



MODULE METHODOLOGIQUE M3

Ajustement linéaire aux moindres carrés

AJUSTEMENT LINEAIRE AUX MOINDRES CARRES

Cécile Mallet

SOMMAIRE

SOMMAIRE.....	2
Statistique Descriptive	3
3. Régression linéaire aux moindres carrés	3
Le critère des moindres carrés : ajustement affine de Y en X.....	5
Qualité de l'ajustement	6
Statistique Inférentielle.....	7
1. Modèle linéaire simple	7
1.1. Calcul des espérances de a et b connaissant les valeurs de x_i	8
1.2 Calcul des variances de a et b connaissant les valeurs de x_i	9
1.3 Intervalle de confiance <u>hypothèse ε suit une loi $N(0, \sigma)$</u>	10
1.4 Test d'hypothèse	10
Test relatif à la corrélation	10
Test relatif à la régression	11
1.5 Intervalle de prévisibilité	11
grands échantillons.....	12
1.6 Evaluation de l'ajustement	12
Interpretation du coefficient de corrélation r.....	13
Interprétation descriptive ou déterministe.....	13
Interprétation inférentielle ou statistique	13



[Licence Creative Commons](https://creativecommons.org/licenses/by-nc-sa/4.0/)

STATISTIQUE DESCRIPTIVE

3. Régression linéaire aux moindres carrés

On considère ici un ensemble d'observations constitué des mesures X et Y de deux grandeurs réelles.

Lorsque les deux grandeurs sont liées par une liaison linéaire (affine), le coefficient de corrélation est un indicateur global de cette relation. Il arrive que l'on souhaite des renseignements plus précis : peut-on attribuer à X une valeur prédictive ou explicative et obtenir des renseignements sur les valeurs possibles de Y connaissant X ? Dans le cas de la régression, contrairement à l'étude de la corrélation, les variables X et Y ne sont pas interchangeables.

X est appelée variable explicative ou prédicteur c'est une variable indépendante.

Y est appelée variable estimée ou expliquée par X c'est une variable dépendante.

Ce ne sont pas les statistiques mais la connaissance du problème qui permet de savoir quelle est la valeur explicative et quelle est la valeur expliquée.

Par exemple si X est une mesure de la température et Y est une mesure de la longueur d'une tige en métal. On va étudier l'influence de la température sur la longueur et non l'influence de la longueur d'une tige métallique sur la température. De même, on peut étudier l'influence de l'ancienneté sur le salaire mais une augmentation de salaire ne fait pas augmenter l'ancienneté. Ou encore si on considère le taux précipitant X et une mesure Y. La pluie influence la réflectivité radar et non le contraire, ...

Si deux variables sont conjointement influencées par une cause commune cette étude ne présente pas d'intérêt. Par exemple, soit X variable relative à l'achat de bijoux et y variable relative à l'achat de matériel vidéo, X et Y sont influencées par le revenu Z mais il n'y a pas de relation de cause à effet.

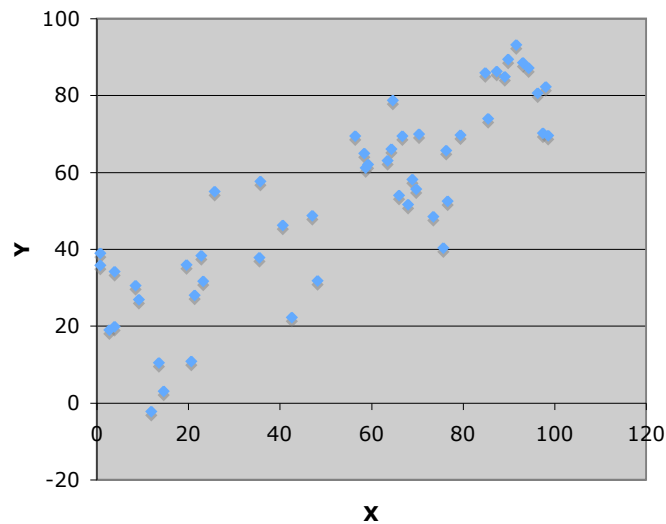
Supposons qu'il existe une relation fonctionnelle du type $Y = aX + b$ entre deux variables X et Y. On peut alors calculer la valeur de l'une connaissant l'autre.

