

Analyse de données

→ Description des données :

Comment décrire mes données ?

je calcule les indicateurs
statistiques

Caractéristique de
tendance centrales:

- Moyenne
- Médiane
- Mode → la valeur du caractère la plus fréquente

Dispersion

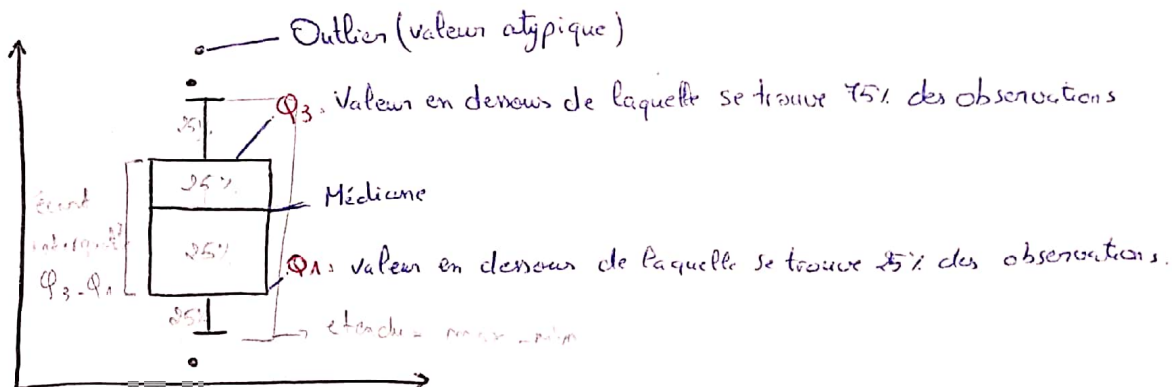
- Étendue (max - min)
- Écart type → $CV = \frac{\sigma}{m}$
- indice de variabilité
- $CV < 0,15$ n'est pas significatif
la moyenne est suffisante

Forme

- Quartile (Q_1, Q_2, Q_3)
- Indice symétrique (skewness)
- Indice d'aplatissement (kurtosis)

⇒ La boîte à moustache est le résumé graphique de la distribution.

Comment lire la boîte à moustache ?



Comment interpréter la boîte à moustache ?

- La dispersion des données : se définit par la longueur de la boîte à moustache.
- L'asymétrie (de la distribution) correspond à la déviation de la ligne médiane (Q_2) 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 points atypiques (Outliers)

→ Régression linéaire (simple / multiple)

C'est un modèle qui cherche à établir une relation linéaire entre 2 variables.

Exemple: Trouver une eqt 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:

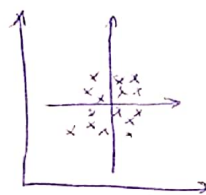
Qst Y a-t-il déjà une relation entre ma (mes) variable(s) exogènes et ma variable endogène ?
↓
(variables explicatives) variable que je cherche à prédire

Solution: 1. Calculer le coeff de corrélation
(ou dans ce cas le lire dans la matrice de corrélation)
2. Observer graphiquement (Nuage des pts)

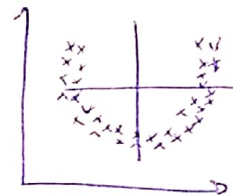
②



Il existe une liaison linéaire
→ je peux tenter une régression linéaire



pas de liaison linéaire



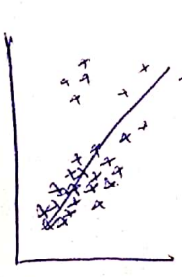
Problème Méthode pas trop fiable dans les cas limites (On ne saura pas distinguer à l'œil)
→ Le coeff de corrélation est plus adapté!

①

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

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

⇒ Je vérifie la matrice de corrélation.



