Analyse statistique - Séance 1

Aubin Poissonnier-Beraud

Villes et environnements urbains - Université Lumière Lyon 2

2025

Programme de la séance

Rappel de la première année

Introduction à la corrélation linéaire

Introduction à la régression linéaire simple

Rappel de la première année

En première année, nous avons présenté :

- L'analyse statistique en sciences sociales et les indicateurs
- Le processus de catégorisation
- La représentation graphique des données
- La notion de représentativité statistique appliquée aux questionnaires
- La notion de dépendance entre deux variables et le test du Khi2

Rappel de la première année

Introduction à la corrélation linéaire

Introduction à la régression linéaire simple

La corrélation

On dit que deux variables sont corrélées quand celles-ci varient ensemble, positivement ou négativement.

- La taille d'un champ de blé est corrélée positivement avec le nombre de kilos de blé qu'il produit
- La vitesse d'une voiture est négativement corrélée au temps de trajet
- Le taux de chômage d'une commune est positivement corrélé à son taux de pauvreté
- Le taux de sélection d'un Master est négativement corrélé au salaire moyen de ses diplômé.es

La notion de corrélation est donc réservée aux variables quantitatives. Elle peut néamoins s'appliquer à des quantités de différentes natures : volumes, taux, moyennes, probabilités etc.

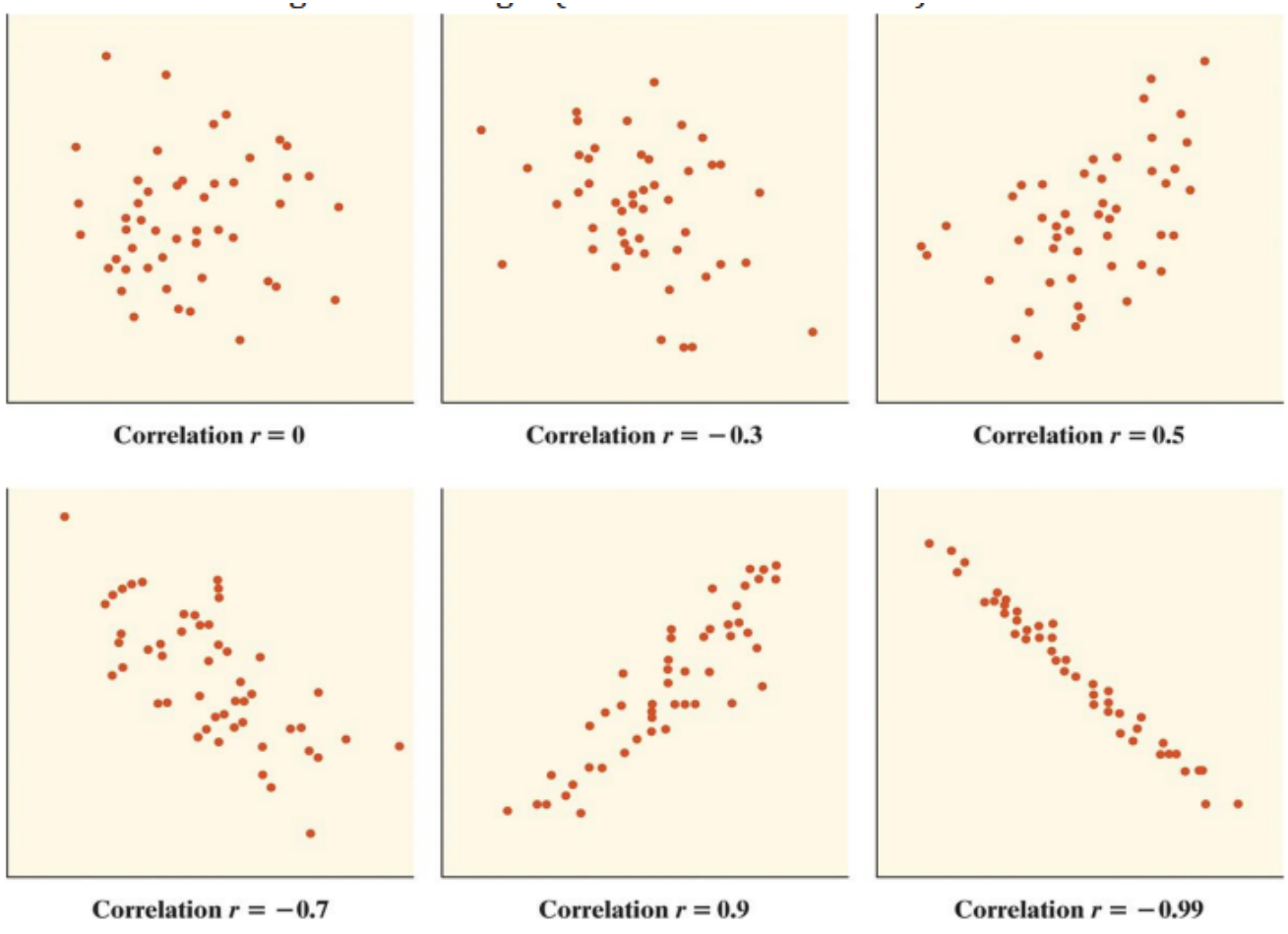


Figure: Exemples de corrélations

Le coefficient de corrélation linéaire de Pearson

Le *coefficient de corrélation de Pearson* r_{XY} est égal la *covariance* de deux variables divisée par le produit de leur *écart-type*.

$$r_{XY} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{1}$$

La covariance est une mesure de la variation simultanée de deux variables. Elle correspond à la moyenne de leur *covariation*.

