Thème 12
Régression linéaire et exponentielle

I. Régression linéaire

En statistiques, lorsque l'on croise deux variables statistiques, on peut être amené à utiliser un modèle pour **prévoir les valeurs en dehors du nuage des points** générés par ces deux variables. On parle alors de statistiques inférentielles.
Concrètement, si l'on note $x_1, \dots, x_n$ les valeurs de la première variable et $y_1, \dots, y_n$ les valeurs associées de la deuxième variable, on cherche à trouver une fonction f qui va vérifier avec le moins d'erreur possible

$$f(x_i) \approx y_i$$

Ainsi, la modélisation la plus simple, mais pas toujours la plus adaptée, est de chercher une **fonction affine** f définie par

$$f(x) = ax + b$$

en choisissant les réels a et b pour que la droite soit au plus proche du nuage de points. On parle alors de **régression linéaire**. Pour cela, on utilise la **méthode des moindres carrés**.

Prenons un exemple. Regardons la vente de téléphones portables dans le monde jusqu'à 2020 (source ZDNet)..

Année	2017	2018	2019	2020
Vente (en millions d'unités)	1 465,5	1 402,6	1 371	1 200

En général, dans ce cas précis, pour manipuler des nombres plus simples, on dira que **2017 est l'année 1**, **2018 est l'année 2**… Ainsi, si l'on reprend les notation précédentes

$$x_1 = 1\ x_2 = 2\ x_3 = 3\ x_4 = 4\ y_1 = 1\,465,5\ y_2 = 1\,402,6\ y_3 = 1\,371\ y_4 = 1\,200$$

Ce que l'on note généralement

x_i	1	2	3	4
y_i	1 465,5	1 402,6	1 371	1 200

On peut alors représenter ce nuage de points et la **droite d'ajustement** qui s'y rapporte.
Cette droite est tracée de façon à minimiser la somme des carrés des résidus (qui peuvent être négatifs). Ainsi, dans le cas présent,

$$r_1^2 + r_2^2 + r_3^2 + r_4^2$$

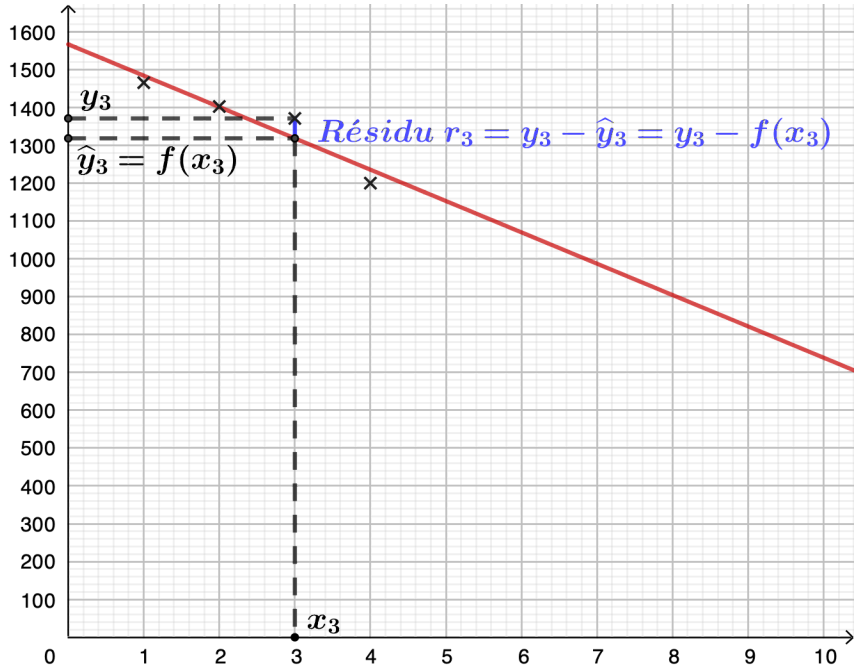
doit être le plus petit possible. On pose

$$r_1 = y_1 - f(x_1) = y_1 - \hat{y}_1$$

$$r_2 = y_2 - f(x_2) = y_2 - \hat{y}_2$$

…

En somme, un **résidu** est la différence entre la valeur observée et la valeur interpolée.
Avec cette contrainte, on peut calculer a et b dans l'expression de la droite d'ajustement.



$$a = \frac{\overline{xy} - \bar{x} \times \bar{y}}{\overline{x^2} - \bar{x}^2}$$

$$b = \bar{y} - a \times \bar{x}$$

$\overline{xy}$ désigne la **moyenne** des produits $x_i \times y_i$

$\bar{x}$ désigne la **moyenne** des valeurs x_i

$\bar{y}$ désigne la **moyenne** des valeurs y_i

$\overline{x^2}$ désigne la **moyenne** des carrés x_i^2

Ainsi, dans l'exemple précédent,

$$\overline{xy} = \frac{x_1 \times y_1 + x_2 \times y_2 + x_3 \times y_3 + x_4 \times y_4}{4}$$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4}{4}$$

$$\bar{y} = \frac{y_1 + y_2 + y_3 + y_4}{4}$$

$$\overline{x^2} = \frac{x_1^2 + x_2^2 + x_3^2 + x_4^2}{4}$$

Après calculs, on trouve $a \approx -82,81$ et $b \approx 1\,566,8$

Remarques

- $\overline{xy} - \bar{x} \times \bar{y}$ est la **covariance** des deux variables.
- $\overline{x^2} - \bar{x}^2$ est la **variance** de la première variable.
- $(\bar{x}; \bar{y})$ est le **point moyen** (qui appartient donc à la droite d'ajustement).

