

Compléments aux leçons 6 et 7

Comment présenter ses calculs. Nous donnons ici une façon précise pour résoudre un problème d'ajustement, linéaire ou non. On suggère d'écrire, dans l'ordre indiqué :

- on rappelle les données (x_i, y_i) , on précise qui de x ou y est la variable *explicative* ou la variable *dépendante*, et on écrit quel type de dépendance est recherché (linéaire, exponentiel...). Une façon synthétique d'écrire tout cela consiste à adopter la formule «on cherche une relation de la forme $y = f(x)$ » (ou $x = f(y)$). On suppose dans la suite que x est la variable explicative.
- si le modèle adopté n'est pas linéaire, on écrit précisément le *changement de variables* auquel on a recours. On doit donc définir \tilde{x} et \tilde{y} en fonction de x et y (à l'aide du tableau donné dans le cours ou, mieux, en réfléchissant un peu). On peut alors rassembler dans un tableau à deux lignes les nouvelles données $(\tilde{x}_i, \tilde{y}_i)$ déduites des données initiales (x_i, y_i) par la transformation choisie.
- une fois effectuée la régression linéaire à partir des nouvelles données (à la main ou avec sa machine), on en reporte les résultats en écrivant «la méthode des moindres carrés donne la relation $\tilde{y} = a\tilde{x} + b$, avec un coefficient de corrélation égal à r ».
- on réécrit cette relation à l'aide de x et y , ce qui donne le résultat attendu (le \hat{y} de la feuille 6).
- on dresse la liste des résidus $\varepsilon_i = y_i - f(x_i)$, et on en donne la représentation suivante : on trace le nuage de points défini par (x_i, ε_i) .

Mise en œuvre de la méthode. On traite ici le problème relatif à la taille des bouleaux codominants de l'exercice 2 de la feuille 7.

On cherche une relation de la forme $y = y_{\max}(1 - \gamma \exp(rt))$, y étant le diamètre de l'arbre et t le temps (en années)¹. On suppose de plus connue la valeur de y_{\max} , égale à 65,43.

Si $y = 65,43(1 - \gamma \exp(rt))$ on a $65,43 - y = 65,43\gamma \exp(rt)$, ce qui amène le changement de variable \tilde{t} et $\tilde{y} = \ln(65,43 - y)$, afin d'avoir la dépendance linéaire de \tilde{y} en fonction \tilde{t} suivante : $\tilde{y} = \ln(65,43\gamma) + r\tilde{t}$.

La droite des moindres carrés relative à ce nuage est la droite $\tilde{y} = -0,0296\tilde{t} + 4,406$, avec un coefficient de corrélation égal à $-0,987$ ². On a alors une estimation \hat{y} de y donnée par $\hat{y} = 65,43 - \exp(-0,0296t + 4,406)$.

La liste des opérations précédentes sont reportés dans le tableau suivant, lequel aboutit au tracé des résidus.

t en années	1	20	40	60	80	100	120
y en cm	1,29	22,14	35,69	49,23	56,88	60,43	63,74
$\tilde{y} = \ln(65,43 - y)$	4,16	3,77	3,39	2,79	2,15	1,61	0,52
$\hat{y} = 65,43 - e^{-0,0296t+4,406}$	-14,12	20,10	40,35	51,56	57,76	61,18	63,08
$\varepsilon = y - \hat{y}$	15,4	2,04	-4,66	-2,33	-0,87	-0,75	0,66

On remarque que malgré un coefficient de corrélation très proche de 1 la taille des résidus est très grande par rapport à la grandeur approchée, surtout au début de la croissance. Pourtant l'ajustement de \tilde{y} en fonction de \tilde{t} donne une suite de résidus assez petits : voir un autre exemple avec le modèle exponentiel plus bas.

¹ on a mis ici une constante supplémentaire par rapport au modèle défini dans le cours. Voir plus bas la note importante à ce sujet.