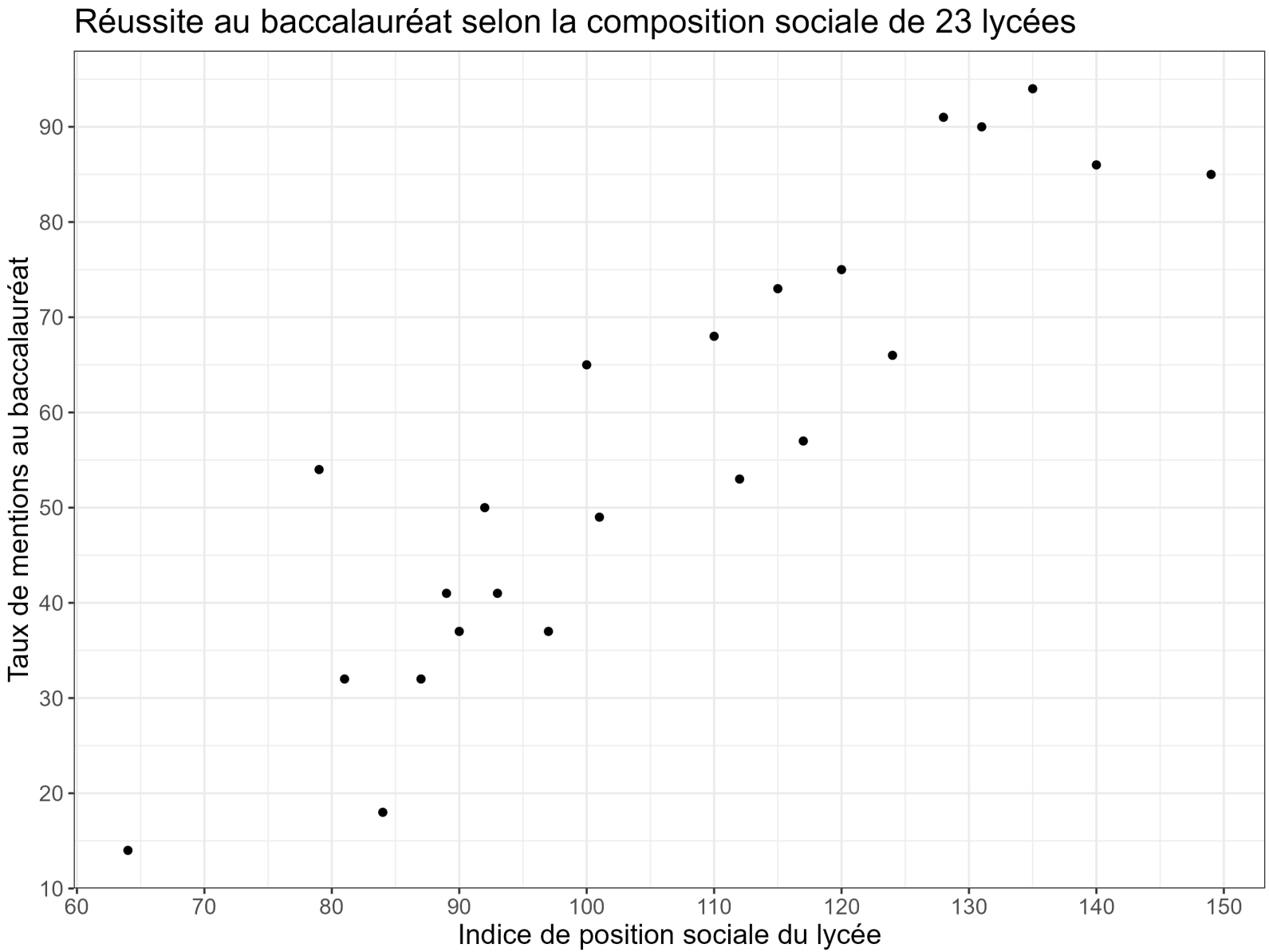
$$Cov(X, Y) = \frac{\sum_{i=1}^{N} (X_i - \overline{X}) \times (Y_i - \overline{Y})}{N} \tag{2}$$

La division par le produit des écarts-types permet de *standardiser* la covariance, c'est-à-dire de la rendre insensible aux différences d'échelles. Elle permet d'obtenir un coefficient toujours compris entre -1 et 1.

