Leçon 3 : Statistique descriptive bivariée (Partie A)

On souhaite maintenant étudier simultanément deux variables X et Y sur une même population. On cherche à caractériser le lien, s'il existe, entre X et Y. Plusieurs cas se présentent selon la nature de ces deux variables :

Partie A. X et Y sont deux variables quantitatives (continues ou discrètes)
Exemple : on peut chercher à savoir s'il y a un lien entre le salaire des hommes et le salaire des femmes au sein d'un foyer.

Partie B. X est qualitative (ou quantitative discrète avec peu de valeurs distinctes) et Y est continue.
Exemple : L'acceptation du crédit X est-elle associée au salaire des hommes, des femmes, ou au revenu global par tête ?

Partie C. X et Y sont deux variables quantitatives.
Exemple : L'acceptation du crédit X est-elle associée au type de contrat de travail Y ?

Deux variables quantitatives : ajustement linéaire

Objectif : On cherche à mettre en évidence l'existence d'une relation entre deux variables quantitatives (continues ou discrètes) X et Y.

La relation fonctionnelle la plus simple est une relation linéaire.

On appelle **nuage de points** l'ensemble des points de coordonnées (x_i, y_i), $i = 1, \cdots, n$.

x_1	x_2	x_3	...	x_n
y_1	y_2	y_3	...	y_n

Exemple

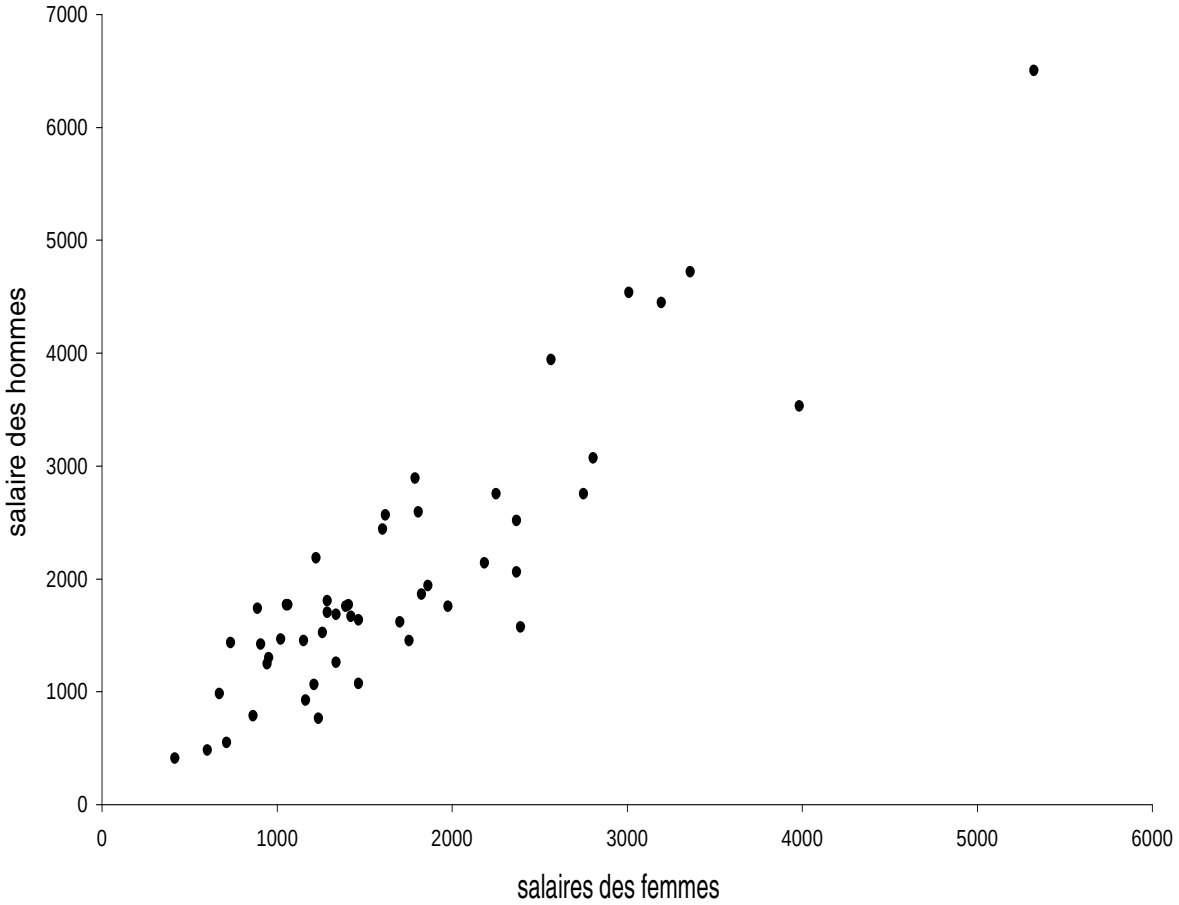
Reprenons l'exemple des crédits et considérons les variables continues X salaire des femmes et Y salaire des hommes.

salaire des femmes x_i	2253	1790	...	1287
salaire des hommes y_i	2752	2893	...	1808

Le nuage de points

La représentation graphique du nuage de points est la première étape essentielle pour déterminer s'il existe ou non une relation entre X et Y.

