

Statistiques - Ajustement de courbes

1 Rappels de Statistiques

1.1 Moyenne, variance, écart-type

Soit une série statistique : x_1, x_2, \dots, x_n (n valeurs)

Moyenne

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Somme des carrés des écarts à la moyenne (sum of squares, SS)

$$SS = \sum_{i=1}^n (x_i - \bar{x})^2$$

Nombre de degrés de liberté (ddl)

ddl = nombre total de valeurs - nombre de valeurs estimées

Pour la somme précédente, on a estimé la moyenne, donc ddl = $n - 1$

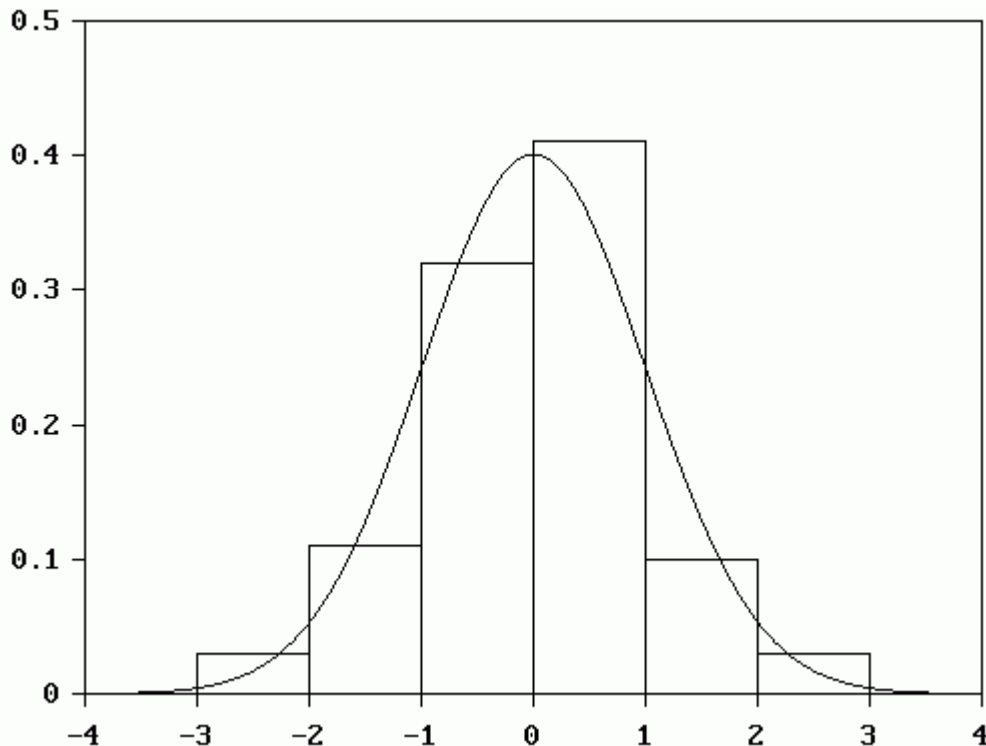
Variance (estimée)

$$\text{Var}(x) = \frac{SS}{ddl} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Ecart-type

$$\sigma(x) = \sqrt{\text{Var}(x)}$$

FIGURE 1 – Histogramme et densité de probabilité



1.2 Histogramme

Si l'on dispose d'un grand nombre de valeurs on peut les regrouper en classes.

Pour chaque classe $[x_i, x_{i+1}]$, on définit :

- l'effectif n_i (nombre de valeurs dans la classe)
- l'amplitude $a_i = x_{i+1} - x_i$
- la fréquence $f_i = n_i/N$ ($N =$ effectif total)
- la densité de fréquence $d_i = f_i/a_i$

L'histogramme peut être tracé en portant la densité de fréquence en fonction des limites de classes (Fig. 1). Dans ces conditions :

- la *surface* de chaque barre est proportionnelle à la fréquence
- l'unité des ordonnées est l'inverse de celle des abscisses
- l'histogramme peut être approché par une courbe continue (densité de probabilité)

1.3 Lois de probabilités

1.3.1 Loi normale (courbe de Gauss)

Caractérisée par sa moyenne μ et son écart-type σ : $\mathcal{N}(\mu, \sigma)$.