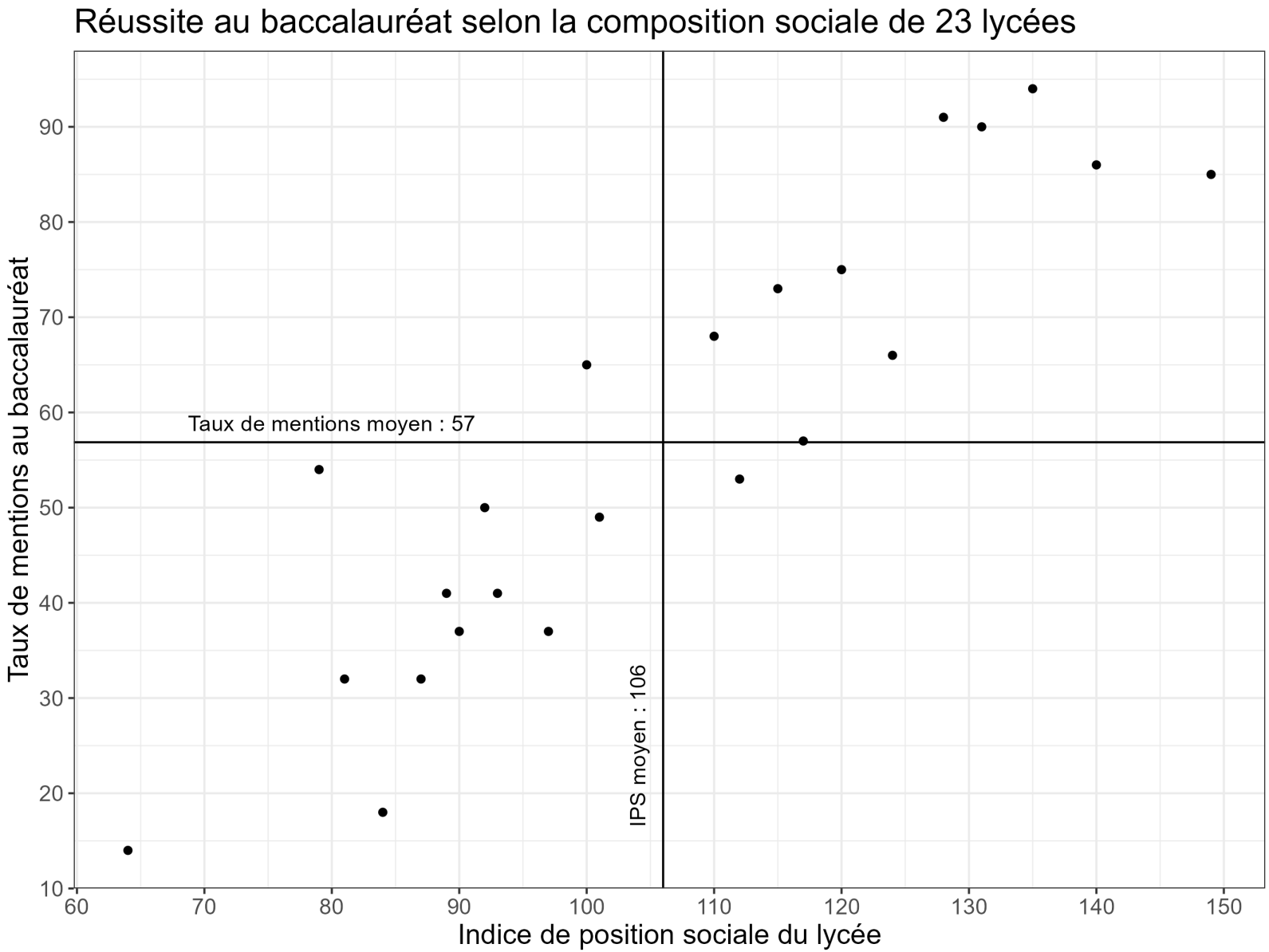
Exemple : composition sociale des lycées et réussite au bac

Indice de position scolaire et taux de mentions au bac		
	IPS	Taux de mentions
1	64	14
2	79	54
3	81	32
4	84	18
5	87	32
6	89	41
7	90	37
8	92	50
9	93	41
10	97	37
11	100	65
12	101	49
13	110	68
14	112	53
15	115	73
16	117	57
17	120	75
18	124	66
19	128	91
20	131	90
21	135	94
22	140	86
23	149	85
Moyenne	106	57

Exemple : composition sociale des lycées et réussite au bac



Exemple : composition sociale des lycées et réussite au bac



Exemple : composition sociale des lycées et réussite au bac

Indice de position scolaire et taux de mentions au bac		
	IPS	Taux de mentions
1	64	14
2	79	54
3	81	32
4	84	18
5	87	32
6	89	41
7	90	37
8	92	50
9	93	41
10	97	37
11	100	65
12	101	49
13	110	68
14	112	53
15	115	73
16	117	57
17	120	75
18	124	66
19	128	91
20	131	90
21	135	94
22	140	86
23	149	85
Moyenne	106	57

Exemple : composition sociale des lycées et réussite au bac

Indice de position scolaire et taux de mentions au bac		
	IPS	Taux de mentions
1	64	14
2	79	54
3	81	32
4	84	18
5	87	32
6	89	41
7	90	37
8	92	50
9	93	41
10	97	37
11	100	65
12	101	49
13	110	68
14	112	53
15	115	73
16	117	57
17	120	75
18	124	66
19	128	91
20	131	90
21	135	94
22	140	86
23	149	85
Moyenne	106	57

Exemple : composition sociale des lycées et réussite au bac

Lycée par lycée, on calcule l'écart de son score IPS à l'IPS moyen; l'écart de son taux de mentions au taux moyen; puis le produit de ces deux écarts.

Indice de position scolaire et taux de mentions au bac

	IPS	Taux de mentions	Écart de l'IPS à la moyenne	Écart du taux de mentions à la moyenne	Produit des écarts
1	64	14	-42	-43	1806
2	79	54	-27	-3	81
3	81	32	-25	-25	625
4	84	18	-22	-39	858
5	87	32	-19	-25	475
6	89	41	-17	-16	272
7	90	37	-16	-20	320
8	92	50	-14	-7	98
9	93	41	-13	-16	208
10	97	37	-9	-20	180
11	100	65	-6	8	-48
12	101	49	-5	-8	40
13	110	68	4	11	44
14	112	53	6	-4	-24
15	115	73	9	16	144
16	117	57	11	0	0
17	120	75	14	18	252
18	124	66	18	9	162
19	128	91	22	34	748
20	131	90	25	33	825
21	135	94	29	37	1073
22	140	86	34	29	986
23	149	85	43	28	1204
Moyenne	106	57	0	0	449

Exemple : composition sociale des lycées et réussite au bac

On remarque que le produit n'est négatif que pour les lycées où l'IPS est supérieur à l'IPS moyen, mais le taux de mentions inférieur au taux moyen, et inversement.