Exemple Salaire des hommes Y en fonction du salaire des femmes X :



Ajustement linéaire par la méthode des moindres carrés

Si l'examen du nuage de points indique qu'il est judicieux de supposer une relation de type linéaire entre Y et X alors on cherche à déterminer l'équation d'une droite

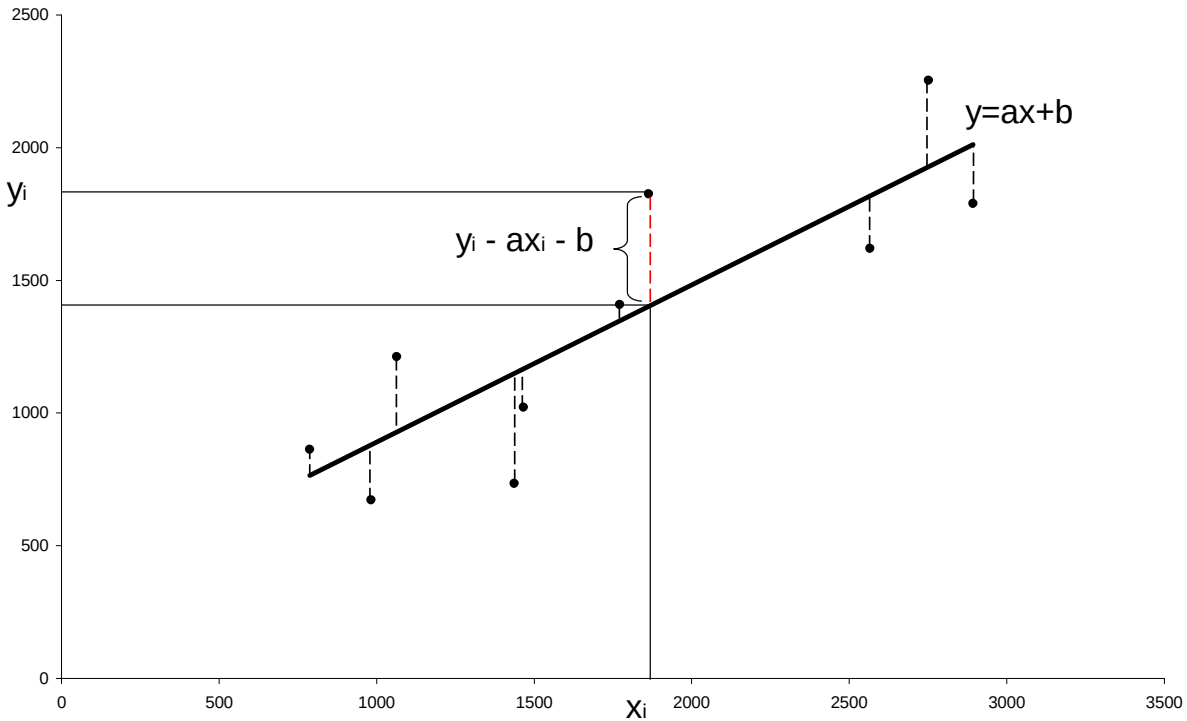
$$y = ax + b \quad \text{où } a \text{ et } b \text{ sont deux réels.}$$

telle que cette droite soit le "plus près" possible du nuage de points. La méthode des moindres carrés précise cette notion de proximité entre la droite (dite "des moindres carrés") et les points du nuage.

Méthode des moindres carrés

Le critère des moindres carrés consiste à déterminer les réels a et b tels que

$$\sum_{i=1}^{n} (y_i - ax_i - b)^2 \quad \text{soit minimale.}$$



Méthode des moindres carrés

Si l'on dispose des couples d'observations (x_i, y_i), pour $i = 1, \cdots, n$ et si on note $\overline{x}$ et $\overline{y}$, les moyennes respectives de X et Y alors, le couple $(\hat{a}, \hat{b})$ solution de ce problème de minimisation est :

$$\begin{cases} \hat{a} = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \\ \\ \hat{b} = \overline{y} - \hat{a}\overline{x} \end{cases}$$

Exemple

On a déjà calculé le salaire moyen des hommes (données brutes) qui vaut $\overline{y} = 2026,44$ euros. Le salaire moyen des femmes est $\overline{x} = 1700,16$ euros.
Ensuite on calcule :

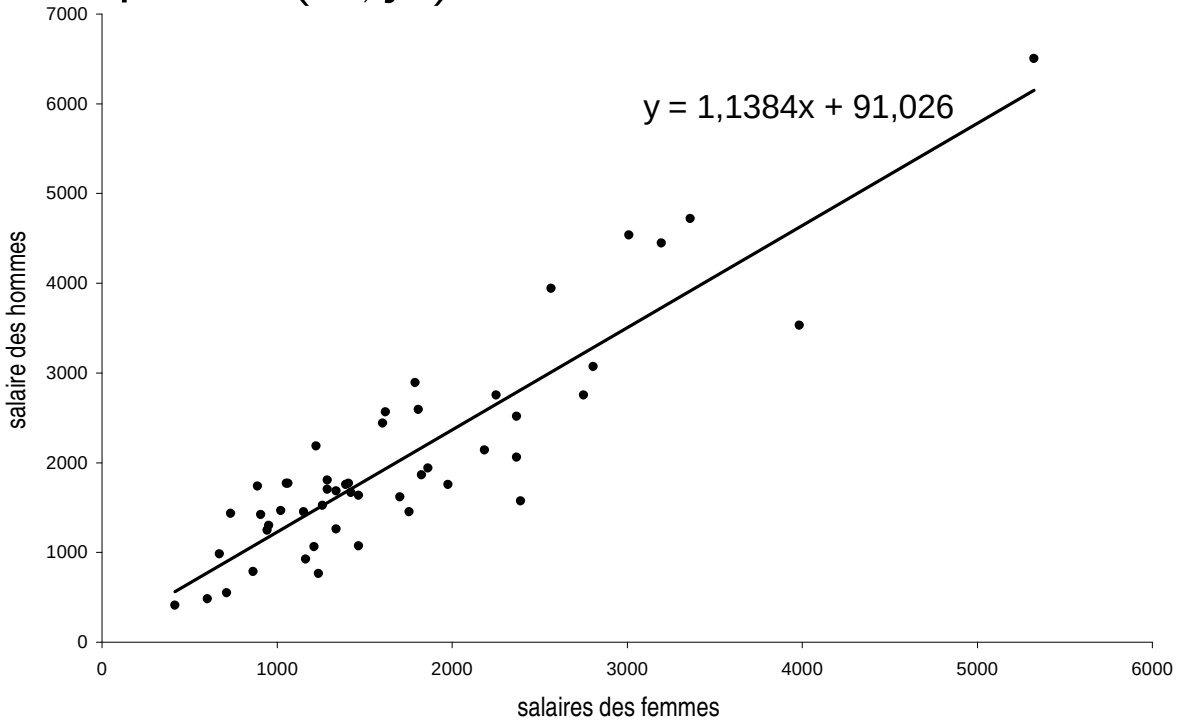
- $\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = (2253 - 1700,16)(2752 - 2026,44) + (1790 - 1700,16)(2893 - 2026,44) + \cdots = 48794684,48.$
- $\sum_{i=1}^{n}(x_i - \overline{x})^2 = (2253 - 1700,16)^2 + (1790 - 1700,16)^2 + \cdots = 42863586,72.$