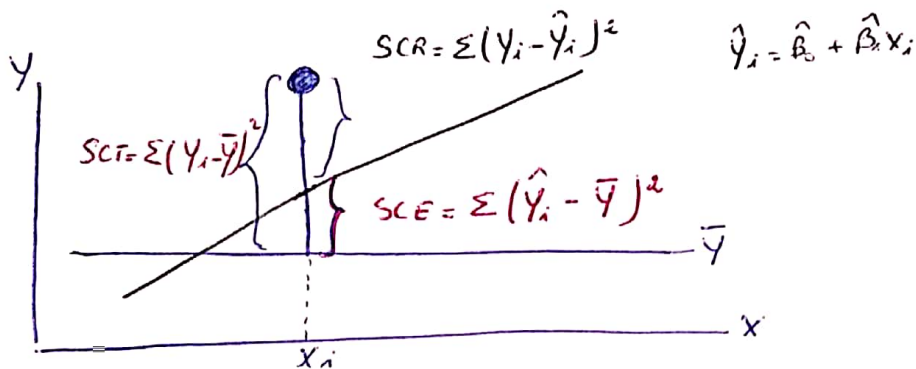
	SAL DEB	SAL ACT
SAL DEB	1	
SAL ACT	0,88	1

• coeff de corrélation ds l'exemple précédent = 0,88
révèle une forte liaison linéaire entre le salaire actuel et le salaire de début.

Modèle de régression

- Modéliser cette relation par une éq: $y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$
- β_i sont les coeff que l'on va chercher.
- ε est l'erreur (ce que notre éq n'a pas pris en compte)
- NB: c'est pas à nous de calculer ces coeff

→ Qualité d'un modèle



SCT: \sum des infos disponibles dans les données de l'éch
SCE: \sum des infos que notre modèle a expliqué
SCR: \sum des infos résiduelles que notre modèle n'a pas expliqué.

Ainsi: $R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \Rightarrow$ coef de détermination: il indique la proportion des infos de l'échantillon que notre modèle a pu expliquer.

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

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

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'esti	Dubin-Watson
1	,916	,839	,806	8,347	1,696

Récapitulatif des modèles

\Rightarrow Le R^2 - 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 significativement 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.

Comparer F avec la loi de Fisher correspondant.

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

$$F(1, n-2)$$

⇒ Modèle significativement global!

Comparer la p -valeur avec le risque α

$$\rightarrow p\text{-valeur} < \alpha$$

⇒ Modèle signif global!

→ probabilité critique : proba que la loi de Fisher dépasse la statistique calculée F .

Tableau d'Anova

modèle	Somme des carrés	ddl	Moy des carrés	D	Sig
1 Régress	12 673,101	7	1810,463	25,985	,000
Résidu	2 438,527	35	69,672		
Total	15 111,628	42			

F

→ p -valeur

⇒ C'est la Sig qu'on compare avec le risque qu'elle nous donne

Il faut aussi s'assurer des coeff de notre éq (β_i)



Dans un premier temps, on obtient une éq, mais il faut chercher si le coeff veut vraiment la valeur qu'on a trouvée ou sinon c'est 0.

Modèle	Coeff non standardisés		Coeff standard	t	sig
	A	Erreur standard	β		
cst	13,951	8,309		1,679	,102
var.1	6,575	1,455	,424	4,520	,000
var.2	9,216	1,959	,389	4,705	,000

Même Boulot,

Notre hypothèse : H_0 : c'est le coeff d'une variable = 0

→ pour la rejeter (et de garder le terme),

il faut que $\text{Sig} < \alpha$ (risque)

Note : Il y a 2 grands courants concernant la cte (certains disent qu'il ne faut pas effectuer le test et d'autres qui disent qu'il faut le faire)

→ Les variables dont le coeff = 0 alors ne figureront pas dans notre éq

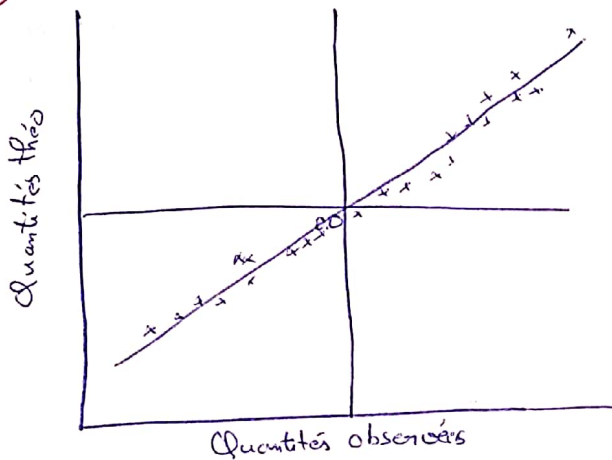
Hypothèses à vérifier

- 1 \hookrightarrow Normalité des résidus
- 2 \hookrightarrow Normalité de y
- 3 \hookrightarrow Homoscédasticité
- 4 \hookrightarrow Non-auto corrélation