La densité de probabilité est donnée par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

La probabilité pour que la variable x soit comprise entre deux valeurs a et b est donnée par l'intégrale (aire sous la courbe) :

$$\text{Prob}(a < x < b) = \int_a^b f(x)dx$$

Exemples :

$$\text{Prob}(\mu - 2\sigma < x < \mu + 2\sigma) \approx 0,95$$

$$\text{Prob}(\mu - 3\sigma < x < \mu + 3\sigma) \approx 0,99$$

La probabilité totale est égale à 1 :

$$\text{Prob}(-\infty < x < +\infty) = 1$$

Cas particulier : loi normale réduite : $\mathcal{N}(0, 1)$ (Fig. 1)

1.3.2 Loi de Student

Dépend du nombre de degrés de liberté (ddl)

La courbe ressemble à celle de la loi normale mais d'autant plus élargie que le ddl est faible (Fig. 2)

Pour $ddl \geq 30$ on retrouve pratiquement la loi normale réduite.

1.3.3 Loi de Fisher-Snedecor

Correspond à la distribution d'un rapport de variances.

Caractérisée par deux ddl

La courbe est bornée à 0 et n'est pas symétrique (Fig. 3)

FIGURE 2 – Loi de Student : $ddl = 30$ (trait plein), 5 (tirets), 2 (pointillés)

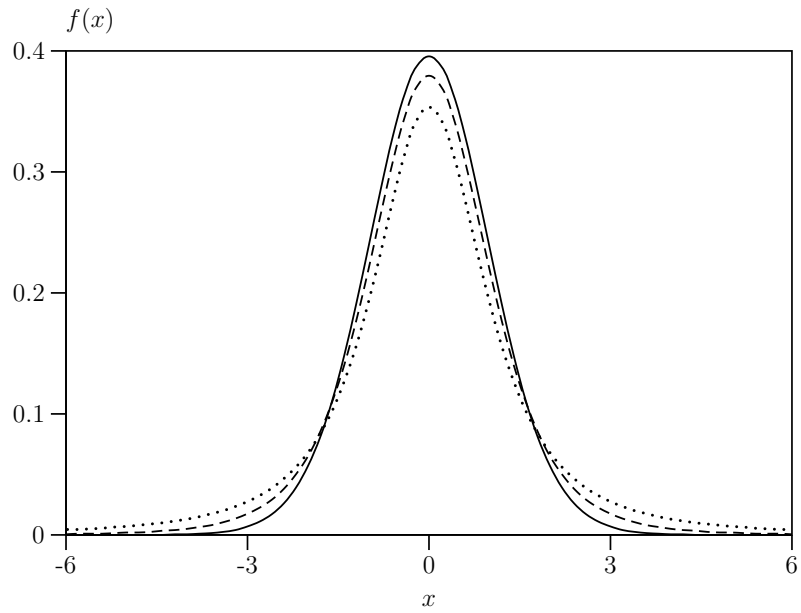
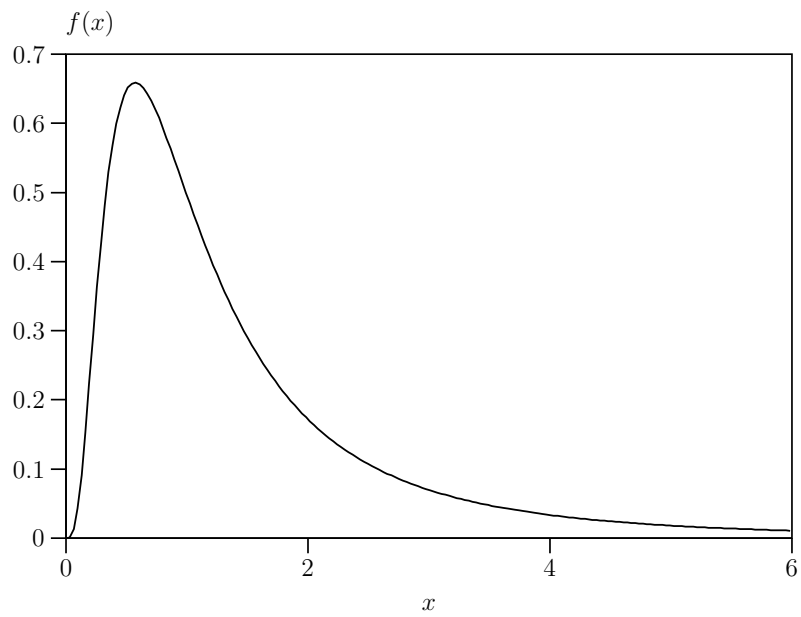