² attention à ne pas en déduire que la régression est bonne : lire absolument plus bas la note à ce sujet.

On peut maintenant représenter les résidus :

Remarques :

sur la définition du modèle «exponentielle saturée» : la définition $y = A(1 - e^{-Bx})$ tirée du livre cité dans la feuille 7 est trop pauvre. On a pu par exemple s'en rendre compte sur l'exemple qu'on vient de traiter : les calculs qui précèdent ont en effet mené à $\hat{y} = 65,43 - \exp(-0,0296t + 4,406) = 65,43(1 - e^{4,406} \exp(-0,0296t))$, qui ne cadre avec le modèle cherché que si on considère que $e^{4,406}$ est proche de 1³ ! On voit donc qu'il est nécessaire d'introduire une nouvelle constante (que nous avons notée γ plus haut), car il n'y a aucune raison que cette constante soit égale à 1.

le coefficient de corrélation n'a pas de sens dans une régression non linéaire. Dans l'exemple traité ci-dessus on a vu qu'en se ramenant à une régression linéaire qui est très bonne (l'examen des résidus de la régression linéaire de \tilde{y} en fonction de t en confirme la qualité), on aboutissait à une régression pas terrible, en tout cas au début (le premier résidu est de l'ordre de 14 tandis que la valeur à approcher est de l'ordre de l'unité, ce qui représente une marge d'erreur de 1400 %). Voilà un autre exemple, dans le modèle exponentiel, à traiter en exercice. On travaille sur les données suivantes

x_i	1	2	3	4	5	6	7	8	9	10
$y_i (\times 10^{10})$	2,649	7,202	19,57	53,21	144,3	393	1070	2903	7895	21470

Tracer le nuage de points correspondant, et constater qu'il est naturel d'attendre une relation de la forme $y = Ae^{Bx}$. Effectuer le changement de variable adéquat, en notant au passage que le coefficient de corrélation est très proche de 1⁴, et vérifier que l'on aboutit à l'estimation $y = e^{22,999752 + 1,000014x}$.

Dresser ensuite la liste $\{\tilde{\varepsilon}_i\}$ des résidus obtenus lors de l'ajustement «auxiliaire» (de $\tilde{y} = \ln(y)$ en fonction de x), et la liste des résidus ε_i . Ce qui permet de comparer la qualité des deux régressions (au point d'abscisse x_i est l'erreur relative, définie comme le rapport ε_i/y_i (resp. $\tilde{\varepsilon}_i/\tilde{y}_i$). Conclure en remarquant que chacun de ces écarts relatifs ε_i/y_i concernant la régression recherchée est environ 25 fois plus grand que l'écart $\tilde{\varepsilon}_i/\tilde{y}_i$ de la régression auxiliaire lui correspondant : on perd donc un ordre de grandeur sur la précision.

On peut donner une formule exprimant le résidu ε_i en fonction de $\tilde{\varepsilon}_i$. Dans ce cas on a en effet $\varepsilon_i = y_i - e^{22,999752 + 1,000014x_i} = e^{\tilde{y}_i} - e^{22,999752 + 1,000014x_i} = e^{\tilde{y}_i} (1 - e^{22,999752 + 1,000014x_i - \tilde{y}_i}) = y_i (1 - e^{-\tilde{\varepsilon}_i})$, qui est très proche de $y_i \tilde{\varepsilon}_i$ lorsque $\tilde{\varepsilon}_i$ est proche de 0 (ce qui doit être le cas si la régression linéaire auxiliaire est bonne !). On peut faire le même type de calculs avec les différents modèles rencontrés.

³une autre façon de faire est d'intégrer cette constante dans un changement de variable $\tilde{t} = t - 4,406/0,0296$; le point est que cela n'a pas de sens dans le cas général : ici t est un âge, et \tilde{t} serait un âge négatif. Le même problème se pose si la variable explicative est la température, une vitesse...

⁴tellement proche que la machine l'arrondi à 1.