Enfin, on peut calculer les valeurs $\hat{a}$ et $\hat{b}$ de la droite des moindres carrés :

$$\begin{cases} \hat{a} = \dfrac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \dfrac{48794684,48}{42863586,72} = 1,1384. \\ \\ \hat{b} = \overline{y} - \hat{a}\overline{x} = 2026,44 - 1,1384 \times 1700,16 = 91,026. \end{cases}$$

Droite des moindres carrés

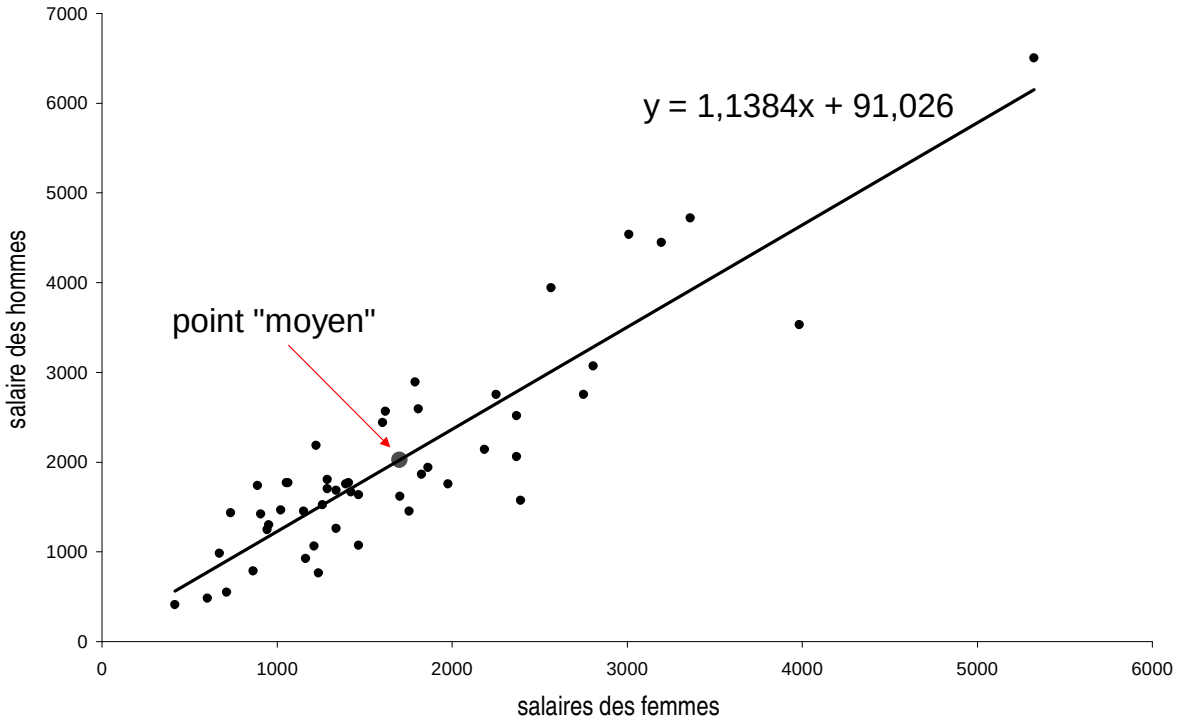
On obtient ainsi la droite des moindres carrés d'équation $y = 1,1384x + 91,026$. On dit qu'on a obtenu un ajustement linéaire des points (x_i, y_i).



Propriété

Le point moyen de coordonnées $(\overline{x}, \overline{y})$ (centre de gravité ou barycentre du nuage de points) appartient à la droite des moindres carrés puisque :

$$\overline{y} = \hat{a}\overline{x} + \hat{b}$$



Le point “moyen” correspond à un couple fictif où la femme aurait un salaire de $\overline{x}$ euros et l'homme un salaire $\overline{y}$.

Un indicateur : le coefficient de corrélation linéaire.

On appelle coefficient de corrélation linéaire le nombre réel

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

On peut montrer que :

$$-1 \leq r \leq 1$$

- Si $r = 0$, on dit que les variables x et y sont non corrélées linéairement.
- Si $|r| = 1$, les points (x_i, y_i), $i = 1, \cdots, n$ sont alignés : la variable y est une fonction linéaire de la variables x.

En pratique, si $|r|$ est proche de 1, on dit qu'il y a corrélation linéaire entre les variables. La corrélation est d'autant plus forte que $|r|$ est très proche de 1.