Plusieurs situations sont envisageables :

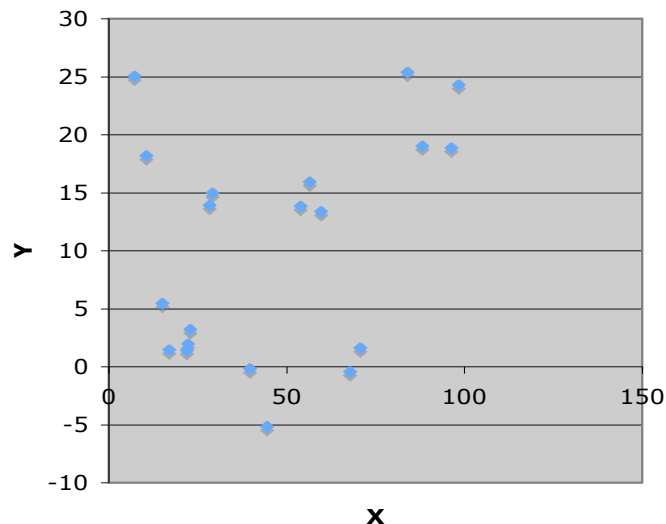
- Les coefficients a et b sont des constantes connues on peut alors calculer Y pour différentes valeurs de la température X. Par exemple : $Y = 1.345 X + 73.6$
- Seule la forme linéaire peut être justifiée théoriquement et les valeurs des coefficients a et b ne sont pas connus. On observe alors pour différentes valeurs de X la valeur correspondante de Y. Les valeurs des coefficients a et b peuvent être déterminées à partir des couples (x_i, y_i) observés.

On obtient ensuite : $Y = aX + b$

- Il existe une relation fonctionnelle non linéaire entre X et Y mais un changement de variable peut la transformer en une relation linéaire. On sait par exemple que la réflectivité radar Z observée par un radar météorologique en présence de précipitation dont le taux précipitant est R est de la forme $Z = a R^b$. Si on pose $Y = \log(Z)$ et $X = \log(R)$ et si on observe pour différentes valeurs du taux précipitant R, la réflectivité Z correspondant, les valeurs des coefficients $\log(a)$ et b peuvent être déterminés à partir des couples $(\log(R_i), \log(Z_i))$ observés. Par la suite comme dans les exemples précédents on pourra déterminer la valeur de la réflectivité à partir du taux précipitant.
- Dans la situation ci-dessous, même si on observe ici une tendance : Il existe une corrélation linéaire entre X et Y: Y croît, en général, lorsque X augmente. Mais il ne s'agit pas d'une relation fonctionnelle et l'on ne peut pas déduire Y de X à partir d'une expression linéaire exacte. On pourra néanmoins utiliser la relation linéaire pour indiquer une tendance et faire certaines prévisions.



- Dans la situation ci-dessous, le nuage de points est très dispersé et il n'y a pas de relation entre les variables statistiques. Il ne semble pas possible de faire passer une courbe simple.



On cherche, dans un ensemble de fonctions, la fonction de X qui approche Y le mieux possible au sens d'un certain critère ; on dit que l'on fait la régression de Y sur X. Si l'on choisit les fonctions affines ($f(X) = aX + b$), on parle de régression linéaire. Le critère le plus usuel est celui des moindres carrés. La régression par rapport à la corrélation présente l'avantage de pouvoir se généraliser à la régression non linéaire. Pour certaines familles de fonctions on va opérer une transformation de manière à se ramener à une régression linéaire. Par exemple :