Indice de position scolaire et taux de mentions au bac

	IPS	Taux de mentions	Écart de l'IPS à la moyenne	Écart du taux de mentions à la moyenne	Produit des écarts
1	64	14	-42	-43	1806
2	79	54	-27	-3	81
3	81	32	-25	-25	625
4	84	18	-22	-39	858
5	87	32	-19	-25	475
6	89	41	-17	-16	272
7	90	37	-16	-20	320
8	92	50	-14	-7	98
9	93	41	-13	-16	208
10	97	37	-9	-20	180
11	100	65	-6	8	-48
12	101	49	-5	-8	40
13	110	68	4	11	44
14	112	53	6	-4	-24
15	115	73	9	16	144
16	117	57	11	0	0
17	120	75	14	18	252
18	124	66	18	9	162
19	128	91	22	34	748
20	131	90	25	33	825
21	135	94	29	37	1073
22	140	86	34	29	986
23	149	85	43	28	1204
Moyenne	106	57	0	0	449

Exemple : composition sociale des lycées et réussite au bac

La moyenne du produit des écarts nous donne la valeur de la covariance de l'IPS et du taux de mentions. Elle vaut ici **449**.

Indice de position scolaire et taux de mentions au bac

	IPS	Taux de mentions	Écart de l'IPS à la moyenne	Écart du taux de mentions à la moyenne	Produit des écarts
1	64	14	-42	-43	1806
2	79	54	-27	-3	81
3	81	32	-25	-25	625
4	84	18	-22	-39	858
5	87	32	-19	-25	475
6	89	41	-17	-16	272
7	90	37	-16	-20	320
8	92	50	-14	-7	98
9	93	41	-13	-16	208
10	97	37	-9	-20	180
11	100	65	-6	8	-48
12	101	49	-5	-8	40
13	110	68	4	11	44
14	112	53	6	-4	-24
15	115	73	9	16	144
16	117	57	11	0	0
17	120	75	14	18	252
18	124	66	18	9	162
19	128	91	22	34	748
20	131	90	25	33	825
21	135	94	29	37	1073
22	140	86	34	29	986
23	149	85	43	28	1204
Moyenne	106	57	0	0	449

Exemple : composition sociale des lycées et réussite au bac

En changeant l'échelle du taux de mentions, comme ici en la divisant par 100, on divise par 100 la covariance.

Indice de position scolaire et taux de mentions au bac

	IPS	Taux de mentions	Écart de l'IPS à la moyenne	Écart du taux de mentions à la moyenne	Produit des écarts
1	64	0.14	-42	-0.43	18.06
2	79	0.54	-27	-0.03	0.81
3	81	0.32	-25	-0.25	6.25
4	84	0.18	-22	-0.39	8.58
5	87	0.32	-19	-0.25	4.75
6	89	0.41	-17	-0.16	2.72
7	90	0.37	-16	-0.20	3.20
8	92	0.50	-14	-0.07	0.98
9	93	0.41	-13	-0.16	2.08
10	97	0.37	-9	-0.20	1.80
11	100	0.65	-6	0.08	-0.48
12	101	0.49	-5	-0.08	0.40
13	110	0.68	4	0.11	0.44
14	112	0.53	6	-0.04	-0.24
15	115	0.73	9	0.16	1.44
16	117	0.57	11	0.00	0.00
17	120	0.75	14	0.18	2.52
18	124	0.66	18	0.09	1.62
19	128	0.91	22	0.34	7.48
20	131	0.90	25	0.33	8.25
21	135	0.94	29	0.37	10.73
22	140	0.86	34	0.29	9.86
23	149	0.85	43	0.28	12.04
Moyenne	106	1	0	0	4.49

Exemple : composition sociale des lycées et réussite au bac

Pour standardiser cette mesure, on calcule l'écart-type (voir support M1 calcul de la variance et de l'écart-type).

Indice de position scolaire et taux de mentions au bac

	IPS	Taux de mentions	Écart de l'IPS à la moyenne	Écart du taux de mentions à la moyenne	Produit des écarts	Écart au carré de l'IPS à la moyenne	Écart au carré du taux de mentions à la moyenne
1	64	14	-42	-43	1806	1764	1849
2	79	54	-27	-3	81	729	9
3	81	32	-25	-25	625	625	625
4	84	18	-22	-39	858	484	1521
5	87	32	-19	-25	475	361	625
6	89	41	-17	-16	272	289	256
7	90	37	-16	-20	320	256	400
8	92	50	-14	-7	98	196	49
9	93	41	-13	-16	208	169	256
10	97	37	-9	-20	180	81	400
11	100	65	-6	8	-48	36	64
12	101	49	-5	-8	40	25	64
13	110	68	4	11	44	16	121
14	112	53	6	-4	-24	36	16
15	115	73	9	16	144	81	256
16	117	57	11	0	0	121	0
17	120	75	14	18	252	196	324
18	124	66	18	9	162	324	81
19	128	91	22	34	748	484	1156
20	131	90	25	33	825	625	1089
21	135	94	29	37	1073	841	1369
22	140	86	34	29	986	1156	841
23	149	85	43	28	1204	1849	784
Moyenne	106	57	0	0	449	467	528

Exemple : composition sociale des lycées et réussite au bac

On obtient ainsi un coefficient de corrélation de :

$$\begin{aligned} r_{XY} &= \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \\ &= \frac{449}{\sqrt{467} \times \sqrt{528}} \\ &= 0.9 \end{aligned} \tag{3}$$

Ce qui correspond à une association extrêmement forte, quasiment parfaite !

