REGRESSION LINEAIRE

Préparation à l'Agrégation Bordeaux 1

Année 2013 - 2014

Jean-Jacques Ruch

Table des Matières

CHAPITRE I

Regression linéaire

Nous allons étudier ici un modèle statistique d'usage fréquent : la régression linéaire. De nombreux modèles se ramènent facilement au modèle linéaire par des transformations simples. Ainsi le modèle $Y = \alpha X^{\beta}$ très utilisé en économétrie, devient linéaire en passant au logarithmes : en posant $Y' = \ln(Y)$, $X' = \ln(X)$ on obtient $Y' = \ln(\alpha) + \beta X'$. Il en va de même pour le modèle à croissance exponentielle : $Y = \alpha \exp(\beta X)$; ou encore pour le modèle logistique, qui rend compte des variations d'un taux de réponse $0 \leq Y \leq 1$ en fonction d'une excitation X : $Y = \frac{\exp(\alpha+\beta X)}{1+\exp(\alpha+\beta X)}$ se linéarise en posant $Y' = \ln(Y/(1-Y))$.
Cependant ce n'est pas toujours possible, ou aussi évident : le modèle $Y = \alpha + \beta X + \gamma X^2$ est linéaire, mais est à deux variables explicatives : si on pose $Z = X^2$ on obtient $Y = \alpha + \beta X + \gamma Z$, c'est de la régression multiple...

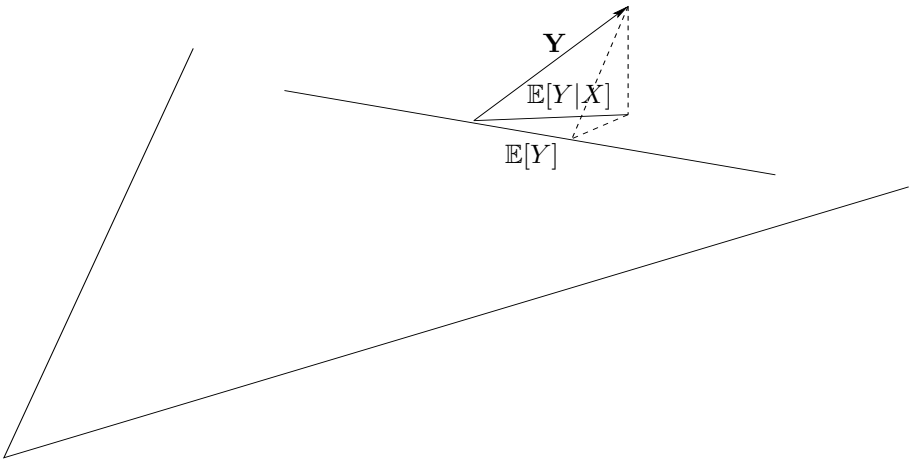
Considérons un couple de variables aléatoires (X, Y). Si X et Y ne sont pas indépendantes, la connaissance de la valeur prise par X change notre incertitude concernant la réalisation de Y : elle la diminue en général, car la distribution conditionnelle sachant $X = x$, a une variance qui est en moyenne inférieure à la variance de Y :

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y|X)] + \mathrm{Var}(\mathbb{E}(Y|X)) \geq \mathbb{E}[\mathrm{Var}(Y|X)]$$

même si l'on peut avoir $\mathrm{Var}(Y|X = x) > \mathrm{Var}(Y)$.
Si la variable X peut servir à prédire Y, on est conduit à rechercher une formule de prévision de Y par X du type $\hat{Y} = f(X)$, sans biais $\mathbb{E}[Y - \hat{Y}] = 0$, ainsi qu'à évaluer l'ordre de grandeur de l'erreur de prévision que l'on mesure par la variance de $\varepsilon = Y - \hat{Y}$. On cherchera bien sûr à minimiser cette variance. Nous étudierons le cas théorique en recherchant la formule de prévision idéale, au sens des moindres carrés, plus spécialement si cette formule est linéaire. Puis nous étudierons le cas usuel où les variables ne sont connues qu'à travers les valeurs d'un échantillon. Dans ce genre de modèle, on dira que X est une *variable explicative* ou un *prédicteur* et Y sera une *variable expliquée* ou un *critère*. Dans un deuxième temps nous nous intéresserons au cas multivarié.