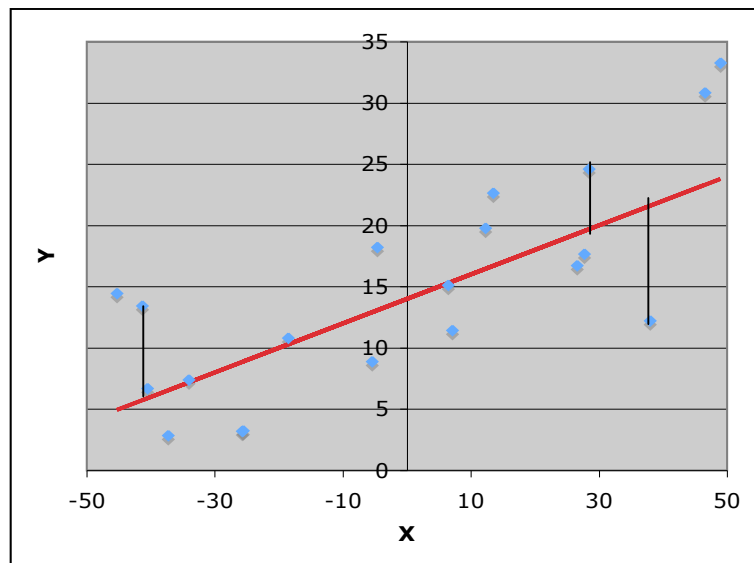
Famille	Fonction	Transformation	Régression linéaire
exponentielle	$Y = a \exp(bX)$	$Y' = \ln(Y)$	$Y' = \ln(a) + bX$
puissance	$Y = a X^b$	$Y' = \ln(Y)$ $X' = \ln(X)$	$Y' = \ln(a) + bX'$
Inverse	$Y = a + b/X$	$X' = 1/X$	$Y = a + bX'$
Logistique	$Y = \frac{1}{1 + e^{-(aX+b)}}$	$Y' = -\ln(1/Y - 1)$	$Y' = aX + b$

Le critère des moindres carrés : ajustement affine de Y en X

On considère que l'éloignement entre un point et une droite est la distance verticale entre le point et la droite. On considère de plus que l'éloignement du nuage de points à la droite est la moyenne de la somme des carrés des éloignements des points du nuage. On cherche à optimiser une relation $Y=aX+b$ à partir des données dont on dispose. On cherche donc la pente a et l'ordonnée à l'origine b pour que la droite résume au mieux le nuage de points. Les couples (x_i,y_i) étant les données et les coefficients a et b les inconnues qui vont être obtenues en minimisant les écarts quadratiques entre les observations y_i et les valeurs ax_i+b :

$$J(a,b) = \sum_{i=1}^n (y_i - (ax_i + b))^2$$

$y_i - (ax_i + b)$ représente la distance verticale du point figurant l'individu ω_i à la droite $y=ax+b$



En écrivant que les coefficients a et b qui minimisent J vérifient les équations normales $\begin{cases} \frac{\partial J}{\partial a} = 0 \\ \frac{\partial J}{\partial b} = 0 \end{cases}$

On obtient la solution unique suivante $\begin{cases} \hat{b} = \bar{y} - \hat{a}\bar{x} \\ \hat{a} = \frac{c_{XY}}{s_X^2} \end{cases}$

On note que $\hat{a} = \frac{c_{XY}}{s_X s_Y} \frac{s_Y}{s_X} = r \frac{s_Y}{s_X}$, dans le cas des variables réduites $\hat{a} = r$

La droite des moindres carrés passe par le barycentre du nuage de points de coordonnées (\bar{x}, \bar{y})

Les valeurs $\hat{y}_i = \hat{a}x_i + \hat{b}$ sont appelées valeurs ajustées ;

Par rapport au nuage de points la droite de régression linéaire de Y en fonction de X minimise la somme des carrés des distances verticales des points à la droite. La droite de régression de X en fonction de Y minimise la somme des carrés des distances horizontales. Les deux droites se coupent au centre de gravité du nuage de points. L'écart entre les deux droites est d'autant plus grand que la corrélation est faible.

Qualité de l'ajustement

Les valeurs $\hat{e}_i = y_i - \hat{y}_i$ sont appelées résidus.

Ils sont de moyenne nulle c'est-à-dire que les valeurs ajustées $\hat{y}_i = \hat{a}x_i + \hat{b}$ ont la même moyenne que les y_i , on a donc $\overline{\hat{y}} = \bar{y}$. La moyenne des résidus est donc nulle.

