Chapitre 6: Regression linéaire simple

1 Introduction

Dans ce chapitre, on s'interesse a des situations ou on observe deux variables X et Y sur chaque unité statistique et on aiemrait pouvoir prédire Y en fonction de X. Y s'appelle la **variable dependante** ou variable réponse. X s'appelle la **variable explicatrice** , variable indépendante, ou regresseur.

On s'interesse à des situations ou Y n'est pas une fonction déterministe de X. Autrement dit, c'est possible d'avoir des unités statistiques qui ont la meme valeur de X mais des valeurs differentes de Y. Cette différence peut etre due à:

- Des erreurs de mesure.
- L'existence d'autres variables dont on ne serait pas entrain de tenir compte.

Exemple 1.1. On cherche a prédire la taille Y en fonction de l'age X. On sait tous que la taille fluctue parmi des personnes de meme age. Mais on sait aussi "qu'en moyenne" la taille augmente avec l'age (jusqu'en debut de l'age adulte).

Soit $\mu_{Y|x}$ la moyenne de Y parmi les unites pour lesquelles $X = x$. On va utiliser le modele suivant pour expliquer la relation entre X et Y:

$$Y = \mu_{Y|x} + \varepsilon, \tag{1.1}$$

ou ε est une variable aleatoire qui rend compte des fluctuations qu'on pourrait observer parmis les unites qui ont la meme valeur de X.

Definition 1.1. La fonction $\mu_{Y|x}$ vue comme fonction de x s'appelle la **courbe de régression** de Y sur X.

Exemple 1.2. Dans la population des hommes, on s'interesse à la taille (Y) en fonction de l'age (X). $\mu_{Y|10}$ serait alors la taille moyenne des garcons de 10ans.

Dans ce chapitre, on s'interesse à l'estimation de fonction de régression. C'est un problème difficile à resoudre en général. On va se limiter à l'estimation de fonctions de régression **linéaire simple**.

Definition 1.2. Une fonction de régression $\mu_{Y|x}$ est dite linéaire simple si

$$\mu_{Y|x} = a + bx. \tag{1.2}$$

L'adjectif "simple" vient du fait qu'on cherche à prédire Y avec seulement **une seule** variable X. Si on laisse Y dependre de plusieurs variables explicatrices $X_1, \dots, X_k$, on parlerait de regression multiple.

On parle de **modèle** lineaire simple lorsqu'on **postule** que la fonction de regression en présence est lineaire simple. Il s'ecrit:

$$Y = a + bX + \varepsilon. \tag{1.3}$$

Objectif:

1. Comment partir d'un echantillon de $(X_1, Y_1), \dots, (X_n, Y_n)$ pour estimer les coefficients a et b d'un modele de regression lineaire simple.
2. Comment predire la valeur de Y connaissant la valeur de X sur une unite.

2 Estimation d'un modèle de regression linéaire simple: la méthode des moindres carrés

La démarche pour estimer une fonction lineaire simple est la suivante:

1. Obtenir un echantillon $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ ou (x_i, y_i) sont les valeurs observées pour X et Y sur l'unité i.

2. Dessiner le nuage des points (x_i, y_i). Si ce nuage à une tendance linéaire, alors on peut raisonnablement **modeliser** la fonction de regression par un modele de regression lineaire simple. On utilise alors la méthode des moindres carrés pour estimer a et b.