Interprétation du coeff. de corrélation linéaire

Exemple

On calcule le coeff. de corrélation linéaire entre le salaire des hommes et celui des femmes :

$$r = \frac{48794684,48}{\sqrt{68840536,32} \times \sqrt{42863586,72}} \simeq 0,898.$$

On peut donc conclure qu'il y a une forte corrélation linéaire entre le salaire des hommes et celui des femmes.
De plus, $r > 0$: la corrélation entre les salaires est positive (pente de la droite des moindres carrés positive) : lorsque le salaire des femmes croit, celui des hommes aussi.

Valeurs prédites

On peut faire des prévisions de la valeur de y pour de nouvelles valeurs de x. La valeur prédite par la relation linéaire si $x = x_0$ est :

$$\hat{y}_0 = \hat{a}x_0 + \hat{b}$$

Exemple

Reprenons les salaires. Si l'on souhaite prédire le salaire d'un homme, connaissant le salaire de la femme $x_0 = 2000$ euros, on utilise l'équation de la droite des moindres carrés :

$$\hat{y}_0 = \hat{a}x_0 + \hat{b} = 1,1384 \times 2000 + 91,026 = 2367,77 \text{ euros.}$$

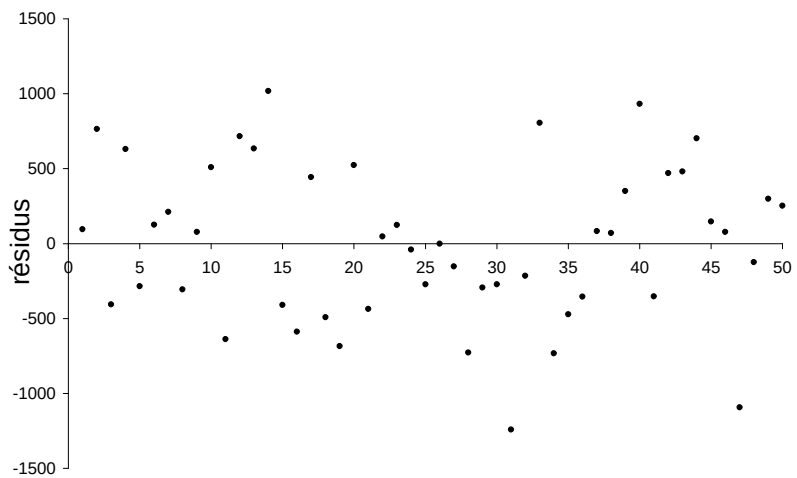
Résidus

On appelle résidus de l'ajustement linéaire, la série de données :

$$e_i = \hat{y}_i - y_i,\ i = 1, \cdots, n.$$

Les résidus e_i mesurent l'erreur commise entre la valeur observée y_i et la valeur donnée par l'ajustement linéaire $\hat{y}_i = \hat{a}x_i + \hat{b}$. Si l'ajustement linéaire est convenable, les résidus ne contiennent plus "d'information" : ils doivent être répartis de façon complètement aléatoire autour de l'axe des abscisses.

salaire y_i	val. prédite $\hat{y}_i$	rés. e_i
2752	2655,77	96,22
2893	2128,71	764,29
1064	1470,73	-406,73
2566	1935,19	630,81
788	1073,44	-285,44



Décomposition de la variance

On peut montrer que la variance de la série (y_i) se décompose de la façon suivante :

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2}_{\text{Variance des } y_i \text{ =Variance totale}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i-\overline{y}\right)^2}_{\text{Variance des } \hat{y}_i \text{ =Variance expliquée par la droite des MC}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}_{\text{Variance des } e_i \text{ =Variance résiduelle}}$$

ou encore

$$s_y^2 = s_{\hat{y}}^2 + s_e^2$$

On peut alors définir la **part ou proportion de variance expliquée** par la droite ajustée :

$$\frac{s^2_{\hat{y}}}{s^2_y} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \overline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

et la **part ou proportion de variance résiduelle**

$$\frac{s^2_e}{s^2_y} = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}$$

L'ajustement est d'autant meilleur que la proportion de variance expliquée par la droite de moindres carrés est proche de 1 (et donc la part de variance résiduelle proche de 0).

Exemple

Reprenons les salaires. La droite des moindres carrés a pour équation $y = 1,1384x + 91,026$ euros. On calcule les valeurs prédites $\hat{y}_i = \hat{a}x_i + \hat{b}$ par cet ajustement linéaire et les résidus $e_i = y_i - \hat{y}_i$:

salaire y_i	val. prédite $\hat{y}_i$	rés. e_i
2752	2655,77	96,22
2893	2128,71	764,29
1064	1470,73	-406,73
2566	1935,19	630,81
788	1073,44	-285,44
...	...	...

Puis, on calcule les variances respectives des séries (y_i), $(\hat{y}_i)$ et (e_i) ...

- $s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{50} \times 68840536,32$
- $s_{\hat{y}}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = \frac{1}{50} \times 55549260,67$
- $s_e^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{50} \times 13294059,36$

donc la part de variance expliquée par la droite ajustée vaut

$$\frac{s_{\hat{y}}^2}{s_y^2} = \frac{55549260,67}{68840536,32} \simeq 0,81$$

soit 81 % de la variance totale du salaire des hommes peut être expliquée par le salaire des femmes. Les salaires sont donc très corrélés linéairement.