1. Régression linéaire simple - modèle théorique

Etant donné deux variables aléatoires X et Y, une fonction f telle que $f(X)$ soit aussi proche que possible de Y en moyenne quadratique est déterminée par l'espérance conditionnelle. En effet, $f(X) = \mathbb{E}[Y|X]$ réalise le minimum de $E[(Y - f(X))^2]$ car $\mathbb{E}[Y|X]$ correspond à la projection orthogonale de Y sur l'espace $L^2(\Omega, \mathcal{F}(X), \mathbb{P}) = L^2_X$



Définition 1. *La fonction qui à une valeur x de X associe $\mathbb{E}[Y|X=x]$ s'appelle* fonction de régression *de Y en X.*

Son graphe est la courbe de régression de Y en X. On peut poser

$$Y = \mathbb{E}[Y|X] + \varepsilon$$

où ε est un résidu aléatoire (pas toujours négligeable). Il a la propriété d'être d'espérance nulle $\mathbb{E}[\varepsilon] = 0$ car $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$. De plus, il est non corrélé linéairement à X et $\mathbb{E}[Y|X]$

$$\mathrm{Cov}(\varepsilon, X) = \mathrm{Cov}(\varepsilon, E[Y|X]) = 0$$

puisqu'il est orthogonal à L^2_X. Enfin sa variance, ou variance résiduelle, vérifie

$$\mathrm{Var}(\varepsilon) = \mathrm{Var}(Y) - \mathrm{Var}(\mathbb{E}[Y|X]) = \mathbb{E}[\mathrm{Var}(Y|X)] = (1 - \eta^2_{Y/X})\mathrm{Var}(Y)$$

où $\eta^2_{Y/X}$ est défini par :

Définition 2. *La qualité de l'approximation de Y par $\mathbb{E}[Y|X]$ est mesurée par le* rapport de corrélation *:*

$$\eta^2_{Y/X} = \frac{\mathrm{Var}(\mathbb{E}[Y|X])}{\mathrm{Var}(Y)} = \frac{\textit{variance expliquée}}{\textit{variance totale}}.$$

Nous allons maintenant supposer que l'on a une régression linéaire, c'est-à-dire que $\mathbb{E}[Y|X] = \alpha + \beta X$. C'est le cas le plus important dans la pratique (par exemple si X et Y suivent une loi normale à deux dimensions). On a donc

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

Nous cherchons alors à déterminer α et β.
En prenant l'espérance des deux membres, on obtient :

$$\mathbb{E}[Y] = \alpha + \beta\mathbb{E}[X] + \mathbb{E}[\varepsilon] = \alpha + \beta\mathbb{E}[X]$$

La droite de régression passe donc par le point $(\mathbb{E}[X], \mathbb{E}[Y])$. On a par conséquent

$$Y - \mathbb{E}[Y] = \beta(X - \mathbb{E}[X]) + \varepsilon$$

et en multipliant par $X - \mathbb{E}[X]$ et en prenant l'espérance :

$$\mathbb{E}[(Y - \mathbb{E}[Y])(X - \mathbb{E}[X])] = \beta\mathbb{E}[(X - \mathbb{E}[X])^2] + \mathbb{E}[\varepsilon(X - \mathbb{E}[X])]$$

ou encore

$$\mathrm{Cov}(X, Y) = \beta\mathrm{Var}(X) + \mathrm{Cov}(\varepsilon, X)$$

Mais, comme ε est non corrélé avec X, il reste

$$\beta = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} = \rho\frac{\sigma_Y}{\sigma_X}$$

où $\rho = \mathrm{Cov}(X,Y)/(\sigma_X\sigma_Y)$ est le *coefficient de corrélation* entre X et Y. On en déduit alors

$$\alpha = \mathbb{E}[Y] - \beta\mathbb{E}[X] = \mathbb{E}[Y] - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\mathbb{E}[X] = \mathbb{E}[Y] - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}\mathbb{E}[X]$$

et donc

Définition 3. *La droite de régression de Y par X est :*

$$\widehat{Y} = \mathbb{E}[Y] + \frac{Cov(X,Y)}{\mathrm{Var}(X)}(X - \mathbb{E}[X]) = \mathbb{E}[Y] + \rho\frac{\sigma_Y}{\sigma_X}(X - \mathbb{E}[X])$$

et $Y = \widehat{Y} + \varepsilon$.

Comme ε est non corrélé avec X, on peut écrire, en prenant la variance des deux membres :

$$\mathrm{Var}(Y) = \rho^2\frac{\sigma_Y^2}{\sigma_X^2}\mathrm{Var}(X) + \mathrm{Var}(\varepsilon) = \rho^2\mathrm{Var}(Y) + \mathrm{Var}(\varepsilon)$$

on obtient que, si la régression est linéaire, alors $\rho^2 = \eta^2_{Y/X}$.

2. Régression linéaire simple - ajustement du modèle sur des données

On dispose de n couples (x_i, y_i) constituant un n-échantillon d'observations indépendantes. On suppose vraie l'hypothèse que la régression est linéaire.