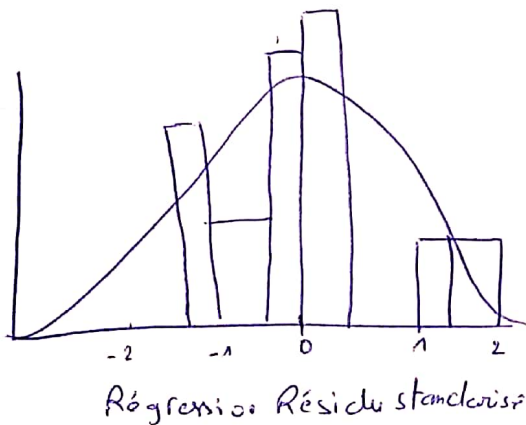
⚠ Attention, il faut les mentionner tous !!!

①

Droite de Henry



\Rightarrow 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)



\Rightarrow D'après qq plot la distribution est normale

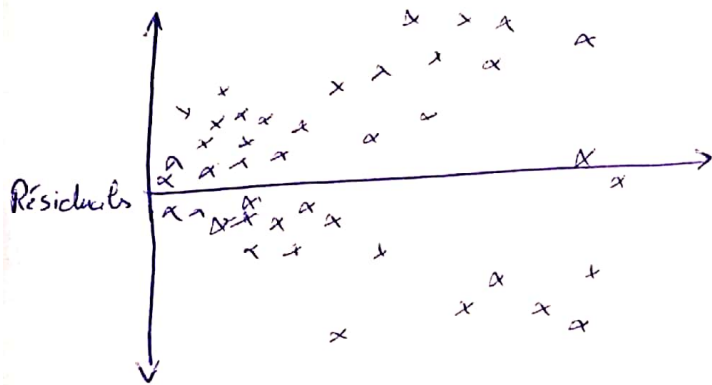
\hookrightarrow L'hypothèse de normalité est retenue

- Métho de graphique : + Histogramme
+ P-P plot
- vérifier si Skewness $AS \approx 0$
- vérifier si Kurtosis $AP \approx 3$
- Test de Kolmogorov - Smirnov

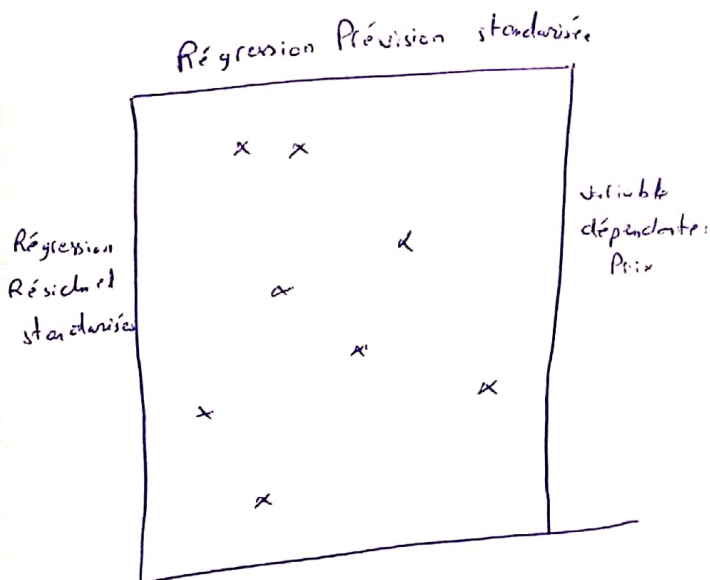
③

$$E(\varepsilon_i) = 0$$

$$V(\varepsilon_i) = \sigma^2$$



⇒ le nuage de pts en forme de cône indique que le modèle ne vérifie pas l'homoscédasticité



⇒ Nuage de pt ne suit aucun pattern : l'homoscédasticité est vérifiée.