Validation de l'ajustement linéaire

Pour valider l'ajustement linéaire, certaines critères doivent être satisfaits :

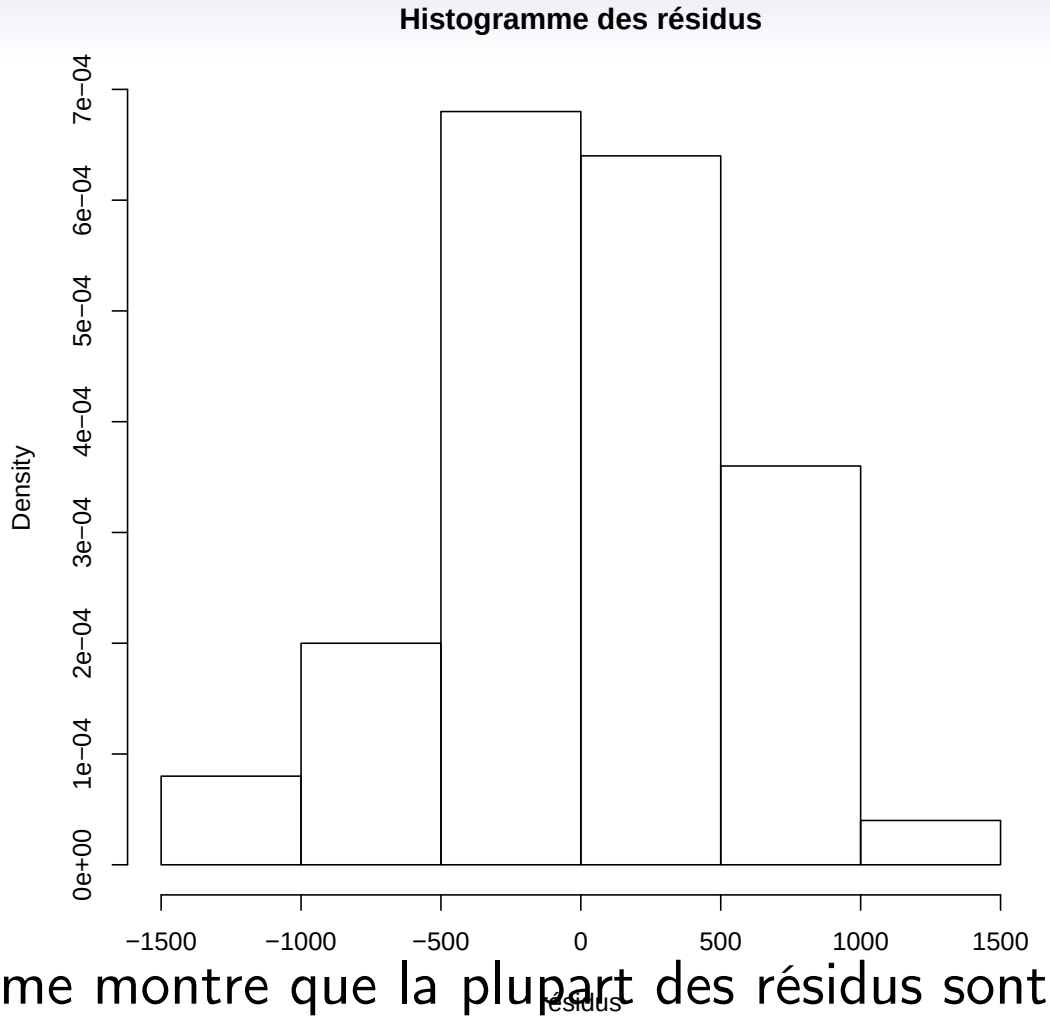
- la forme du nuage de points.
- la valeur du coefficient de corrélation linéaire r et le pourcentage de variance expliquée par la droite des moindres carrés.
- l'examen des résidus : ils doivent fluctuer autour de l'axe des abscisses de façon aléatoire, et être de faible amplitude (leur écart-type doit être très inférieur à l'ecart-type de la série y_i).

Récapitulatif pour l'exemple des salaires

- Le nuage de points des salaires des hommes en fonction de celui des femmes nous indique qu'il y a probablement un relation linéaire entre les deux.
- Le coeff. de corrélation linéaire r vaut $0,898$ et le pourcentage de variance expliquée par la droite des moindres carrés est de 81%.
- l'examen des résidus montre que leur écart-type vaut

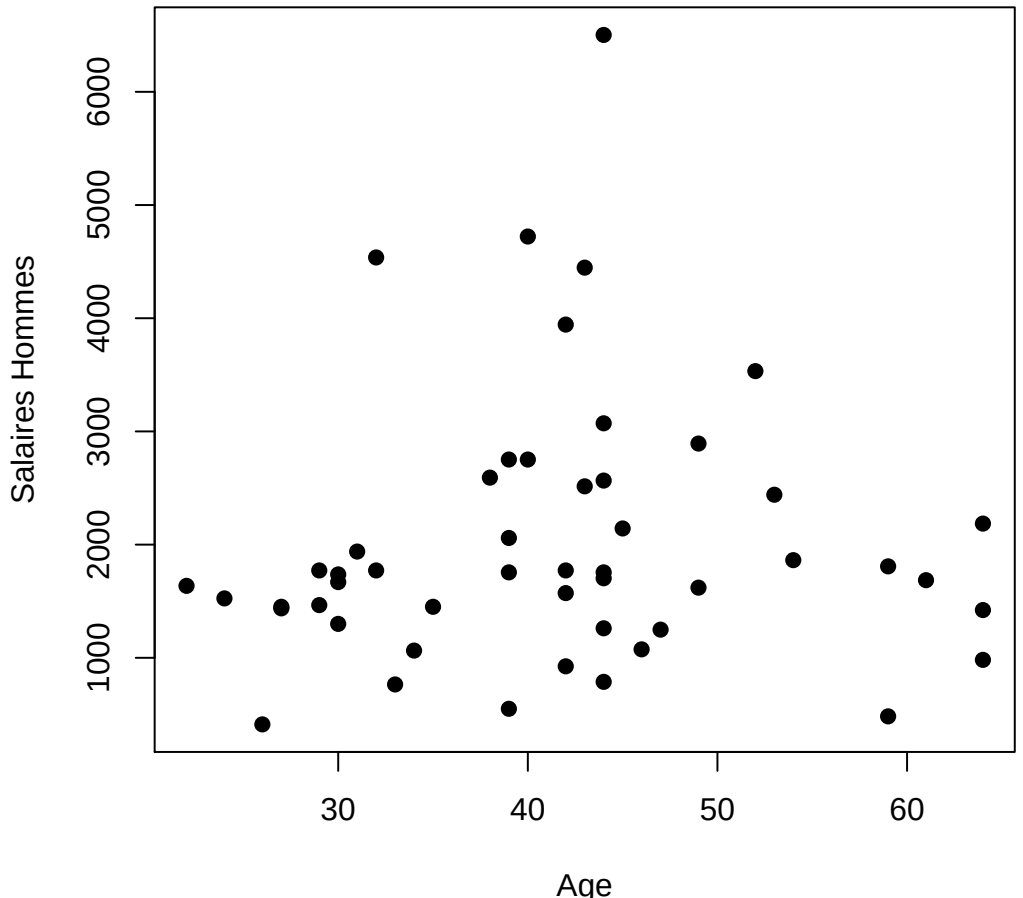
$$s_e = \sqrt{\frac{13294059,36}{50}} = 515,64$$

