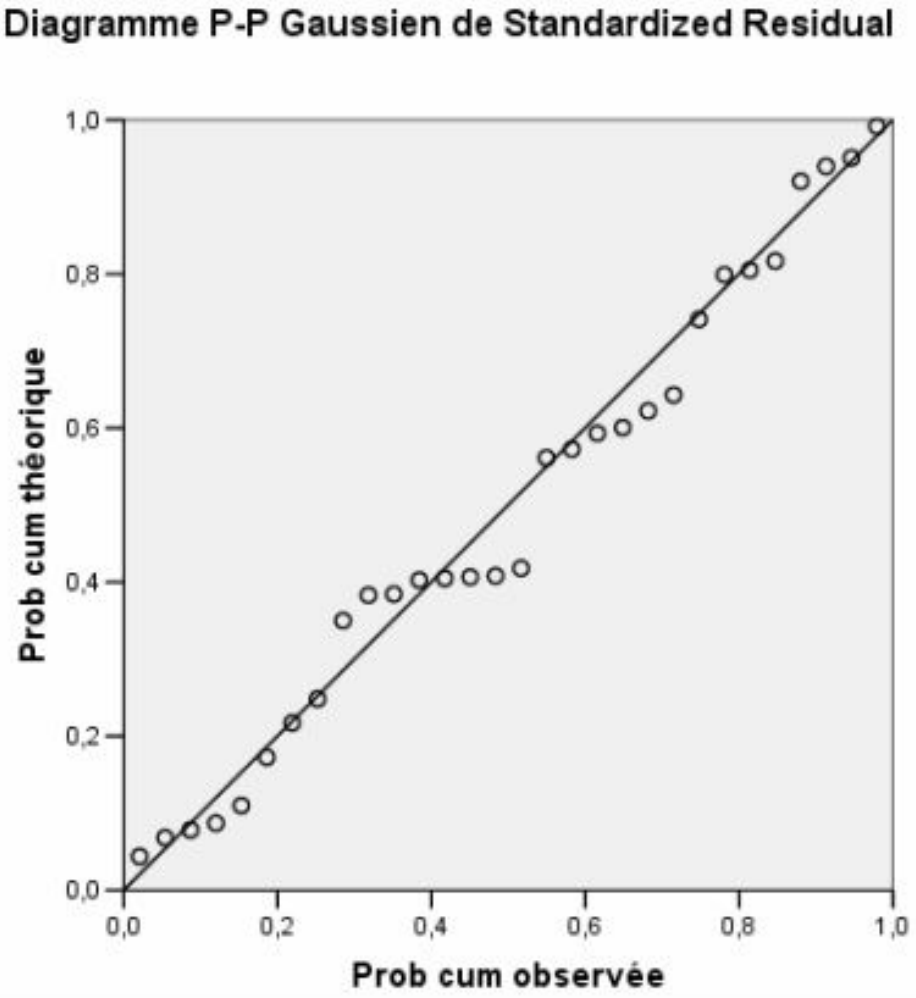


La distribution des résidus est proche de la droite
(vu que plus les résidus se rapprochent de la droite, plus on dit que leur distrib est normale)

D'après le Q-Q plot la distribution est normale

Distribution des résidus





Là encore l'hypothèse de normalité peut être retenue.

NORMALITÉ DE Y

Normalité de Y?