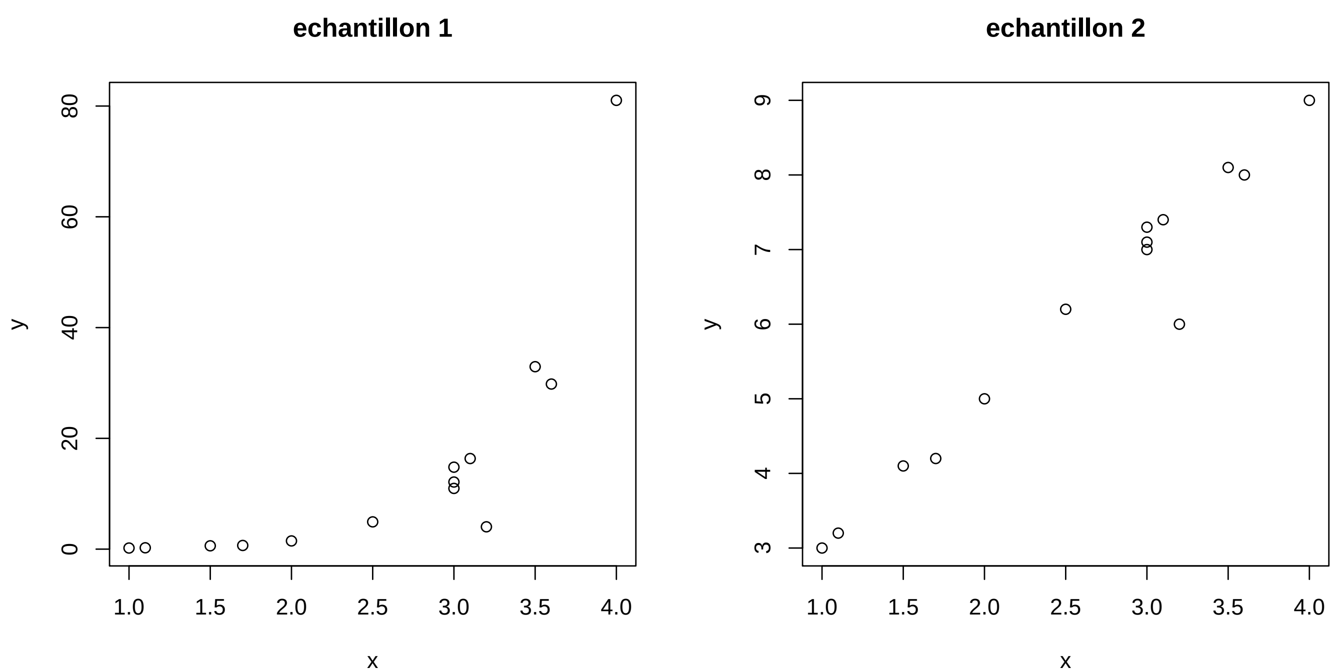
3. Si la tendance n'est pas linéaire, le modele de regression lineaire simple n'est donc approprié.

Exemple 2.1. Soit deux echantillons de points (X, Y).

x	1.0	1.1	1.5	1.7	2.0	2.5	3.0	3.0	3.0	3.1	3.2	3.5	3.6	4.0
y	0.2	0.24	0.60	0.66	1.48	4.92	14.80	10.96	12.11	16.35	4.03	32.94	29.80	81.03

x	1.0	1.1	1.5	1.7	2.0	2.5	3.0	3.0	3.0	3.1	3.2	3.5	3.6	4.0
y	3.0	3.2	4.1	4.2	5.0	6.2	7.3	7.0	7.1	7.4	6.0	8.1	8.0	9.0

Le graphe des nuages de points donne:



On voit que le nuage 1 n'a pas une tendance lineaire, mais une tendance plutot exponentielle. On ne peut donc pas utiliser le modele lineaire simple dans ce cas. Mais le nuage 2 a bien une tendance lineaire. Dans ce cas, il parait raisonable de modeliser sa courbe de regression par l'equation 1.2.

2.1 Méthodes des moindres carrés

Le modèle de régression lineaire simple s'ecrit:

$$Y_i = a + bx_i + \varepsilon_i, \tag{2.1}$$

ou chaque $\varepsilon_i \sim N(0, \sigma^2)$. On cherche a estimer a et b à partir des données $(x_1, y_1), \ldots, (x_n, y_n)$.

Le principe des moindres carrés revient à estimer a et b par les valeurs qui minimisent:

$$\sum_{i=1}^{n} (y_i - a - bx_i)^2 . \tag{2.2}$$

On montre facilement le

Theoreme 2.1. *Posons*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \textit{et} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \tag{2.3}$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 , \tag{2.4}$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 , \tag{2.5}$$

and

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) . \tag{2.6}$$

Alors les valeurs de a et b qui minimisent $\sum_{i=1}^{n} (y_i - a - bx_i)^2$ sont données par:

$$\hat{a} = \bar{y} - b\bar{x} \quad \textit{et} \quad \hat{b} = \frac{S_{xy}}{S_{xx}}. \tag{2.7}$$