ce qui est 2 fois inférieur à l'écart-type des salaires des hommes $s_y = 1173,38$.



L'histogramme montre que la plupart des résidus sont concentrés dans les classes [-500,0[et [0,500[, donc les points du nuage (x_i, y_i) sont proches de la droite ajustée.

Un ajustement linéaire n'est pas toujours possible. Par exemple, intéressons nous au salaire des hommes en fonction de leur âge. Pour cela, on représente le nuage de points :



De plus, le coeff. de corrélation linéaire vaut 0,058694 ! ! !

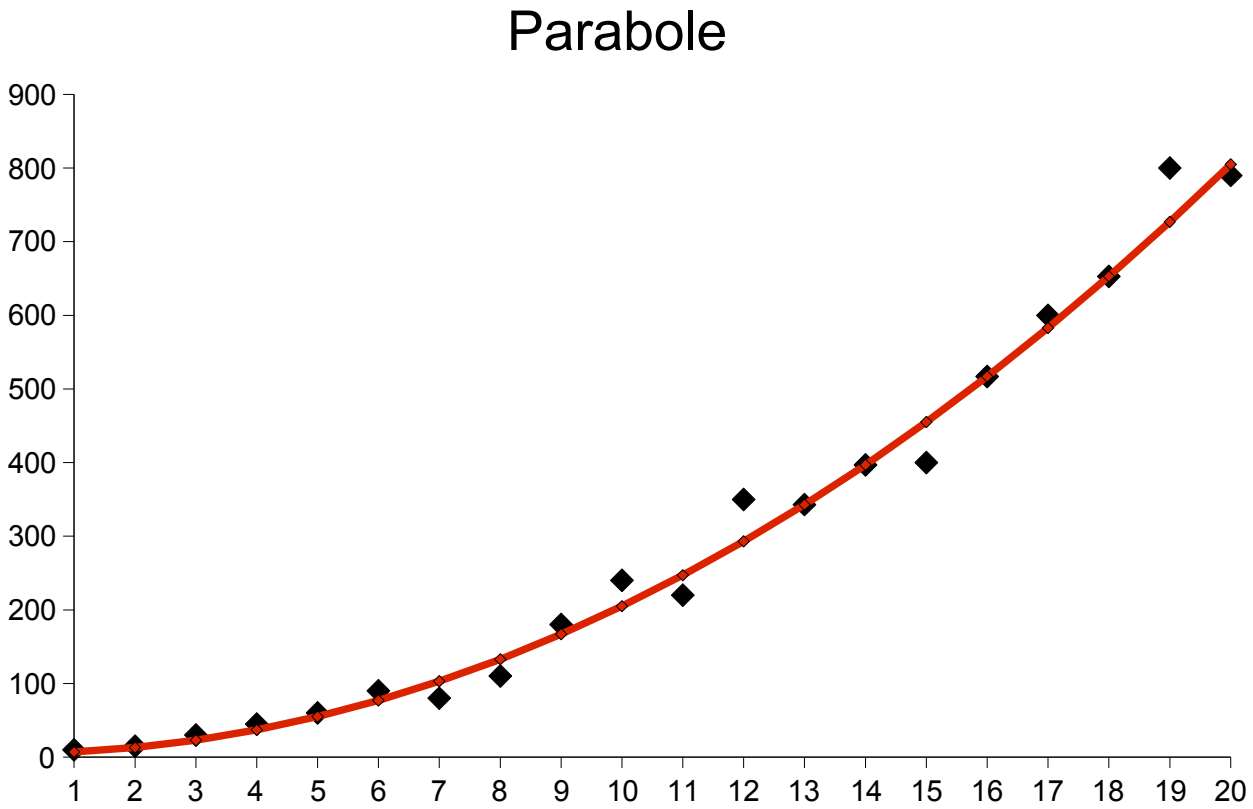
Ajustement non-linéaire par changement de variables

Principe : Se ramener à un ajustement linéaire grâce à un simple changement de variables.