La variance des $e_i = \frac{1}{n} J(\hat{a}, \hat{b})$ est la grandeur qui a été minimisée c'est la variance résiduelle de la régression

$y_i - \bar{y}$ l'écart total de la donnée à la moyenne des données

$\hat{y}_i - \bar{y}$ écart expliqué (par les variations de x)

$y_i - \hat{y}_i$ écart inexpliqué ou résiduel

On a l'écart total qui est la somme de l'écart expliqué et inexpliqué (résidus)

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

On peut montrer que la variance totale de Y est la somme de la variance expliquée et de la variance inexpliquée (résiduelle)

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Le **coefficient de détermination** R^2 est le rapport entre variance expliquée et variance totale, il mesure la proportion de la variable Y qui s'explique par les variations de X

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Si on note r le coefficient de corrélation entre y et x ou entre y et \hat{y} on a $r = \pm(R^2)^{0.5}$

Plus r est proche de 1 en valeur absolue plus la relation linéaire est forte et plus r est proche de 0 plus la relation linéaire est faible.

STATISTIQUE INFÉRENTIELLE

1. Modèle linéaire simple

On considère que les observations x_i y_i sont les réalisations d'un couple X, Y de variables aléatoires qui sont en liaison linéaire (exacte ou approchée)

$$Y = \alpha X + \beta + \varepsilon$$

α et β sont des constantes. ε est une variable aléatoire écart que l'on cherche à minimiser.

X : variable déterministe explicative

Y : variable aléatoire réelle à expliquer

Les coefficients de la droite de régression des moindres carrés minimisent l'écart 'en moyenne quadratique' :

$$E(\varepsilon) = 0$$

$$V(\varepsilon) = \sigma^2$$

D'où $E(Y) = \alpha X + \beta$ est une fonction affine de X

On suppose que les x_i sont connus sans erreur les observations y_i sont les réalisations de Y ,

$$y_i = \alpha x_i + \beta + \varepsilon_i$$

les y_i sont les valeurs prises par n variables aléatoires indépendantes

$$\text{d'espérance} \quad E(y_i) = E(\alpha x_i + \beta + \varepsilon_i) = \alpha x_i + \beta$$

$$\text{et de variance} \quad V(y_i) = V(\alpha x_i + \beta + \varepsilon_i) = \sigma^2$$

Il est alors possible de considérer que a et b obtenus à partir des équations normales (chapitre 1) sont des estimateurs de α et β dont on va pouvoir déterminer à partir des statistiques l'espérance et la variance. Si on ajoute l'hypothèse que les ε_i suivent une loi gaussienne, les intervalles de confiance pourront être définis.

$$a = \hat{\alpha}$$

$$b = \hat{\beta}$$

1.1. Calcul des espérances de a et b connaissant les valeurs de x_i

Hypothèse $E(\varepsilon_i) = 0$

$$a = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$E(a) = \frac{\sum_{i=1}^n (E(y_i) - E(\bar{y}))(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$E(y_i) = E(\alpha x_i + \beta + \varepsilon_i) = \alpha x_i + \beta + E(\varepsilon_i) = \alpha x_i + \beta$$

$$E(\bar{y}) = \frac{1}{n} \sum_{i=1}^n E(y_i) = \alpha \frac{1}{n} \sum_{i=1}^n x_i + \beta = \alpha \bar{x} + \beta$$

$$E(y_i) - E(\bar{y}) = \alpha(x_i - \bar{x})$$

$$E(a) = \frac{\sum_{i=1}^n \alpha(x_i - \bar{x})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \alpha$$

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

$$E(b) = E(\bar{y}) - E(a)\bar{x} = \alpha\bar{x} + \beta - \alpha\bar{x} = \beta$$

a et b sont des estimateurs non biaisés de α et β

Autrement dit si l'on recommence plusieurs fois la même expérience avec les mêmes valeurs de x_i et que l'on observe les différentes réalisations de y_i . Les moyennes de a et b obtenues pour les différentes expériences permettent d'approcher les vraies valeurs de α et β dont qui lient les variables