- $\hat{a}$ et $\hat{b}$ s'appellent les estimateurs des moindres carrés de a et b.
- $\hat{y} = \hat{a} + \hat{b}x$ s'appelle la valeur estimée ou la prediction de y.
- la droite d'equation $y = \hat{a} + \hat{b}x$ s'appelle la droite de regression estimée de Y sur X.
- $e_i = y_i - \hat{y}_i$ s'appelle le residu de l'orbservation i.
- La somme des carrés des résidus est

$$\begin{aligned} SSE &= \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2 \\ &= S_{yy} - \frac{S_{xy}^2}{S_{xx}}. \end{aligned}$$

Exemple 2.2. On a les données suivantes:

x	12.8	13.9	14.1	14.7	15.0	15.1	16.0	16.5	16.6	17.2	17.9
y	110	54	98	50	67	58	52	50	43	15	28

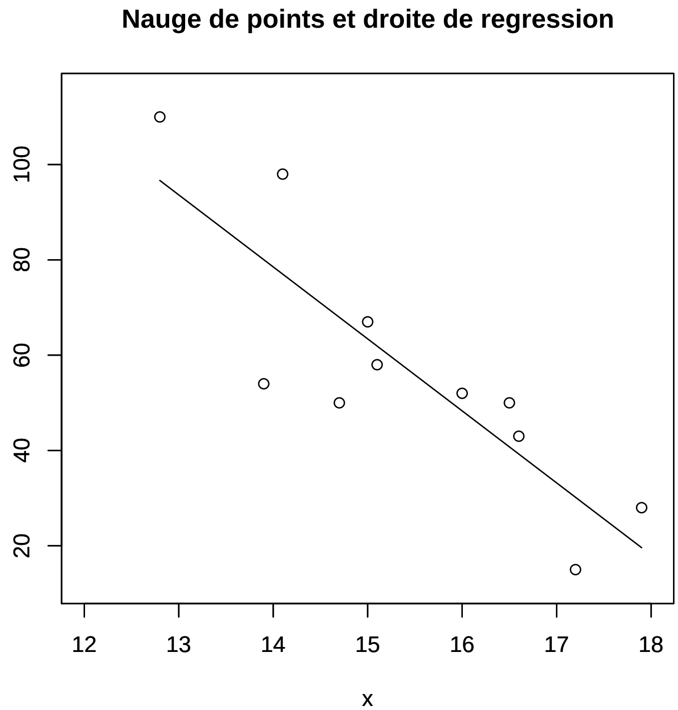
On cherche la droite de regression lineaire de y sur x.

On a besoin de calculer: $\bar{x}$, $\bar{y}$, S_{xx} et S_{xy}. A partir des données on a:

$\sum x_i = 169.8$, $\sum y_i = 625$, $\sum x_i y_i = 9286.2$, $\sum x_i^2 = 2645.02$. D'ou: $S_{xx} = \sum x_i^2 - (\sum x_i)^2/n = 2645.02 - \frac{169.8^2}{11} = 23.92.$, $S_{xy} = \sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i) = 9286.2 - \frac{169.8 * 625}{11} = -351.53$. Et $\hat{b} = \frac{S_{xy}}{S_{xx}} = \frac{-351.53}{23.92} = -15.11$. et

$\hat{a} = \bar{y} - \hat{b}\bar{x} = \frac{625}{11} - (-15.11)\frac{169.8}{11} = 290.06.$

La droite de regression lineaire est dont $y = 290.06 - 15.11x$. Le graphe ci-dessous donne le nuage de points (x, y) ainsi que la droite de regression estimée.



3 Inference sur le modele

On voudrait d'abord savoir comment estimer la variabilité qu'il y a sur la perturbation ε.

Theoreme 3.1. *On peut estimer σ^2 par:*

$$\begin{aligned} s_e^2 &= \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - (\hat{a} + \hat{b}x_i)\right)^2 \\ &= \frac{S_{yy} - (S_{xy})^2/S_{xx}}{n-2}. \end{aligned}$$

Theoreme 3.2. *Dans le modele lineaire simple, les variables aleatoires*

$$t_a = \frac{\hat{a} - a}{s_e\sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}}, \tag{3.1}$$

et

$$t_b = \frac{\hat{b} - b}{s_e/\sqrt{S_{xx}}}, \tag{3.2}$$

suivent une distribution de student à $n-2$ degre de liberté.