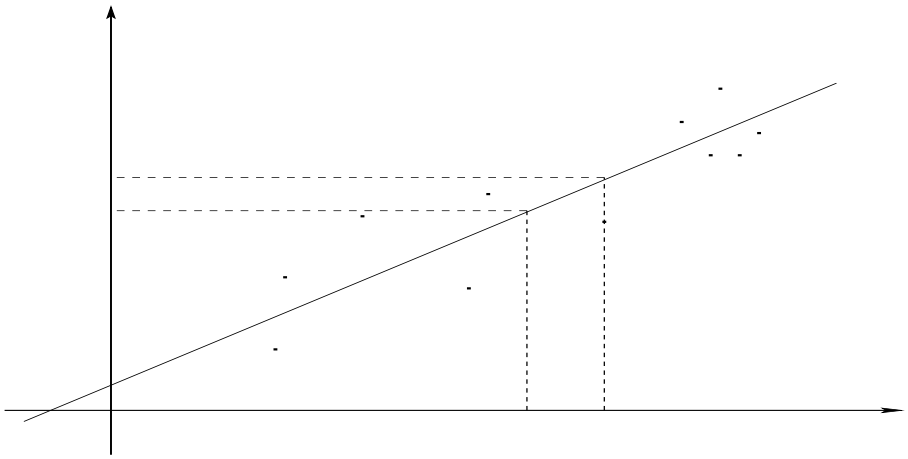
Le problème ici va consister à estimer $\alpha,\ \beta$ ainsi que la variance σ^2 de ε.

On suppose en fait que pour chaque observation on a

$$y_i = a + bx_i + \varepsilon_i$$

où les ε_i sont des réalisations indépendantes d'une variable ε d'espérance nulle et de variance constante σ^2, quel que soit x_i.

Nous allons nous intéresser à l'estimation de α, β et σ^2 par la méthode des moindres carrés. Cette méthode due à Gauss reprend sur l'échantillon la propriété que $\mathbb{E}[Y|X] = \alpha + \beta X$ est la meilleure approximation de Y par X en moyenne quadratique. On cherche donc à ajuster au nuage de points (x_i, y_i) une droite d'équation $y^* = a + bx$ de telle sorte que $\sum(y_i - y_i^*)^2$ soit minimal. On étudiera ensuite les propriétés de a et b en tant qu'estimations de α et β ainsi que de l'estimation $\hat{\sigma}^2$ de σ^2.



La méthode élémentaire de détermination de a et b est la suivante :

$$\sum_{i=1}^{n}(y_i - y_i^*)^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2 = F(a,b)$$

Le minimum est atteint pour $\frac{\partial F}{\partial a} = \frac{\partial F}{\partial b} = 0$, ce qui donne les deux équations

$$\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \iff \overline{y} = a + b\overline{x}$$

$$\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0$$

On obtient

$$b = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\mathrm{Cov}(x,y)}{\mathrm{Var}(x)} = r\frac{s_y}{s_x}$$

et donc

$$y^* = \overline{y} + r\frac{s_y}{s_x}(x - \overline{x})$$

La droite des moindres carrés passe donc par le centre de gravité du nuage $(\overline{x}, \overline{y})$ et sa pente est l'analogue empirique de la pente de la droite de régression $\rho\frac{\sigma_y}{\sigma_x}$.

Théorème 4.

Les quantités a, b et y^* sont des estimations sans biais de α, β et $\mathbb{E}[Y|X = x] = \alpha + \beta x$.

DÉMONSTRATION. La quantité b est une réalisation de la variable aléatoire

$$B = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Nous allons montrer que

$$\mathbb{E}[B|X_1 = x_1, \ldots, X_n = x_n] = \beta.$$

On a

$$\mathbb{E}[B|X_1 = x_1, \ldots, X_n = x_n] = \frac{\sum_{i=1}^{n}(x_i - \overline{x})\mathbb{E}[(Y_i - \overline{Y})|X_1 = x_1, \ldots, X_n = x_n]}{\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Or, d'après l'hypothèse de régression linéaire

$$\mathbb{E}[Y_i|X_1 = x_1, \ldots, X_n = x_n] = \alpha + \beta x_i$$

et donc aussi

$$\mathbb{E}[\overline{Y}|X_1 = x_1, \ldots, X_n = x_n] = \alpha + \beta \overline{x}.$$