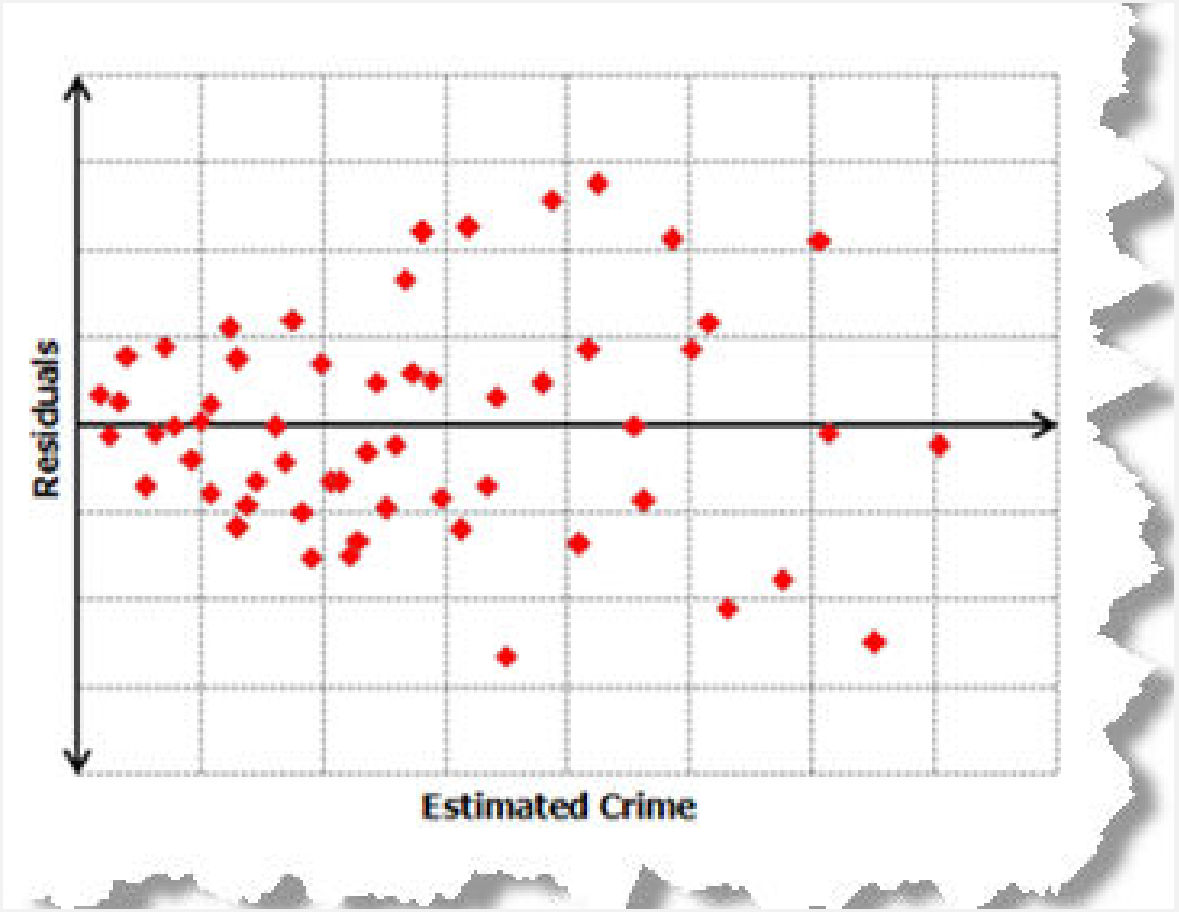
- Méthode graphique:
  - ✓ Histogramme
  - ✓ P-P plot

- Vérifier si Skewness AS ≈ 0
- Vérifier si Kurtosis AP ≈ 3
- Test de Kolmogorov – Smirnov