FIGURE 3 – Loi de Fisher-Snedecor : $ddl_1 = 10, ddl_2 = 5$



2 Régression linéaire

2.1 Ajustement d'une droite

Le problème consiste à déterminer l'équation de la droite qui passe le plus près possible d'un ensemble de points.

Le modèle est défini par l'équation :

$$y = \beta_0 + \beta_1 x$$

- x est la variable indépendante (ou « explicative »), supposée connue sans erreur
- y est la variable dépendante (ou « expliquée »), entachée d'une erreur de mesure σ
- β_0 et β_1 sont les paramètres du modèle (valeurs théoriques)

Supposons que les n points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ soient parfaitement alignés (Fig. 1A), de sorte que chacun d'eux vérifie l'équation de la droite :

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_1 \\ y_2 &= \beta_0 + \beta_1 x_2 \\ \dots &\dots \dots \dots \dots \\ y_n &= \beta_0 + \beta_1 x_n \end{aligned}$$

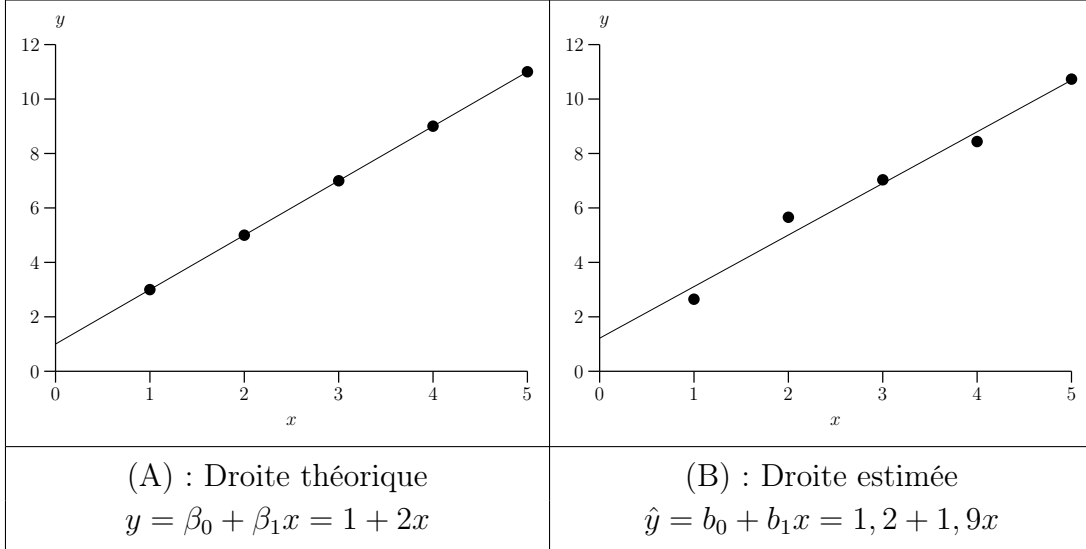
ou, sous forme matricielle :

$$\mathbf{y} = \mathbf{X}\beta$$

avec :

