Statistiques et informatique

appliquées aux sciences sociales

Traiter des données numériques 2

Mardi 14/11/2023 15h30-17h Censier D2

Licence de sciences sociales 3e année
Université Paris 1 Panthéon Sorbonne

Plan de la séance

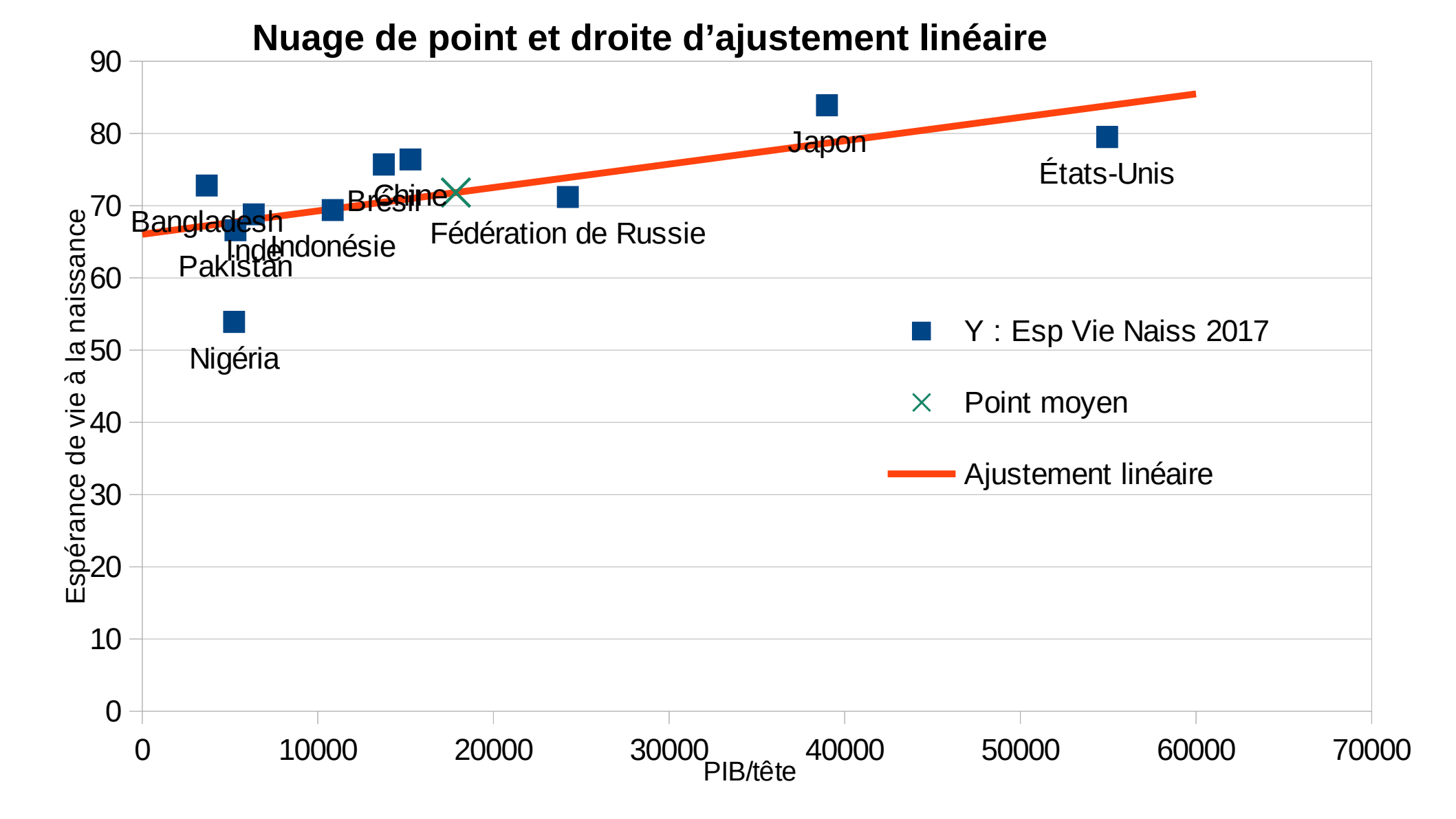
- Indicateurs du lien entre deux variables
- Droite de régression
- Variance des résidus et analyse de la variance
- Coefficient de détermination

