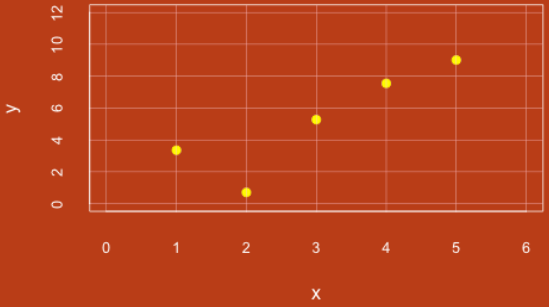
MOOC Statistique pour ingénieur
Thème 4 : Régression linéaire
Vidéo 2 : Le modèle linéaire et ses hypothèses

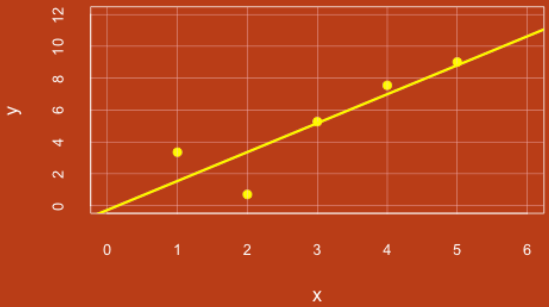
Anca Badea

Institut Mines-Télécom
Mines Saint-Étienne

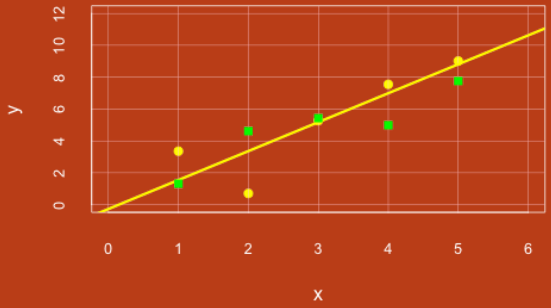
Sommaire

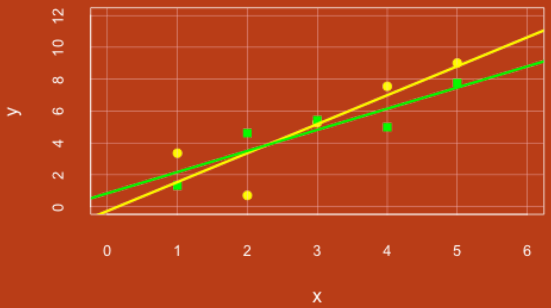
Et si on avait plusieurs jeux de données ?





Et si on avait plusieurs jeux de données ?

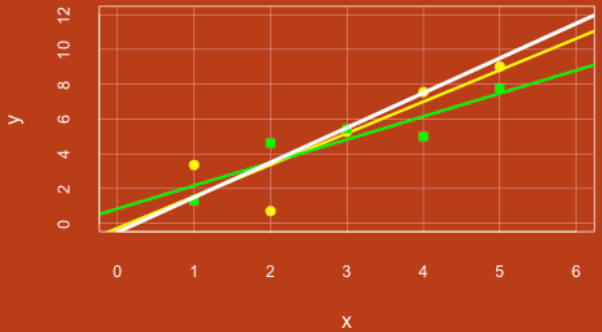




$\widehat{\beta}_0 = -0,27; \widehat{\beta}_1 = 1,82$

$\widehat{\beta}_0 = 0,84; \widehat{\beta}_1 = 1,33$

Et si on avait plusieurs jeux de données ?



$\widehat{\beta}_0 = -0,27; \widehat{\beta}_1 = 1,82$ $\widehat{\beta}_0 = 0,84; \widehat{\beta}_1 = 1,33$ $\beta_0 = -0,5; \beta_1 = 2$

La *modélisation* précédente ne prenait pas en compte cette variabilité...

Hypothèses

- la variable expliquée Y est une v.a.r.
- la variable explicative X est une v.a.r.
- *l'hypothèse* :
  - en moyenne
  - et conditionnellement aux observations de la variable explicative,
  - la variable expliquée est une fonction affine de celle-ci

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$$

ou bien

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

simplification : X déterministe

$$\mathbb{E}(Y) = \beta_0 + \beta_1 x$$

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Hypothèses

- l'erreur ε est une v.a.r. $\mathbb{E}(\varepsilon) = 0$, $\mathbb{V}(\varepsilon) = \sigma^2$

  σ paramètre à estimer
  en plus de β_0, β_1

- hypothèse supplémentaire $\boxed{\varepsilon \sim \mathcal{N}(0, \sigma^2)}$

- questions :
  - comment se *propage* cette hypothèse ?
  - quels estimateurs pour les paramètres à estimer ?