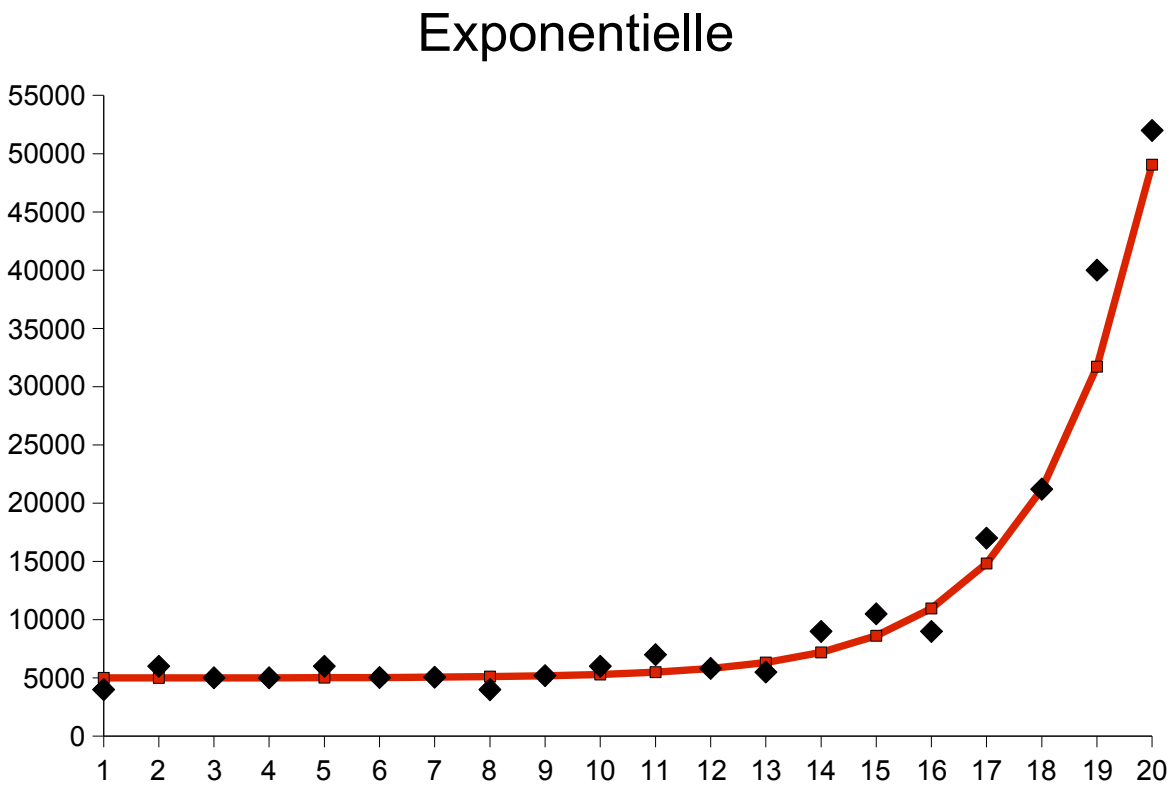
Cette approche est possible pour des ajustements de la forme :

- $y = 1/(ax + b)$,
- $y = b\exp(ax)$,
- $y = \exp(ax + x)$,
- $y = \ln(ax + b)$,
- $= ax^2 + b$,
- ...

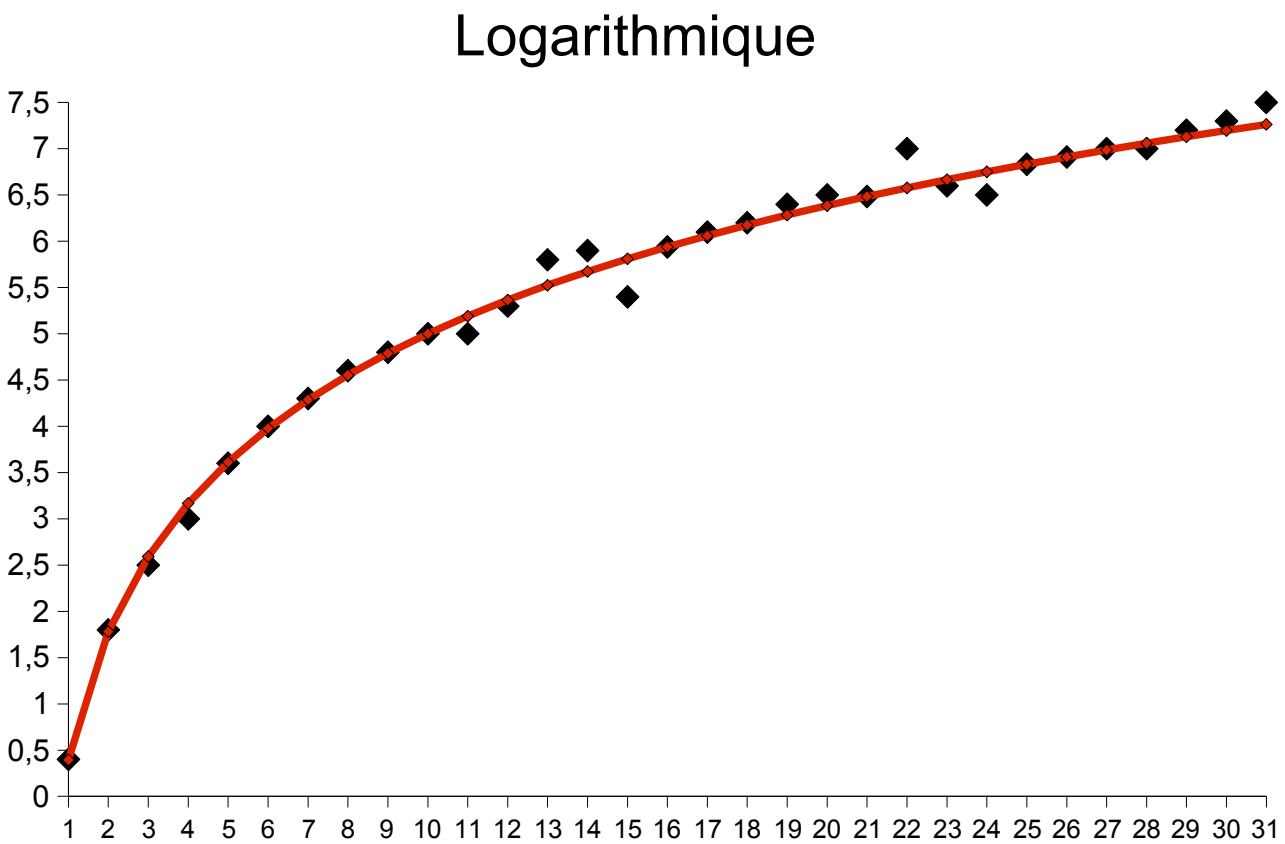
Exemple


Parabole
900
800
700
600
500
400
300
200
100
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