Lien entre deux variables

- On se ramène en général à une liaison linéaire entre deux variables numériques
- Celle-ci est bâtie sur la covariance, moyenne des produits des écarts aux moyennes des deux variables
- La covariance est majorée, en valeur absolue, par le produit des écarts-types
- La corrélation entre les deux variables s'obtient en divisant la covariance par ce majorant
- Une corrélation proche de 0 traduit une absence de corrélation *linéaire* et en général un nuage de point relativement symétrique
- Une corrélation proche de 1 correspond à une relation parfaitement linéaire et en général un effet mécanique

Droite de régression

- Nuage de points
- Retour aux données
- Récapitulatif
- Interprétation des résultats
- Limite de la méthode


Nuage de point et droite d'ajustement linéaire
Espérance de vie à la naissance
PIB/tête
0
10
20
30
40
50
60
70
80
90
0
10000
20000
30000
40000
50000
60000
70000
Bangladesh
Pakistan
Inde
Indonésie
Nigéria
Brésil
Chine
Fédération de Russie
Japon
États-Unis
Y : Esp Vie Naiss 2017
Point moyen
Ajustement linéaire

Pays	X : PIB/tête (2017)	Y : Esp Vie Naiss 2017	U : Résidus
Nigéria	5231	53,9	-13,8267569948
Pakistan	5311	66,6	-1,1526855451
Bangladesh	3677	72,8	5,5769050949
Inde	6353	68,8	0,70959508708
Indonésie	10846	69,4	-0,14661711963
Chine	15270	76,4	5,4195340483
Brésil	13755	75,7	5,2105559697758
Fédération de Russie	24233	71,2	-2,6854359069
Japon	38986	83,9	5,2330153093
États-Unis	54941	79,5	-4,3381099429
Moyenne	17860,3	71,82	*0*
Variance(P)	258521646,61	60,5636	33,407124109
Covariance(avec X)	=Variance	83788,644	*0*
Écart-type	16078,6	7,7823	5,7798896278
Corrélation	1	0,6696	*0*

$$Y = a \cdot X + b + U$$

$$a = \frac{cov(X, Y)}{var(X)}$$

$$b = \bar{Y} - a \cdot \bar{X}$$

$$Var(Y) = a^2 \cdot Var(X) + Var(U)$$

$$R^2 = \frac{Variance\ expliquée}{Variance\ totale}$$

R²=44 % ici

Bases des régressions linéaires

- Principe général : évaluer comment un écart à la moyenne de la variable explicative se répercute sur un écart à la moyenne de la variable expliquée, en les supposant proportionnels
- Choix de la variable expliquée/explicative (traitement asymétrique)
- Notion de valeur prédite par le modèle linéaire
- Notion de résidus (ou erreurs) : différences entre valeurs observées et valeurs prédites
- Méthode des *moindres carrés (least squares)*: on minimise la moyenne des carrés des résidus
- Le coefficient directeur de la droite de régression s'obtient en divisant la covariance des deux variables par la variance de la variable expliquée
- L'ordonnée à l'origine s'obtient en vérifiant que les résidus sont de moyenne nulle et (de manière équivalente) que la droite passe par le *point moyen* du nuage
- Ceci conduit à une solution dans laquelle les résidus sont de moyenne nulle et indépendants linéairement de la variable explicative (c'est à dire que leur covariance avec l'explicative est nulle)
- Le modèle réalise une analyse de la variance de la variable expliquée en la décomposant en deux éléments : la variance expliquée par le modèle et la variance des résidus
- Pour résumer la qualité de l'ajustement, on utilise le *coefficient de détermination*
- On peut introduire plus d'une variable explicative.