Sommaire

1. Formalisation
2. Estimation / estimateurs
3. Exemple

Méthodes

- moindres carrés
- maximum de vraisemblance

pour $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \ldots, n$
$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ **v. a. i. i. d.**

conduisent aux mêmes estimations / estimateurs pour β_0, β_1

de plus
$\mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i$

$\mathbb{V}(Y_i) = \mathbb{V}(\varepsilon_i) = \sigma^2$

$$\boxed{Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)}$$

Estimateur / estimation de σ^2

pour $\sigma^2 = \mathbb{E}(\varepsilon_i^2)$ l'estimateur obtenu par la méthode du maximum de vraisemblance est l'estimateur classique de l'espérance

(i.e. la moyenne empirique)

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

dont l'espérance est égale à $\mathbb{E}(\widehat{\sigma}^2) = \frac{n-2}{n}\sigma^2$

et alors on peut définir un estimateur non-biaisé de σ^2 comme

$$\widehat{\sigma}^{*2} = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

on utilisera la même notation pour l'estimation correspondante

Estimateurs / estimations de $\beta_0,\ \beta_1$

$$\begin{cases} \widehat{\beta}_1 &= \dfrac{1}{n s_x^2} \displaystyle\sum_{i=1}^{n} (x_i - \overline{x}) Y_i = \sum_{i=1}^{n} c_i Y_i \\ \\ \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{x} \end{cases}$$

calculons leurs espérances

$$\mathbb{E}(\widehat{\beta}_1) = \sum_{i=1}^{n} c_i \mathbb{E}(Y_i) = \beta_0 \sum_{i=1}^{n} c_i + \beta_1 \sum_{i=1}^{n} c_i x_i = \beta_1$$

en utilisant $\displaystyle\sum_{i=1}^{n} c_i = 0$ et $\displaystyle\sum_{i=1}^{n} c_i x_i = 1$

$$\mathbb{E}(\widehat{\beta}_0) = \beta_0$$

Estimateurs / estimations de $\beta_0,\ \beta_1$

leurs variances

$$\mathbb{V}(\widehat{\beta}_1) = \frac{\sigma^2}{n s_x^2} \qquad \mathbb{V}(\widehat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{n s_x^2} \right)$$

$$\mathbb{C}ov(\widehat{\beta}_0, \widehat{\beta}_1) = -\sigma^2 \frac{\overline{x}}{n s_x^2}$$

Théorème (Gauss-Markov)

Pour le modèle de régression $Y = \beta_0 + \beta_1 x + \varepsilon$ et sous les hypothèses précédentes pour ε, les estimateurs MC $\widehat{\beta}_0$, $\widehat{\beta}_1$ sont

- *des combinaisons linéaires des Y_i,*
- *sans biais,*
- *de variance minimale (comparés à tous les autres estimateurs sans biais).*

Distributions

des estimateurs $\widehat{\beta}_0,\ \widehat{\beta}_1$

$\widehat{\beta}_i \sim \mathcal{N}(\beta_i, \mathbb{V}(\widehat{\beta}_i))$

$\sigma^2 \to \widehat{\sigma}^{*2}$

$\mathbb{V}(\widehat{\beta}_i) \to s^2_{\widehat{\beta}_i}$

$s^2_{\widehat{\beta}_1} = \frac{\widehat{\sigma}^{*2}}{n s^2_x};\ s^2_{\widehat{\beta}_0} = \widehat{\sigma}^{*2}\left(\frac{1}{n} + \frac{\bar{x}^2}{n s^2_x}\right)$

$$\frac{\widehat{\beta}_i - \beta_i}{s_{\widehat{\beta}_i}} \sim \mathcal{T}(n-2)$$

loi de Student à $n-2$ degrés de libertés

de l'estimateur de σ^2

$$\frac{n-2}{\sigma^2}\widehat{\sigma}^{*2} \sim \chi^2_{n-2}$$

loi du χ^2 à $n-2$ degrés de libertés

Estimations par intervalle de confiance

$$\mathbb{P}\left(-t_{\alpha/2} \leq \frac{\widehat{\beta}_i - \beta_i}{s_{\widehat{\beta}_i}} \leq t_{\alpha/2}\right) = 1 - \alpha$$