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

FIGURE 4 – Régression linéaire



En général, les points ne sont pas parfaitement alignés (Fig. 1B), de sorte que :

$$\begin{aligned}
 y_1 &= \beta_0 + \beta_1 x_1 + \epsilon_1 \\
 y_2 &= \beta_0 + \beta_1 x_2 + \epsilon_2 \\
 \dots &\dots \dots\dots\dots\dots\dots\dots\dots \\
 y_n &= \beta_0 + \beta_1 x_n + \epsilon_n
 \end{aligned}$$

Il y a n équations et $(n+2)$ inconnues (β_0, β_1 et les n écarts ϵ_i) : le système est donc indéterminé.

On va estimer le vecteur β par un vecteur \mathbf{b} . On aura donc une droite estimée :

$$\hat{y} = b_0 + b_1 x$$

Les ϵ_i définissent un vecteur de résidus :

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \dots \\ \epsilon_n \end{bmatrix}$$

On calcule \mathbf{b} pour que $\|\epsilon\|$ soit minimal (*critère des moindres carrés*).

$$\|\epsilon\|^2 = \epsilon_1^2 + \epsilon_2^2 + \dots + \epsilon_n^2 = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SS_r$$

où SS_r est la *somme des carrés des écarts résiduelle*.

On montre que \mathbf{b} est solution du système :

$$\mathbf{A}\mathbf{b} = \mathbf{c}$$

avec :

$$\mathbf{A} = \mathbf{X}^\top \mathbf{X} \quad \mathbf{c} = \mathbf{X}^\top \mathbf{y}$$

soit :

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{y})$$

2.2 Analyse de la variance

L'équation suivante est vérifiée :

$$SS_t = SS_e + SS_r \tag{1}$$

avec :

$$SS_t = \sum_{i=1}^n (y_i - \bar{y})^2 \quad SS_e = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad SS_r = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

– \bar{y} est la moyenne des valeurs de y :

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

- SS_t est la *somme des carrés des écarts totale* ; elle possède $(n-1)$ degrés de liberté
- SS_e est la *somme des carrés des écarts expliquée* ; elle possède 1 degré de liberté.
- SS_r est la *somme des carrés des écarts résiduelle* ; elle possède $(n-2)$ degrés de liberté

Notons que les degrés de liberté (d.d.l.) s'additionnent, tout comme les sommes de carrés d'écarts :

$$(n-1) = 1 + (n-2)$$

Les *variances* s'obtiennent en divisant chaque somme de carrés d'écart par le nombre de d.d.l. correspondants :

$$V_t = \frac{SS_t}{n-1} \quad V_e = SS_e \quad V_r = \frac{SS_r}{n-2}$$

Ce sont respectivement les variances totale, expliquée, et résiduelle. (Ces variances ne s'additionnent pas !)

On en déduit les quantités suivantes :

- le **coefficient de détermination** r^2

$$r^2 = \frac{SS_e}{SS_t}$$

r^2 représente le pourcentage des variations de y qui est « expliqué » par la variable indépendante. Il est toujours compris entre 0 et 1. Une valeur de 1 indiquerait un ajustement parfait.

- le **coefficient de corrélation** r

C'est la racine carrée du coefficient de détermination, affecté du signe de la pente b_1 . Il est toujours compris entre -1 et 1.

- l' **écart-type résiduel** s_r

C'est la racine carrée de la variance résiduelle ($s_r = \sqrt{V_r}$). C'est une estimation de l'erreur faite sur la mesure de la variable dépendante y . Une valeur de 0 indiquerait un ajustement parfait.