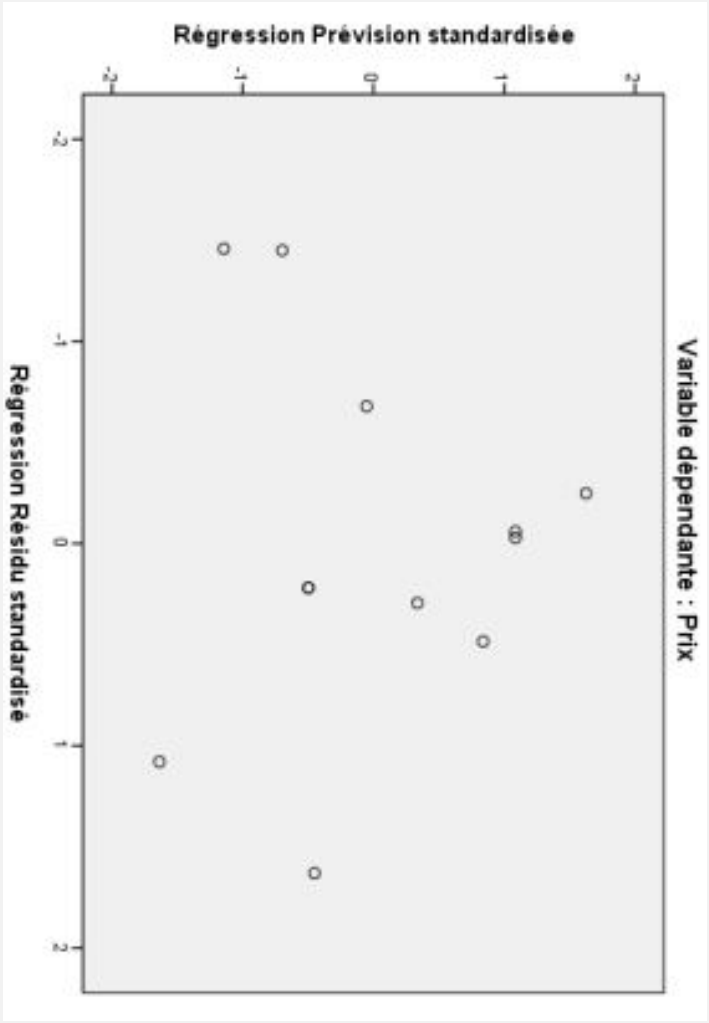
$E(\varepsilon i) = 0$

$V(\varepsilon i) = \sigma 2$

HOMOSCÉDASTICITÉ



Le nuage de points en forme de cône indique que le modèle ne vérifie pas l'homoscédasticité



Nuage de point ne suit aucun pattern : l'homoscédasticité est vérifiée

NON AUTOCORRÉLATION

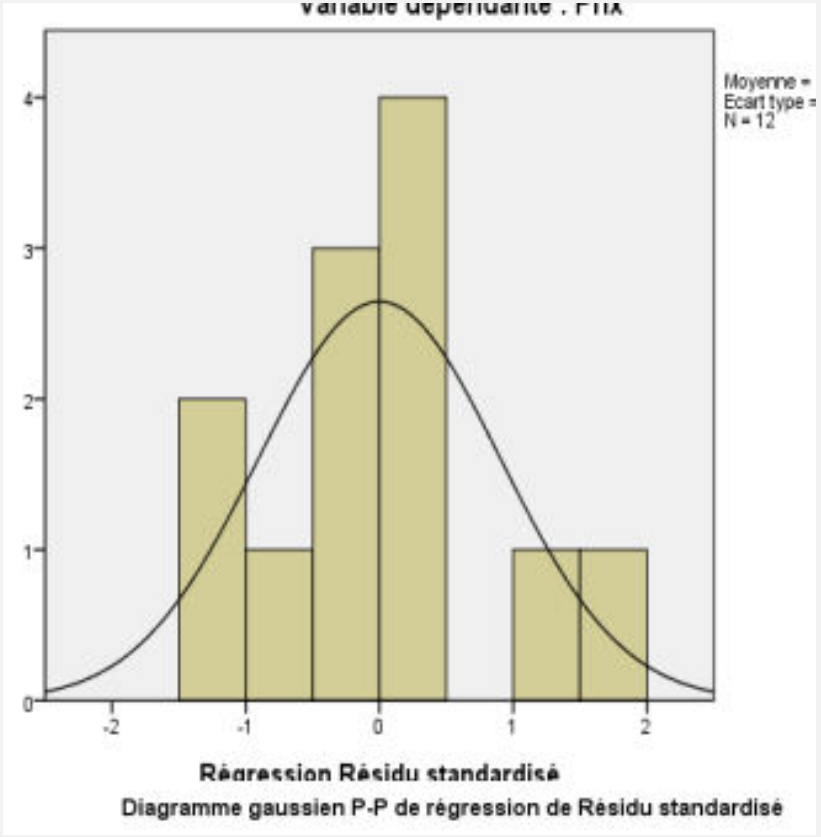
→Coefficient de Durbin-Watson (doit être proche de 2) pour que la NON-autocorrélation soit vérifiée.

1,625 est proche de 2, donc c'est vérifié !