avec $t_{\alpha/2}$ la valeur telle que $\mathbb{P}\left(T \leq t_{\alpha/2}\right) = 1 - \alpha/2$ et T de loi de Student $\mathcal{T}(n-2)$

$$Ic_{1-\alpha}(\beta_i) = \left[\widehat{\beta}_i - t_{\alpha/2} \times s_{\widehat{\beta}_i},\ \widehat{\beta}_i + t_{\alpha/2} \times s_{\widehat{\beta}_i}\right]$$

Estimations par intervalle de confiance

$$\mathbb{P}\left(\chi_1^2 \leq \frac{(n-2)\widehat{\sigma}^{*2}}{\sigma^2} \leq \chi_2^2\right) = 1-\alpha$$ avec χ_1^2 et χ_2^2 les valeurs telles que

$\mathbb{P}(Z \leq \chi_1^2) = \alpha/2$ et $\mathbb{P}(Z \leq \chi_2^2) = 1-\alpha/2$ et Z de loi χ_{n-2}^2

$$Ic_{1-\alpha}(\sigma^2) = \left[\frac{(n-2)\widehat{\sigma}^{*2}}{\chi_2^2}, \frac{(n-2)\widehat{\sigma}^{*2}}{\chi_1^2}\right]$$

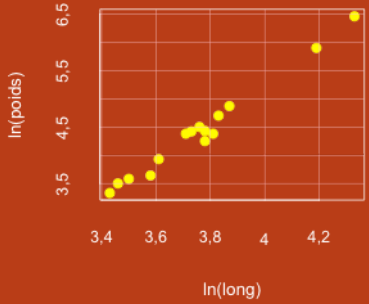
Sommaire

Exemple

mesures sur 15 alligators [1]

- le poids (en livres)
- la distance entre l'arrière de la tête à l'extrémité du nez (en pouces)
- échelle logarithmique

	x ln (long)	y ln(poids)
1	3,87	4,87
2	3,61	3,93
3	4,33	6,46
4	3,43	3,33
5	3,81	4,38
6	3,83	4,70
7	3,46	3,50
8	3,76	4,50
9	3,50	3,58
10	3,58	3,64
11	4,19	5,90
12	3,78	4,43
13	3,71	4,38
14	3,73	4,42
15	3,78	4,25



1. Mendenhall, Wackerly, Scheaffer *Mathematical Statistics with Applications* (1990)

Exemple

$n = 15$; $\bar{x} \approx 3,76$; $\bar{y} \approx 4,42$; $s_x^2 \approx 0,06$; $s_y^2 \approx 0,68$; $\mathbb{C}\text{ov}_{xy} \approx 0,2$

$$\widehat{\beta}_1 = \frac{\mathbb{C}\text{ov}_{xy}}{s_x^2} \approx 3,43 \; ; \qquad \widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x} \approx -8,48$$

$$\widehat{\sigma}^{*2} = \frac{1}{13} \sum_{i=1}^{15} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 \approx 0,02 \; ; \; s_{\widehat{\beta}_1} \approx 0,13 \; ; \; s_{\widehat{\beta}_0} \approx 0,5$$

$t_{0,025} = 2,16$; $\chi^2_1 = 5,01$; $\chi^2_2 = 24,74$

$$Ic_{0,95}(\beta_0) \approx [-8,48 - 2,16 \times 0,5;\; -8,48 + 2,16 \times 0,5] \approx [-9,56;\; -7,4]$$

$$Ic_{0,95}(\beta_1) \approx [3,43 - 2,16 \times 0,13;\; 3,43 + 2,16 \times 0,13] \approx [3,15;\; 3,72]$$

$$Ic_{0,95}(\sigma^2) \approx \left[\frac{13 \times 0,02}{24,74}, \; \frac{13 \times 0,02}{5,01} \right] \approx [0,01;\; 0,05]$$

Exemple

Variable	Coefficient	Ecart-type
Intercept	$-8,48$	$0,5$
lnLength	$3,43$	$0,13$

$\widehat{\sigma}^*$	ddl	R^2	R^2_{adj}
$0,12$	13	$0,9808$	$0,9794$


INSTITUT
Mines-Télécom