3.1 Intervalle de confiance pour a et b

Soit $t_{\alpha/2}$ tel que $P(T_{n-2} > t_{\alpha/2}) = \alpha/2$. Le theoreme 3.2 implique que:

$\hat{a} \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{S_{xx}}}$ est un I.C. à $(1-\alpha)$ pour a.

$\hat{b} \pm t_{\alpha/2} \frac{s_e}{\sqrt{S_{xx}}}$ est un I.C. à $(1-\alpha)$ pour b.

Exemple 3.1. Dans l'example precedent, construire un IC a 95% pour a et b.
Solution: $n = 11$ donc $t_{\alpha/2} = 2.262$.
$S_{yy} = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = 7523.64$.
$s_e^2 = \frac{S_{yy} - (S_{xy})^2/S_{xx}}{n-2} = 261.95$. D'ou $s_e = 16.18$.
A partir de la formule: $IC(a) = 290.06 \pm 15.57$.
et $IC(b) = -15.11 \pm 7.48$.

3.2 Test sur b

Tester si $b = 0$ ou pas est important en pratique parce que $b = 0$ signifie que statistiquement Y ne varie pas avec X. Donc on veut tester:
$H_0: \ b = 0$ contre $H_1: \ b \neq 0$, au seuil α.

La statistique de test est $t_b = \frac{\hat{b}}{s_e/\sqrt{S_{xx}}}$. On rejette H_0 si $t_b < -t_{\alpha/2}$ ou si $t_b > t_{\alpha/2}$, ou $t_{\alpha/2}$ est tel que $P(T_{n-2} > t_{\alpha/2}) = \alpha/2$.

Exemple 3.2. Dans l'exemple precedent, tester $H_0: \ b = 0$ contre $H_1: \ b \neq 0$ à 5%.
Solution: Comme $n = 11$, on trouve $t_{\alpha/2} = 2.262$. La statistique de test vaut $t_b = \frac{-15.11}{16.18/\sqrt{23.92}} = -4.56 < -2.262$. Donc on rejette H_0.

En fait, on pouvait prendre la meme decision rien qu'en regardant l'IC sur b qui ne contient pas 0.

3.3 Intervalle de confiance sur la moyenne de Y a $X = x_0$

Supposons qu'on connait une valeur de X, disons x_0. La valeur moyenne de Y sachant $X = x_0$ est $a + bx_0$ qu'on peut naturellement estimer par $\hat{a} + \hat{b}x_0$. On peut montrer que l'IC a $(1-\alpha)$ pour $a + bx_0$ est donne par:

$$IC(a + bx_0) = \hat{a} + \hat{b}x_0 \pm t_{\alpha/2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}. \tag{3.3}$$

Exemple 3.3. Dans l'exemple precedent, trouver un intervalle de confiance a 95% pour la reponse moyenne Y a $X = 15.5$.
Solution: L'estime de la moyenne de Y a $X = 15.5$ est $290.06 - 15.11 * 15.5 = 55.85$.

$n = 11$, $t_{\alpha/2} = 3.25$. D'ou par la formule $IC(a + bx_0) = 55.85 \pm 11.04$.

3.4 Intervalle de prediction pour Y sachant $X = x_0$

On se rappelle que d'apres le modele de regression lineaire simple, $Y_i = a + bx_i + \varepsilon_i$, ou $\varepsilon_i \sim N(0, \sigma^2)$. Si on cherche a predire Y sachant que $X = x_0$ il faut tenir de l'erreur commise en estimant a par $\hat{a}$, b par $\hat{b}$ mais aussi de la variabilité de ε_i.

On peut montrer que l'intervalle de prediction a $(1-\alpha)$ pour Y lorsque $X = x_0$ est:

$$IC(Y|X = x_0) = \hat{a} + \hat{b}x_0 \pm t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}. \tag{3.4}$$

Exemple 3.4. Trouver un intervalle de prediction a 95% pour Y lorsque $X = 15.5$.
Solution: Avec la formule on trouve $IC(Y|X = 15.5) = 55.85 \pm 38.23$.