Rappel de la première année

Introduction à la corrélation linéaire

Introduction à la régression linéaire simple

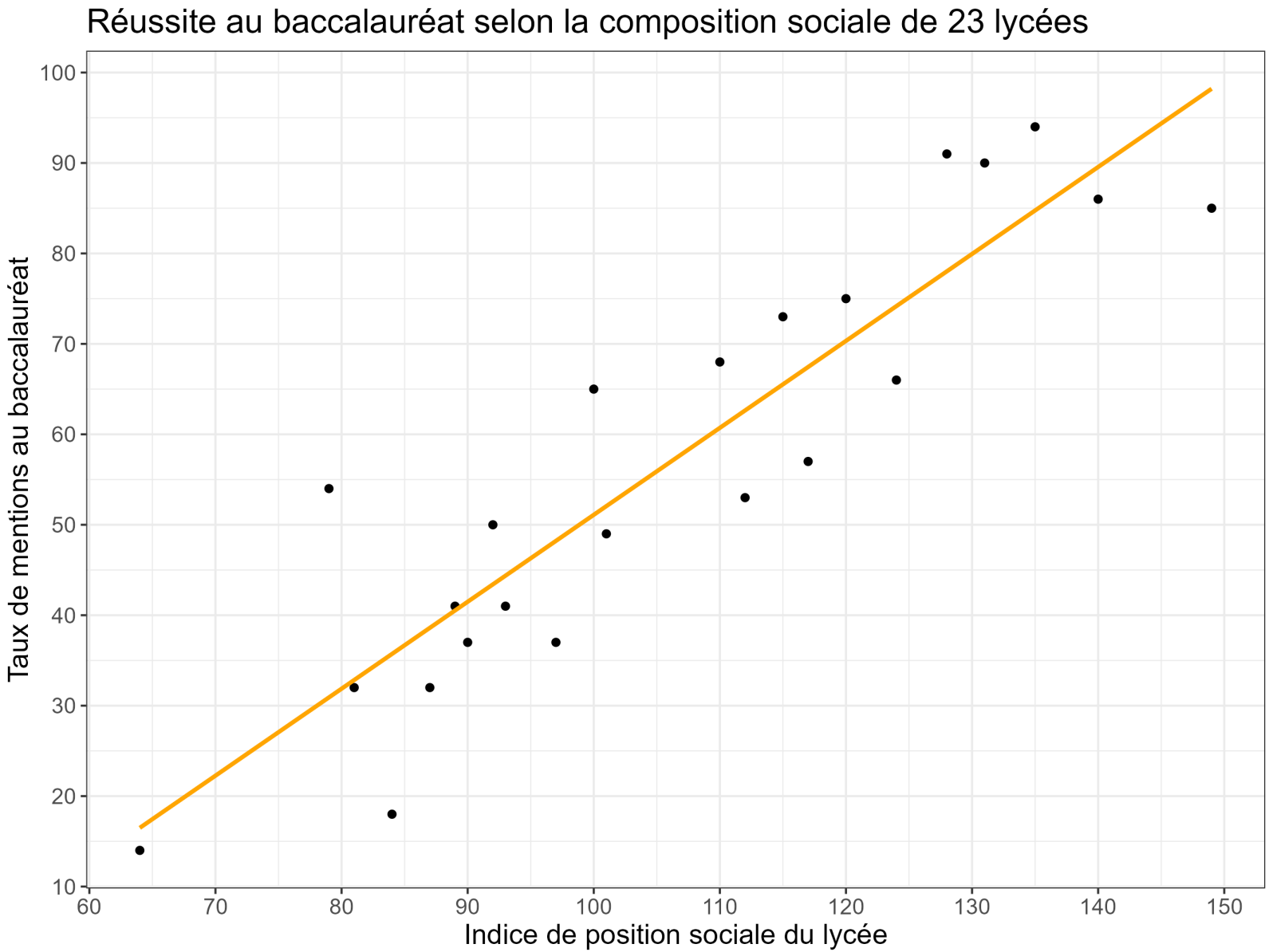
La régression linéaire simple

La *régression linéaire simple* est une technique économétrique qui permet de résumer une relation entre deux variables et de s'interroger sur sa significativité statistique. Le nuage de points y est résumé par une droite, appelée droite de régression.
Ce modèle tente de lier les variations d'une variable *expliquée* (ou dépendante) à celles d'une variable *explicative* (ou indépendante, covariable, régresseur). Il cherche donc à mesurer les *relations de dépendance* entre plusieurs grandeurs ou dimensions du monde social.

L'équation de régression

Les modèles de régression linéaires s'inspirent des équations linéaires de la forme $y = ax + b$ où b est l'ordonnée à l'origine, soit la valeur de y quand $a = 0$, et a la pente, soit le nombre d'unités d'augmentation de y quand x augmente d'une unité.