- le **rapport de variance** F

C'est le rapport de la variance expliquée à la variance résiduelle ($F = V_e/V_r$). Il serait infini dans le cas d'un ajustement parfait.

2.3 Précision des paramètres

La matrice :

$$\mathbf{V} = V_r \cdot \mathbf{A}^{-1} = V_r \cdot (\mathbf{X}^\top \mathbf{X})^{-1}$$

est appelée **matrice de variance-covariance** des paramètres. C'est une matrice symétrique dont la structure est la suivante :

$$\mathbf{V} = \begin{bmatrix} \text{Var}(b_0) & \text{Cov}(b_0, b_1) \\ \text{Cov}(b_0, b_1) & \text{Var}(b_1) \end{bmatrix}$$

Les termes diagonaux sont les variances des paramètres, à partir desquelles on calcule les écart-types :

$$s_0 = \sqrt{\text{Var}(b_0)} \quad s_1 = \sqrt{\text{Var}(b_1)}$$

Le terme non-diagonal est la covariance des deux paramètres, d'où l'on tire le coefficient de corrélation r_{01} :

$$r_{01} = \frac{\text{Cov}(b_0, b_1)}{s_0 s_1}$$

2.3.1 Présentation des résultats

Les résultats sont habituellement présentés sous la forme :

$$y = (b_0 \pm s_0) + (b_1 \pm s_1)x$$

Les écart-types sont donnés avec 1 ou 2 chiffres significatifs. Chaque paramètre est donné avec autant de décimales que son écart-type.

Rappel : les chiffres significatifs sont comptés à partir du premier chiffre non nul. (Ex. : $1,234 \pm 0,012$ ou $1,23 \pm 0,01$)

2.4 Interprétation probabiliste

On suppose que les résidus $\epsilon_i = (y_i - \hat{y}_i)$ sont identiquement et indépendamment distribués selon une distribution normale de moyenne 0 et d'écart-type σ (estimé par s_r).

On montre alors que les paramètres (b_0, b_1) sont distribués selon une loi de Student à $(n - 2)$ d.d.l.

On peut dès lors calculer un intervalle de confiance pour chaque paramètre, par exemple :

$$\left[b_0 - t_{1-\alpha/2} \cdot s_0 \quad , \quad b_0 + t_{1-\alpha/2} \cdot s_0 \right]$$

où $t_{1-\alpha/2}$ est la valeur de la variable de Student correspondant à la probabilité choisie (habituellement $\alpha = 0,05$). Cet intervalle a une probabilité $(1 - \alpha)$ de contenir la valeur théorique β_0 .

On peut aussi calculer une valeur « critique » $F_{1-\alpha}$ à partir de la distribution de Fisher-Snedecor à 1 et $(n - 2)$ d.d.l. L'ajustement peut être considéré comme satisfaisant si le rapport de variance F excède 4 fois cette valeur critique.

Note : pour une droite de régression, $F_{1-\alpha} = (t_{1-\alpha/2})^2$

2.5 Tests d'hypothèses

- On pose une *hypothèse nulle* (H_0) concernant la droite théorique
- On calcule, sous H_0 , la probabilité α d'obtenir les paramètres observés
- Si cette probabilité est suffisamment faible, on rejette H_0 au risque α

2.5.1 Test de l'ordonnée à l'origine

$H_0 : \beta_0 = 0$ (la droite théorique passe par l'origine)

Sous H_0 , la variable :

$$b_0^* = \frac{b_0 - \beta_0}{s_0} = \frac{b_0}{s_0}$$

suit une loi de Student à $(n - 2)$ d.d.l.

On calcule, sous H_0 , $\alpha = \text{Prob}(|T| \geq b_0^*)$, T étant la variable de Student.

2.5.2 Test de la pente

$H_0 : \beta_1 = 0$ (la droite théorique est horizontale : y ne dépend pas de x)

Sous H_0 , la variable :

$$b_1^* = \frac{b_1 - \beta_1}{s_1} = \frac{b_1}{s_1}$$

suit une loi de Student à $(n - 2)$ d.d.l.