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation	Durbin-Watson
1	,989[a]	,977	,956	2,2868	1,625

- Mais bon, là encore il faut vérifier les hypothèses suivantes :

Normalité des Résidus (P-P Plot ou Q-Q Plot) :



Test de Normalité : P-P plot

Test de Normalité : Q-Q plot

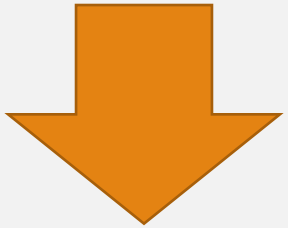
(

- Normalité des Y :

  - Graphiquement, comme pour les résiduels
  - Vérifier si Skewness AS ≈ 0
  - Vérifier si Kurtosis AP ≈ 3

- Autoscédascticité
- Auto-corrélation

Maintenant, vous remplacez les valeurs des coefficient des variables que vous avez retenu dans l'équation

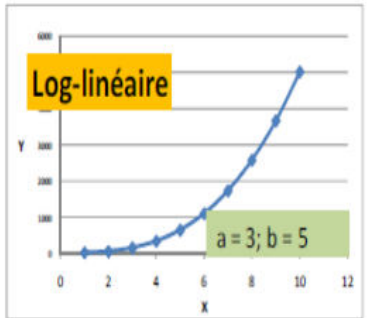


Votre équation est prête pour prédire de nouvelles valeurs !!

Cas où notre équation n'est pas linéaire (mais on peut la linéariser)

Modèle log-linéaire

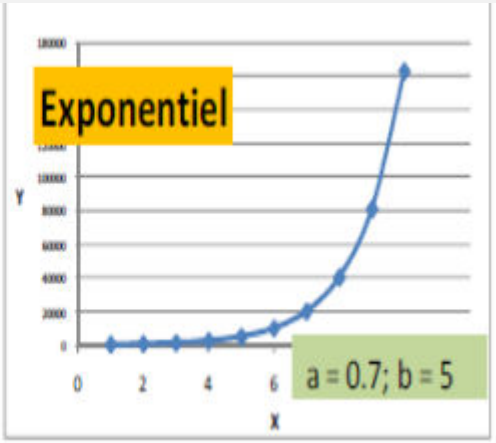
$$Y = bX^a$$



→ Linéarisation : ln(y) = a ln(x) + ln(b)

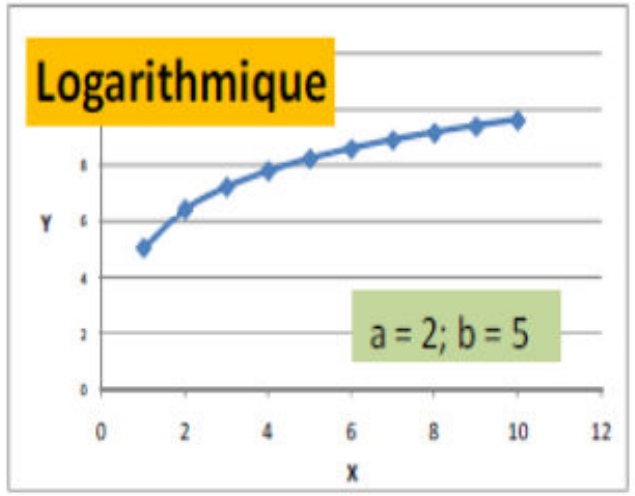
Modèle exponentiel

$$Y = e^{aX+b}$$



→ Linéarisation : ln(y) = a x + ln(b)

Modèle logarithmique



$$Y = a\ln(X) + b$$

Pas besoin de linéarisation, il suffit de considérer X'=ln(X) la nvlle variable

ACP

DÉFINITION DE L'ACP

But de l'ACP → Trouver de nouvelles structures (classes) dans nos données

Comment ? → Graphiquement, en projettant les données sur un espace à dimension plus réduite

PRINCIPE*

- **Certaines variables ont une part d'information commune entre elles, donc on va essayer de réduire le nbre de ces variables pour pouvoir se contenter d'un nbre d'axe réduit sur lequel on va projeter nos points.**

Par exemple si les Variables sont totalement indépendantes, ACP n'est pas faisable

* d'une façon très très très très vague, c'est de l'algèbre linéaire approfondi …

ÉTAPES D'UNE ACP

1. Centrer et Réduire :

- Si on n'a pas la même grandeur ⇨ Il faut centrer
- Si on n'a pas la même unité ⇨ Il faut réduire

ÉTAPES D'UNE ACP(2)