On obtient donc

$$\mathbb{E}[B|X_1 = x_1, \ldots, X_n = x_n] = \beta \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2} = \beta$$

et en prenant l'espérance $\mathbb{E}[B] = \beta$.
Comme $a = \overline{y} - b\overline{x}$, a est une réalisation de $A = \overline{Y} - B\overline{X}$, et par le même procédé

$$\begin{aligned} \mathbb{E}[A|X_1 = x_1, \ldots, X_n = x_n] &= \mathbb{E}[\overline{Y}|X_1 = x_1, \ldots, X_n = x_n] - \overline{x}\mathbb{E}[B|X_1 = x_1, \ldots, X_n = x_n] \\ &= \alpha + \beta\overline{x} - \overline{x}\beta = \alpha \end{aligned}$$

et donc $\mathbb{E}[A] = \alpha$.
Puisque $\mathbb{E}[Y|X = x] = \alpha + \beta x$, $y^* = a + bx$ est une estimation de $\mathbb{E}[Y|X = x] = \alpha + \beta x$, dont l'estimateur associé est sans biais. □

On peut montrer que parmi les estimateurs sans biais de α et β, A et B sont ceux de variance minimale.

Pour estimer $\sigma^2 = \mathrm{Var}(\varepsilon)$ il est naturel d'utiliser la variance des résidus $e_i = y_i - y_i^*$, c'est-à-dire la quantité que l'on a minimisée $\sum_{i=1}^n (y_i - y_i^*)^2$. On montre alors

Théorème 5.

La quantité $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^n (y_i - y_i^*)^2}{n-2}$ *est une estimation de* σ^2*, dont l'estimateur associé est sans biais .*

3. Le cas du modèle linéaire gaussien

Jusquà présent nous n'avions fait que l'hypothèse que la régression est linéaire. Nous allons en plus faire l'hypothèse suivante :

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

Théorème 6.

Si ε *suit une loi normale centrée de variance* σ^2 *alors :*
a) la loi conditionnelle de $Y|X = x$ *est* $\mathcal{N}(\alpha + \beta x, \sigma^2)$ *;*
b) si les x_i *sont fixés, les lois de* B*,* A *et* Y^* *sont gaussiennes :*

$$\begin{aligned} B &\sim \mathcal{N}\left(\beta, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) \\ A &\sim \mathcal{N}\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right) \\ Y^* &\sim \mathcal{N}\left(\alpha + \beta x, \sigma^2\left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right) \end{aligned}$$

Les variables aléatoires A, B et Y^* sont des estimateurs de variance minimale de α, β et σ^2.

Dans ces conditions on peut vérifier que

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^n (y_i - y_i^*)^2}{\sigma^2} = \frac{ns_{y/x}^2}{\sigma^2}$$

est une réalisation d'une variable χ^2_{n-2} (on a du estimer deux paramètres).

Les lois des variables aléatoires A et B supposent que la variance σ^2 est connue. Lorsqu'elle est inconnue on utilise un estimateur. Ainsi, puisque

$$\frac{(B-\beta)\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}{\sigma} = \frac{(B-\beta)\sqrt{ns_x^2}}{\sigma}$$

et

$$\frac{(A-\alpha)}{\sigma\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^n (x_i-\overline{x})^2}\right)}} = \frac{(A-\alpha)}{\sigma\sqrt{\frac{1}{n}\left(1 + \frac{\overline{x}^2}{s_x^2}\right)}}$$

suivent une loi $\mathcal{N}(0,1)$ et que $\frac{ns_{y/x}^2}{\sigma^2}$ suit une loi χ^2_{n-2}, et sont indépendantes on a : les variables aléatoires

$$\frac{(B-\beta)\sqrt{ns_x^2}}{\sigma} \;/\; \sqrt{\frac{ns_{y/x}^2}{\sigma^2} \;/\; (n-2)} = (B-\beta)\sqrt{\frac{s_x^2(n-2)}{s_{y/x}^2}}$$

et

$$\frac{(A-\alpha)}{\sqrt{\frac{1}{n}\left(1 + \frac{\overline{x}^2}{s_x^2}\right)}\sigma} \Big/ \sqrt{\frac{ns_{y/x}^2}{\sigma^2} \;/\; (n-2)} = \frac{(A-\alpha)}{\sqrt{s_{y/x}^2}}\sqrt{\frac{n-2}{1+\frac{\overline{x}^2}{s_x^2}}}$$

suivent une loi de Student T_{n-2} à $n-2$ degrés de liberté. On peut alors en déduire des intervalles de confiance pour les coefficients α et β.

4. Tests dans le modèle linéaire gaussien

On suppose toujours que ε suit une loi normale.

Test du caractère significatif de la régression

On commence par tester le caractère significatif de la régression, autrement dit on veut tester si l'hypothèse de régression linéaire est justifiée ou non. Pour ce faire on teste l'hypothèse

$$H_0 : \beta = 0 \text{ contre } H_1 : \beta \neq 0.$$

On a donc sous H_0, $\frac{Bs_x}{s_{y/x}}$ suit une loi de Student T_{n-2}. On obtient la région de rejet avec un risque δ

$$]-\infty, t_{n-2,1-\delta/2}[\cup]t_{n-2,1-\delta/2}, \infty[.$$

On peut construire un autre test utilisant une autre statistique qui évite de calculer B.