1.2 Calcul des variances de a et b connaissant les valeurs de x_i

variance de a (hypothèse $E(\varepsilon_i) = 0$
 $V(\varepsilon_i) = \sigma^2$)

$$a = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$V(a) = \frac{V\left(\sum_{i=1}^n y_i (x_i - \bar{x})\right)}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2}$$

les y_i sont des observations indépendantes (ε_i sont indépendantes)

$$V\left(\sum_{i=1}^n y_i (x_i - \bar{x})\right) = \sum_{i=1}^n V(y_i (x_i - \bar{x})) = \sum_{i=1}^n (x_i - \bar{x})^2 V(y_i)$$

$$V(y_i) = V(\alpha x_i + \beta + \varepsilon_i) = V(\varepsilon_i) = \sigma^2$$

$$V(a) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \sigma^2}{\left(\sum_{i=1}^n (x_i - \bar{x})^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

variance de b (hypothèse $E(\varepsilon_i) = 0$
 $V(\varepsilon_i) = \sigma^2$)

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

$$b = \bar{y} - \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{A} \bar{x} = \frac{1}{A} (A\bar{y} - \sum_{i=1}^n y_i (x_i - \bar{x})\bar{x}) = \frac{1}{A} \sum_{i=1}^n y_i \left(\frac{A}{n} - (x_i - \bar{x})\bar{x}\right)$$

$$V(b) = \frac{1}{A^2} \sum_{i=1}^n \left(\frac{A}{n} - (x_i - \bar{x})\bar{x}\right)^2 V(y_i) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{A}\right)$$

$$V(b) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

Pour le calcul de $V(a)$ et $V(b)$ on utilise l'erreur type d'estimation

$$s = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (e_i)^2}$$

qui est une estimation non biaisée de σ .

Si l'ensemble des observations est tel que A proche de 0 Tous les points sont proches du barycentre et le phénomène que l'on cherche à modéliser (dépendance de y avec x) est peu échantillonné. Les variances de a et b peuvent être très grandes. N'importe quelle droite de régression peut convenir

1.3 Intervalle de confiance hypothèse ε suit une loi $N(0, \sigma)$

L'estimateur a suit une loi normale $N(\alpha, \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}})$

L'estimateur b suit une loi normale $N(\beta, \sigma \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}})$

Où σ est estimé par s

Il est alors possible d'obtenir un intervalle de coefficient de confiance $1-\gamma$ en calculant avec une table de Student à n-2 degrés de liberté $t_{g,n-2}$

$$a - t_{\gamma,n-2} \sqrt{V(a)} < \alpha < a + t_{\gamma,n-2} \sqrt{V(a)}$$

$$b - t_{\gamma,n-2} \sqrt{V(b)} < \beta < b + t_{\gamma,n-2} \sqrt{V(b)}$$

Exemple :

Si $n=25$ et $\gamma=0,05$ soit 95% de confiance $t_{g,n-2}=2,069$

Si n grand $n>120$ $\gamma=0,05$ soit 95% de confiance $t_{g,n-2}=1,96$

Si n grand $n>120$ $\gamma=0,04$ soit 96% de confiance $t_{g,n-2}=2$

$$a - 2\sqrt{V(a)} < \alpha < a + 2\sqrt{V(a)}$$

$$b - 2\sqrt{V(b)} < \beta < b + 2\sqrt{V(b)}$$

1.4 Test d'hypothèse

Test relatif à la corrélation

On test l'hypothèse $H_0: \rho=0$

Si r est l'estimation calculée à partir d'un échantillon de taille n du coefficient de corrélation ρ

$$|r| = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}} = \sqrt{R^2} = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

$t = \frac{|r|}{\sqrt{1-r^2}} \sqrt{n-2}$ est une variable aléatoire de Student à n-2 degrés de liberté. On rejette

l'hypothèse H_0 si t est supérieure au $t_{g,n-2}$ des tables de la loi de Student.