2. Vérifier si l'ACP est pertinente :

➡ En utilisant la matrice de corrélation, dire que comme quoi il y'a pas mal de coefficient élevés proche de 1,

➡ En utilisant le test de KMO (si KMO est proche de 0,7 c'est bon, proche de 0 pas la peine)

Indice KMO et test de Bartlett

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.		,780
Test de sphéricité de Bartlett	Khi-deux approximé	240,823
	ddl	55
	Signification de Bartlett	,000

En utilisant la matrice anti-image

	CYLINDRE	LARGEUR	LONGUEUR	POIDS	PUISSANC	VITESSE
CYLINDRE	.93403					
LARGEUR	.00453	.66050				
LONGUEUR	-.26598	-.76061	.65758			
POIDS	-.32958	.39762	-.59316	.65060		
PUISSANC	-.21941	-.41220	.53269	-.65901	.62545	
VITESSE	-.06258	.46028	-.52436	.70642	-.89849	.52070

Measures of Sampling Adequacy (MSA) are printed on the diagonal.

Plus msa_i est élevé et proche de 1, plus la variable correspondante contribue fortement dans la construction des facteurs.

En utilisant le test de sphéricité de Barlett

Le But du test est de prouver que la matrice de corrélation est différente de la matrice d d'identité

Comme pour les autres tests d'identité, l'hypothèse est que la matrice des corrélations est égale à la matrice identité.

Indice KMO et test de Bartlett

Mesure de précision de l'échantillonnage de Kaiser-Meyer-Olkin.		,498
Test de sphéricité de Bartlett	Khi-deux approximé	41,782
	ddl	21
	Signification de Bartlett	,004

Comparer la sig avc le risque alpha

3. Trouver les facteurs qu'on doit garder :

Après avoir fait des calculs, on a obtenu un certain nombre de facteurs (axes), on doit choisir le nombre qui convient

Pour ça, on va prendre en compte 4 critère :

- % d'informations obtenu par un facteur
- Diagramme des valeurs propres
- Corrélation reproduites
- Qualité de représentation

- % d'informations obtenu par un facteur
  (pourcentage d'inertie)

Variance totale expliquée

Composante	Valeurs propres initiales		
	Total	% de la variance	% cumulés
1	7,451	67,735	67,735
2	1,098	9,985	77,720
3	,883	8,025	85,745
4	,505	4,591	90,336
5	,337	3,067	93,404
6	,271	2,462	95,866
7	,173	1,575	97,440
8	,140	1,272	98,713
9	,081	,738	99,451
10	,045	,408	99,858
11	,016	,142	100,000

Méthode d'extraction : Analyse en composantes principales.

J'estime que 4 facteurs sont bien, j'ai déjà 90,336% et de toute façon le facteur 5 ne me va ajouter que 3,.