Exemple : composition sociale des lycées et réussite au bac



L'équation de régression

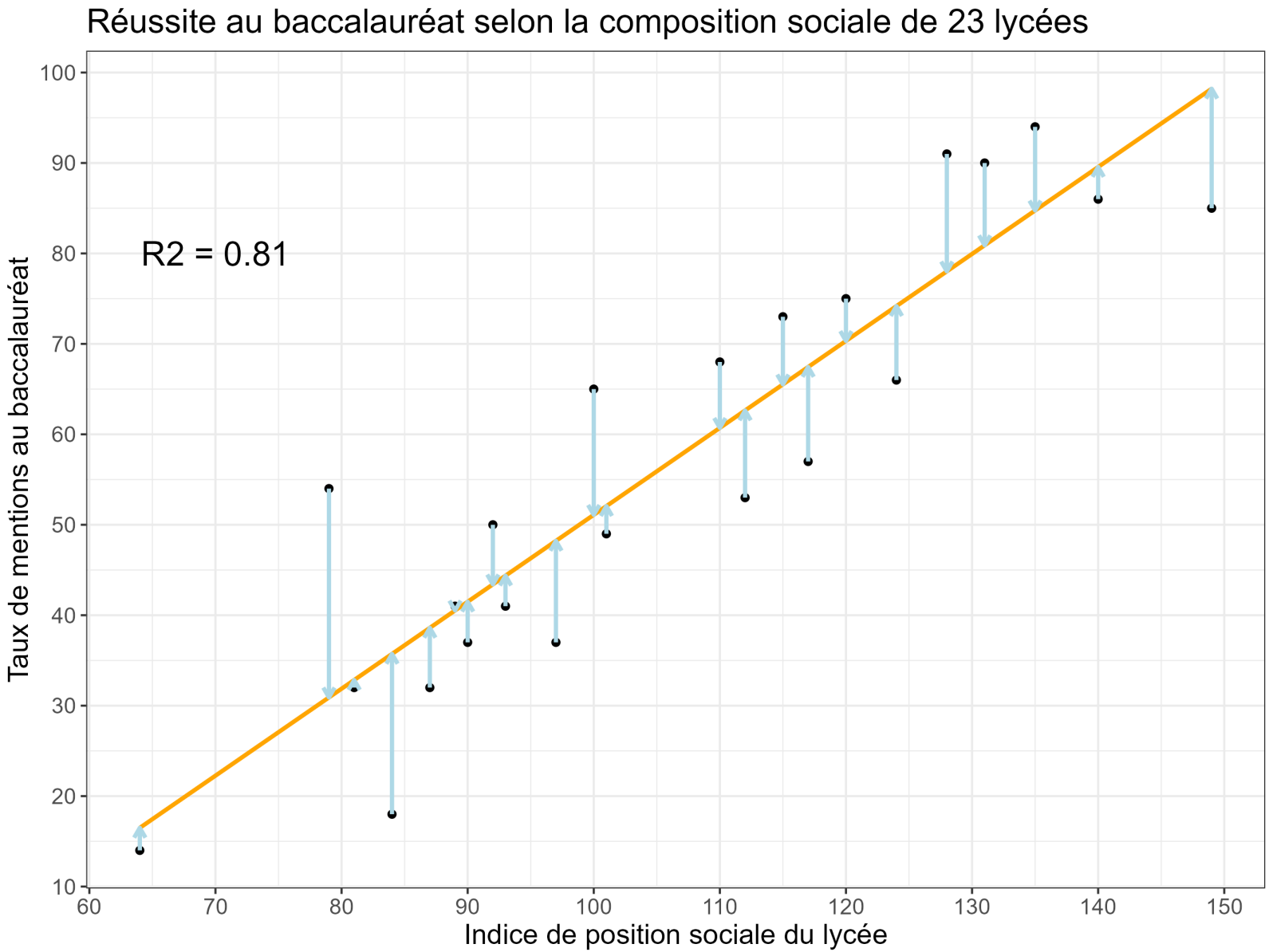
Le modèle s'écrit sous la forme d'une équation qui, pour chaque individu i, prédit la valeur d'une variable Y à partir de la valeur de la variable X :

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ \textit{Taux de mentions}_i &= \beta_0 + \beta_1 \textit{IPS}_i + \varepsilon_i \end{aligned} \tag{4}$$

Les moindres carrés ordinaires

Estimer un modèle consister à déterminer la valeur des paramètres β_0 et β_1 de manière à *maximiser* l'ajustement du modèle aux données. Cela revient à chercher β_0 et β_1 tels que, à partir de l'IPS d'un lycée, on soit en mesure de déterminer son taux de mentions en se trompant en moyenne le moins possible. La méthode des moindres carrés ordinaires maximise l'ajustement en minimisant la somme des termes résiduels ε_i. R_2 correspond à la part de la variance expliquée par notre modèle, c'est une mesure de sa qualité.

Les moindres carrés ordinaires



Lire un modèle de régression

Table

	Dependent variable:
	Taux de mentions
IPS	0.96***
	(0.10)
Constant	−45.04***
	(10.74)
Observations	23
R^2	0.82
Adjusted R^2	0.81
Residual Std. Error	10.29 (df = 21)
Note:	*p<0.1; **p<0.05; ***p<0.01

La constante vaut -45. Le modèle permet d'estimer qu'en moyenne, le taux de mentions d'un lycée dont l'IPS vaudrait 0 serait de -45% de mentions.
Le coefficient pour l'âge est significatif au seuil de 1%. Il y a donc moins d'1% de chance de se tromper en affirmant qu'il est différent de 0. Il vaut 0.96. Le modèle permet d'estimer qu'un point d'IPS augmente en moyenne le taux de mentions de 0.96 points de pourcentage.