Exemple :

Si $n=27$ $r=0,8$ $t=6,666 > t_{g,n-2}$ quelque soit γ l'hypothèse de non corrélation est rejetée

Si $n=18$ $r=0,6$ $t=3$ $\gamma=0,05$ soit 95% de confiance $t_{g,n-2}=2,12$

l'hypothèse de non corrélation est rejetée avec une confiance de 95%

Si $n=18$ $r=0,6$ $t=3$ $\gamma=0,001$ soit 99,9% de confiance $t_{g,n-2}=4,015$

l'hypothèse de non corrélation est rejetée avec une confiance de 99,9%.

Test relatif à la régression

On teste l'hypothèse $H_0: \alpha = 0$

Si a est l'estimation calculée à partir d'un échantillon de taille n du coefficient de corrélation α

$t = \frac{|a - \alpha|}{\sqrt{V(a)}}$ est une variable aléatoire de Student à $n-2$ degrés de liberté. On rejette l'hypothèse H_0 si t est supérieure au $t_{\gamma, n-2}$ des tables de la loi de Student. Autrement dit si on considère l'intervalle de confiance $I = \left[a - t_{\gamma, n-2} \sqrt{V(a)}; a + t_{\gamma, n-2} \sqrt{V(a)} \right]$ si $0 \notin I$ avec un risque γ de se tromper, il existe une relation linéaire entre x et y (le coefficient α n'est pas nul)

Exemple du test du rendement cours 1 p19

Si $n=8$ $\gamma=0,05$ soit 95% de confiance $t_{\gamma, n-2}=2,447$ $a=5,1389$ $(V(a))^{0.5}=0,283$

$5,1389 - 2,447 * 0,283 < \alpha < 5,1389 + 2,447 * 0,283$

$4,4464 < \alpha < 5,8314$

avec une confiance de 95% (ou un risque de 5%) il existe une relation linéaire entre x et y Le coefficient α n'est pas nul.

1.5 Intervalle de prévisibilité

L'équation de régression aux moindres carrés peut être utilisée pour prévoir la valeur de la variable dépendante y correspondant à une valeur de x_0 choisie. On obtient $\hat{y}_0 = ax_0 + b$ alors que $y_0 = \alpha x_0 + \beta + \varepsilon$ La variabilité de la prédiction de y_0 correspondant à x_0 est donnée par

$$V(\hat{y}_0 - y_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

σ est estimé par l'erreur type d'estimation s .

Cet intervalle est d'autant plus large que x_0 est loin du centre du nuage de points.

Si on s'intéresse à la réponse moyenne obtenue pour la valeur x_0 $\hat{\mu}_0 = ax_0 + b$ alors que la moyenne réelle est donnée par $\mu_0 = \alpha x_0 + \beta$ La variabilité de la prédiction de μ_0 correspondant à x_0 est donnée par

$$V(\hat{\mu}_0 - \mu_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

On remarque que $\hat{\mu}_0 = ax_0 + b$ est égale à $\hat{y}_0 = ax_0 + b$ cependant la variance est plus faible. En effet il y a une plus grande incertitude dans la prédiction d'une observation particulière que dans la prédiction d'une réponse moyenne.

grands échantillons

Si on a un grand échantillon $n > 30$

alors 68% des points du diagramme de dispersion se situent à moins de σ au dessus ou au dessous de la droite de régression

alors 96% des points du diagramme de dispersion se situent à moins de 2σ au dessus ou au dessous de la droite de régression