On calcule, sous H_0 , $\alpha = \text{Prob}(|T| \geq b_1^*)$, T étant la variable de Student.

2.5.3 Test de F

$H_0 : V_e = V_r \iff F = 1$

Sous H_0 , F suit une loi de Fisher-Snedecor à 1 et $(n - 2)$ d.d.l.

On calcule, sous H_0 , $\alpha = \text{Prob}(F \geq F_{obs})$

3 Régression multilinéaire

3.1 Equations normales

Le modèle de régression est :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$$

où les x_i sont m variables (en principe indépendantes).

La méthode des équations normales est encore valable avec :

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \cdots \\ b_m \end{bmatrix}$$

$$\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{y})$$

Il y a $p = m + 1$ paramètres. Le nombre d'observations n doit être $> p$.

Cas particulier : Les x_i peuvent être des fonctions d'une autre variable x , pour autant que ces fonctions ne contiennent pas de paramètres.

Exemples :

- Polynôme : $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots$
- Série de Fourier : $y = \beta_0 + \beta_1 \sin x + \beta_2 \sin 2x + \cdots$

3.2 Analyse de la variance

L'équation $SS_t = SS_e + SS_r$ peut être appliquée avec les modifications suivantes :

- la somme des carrés des écarts expliquée SS_e a $(p-1)$ degrés de liberté.
- la somme des carrés des écarts résiduelle SS_r a $(n-p)$ degrés de liberté.

Notons que les degrés de liberté s'ajoutent encore :

$$(n-1) = (p-1) + (n-p)$$

Les variances expliquée et résiduelle deviennent :

$$V_e = \frac{SS_e}{p-1} \quad V_r = \frac{SS_r}{n-p}$$

Les quantités r^2, s_r, F s'en déduisent comme dans le chapitre précédent, sauf qu'ici le coefficient de corrélation r est toujours positif.

En régression multilinéaire, l'utilisation de r^2 peut être trompeuse car il est toujours possible d'en augmenter artificiellement la valeur en ajoutant des variables ou en utilisant un polynôme de plus haut degré. Pour surmonter ce problème, on a défini le **coefficient de détermination ajusté** :

$$r_a^2 = 1 - (1 - r^2) \frac{n-1}{n-p}$$

3.3 Précision des paramètres

La matrice de variance-covariance $\mathbf{V} = V_r \mathbf{A}^{-1}$ (avec $\mathbf{A} = \mathbf{X}^\top \mathbf{X}$) est une matrice symétrique $p \times p$ telle que :

- le terme diagonal V_{ii} est la variance du paramètre b_i
 - le terme non-diagonal V_{ij} est la covariance des paramètres b_i et b_j
- Le coefficient de corrélation des paramètres b_i et b_j se calcule par :

$$r_{ij} = \frac{V_{ij}}{\sqrt{V_{ii}V_{jj}}}$$

3.4 Interprétation probabiliste

Si l'on suppose que les résidus sont identiquement et indépendamment distribués selon une distribution normale, les paramètres de la régression suivent une distribution de Student à $(n-p)$ d.d.l. Les intervalles de confiance et les tests d'hypothèses s'en déduisent comme pour la régression linéaire.

La valeur « critique » $F_{1-\alpha}$ se déduit de la distribution de Fisher-Snedecor à $(p-1)$ et $(n-p)$ d.d.l. Cependant, la relation $F_{1-\alpha} = (t_{1-\alpha/2})^2$ n'est plus vérifiée si $p > 2$.

4 Régression non linéaire

4.1 Introduction

On considère ici des modèles qui sont non linéaires *par rapport aux paramètres*. Par exemple, le modèle exponentiel $y = ae^{-bx}$ est non-linéaire par rapport au paramètre b .