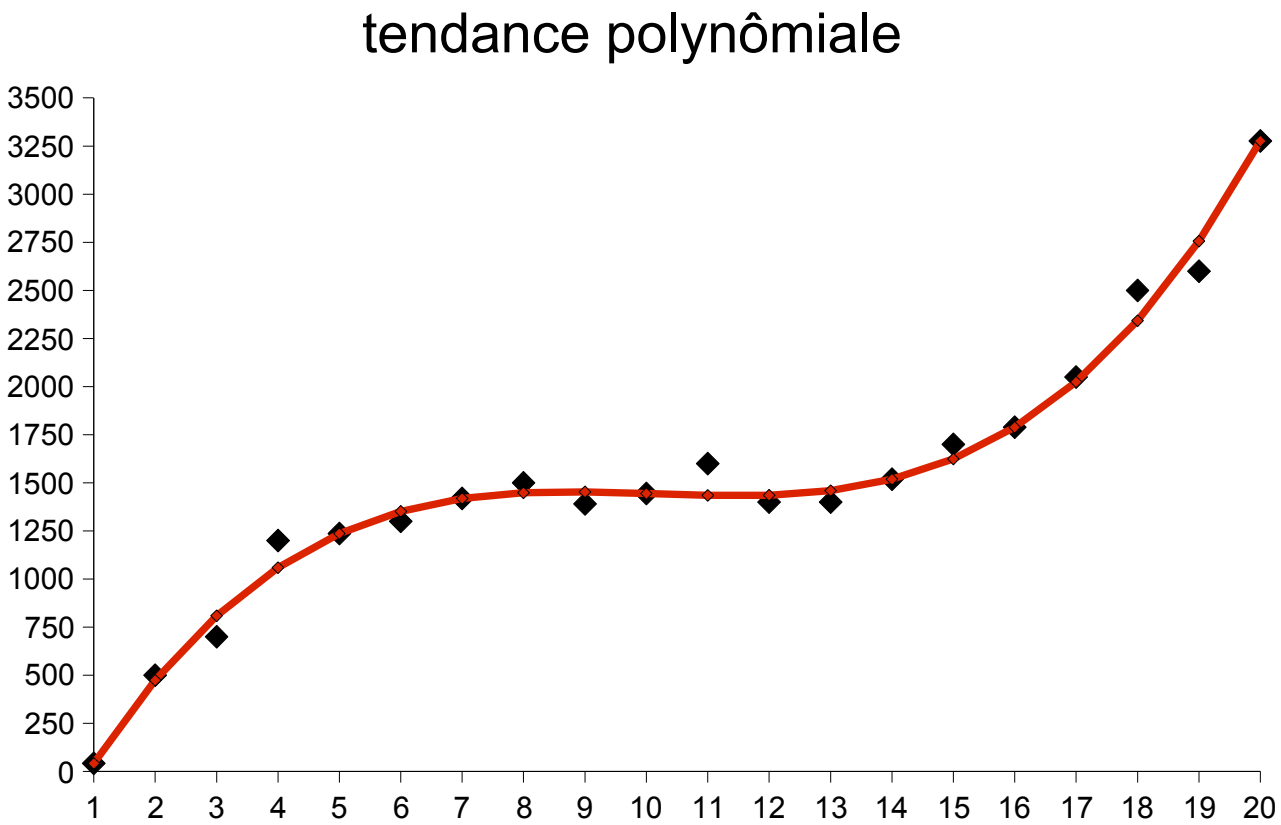
Exemple


Exponentielle
55000
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Exemple


Logarithmique
7,5
7
6,5
6
5,5
5
4,5
4
3,5
3
2,5
2
1,5
1
0,5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Exemple


tendance polynômiale
3500
3250
3000
2750
2500
2250
2000
1750
1500
1250
1000
750
500
250
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20

Exemple

si le nuage de points suggère un ajustement de la forme

$$y = \frac{1}{ax+b}$$

on tranforme les données en posant

$$z_i = \frac{1}{y_i}\ ,\ i = 1, \cdots, n$$

Puis on fait un ajustement linéaire pour le nuage de points (x_i, z_i) de la forme $z = \hat{a}x + \hat{b}$.
Enfin, pour obtenir l'ajustement de la série initiale (y_i), il suffit de prendre :

$$y = \frac{1}{\hat{a}x + \hat{b}}.$$

Exemple

si le nuage de points suggère un ajustement de la forme

$$y = b\,\exp(a\,x),$$

on tranforme les données en posant

$$z_i = \ln(y_i)\,,\ i = 1, \cdots, n$$

Puis on fait un ajustement linéaire pour le nuage de points (x_i, z_i) de la forme $z = \hat{A}x + \hat{B}$.
Si $z = \ln(y)$, on en déduit que
$y = \exp(z) = \exp(\hat{A}x + \hat{B}) = \exp(\hat{B})\exp(\hat{A}x)$
Enfin, pour obtenir l'ajustement de la série initiale (y_i), il suffit de prendre :

$$\hat{b} = \exp\hat{B}\,, \quad \hat{a} = \hat{A} \text{ et } y = \hat{b}\exp(\hat{a}x).$$

Remarque

- La méthode du changement de variable n'est pas toujours possible.
- C'est la représentation graphique du nuage de points qui va nous guider pour le choix d'un changement de variable.