On a $\sum_{i=1}^n (Y_i - Y_i^*)^2/\sigma^2$ suit une loi du χ^2_{n-2}. De plus, si l'hypothèse H_0 de non régression linéaire est satisfaite, i.e. si $\beta = 0$, alors

$$\frac{B\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2}}{\sigma} \;\sim\; \mathcal{N}(0,1)$$

et donc

$$\frac{\sum_{i=1}^n (Y_i^* - \overline{Y})^2}{\sigma^2} = \frac{B^2 \sum_{i=1}^n (X_i - \overline{X})^2}{\sigma^2} \;\sim\; \chi_1^2$$

On en déduit (par le théorème de Cochran) que $\sum_{i=1}^n (Y_i - Y_i^*)^2$ et $\sum_{i=1}^n (Y_i^* - \overline{Y})^2$ sont indépendante et

$$\frac{\sum_{i=1}^n (Y_i^* - \overline{Y})^2}{\sum_{i=1}^n (Y_i - Y_i^*)^2}(n-2) \;\sim\; F(1, n-2)$$

En pratique $y_i - \overline{y} = y_i - y_i^* + y_i^* - \overline{y}$ et

$$\sum_{i=1}^{n}(y_i^* - \overline{y})(y_i - y_i^*) = 0.$$

On obtient donc que

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - y_i^*)^2 + \sum_{i=1}^{n}(y_i^* - \overline{y})^2$$

c'est-à-dire : somme des carrés totale=somme des carrés résiduelle+somme des carrés expliquée.

Comme on sait que

$$y^* = \overline{y} + r\frac{s_y}{s_x}(x - \overline{x})$$

on en déduit que

$$\sum_{i=1}^{n}(y_i^* - \overline{y})^2 = r^2 s_y^2 = r^2 \sum_{i=1}^{n}(y_i - \overline{y})^2$$

et que

$$\sum_{i=1}^{n}(y_i - y_i^*)^2 = n s_{y/x}^2 = (1 - r^2)\sum_{i=1}^{n}(y_i - \overline{y})^2 = (1 - r^2)s_y^2.$$

La statistique testée ci-dessus correspond donc à $\dfrac{r^2}{1 - r^2}(n - 2)$.

Construction d'un intervalle de confiance pour une valeur donnée du jeu de valeurs

On suppose que l'on souhaite donner une réponse moyenne à la variable explicative de départ.

On peut d'abord remarquer que loi de Y^* est la loi normale :

$$\mathcal{N}\left(\alpha + \beta x, \sigma^2\left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)\right)$$

Ensuite comme on sait que

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(y_i - y_i^*)^2}{\sigma^2} = \frac{n s_{y/x}^2}{\sigma^2}$$

est une réalisation d'une variable χ^2_{n-2} on en déduit que :

$$\frac{Y^* - \alpha - \beta x}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}} \Big/ \sqrt{\frac{\frac{n s_{y/x}^2}{\sigma^2}}{n-2}} = \frac{Y^* - \alpha - \beta x}{\sqrt{s_{y/x}^2\left(1 + \frac{(x-\overline{x})^2}{s_x^2}\right)}}\sqrt{n-2}$$

suit une loi de Student T_{n-2} à $n - 2$ degrés de liberté. On obtient donc un intervalle de confiance, au niveau $1 - \delta$

$$\left]\alpha + \beta x - t_{\delta/2}\frac{\sqrt{s_{y/x}^2\left(1 + \frac{(x-\overline{x})^2}{s_x^2}\right)}}{\sqrt{n-2}}, \alpha + \beta x + t_{\delta/2}\frac{\sqrt{s_{y/x}^2\left(1 + \frac{(x-\overline{x})^2}{s_x^2}\right)}}{\sqrt{n-2}}\right[$$

Construction d'un intervalle de prédiction

On suppose que l'on souhaite prévoir la valeur de Y pour une nouvelle valeur x_0 de X. Le choix naturel est $y_0^* = \alpha + \beta x_0$. On détermine alors un intervalle de confiance en remarquant que la loi de Y_0^* est la loi normale :

$$\mathcal{N}\left(\alpha + \beta x, \sigma^2\left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)\right)$$

et la loi conditionnelle de $Y|X = x_0$ est $\mathcal{N}(\alpha + \beta x_0, \sigma^2)$.