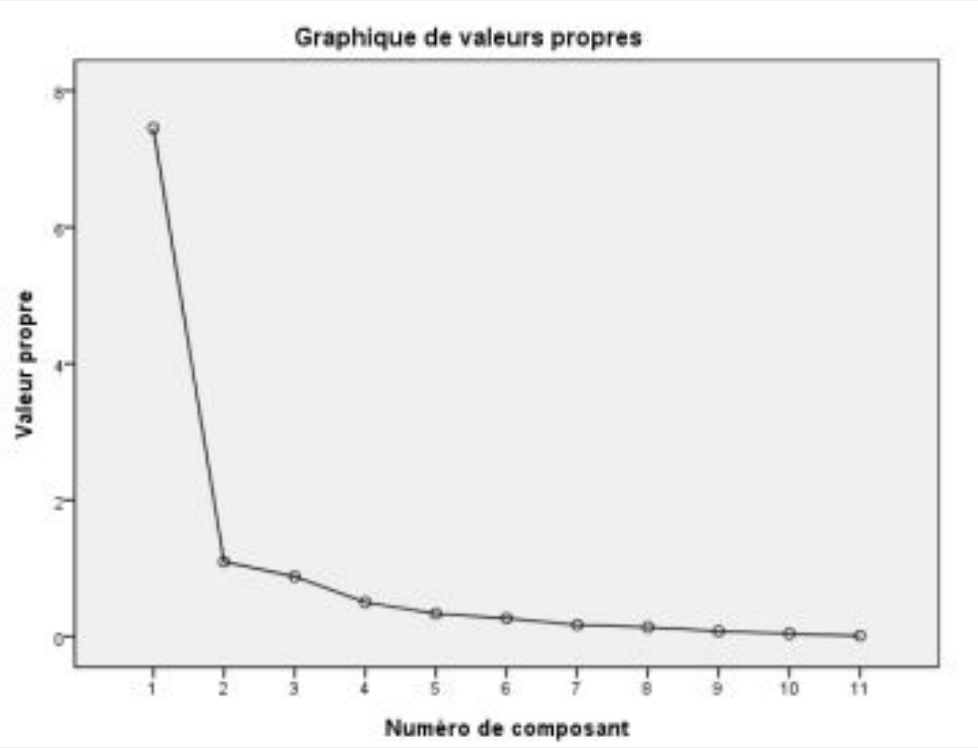
Diagramme des valeurs propres

Valeurs propres

Variance totale expliquée

Composante	Valeurs propres initiales		
	Total	% de la variance	% cumulés
1	7,451	67,735	67,735
2	1,098	9,985	77,720
3	,883	8,025	85,745
4	,505	4,591	90,336
5	,337	3,067	93,404
6	,271	2,462	95,866
7	,173	1,575	97,440
8	,140	1,272	98,713
9	,081	,738	99,451
10	,045	,408	99,858
11	,016	,142	100,000

Méthode d'extraction : Analyse en composantes principales.



Dire que la valeur propre doit être plus grande que 1, ou au moins proche de 1,

• Qualité de représentation

Qualité de représentation

	Initial	Extraction
Mass	1,000	,979
Fore	1,000	,918
Bicep	1,000	,934
Chest	1,000	,890
Neck	1,000	,921
Shoulder	1,000	,912
Waist	1,000	,958
Height	1,000	,976
Calf	1,000	,953
Thigh	1,000	,841
Head	1,000	,993

Méthode d'extraction : Analyse en composantes principales.

On a pu extraire un % considérable de chaque variable , donc c bien

Si on remarque que le premier axe est celui sur lequel les coefficients de corrélation sont les plus élevés., nous sommes en présence **d'un effet taille**

Et on doit envisager une rotation

Matrice des composantes[a]

	Composante			
	1	2	3	4
Mass	,956	,187	-,073	,099
Fore	,935	,014	-,090	,033
Bicep	,854	-,401	-,094	-,124
Chest	,890	-,279	-,014	-,141
Neck	,897	-,044	-,003	-,338
Shoulder	,900	,154	-,041	-,260
Waist	,870	,065	-,047	,090
Height	,429	,743	,474	-,113
Calf	,831	,139	-,033	,424
Thigh	,862	-,009	-,176	,245
Head	,391	-,473	,774	,129

Méthode d'extraction : Analyse en composantes principales.

Après rotation

Matrice des composantes après rotation[a]

	Composante			
	1	2	3	4
Mass	,578	,729	,311	,042
Fore	,652	,647	,172	,101
Bicep	,801	,424	-,151	,265
Chest	,792	,434	-,001	,271
Neck	,861	,329	,237	,123
Shoulder	,770	,423	,362	,006
Waist	,549	,644	,208	,108
Height	,105	,194	,959	,066
Calf	,288	,866	,197	,139
Thigh	,486	,769	,050	,069
Head	,181	,101	,069	,972

INTERPRÉTATION

- On utilise la matrice des composantes

Matrice des composantes après rotation[a]

	Composante			
	1	2	3	4
Mass	,578	,729	,311	,042
Fore	,652	,647	,172	,101
Bicep	,801	,424	-,151	,265
Chest	,792	,434	-,001	,271
Neck	,861	,329	,237	,123
Shoulder	,770	,423	,362	,006
Waist	,549	,644	,208	,108
Height	,105	,194	,959	,066
Calf	,288	,866	,197	,139
Thigh	,486	,769	,050	,069
Head	,181	,101	,069	,972