Lire un modèle de régression

L'équation de régression estimée à partir des 23 lycées est donc :

$$\textit{Taux de mentions}_i = -45 + 0.96 \textit{IPS}_i + \varepsilon_i \tag{5}$$

Pour un lycée dont l'IPS vaudrait 60, le taux de mentions estimé par le modèle est de $-45 + 0.96 \times 60 = 12,6\%$.
Pour un lycée dont l'IPS vaudrait 100, le taux de mentions estimé par le modèle est de $-45 + 0.96 \times 100 = 51\%$.

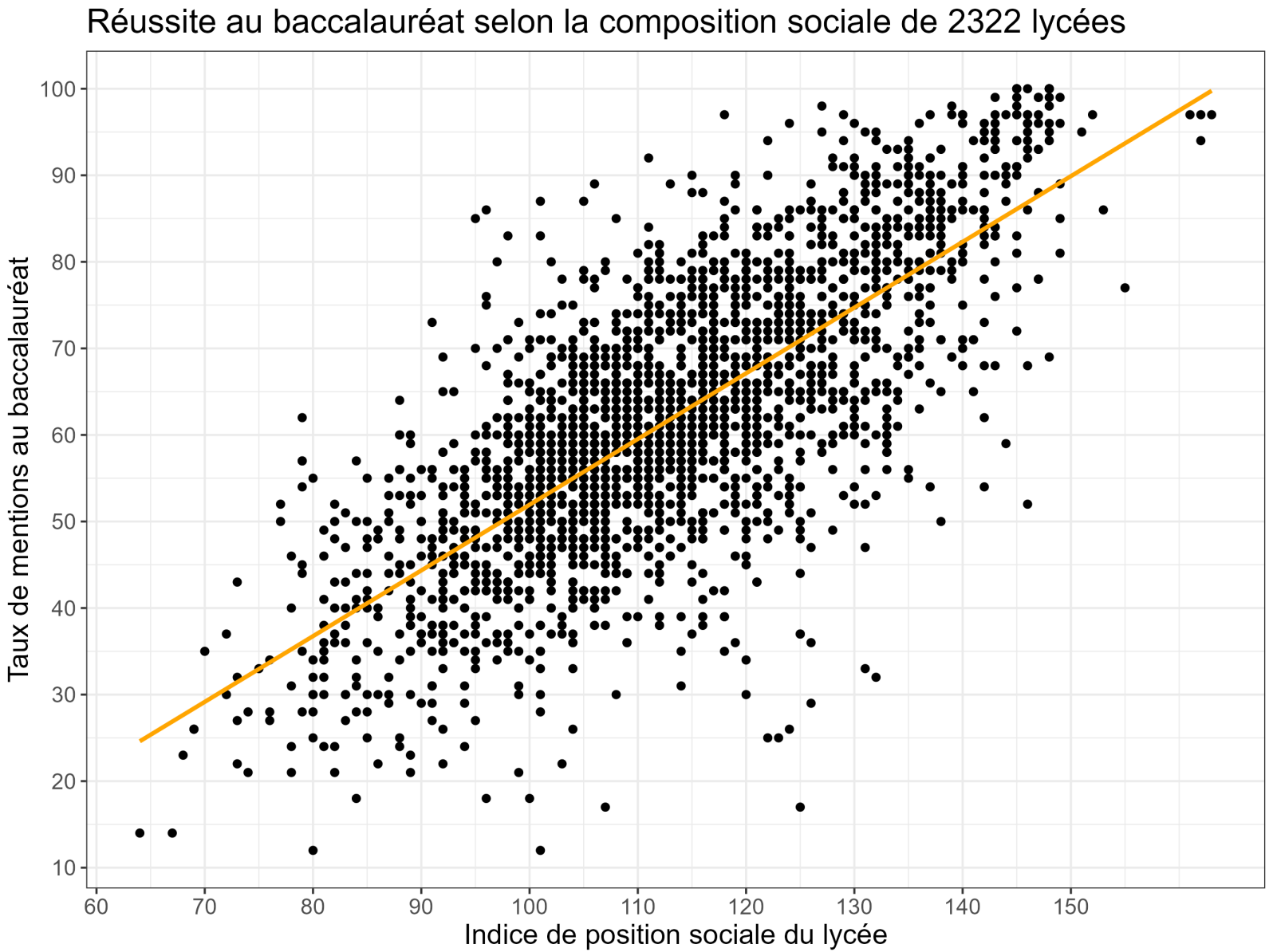
Le modèle sur l'ensemble des lycées

Table

	Dependent variable:
	Taux de mentions
IPS	0.76***
	(0.01)
Constant	−23.99***
	(1.63)
Observations	2,322
R^2	0.55
Adjusted R^2	0.55
Residual Std. Error	10.71 (df = 2320)
Note:	*p<0.1; **p<0.05; ***p<0.01

Le modèle construit sur les 2 322 lycées aboutit à un coefficient β_1 inférieur au premier modèle (0.76 contre 0.96). Il n'explique plus que 55% de la variance des taux de mentions grâce à l'IPS, contre 82% dans le premier modèle.