En remarquant que Y_0 et Y_0^* sont indÓpandantes on a

$$Y_0 - Y_0^* \ \sim\ \mathcal{N}\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)\right)$$

et donc

$$\frac{Y_0 - Y_0^*}{\sqrt{\hat{\sigma}^2\left(1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2}\right)}} = \frac{Y_0 - Y_0^*}{\sqrt{n s_{y/x}^2\left(1+\frac{1}{n}+\frac{(x_0-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2}\right)}}\sqrt{n-2} \ \sim \ T_{n-2}.$$

D'où l'on peut dénduire l'intervalle de prédiction au niveau $1-\delta$

$$\left]\alpha+\beta x_0 - t_{\delta/2}\frac{\sqrt{s_{y/x}^2\left(n+1+\frac{(x-\overline{x})^2}{s_x^2}\right)}}{\sqrt{n-2}}, \alpha+\beta x_0 + t_{\delta/2}\frac{\sqrt{s_{y/x}^2\left(n+1+\frac{(x-\overline{x})^2}{s_x^2}\right)}}{\sqrt{n-2}}\right[$$

5. Exemple

Nous disposons d'un échantillon de 24 offres de vente d'appartements situés dans le V^e et le VI^e arrondissement de Paris en 1975.

Y Prix en millier de francs	130	280	800	268	790	500	320	250	378	250	350	300
X Surface en m^2	28	50	196	55	190	110	60	48	90	35	86	65

Y Prix en millier de francs	155	245	200	325	85	78	375	200	270	295	85	495
X Surface en m^2	32	52	40	70	28	30	105	52	80	60	20	100

Une représentation des couples (X_i, Y_i), donne un nuage de points dont la forme autorise un ajustement linéaire. On pose donc le modèle $Y = \alpha + \beta X + \varepsilon$ et on supposera $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. Des calculs élémentaires conduisent aux résultats suivants :

$$\overline{x} = 70,08333\, m^2 \qquad \overline{y} = 309331\, F$$

$$s_x = \sqrt{\frac{1}{24}\sum_{i=1}^{24}(x_i-\overline{x})^2} = 44,6915\, m^2 \qquad s_y = \sqrt{\frac{1}{24}\sum_{i=1}^{24}(y_i-\overline{y})^2} = 182950,5\, F$$

et

$$r = \frac{\mathrm{Cov}(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{24}(x_i-\overline{x})(y_i-\overline{y})}{24 s_x s_y} = 0,9733$$

Il y a une forte corrélation qui signifie qu'à $97,33\%$ le prix d'un appartement est expliqué par sa surface. On en déduit les estimations a et b de α et β

$$a = 30,0921 \qquad b = 3,9844$$

et l'équation de la droite d'ajustement

$$y^* = 30,0921 + 3,9844x$$

La variance résiduelle $s_{y/x}^2$ s'obtient directement par la formule $s_{y/x}^2 = (1-r^2)s_y^2$, soit

$$s_{y/x}^2 = 1762,1816 \qquad s_{y/x} = 41,98$$

et on en déduit

$$\hat{\sigma}^2 = \frac{n}{n-2}s_{y/x}^2 = 1922,38 \qquad \hat{\sigma} = 43,84$$

L'hypothèse de normalité de ε permet alors de donner des intervalles de confiance. Ainsi, pour σ^2, on a $n s_{y/x}^2/\sigma^2$ qui suit une loi du χ_{n-2}^2. Ici $n = 24$, on a donc un χ^2 à 22 degrés de liberté, dont la table fournit les bornes 11 et 36.8 pour un intervalle symetrique de niveau 0.95. L'intervalle de confiance pour σ^2 est donc

$$\frac{24 s_{y/x}^2}{36.8} = 1149,25 \leq \sigma^2 \leq \frac{24 s_{y/x}^2}{11} = 3844,76$$

ou encore

$$33,90 \leq \sigma \leq 62,01$$

Le test de signification de la régression peut être effectué par l'analyse de la variance .

Variation	Somme des carrés	Degrés de liberté	Carré moyen
Expliquée	761009	1	761009
Résiduelle	42292	22	1922,4
Totale	803301	23	

La valeur obtenue est donc $761009/1922,4 = 396$ alors que la valeur lu dans la table est $4,30$. Donc le résultat est très significatif.
Supposons maintenant que l'on désire prévoir à l'aide du modèle la valeur naturelle de Y pour une valeur non observés x_0 de X. La prévision naturelle est $Y_0^* = a + bx_0$. Afin d'encadrer cette valeur cherchons un intervalle de prédiction.
On a vu que Y_0^* est distribué selon la loi

$$\mathcal{N}\left(\alpha + \beta x_0, \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right)$$