4.2 Linéarisation

Une transformation de variable permet souvent de linéariser le modèle. Par exemple, pour le modèle exponentiel :

$$\ln y = \ln a - bx$$

On peut dès lors appliquer la régression linéaire avec comme paramètres $\ln a$ et b .

Mais cette transformation modifie l'écart-type de la variable indépendante :

$$\sigma(\ln y) \approx d \ln y = \frac{dy}{y} \approx \frac{\sigma(y)}{y}$$

L'utilisation de la régression linéaire n'est donc justifiée en toute rigueur que si l'écart-type de la variable transformée est constant. Dans l'exemple choisi, cela suppose que $\sigma(\ln y)$ est constant, donc que $\sigma(y)$ est proportionnel à y .

Dans le cas contraire, les paramètres déterminés par linéarisation ne peuvent être considérés que comme des estimations préliminaires qu'il convient d'affiner par approximations successives.

4.3 Affinement des paramètres

Dans le cas général, le modèle de régression est de la forme :

$$y = f(x; a, b \dots)$$

où f est une fonction non linéaire des paramètres $a, b \dots$

Supposons que l'on ait une première estimation $(a^0, b^0 \dots)$ des paramètres, obtenue par exemple par linéarisation. Ecrivons le développement limité de y au voisinage de cette estimation :

$$y = y^0 + y'_a \cdot (a - a^0) + y'_b \cdot (b - b^0) + \dots$$

où :

$$\begin{aligned} y^0 &= f(x; a^0, b^0 \dots) \\ y'_a &= \frac{\partial f}{\partial a}(x; a^0, b^0 \dots) \\ y'_b &= \frac{\partial f}{\partial b}(x; a^0, b^0 \dots) \\ &\dots \end{aligned}$$

L'équation peut se mettre sous la forme :

$$y - y^0 = y'_a \cdot (a - a^0) + y'_b \cdot (b - b^0) + \dots$$

qui correspond au problème de régression multilinéaire :

$$\mathbf{z} = \mathbf{J} \cdot \delta$$

avec :

$$\mathbf{z} = \begin{bmatrix} y_1 - y_1^0 \\ y_2 - y_2^0 \\ \dots \\ y_n - y_n^0 \end{bmatrix} \quad \mathbf{J} = \begin{bmatrix} y'_{a1} & y'_{b1} & \dots \\ y'_{a2} & y'_{b2} & \dots \\ \dots & \dots & \dots \\ y'_{an} & y'_{bn} & \dots \end{bmatrix} \quad \delta = \begin{bmatrix} a - a^0 \\ b - b^0 \\ \dots \end{bmatrix}$$

où \mathbf{J} est la *matrice Jacobienne*, telle que $y'_{ai} = \partial f(x_i; a^0, b^0 \dots) / \partial a$ etc.

L'application des formules de la régression linéaire conduit à :

$$\delta = (\mathbf{J}^\top \mathbf{J})^{-1} (\mathbf{J}^\top \mathbf{z}) \quad (2)$$

Connaissant le vecteur des corrections δ , il est possible de calculer de nouvelles estimations $a, b \dots$ des paramètres. Le processus est répété jusqu'à convergence des estimations.

La méthode ainsi décrite est connue sous le nom de *méthode de Gauss-Newton*.

4.4 Analyse de la variance

La variance résiduelle est :

$$V_r = \frac{SS_r}{n - p}$$

où p désigne le nombre de paramètres du modèle.

L'équation d'analyse de la variance n'est pas vérifiée pour les modèles non linéaires :

$$SS_t \neq SS_e + SS_r$$

Il est toujours possible de calculer r^2 et F , ainsi que les intervalles de confiance des paramètres, mais leur interprétation est plus délicate. En particulier, r^2 peut être > 1 ! Par ailleurs, la distribution des paramètres ne suit qu'approximativement la loi de Student.