Le modèle sur l'ensemble des lycées

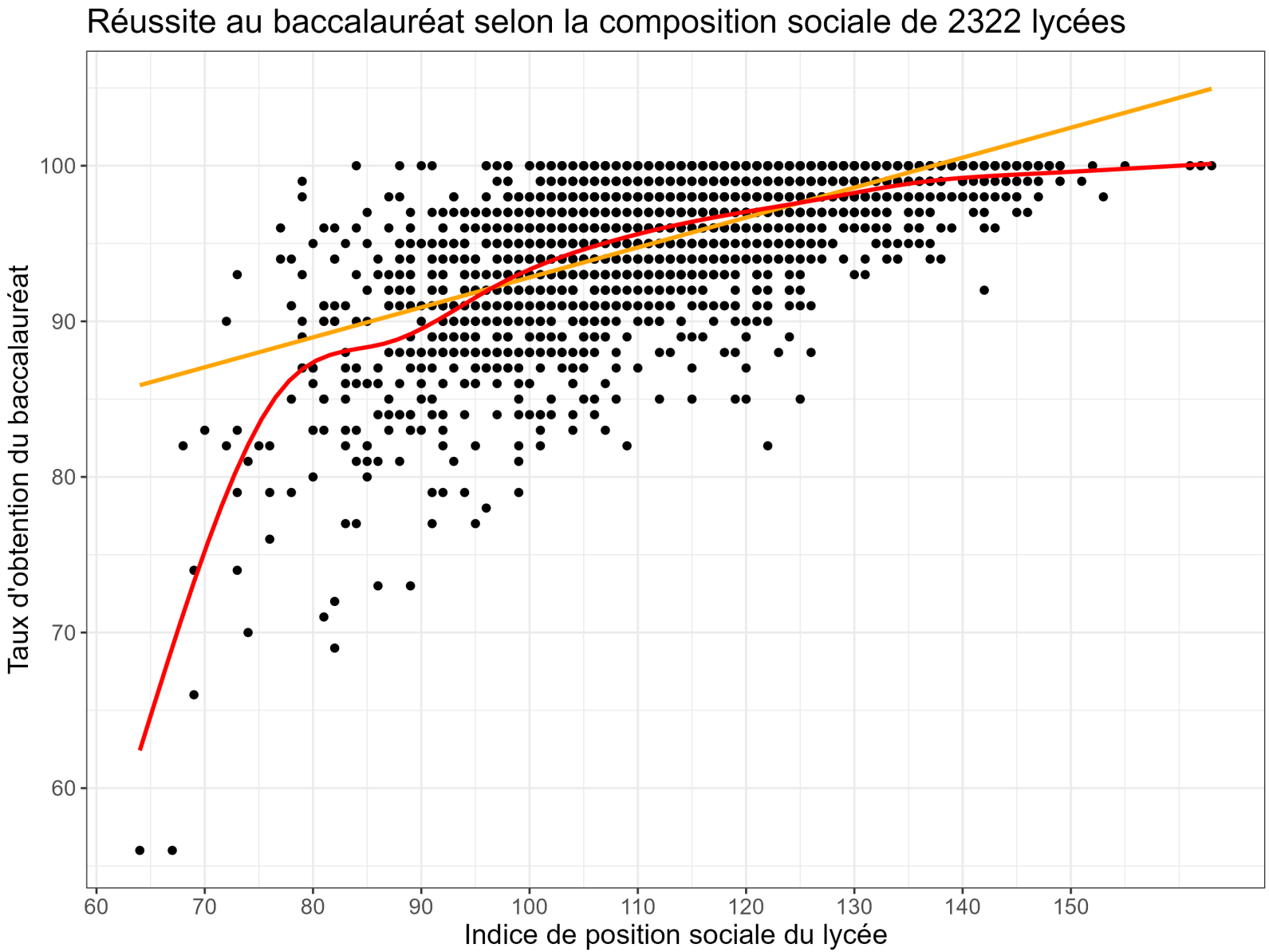


Dans quels cas la régression linéaire est-elle pertinente ?'

La régression linéaire simple est une modélisation qui résume une relation entre deux séries d'observations potentiellement très nombreuses – les valeurs observées de X et les valeurs de Y – à partir de deux coefficients. Elle opère nécessairement une simplification de la réalité, plus ou moins grossière selon l'ajustement du modèle.

Il est donc essentiel de s'assurer qu'il est pertinent de modéliser une relation linéaire pour rendre compte du lien entre deux variables. Le moyen le plus sommaire pour le vérifier consiste à construire un nuage de points pour vérifier graphiquement la forme du nuage. D'autres hypothèses doivent par ailleurs être vérifiées pour s'assurer que l'estimation des coefficients ainsi que les tests menés sur ces estimations soient correctes. Elles seront étudiées à la prochaine séance.

Exemple d'une régression linéaire imparfaitement ajustée



Conclusion sur la corrélation : coefficient de corrélation et régression linéaire

Quel est le lien entre le coefficient de corrélation et la relation linéaire ?

Bien que liés, le coefficient de corrélation linéraire et les coefficients de la régression linéaire ne sont pas équivalents.

- Le coefficient de corrélation mesure la **force** (proche de 0 ou proche de 1) et la **direction** (négatif ou positif) d'une relation entre deux variables, en partant du principe que cette relation est symétrique. On pense que X et Y sont interchangeables.
- Les coefficients de régression mesurent **l'effet moyen** d'une variable explicative pensée comme fixe sur une variable expliquée pensée comme variante, en partant du principe que cette relation est asymétrique. On pense que X et Y ne sont pas interchangeables.