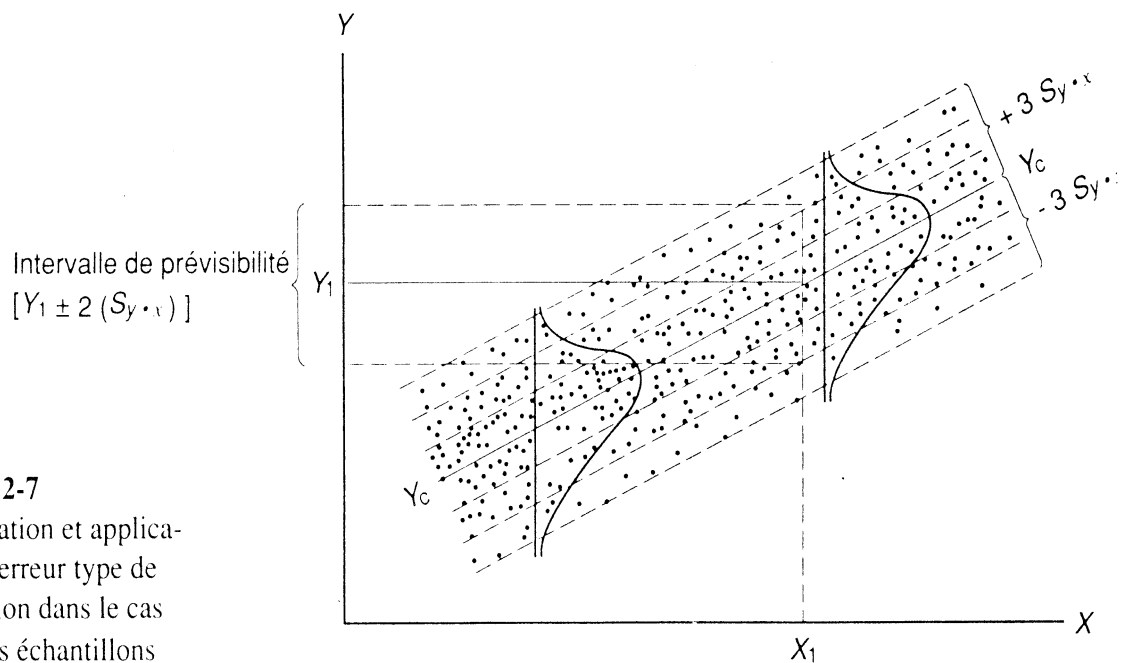


Figure 12-7

Interprétation et application de l'erreur type de l'estimation dans le cas de grands échantillons

1.6 Evaluation de l'ajustement

Seule la visualisation des données permet d'interpréter correctement le coefficient de corrélation et de s'assurer qu'il existe une relation linéaire entre x et y . D'autre part, les intervalles de confiance sont obtenus avec l'hypothèse que ε suit une loi normale. Il est donc fondamental de visualiser l'allure des données et des résidus obtenus. Pour détecter les défaillances du modèle $Y = \alpha X + \beta + \varepsilon$ avec ε suit une loi $N(0, \sigma)$

Les valeurs $\hat{e}_i = y_i - \hat{y}_i$ sont appelées résidus $\hat{e}_{si} = \frac{y_i - \hat{y}_i}{s}$ résidu réduit

$$E(\hat{e}_{si}) = 0$$

$$V(\hat{e}_{si}) = \frac{V(e_i)}{s^2} = \frac{\sigma^2}{s^2} \approx 1$$

Si n est assez grand les résidus réduits suivent une loi $N(0,1)$, si le modèle est correct on doit donc observer 68% des résidus réduits entre ± 1 et 96% entre ± 2

Un graphe des résidus est un élément essentiel de toute analyse de régression. On analysera le graphe des résidus en fonction

- des valeurs estimées y_i
- des valeurs du régresseur x_i
- de l'ordre d'acquisition des données i

INTERPRETATION DU COEFFICIENT DE CORRELATION R

Interprétation descriptive ou déterministe

Le coefficient de détermination R^2 égale à r au carré indique la proportion de variance due à la relation linéaire entre Y et X et $1 - R^2$ la proportion de variance résiduelle. Il s'agit d'une caractéristique propre à l'échantillon considéré. Un r fort n'indique pas une corrélation significative. Il peut avoir été obtenu avec très peu de points. La corrélation observée sur l'échantillon peut être fortuite et non due à une corrélation réelle des variables dans la population globale.

Interprétation inférentielle ou statistique

Le coefficient de corrélation r prouve une corrélation significative pour un risque donné α en fonction du nombre de degré de liberté ($n - 2$) ayant servi à le calculer. Une valeur significative de r n'implique pas qu'une grande partie de la variance totale d'une variable est expliquée par sa liaison avec l'autre. Elle prouve seulement qu'avec le risque choisi, ce n'est par un hasard si l'échantillon observé semble montrer une tendance.