$$Y = a \cdot X + b + U$$

$$a = \frac{cov(X, Y)}{var(X)}$$

$$b = \bar{Y} - a \cdot \bar{X}$$

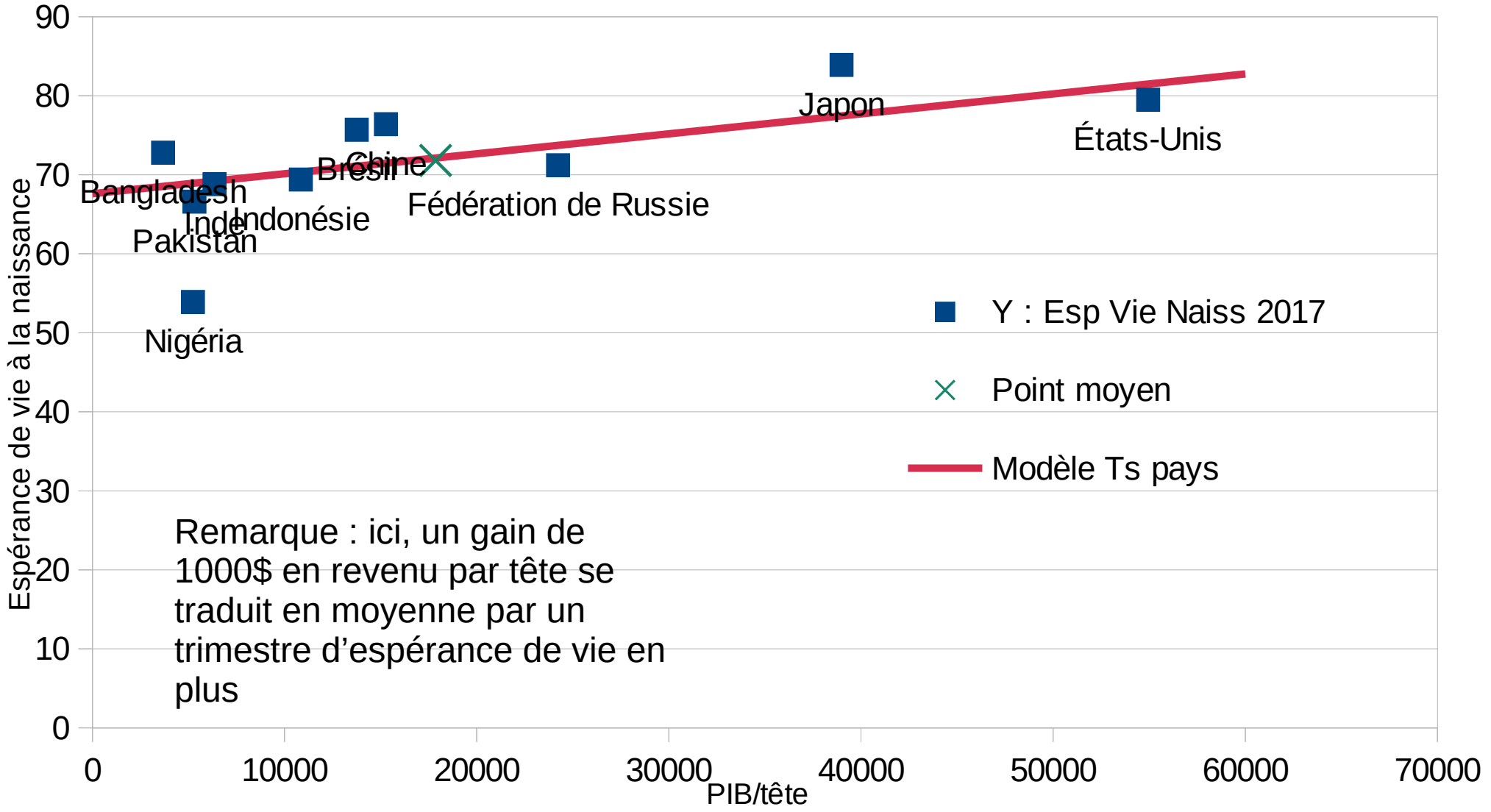
$$Var(Y) = a^2 \cdot Var(X) + Var(U)$$

$$R^2 = \frac{Variance\ expliquée}{Variance\ totale}$$

R²=44 % ici

Interprétation des résultats

- L’espérance de vie apparaît fortement déterminée par le niveau de richesse (un résultat attendu)
- En effet, la corrélation entre les deux variables de l’ordre de 2/3 (un lien linéaire particulièrement fort) et les écarts de richesse entre pays expliquent 44 % des différences d’espérance de vie (coefficient de détermination)
- Par rapport à la moyenne entre pays les plus peuplés, un gain de 1000$ de Revenu par tête se traduit par un peu moins de 4 mois de vie en plus, en moyenne (interprétation du coefficient directeur)
- Ce résultat n’est pas un artefact : il est confirmé lorsque l’on construit la droite de régression pour tous les pays ayant des données (corrélation : 0,64 ; R^2 : 0,42)


90
80
70
60
50
40
30
20
10
0
Espérance de vie à la naissance
0
10000
20000
30000
40000
50000
60000
70000
PIB/tête
Bangladesh
Inde
Pakistan
Nigéria
Indonésie
Brésil
Chine
Fédération de Russie
Japon
États-Unis
Y : Esp Vie Naiss 2017
Point moyen
Modèle Ts pays
Remarque : ici, un gain de 1000$ en revenu par tête se traduit en moyenne par un trimestre d'espérance de vie en plus

Tests sur les liens entre deux variables

- Corrélation r entre deux variables (Pearson) : on se ramène à une loi de Student à n-2 degrés de liberté sous l'hypothèse d'indépendance entre les variables

$$r \cdot \sqrt{\frac{n-2}{1-r^2}} \sim T(n-2)$$