Concrètement, le calcul à la main des coefficients a et b est très fastidieux (il n'est d'ailleurs pas au programme de l'enseignement scientifique d'en connaître les formules). **On utilisera donc systématiquement la calculatrice**.

🖥️ **Pour les calculatrices TI** **https://youtu.be/lC1lfgYE51s**

🖥️ **Pour les calculatrices NumWorks** **https://youtu.be/5BhuuU2Gouk**

Une fois trouvée l'expression de la droite d'ajustement, ici

$$f(x) = -82,81x + 1\,566,8$$

On peut alors **interpoler** les valeurs pour les années suivantes. Dans l'exemple, l'année 2021 correspond à $x = 5$ et l'année 2022 à $x = 6$

Or, $f(5) = -82,81 \times 5 + 1\,566,8 = 1\,152,75$

Et $f(6) = -82,81 \times 6 + 1\,566,8 = 1\,069,94$

On peut donc estimer, si l'on suit ce modèle, que la production de téléphones portables sera 1 152,75 millions en 2021 et de 1 069,94 millions en 2022.

Comment vérifier la validité du modèle ?

Les calculs précédents ne nous permettent pas d'estimer la pertinence du modèle. Un outil permet de la mesurer, c'est la **corrélation** entre les deux variables, elle se calcule à l'aide de la formule

$$\rho = \frac{\overline{xy} - \bar{x} \times \bar{y}}{\left(\overline{x^2} - \bar{x}^2\right)\left(\overline{y^2} - \bar{y}^2\right)}$$

Cette corrélation est comprise entre -1 et 1. Si elle est égale à -1 ou 1, les points sont alignés et plus elle se rapproche de 0 et moins la corrélation est bonne. On utilise souvent en statistiques le **coefficient de détermination** noté R^2 qui, dans le cas d'un ajustement linéaire, égale ρ^2.

Plus R^2 est proche de 1, plus la régression est adaptée.
Plus R^2 est proche de 0, moins la régression est adaptée.

Dans l'exemple précédent, $R^2 \approx 0,887$ est proche de 1 donc il s'agit plutôt d'une bonne régression et l'on peut être tenté de se fier aux prédictions du modèle. Cela dit, le modèle est adapté si le contexte (économique, social, démographique, climatique…) reste identique. Dans le domaine économique, ce contexte peut varier énormément d'un mois à l'autre. Dans l'exemple précédent, les ventes de téléphones portables ont été de 1,35 milliards en 2021, ce qui est contraire aux prédictions du modèle qui implique une diminution des ventes par rapport à 2020 (De 1,2 à 1,15 milliards).

II. Régression exponentielle

Dans certains cas, la régression linéaire n'est pas performante ou peut être encore améliorée. On utilise alors une **régression exponentielle**. La fonction f qui interpole les valeurs est alors du type

$$f(x) = ae^{bx}$$

Pour calculer les coefficients a et b, il faut alors faire une interpolation linéaire entre les valeurs $x_1, \dots, x_n$ et les valeurs $\ln ln(y_1), \dots, \ln ln(y_n)$ où ln désigne le **logarithme néperien**.

Reprenons l'exemple du paragraphe 1. On obtient alors le tableau suivant.

x_i	1	2	3	4
$\ln ln(y_i)$	7,29	7,246	7,223	7,09

A l'aide de la calculatrice, on trouve alors

$$\ln ln(y) = -0,0623x + 7,368$$

Avec un coefficient de détermination $R^2 \approx 0,872$ proche de celui trouvé auparavant (et même moins proche de 1, ce qui laisse suggérer ici que la régression exponentielle est moins adaptée).

On utilise ensuite la propriété suivante, soit y un nombre strictement positif,

$$e^{\ln ln(y)} = y$$

pour écrire

$$e^{\ln ln(y)} = e^{-0,0623x+7,368}$$

On utilise ensuite la propriété des puissances, soient a et b deux nombres réels,

$$e^{a+b} = e^a \times e^b$$

On en déduit que

$$y = e^{7,368} \times e^{-0,0623x}$$

Donc

$$f(x) = 1\,584,46e^{-0,0623x}$$

Or, $f(5) = 1\,584,46e^{-0,0623\times5} = 1\,152,75$
Et $f(6) = 1\,584,46e^{-0,0623\times6} = 1\,090,63$
On peut donc estimer, si l'on suit ce modèle, que la production de téléphones portables sera 1 152,75 millions en 2021 et de 1 090,63 millions en 2022 (on remarquera que là encore l'interpolation encore plus éloignée de la réalité, ce que confirme le coefficient de détermination).

Remarques

- Comme pour la régression linéaire, la calculatrice permet de calculer directement l'expression de la fonction f sans passer par le logarithme. Pour la **calculatrice TI**, le procédé est le même que pour la régression linéaire mais il faut choisir **RégExp** à la place de **Réglin(ax+b)**. Pour la **calculatrice NumWorks**, le procédé est le même que pour la régression linéaire. On peut simplement choisir le type de régression dans l'onglet **Graphique** puis dans l'onglet **régression**. Pour obtenir le **coefficient de détermination** R^2, il suffit, après avoir choisi le type de régression de retourner dans l'onglet **Stats**.
- La calculatrice **NumWorks** permet aussi de calculer directement une prédiction du modèle : elle est aussi disponible dans l'onglet **régression**.
- La calculatrice **NumWorks** permet 10 types de régressions différentes. Rien n'empêche de changer de modèle de régression pour obtenir une meilleure prédiction (c'est-à-dire avec un coefficient de détermination plus proche de 1).