D'autre part, on sait que $Y_0 = Y_{/X=x_0}$ suit une loi $\mathcal{N}\left(\alpha + \beta x_0, \sigma^2\right)$. En utilisant l'indépendance de Y_0 et de Y_0^*, on obtient que

$$Y_0 - Y_0^* \ \sim \ \mathcal{N}\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right)$$

et donc

$$\frac{Y_0 - Y_0^*}{\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}} \ \sim \ T_{n-2}$$

En remplaçant x_0 par sa valeur et Y_0^* par $a + bx_0$, on peut donc obtenir un intervalle probable pour Y_0. Cet intervalle sera d'autant plus grand que x_0 sera éloigné de $\overline{x}$.
Ainsi, pour notre exemple, on trouve dans la table que :

$$\mathbb{P}(|T_{n-2}| < 2,074) = 0,95.$$

En prenant $x_0 = 100$, on a $y_0^* = 428,53$,

$$\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)} = 45,15$$

d'où, l'intervalle de précision à 95% est :

$$334,89 < Y_0 < 522,17$$

ce qui n'est pas très précis.

6. Régression linéaire multiple

On s'intéresse à une variable aléatoire réelle Y qui dépend d'une vecteur aléatoire X de $\mathbb{R}^p$ et on cherche à expliquer la variation de Y à partir des variations de X. Pour cela on se place dans le cadre d' un *modèle linéaire* où $\mathbb{E}[Y|X]$ est linéaire. On notera ε la variable aléatoire représentant le bruit associé au modèle. Comme dans la régression simple on a $\mathbb{E}[\varepsilon|X] = 0$ et $\mathbb{E}[\varepsilon] = 0$. On supposera de plus, que la variance σ^2 de ε ne dépend pas de X.

Si X est un vecteur aléatoire de $\mathbb{R}^p$ et Y une variable aléatoire réelle, l'espérance conditionnelle de Y sachant X est donnée par :

$$\forall\ x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}, \quad \mathbb{E}[Y \mid X = x] = a + b_1 x_1 + \cdots + b_p x_p + \varepsilon.$$

Afin d'estimer le paramètre inconnu $\theta = \begin{pmatrix} a \\ b \end{pmatrix}$ avec $b = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix}$, on effectue une suite de n observations avec $n > p+1$ données par :

$$\forall\ i,\ 1 \leq i \leq n, \quad y_i = a + b_1 x_{i1} + \cdots + b_p x_{ip} + \varepsilon_i$$

où pour $(x_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ est une suite de réels connus, et où $(\varepsilon_i)_{1\leq i\leq n}$ est une suite de variables aléatoires telles que

$$\mathbb{E}[\varepsilon_i] = 0, \quad \mathrm{Var}(\varepsilon_i) = \sigma^2 \quad \text{et} \quad \mathbb{E}[\varepsilon_i \varepsilon_j] = 0 \ \text{ si } i \neq j.$$

On peut écrire la *régression linéaire* sous forme matricielle

$$Y = X\theta + \varepsilon$$

avec

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \quad \text{et} \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

X étant une matrice rectangulaire de dimension $n \times (p+1)$.

Définition 7. *On dit que le modèle ci-dessus est identifiable, s'il existe un unique vecteur $\theta \in \mathbb{R}^{p+1}$ tel que $\mathbb{E}[Y] = X\theta$.*

Par la suite nous supposerons toujours que cette condition est vérifiée. Il est clair que ceci est équivalent à ce que les vecteurs colonnes de X soient linéairement indépendants ou encore que X soit de rang $p+1$. Notons $\mathrm{Vect}(X)$ le sous-espace vectoriel de $\mathbb{R}^n$ de dimension $p+1$ engendré par les vecteurs colonnes de X,

$$\mathrm{Vect}(X) = \{\alpha 1_n + \beta_1 x_1 + \cdots + \beta_p x_p,\ \alpha, \beta_1, \ldots, \beta_p \in \mathbb{R}\}$$

avec

$$1_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}, \quad x_1 = \begin{pmatrix} x_{11} \\ \vdots \\ x_{n1} \end{pmatrix}, \quad \ldots \quad , \ x_p = \begin{pmatrix} x_{1p} \\ \vdots \\ x_{np} \end{pmatrix}$$

Définition 8. *On appelle estimateur des moindres carrés de θ, la valeur $\hat{\theta}$ qui minimise :*

$$\theta \mapsto \Delta(\theta) = ||Y - X\theta||^2.$$

Proposition 9. *Avec les notations précédentes on a :*
- *la matrice X^tX est inversible et la matrice de projection orthogonale sur Vect(X) est $H = X(X^tX)^{-1}X^t$;*
- *l'estimateur $\hat{\theta}$ de θ satisfait*

$$\hat{\theta} = (X^tX)^{-1}X^tY$$

DÉMONSTRATION. Comme X est de rang $p+1$, X^tX est une matrice inversible de $\mathcal{M}_{p+1}(\mathbb{R})$ (on peut par exemple voir que le noyau de la matrice X^tX est le même que celui de X, et que ce dernier est réduit au vecteur nul).
Des considérations d'algèbre linéaire montrent que la matrice de la projection orthogonal sur $\mathrm{Vect}(X)$ est

$$X(X^tX)^{-1}X^t$$

En effet, notons H cette matrice. On veut que pour tout $v \in \mathbb{R}^n$, $v - Hv \perp \mathrm{Vect}(X)$. Or, les éléments de $\mathrm{Vect}(X)$ sont de la forme Xu, donc en particulier $Hv = Xu_0$.