④

⇒ Coef de Durbin-Watson. (doit être proche de 2) pour que la Non-collinéarité soit vérifiée

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

proche de 2
de vérifié.

↳ Mais bon, là encore il faut vérifier les hypothèses suivantes:
⇒ Normalité des Résidus (P-P plot ou Q-Q plot).



Maintenant, vous remplacez les valeurs des Coeff des variables que vous avez retenu dans l'éq



Votre eq est prête à prédire de nouvelles valeurs!!

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

1) Modèle log-linéaire

$$\ln(y) = a \ln(x) + \ln(b) \hookrightarrow y = bx^a$$

2) Modèle exponentiel

$$\ln(y) = ax + \ln(b) \hookrightarrow y = e^{ax+b}$$

3) Modèle logarithmique.

$$y = a \ln(x) + b \rightarrow \text{pas besoin de linéarisation, il suffit de } X' = \ln(x)$$

ACP

But de l'ACP \rightarrow Trouver de nouvelles structures (classes) de nos données

Comment ? \rightarrow Graphiquement, en projetant les données sur un espace à dimension plus réduite

- Certaines variables ont une part d'info commune entre elles, donc on va essayer de réduire le nombre de ces variables pour pouvoir se concentrer d'un nombre d'axe réduit sur lequel on va projeter nos pts.
- si les variables sont totalement indépendantes, ACP n'est pas faisable

1° Centrer et Réduire

- + Si on n'a pas la même grandeur \rightarrow il faut centrer
- + Si on n'a pas la même unité \rightarrow il faut réduire

2° Vérifier si l'ACP est pertinente

- ① \rightarrow En utilisant la matrice de corrélation, dire comme quoi il y'a pas mal de coeff élevés ≈ 1
- ② \rightarrow Le test de KMO (si $KMO \approx 0,7$ c'est bon,

KMO	,780
Test de sphéricité	Sig ,000

indice KMO et
test de Bartlett

→ En utilisant la matrice anti-image

	Cylind	Longe	Long	Poids	...
Cylind	1,93603				
Longe		,66050			
Long			,65060		
Poids				0,62515	
T					

→ plus msa_i est élevé et ≈ 1 , plus la variable correspondante contribue fortement à la construction des facteurs

④ → En utilisant le test de sphéricité de Bartlett

• Le but du test est de prouver que la matrice de corrélation est diff. de la matrice d'identité

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

Comparer sig avec α

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 critères :

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

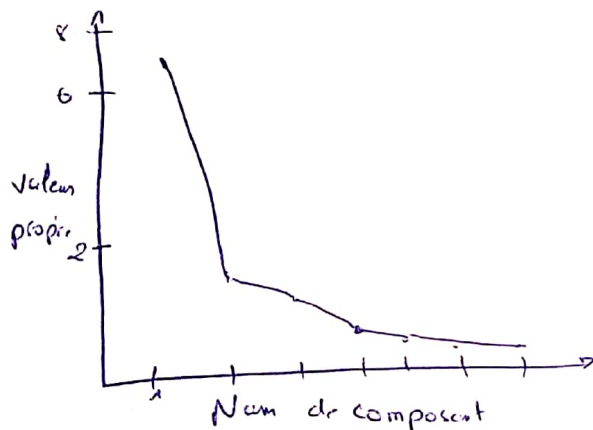
① % d'infos obtenu par un facteur (pourcentage d'inertie)

Valeurs propres

Composant	Valeurs propres initiales		
	Total	% de l'inertie	% cumulés
1	7,451	67,735	67,735
2	1,098	9,985	77,720
3	,883	8,020	85,740
4	,505	4,591	90,331
5			93,404
6			95,866
7			97,444
8			98,111

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

② Diagramme des valeurs propres.



⇒ dire que la valeur propre doit être plus grande que 1 ou au moins proche de 1

③ Qualité de représentation

	Initiale	Extraction
Mars	1,000	,979
Fore	1,000	,918
Biceps	1,000	,934
...

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

✓ \Rightarrow si on remarque qu'on peut réduire les axes (selon coeff de corrélation les plus élevés) nous ~~avons besoin~~ ^{avons besoin} d'une rotation

On utilise la matrice des composantes pour regrouper les éléments selon le max de chaque var