Pour chaque variable on détermine le max

Mass = Axe2

Fore = Axe1

Bicep= Axe1

Chest= Axe1

Neck= …

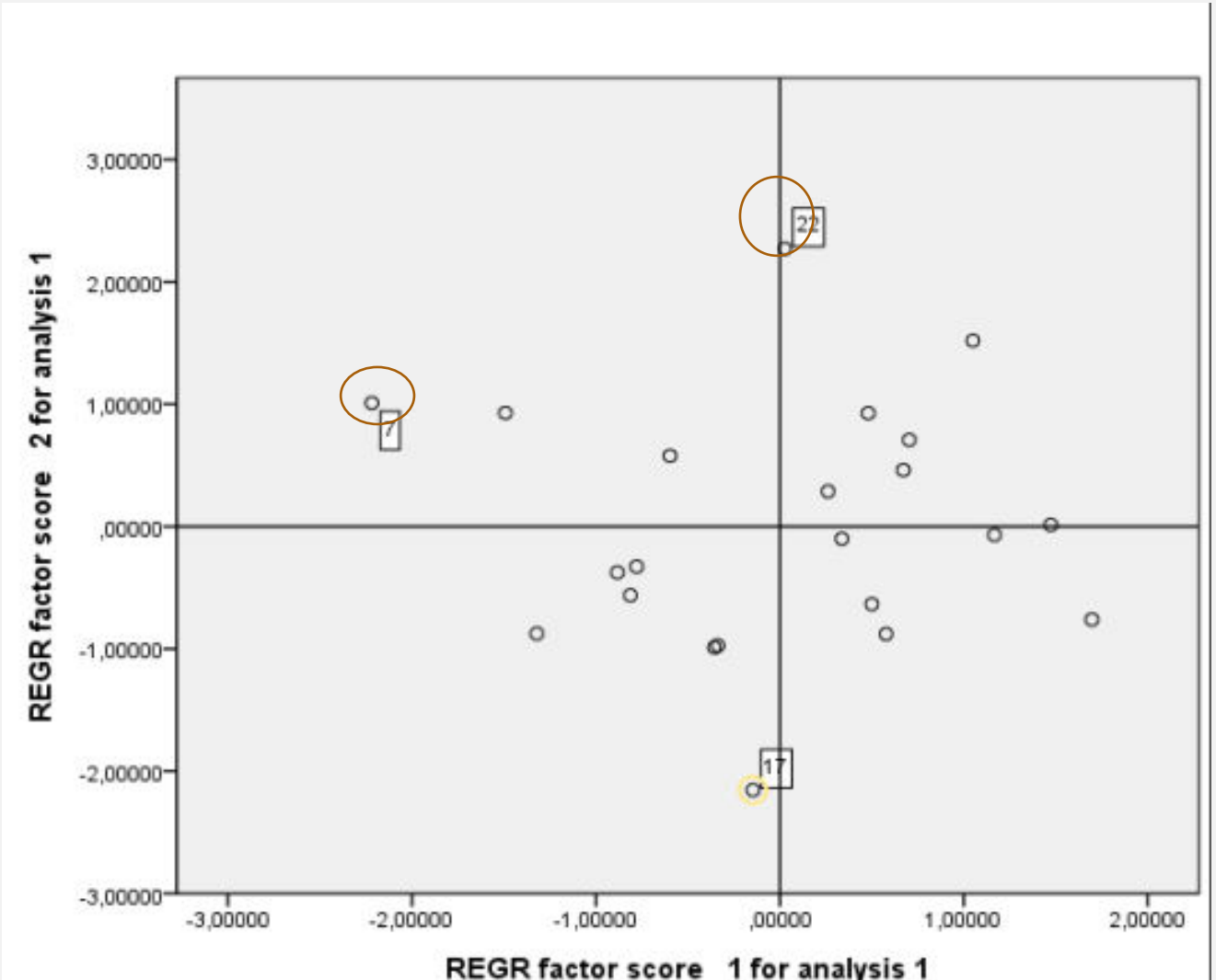
…Height=Axe3…

Head=Axe4

-> le 1er axe est représenté par Fore, Bicep, Chest

-> On remarque que Head et Height sont toutes seules représenté par axe3 et 4, donc elle est indépendantes des autres

- Généralement, les points qu'elle demande de décrire sont des outliers (points atypiques)



§ FIN §

BONNE CHANCE !