Ici, on trouve une probabilité p=0,03417 d'observer une corrélation de 0,6696 entre ces variables : on rejette donc l'hypothèse d'indépendance linéaire (malgré le petit nombre d'observations!)

- Sous l'hypothèse que les résidus suivent une loi normale et sont indépendants, les coefficients calculés lors d'une régression suivent également une loi normale : on peut tester leur nullité et donner un intervalle de confiance

On retrouve une probabilité p=0,03417 pour le test de l'hypothèse d'une nullité du coefficient directeur (qui revient à une indépendance linéaire des deux variables)

Récapitulatif (1)

- Nous reviendrons sur le modèle linéaire à partir de la prochaine séance, cette fois à partir d’un nombre quelconque de variables explicatives
- Nous reprenons à présent dans cette séance les différentes méthodes de calcul mobilisées sur les variables numériques

Indicateur	IDH 2017	PIB/tête (2017)	Esp Vie 2017	Ann édu (2017)	Population totale (2017)	Ratio Palma
Somme	7,196	178603	718,2	131,5	4370,6	18,4
Moyenne par pays	0,7196	17860,3	71,82	13,15	437,06	1,84
Moyenne par hab.	*0,7104*	*14806,3*	*72,33*	*13,16*	*pas de sens*	*1,85*
Médiane	0,723	12300,5	72	13,3	203,15	1,7
Q1	0,616	5571,5	68,95	11,625	171,25	1,35
Q3	0,80175	21992,25	76,225	15,35	309,375	2,075
Variance	0,0169	258521647	60,5636	6,0965	222810,5	0,4264
Écart-type	0,130	16079	7,782	2,469	472,028	0,653
Coeff. de variation	0,181	0,900	0,108	0,188	1,080	0,355
Q3-Q1	0,18575	16420,75	7,275	3,725	138,125	0,725
relatif	0,257	1,335	0,101	0,280	0,680	0,426
Q3/Q1	1,3015	3,948	1,106	1,320	1,807	1,537
D9/D1	1,629	7,995	1,224	1,582	9,457	1,942