Si on prend $(u_1, \dots, u_{p+1})$ une base de $\mathbb{R}^{p+1}$, alors pour $1 \leq i \leq p+1$ on a

$$u_i^t X^t (v - Hv) = 0$$

donc $X^t v = X^t H v = X^t X u_0$. Or, comme $X^t X$ est une matrice inversible on a

$$u_0 = (X^t X)^{-1} X^t v$$

c'est-à-dire

$$Hv = X u_0 = X(X^t X)^{-1} X^t v$$

d'où $H = X(X^t X)^{-1} X^t$.
La quantité $||Y - X\theta||^2$ est minimale si $X\theta$ est égale à la projection orthogonale de Y sur $\text{Vect}(X)$. On en déduit donc

$$X\theta = X(X^t X)^{-1} X^t Y$$

et comme X est est de rang $p+1$ (injectivité) on a donc $\theta = (X^t X)^{-1} X^t Y$ □

A partir de l'estimateur des moindres carrés $\hat{\theta}$, on peut prédire Y par $\hat{Y} = X\hat{\theta}$ et ε par $\hat{\varepsilon} = Y - X\hat{\theta}$.
Soit $\text{Vect}(X)^\perp$ le sous-espace vectoriel de $\mathbb{R}^n$ orthogonal à $\text{Vect}(X)$ et soit $L = I_n - X(X^t X)^{-1} X^t$ la matrice de la projection orthogonale sur $\text{Vect}(X)^\perp$. On a

$$\hat{Y} = X(X^t X)^{-1} X^t Y \quad \text{et} \quad \hat{\varepsilon} = LY$$

donc ce sont les projections orthogonales de Y respectivement sur $\text{Vect}(X)$ et $\text{Vect}(X)^\perp$. Il est alors naturel d'estimer la variance σ^2 par

$$\hat{\sigma}^2 = \frac{||Y - \hat{Y}||^2}{n-p-1} = \frac{1}{n-p-1} \sum_{i=1}^{n} (Y_i - a - b_1 x_{i1} - \cdots - b_p x_{ip})^2$$

où le dénominateur $(n-p-1)$ correspond à la dimension de $\text{Vect}(X)^\perp$.

Théorème 10.

L'estimateur des moindres carrés $\hat{\theta}$ de θ est sans biais. De plus, on a

$$\text{Var}(\hat{\theta}) = \sigma^2 (X^t X)^{-1}.$$

Il est de variance minimale parmi les estimateurs linéaires sans biais.
De plus, $\hat{\sigma}^2$ est un estimateur sans biais de σ^2.

Démonstration. On a

$$\mathbb{E}[Y] = X\theta \quad \text{et} \quad \text{Var}(Y) = \sigma^2 I_n$$

Or, on sait que $\hat{\theta} = MY$ avec $M = (X^t X)^{-1} X^t$. On a

$$\mathbb{E}[\hat{\theta}] = M\mathbb{E}[Y] = MX\theta = \theta$$

et

$$\text{Var}(\hat{\theta}) = M\sigma^2 I_n M^t = \sigma^2 M M^t = \sigma^2 (X^t X)^{-1}$$

Soit alors $\tilde{\theta} = M'Y$ un autre estimateur linéaire sans biais de θ. On a $\theta = \mathbb{E}[\tilde{\theta}] = M'\mathbb{E}[Y] = M'X\theta$, d'où $M'X = I_{p+1}$. Comme $M = (X^t X)^{-1} X^t$, il en découle que

$$M' M^t = (X^t X)^{-1}.$$

Par conséquent, si on pose $N = M' - M$, alors $NM^t = MN^t =$. La variance de $\tilde{\theta}$ se déduit alors,

$$\text{Var}(\tilde{\theta}) = \sigma^2 M' M'^t = \sigma^2 (M+N)(M+N)^t = \sigma^2 M M^t + \sigma^2 N N^t \geq \sigma^2 M M^t = \text{Var}(\hat{\theta})$$

Pour l'estimateur de la variance, on procède de la même manière que ce qui a été fait pour la régression simple. □