Statistique : Modèle de régression linéaire : formulation, résolution, analyse de performance

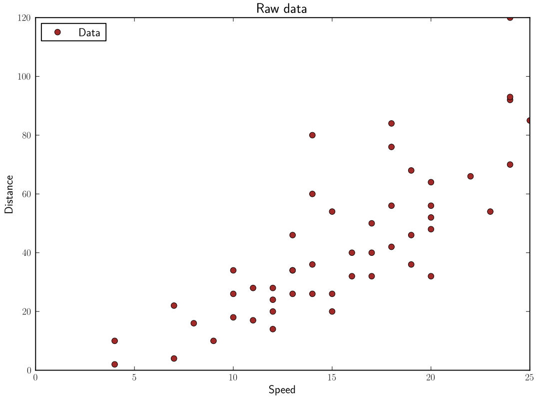
Joseph Salmon

Septembre 2014

Plan du cours

Exemple en dimension deux

Exemple : distance de freinage d'une voiture en fonction de la vitesse (50 mesures)

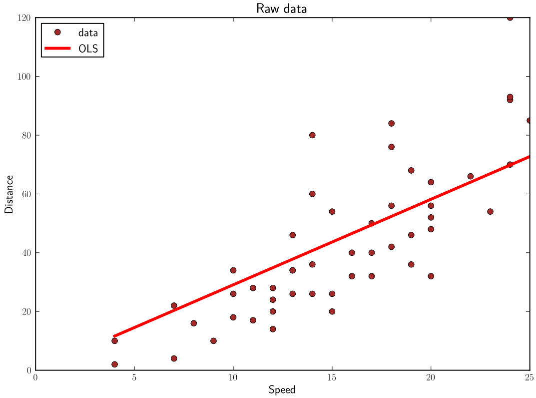


Dataset *cars* :

https://forge.scilab.org/index.php/p/rdataset/source/file/master/csv/datasets/cars.csv

Exemple en dimension deux

Exemple : distance de freinage d'une voiture en fonction de la vitesse (50 mesures)



Dataset *cars* :

`https://forge.scilab.org/index.php/p/rdataset/source/file/master/csv/datasets/cars.csv`

Modélisation I

Statistique : Modèle de régression linéaire : formulation, résolution, analyse de performance

Introduction régression

Moindres carrés uni-dimensionnels

Moindres carrés multi-dimensionnels

Joseph Salmon

Jeu d'observations : (y_i, x_i), pour $i = 1, \dots, n$

Hypothèse de modèle linéaire ou de régression linéaire :

$$\boxed{y_i \approx \theta_0^* + \theta_1^* x_i}$$

- θ_1^* coefficient directeur
- θ_0^* ordonnée à l'origine

Rem: Les deux paramètres sont inconnus du statisticien

Définition

- y est une **observation** ou une variable à expliquer
- x est une variable **explicative** ou *feature* en anglais

Interprétation des notations

Exemple : dataset *cars*

- $n = 50$
- y_i : temps de freinage de la voiture i
- x_i : vitesse de la voiture i
- x : l'observation est le temps de freinage
- y : la variable explicative est la vitesse
- l'hypothèse de la régression linéaire/modèle linéaire revient à postuler que le temps de freinage d'une voiture est proportionnel à sa vitesse (ou plutôt affine)

McKinney (2012) : `python` pour les statistiques

Modélisation II

On donne un sens au symbole $\approx$ de la manière suivante :

Modèle probabiliste

$$
\begin{aligned}
& y_i = \theta_0^* + \theta_1^* x_i + \varepsilon_i, \\
& \varepsilon_i \overset{i.i.d}{\sim} \varepsilon, \text{ pour } i = 1, \dots, n \\
& \mathbb{E}(\varepsilon) = 0
\end{aligned}
$$

Interprétation

$\varepsilon_i = y_i - \theta_0^* - \theta_1^* x_i$: erreur(s) entre le modèle théorique et les observations, représentées par des variables aléatoires ε_i centrées (on parle aussi de **bruit blanc**).

Rem: L'aspect aléatoire peut avoir diverses causes : bruit de mesures, bruit de transmission, variabilité dans une population, etc.

Modélisation III

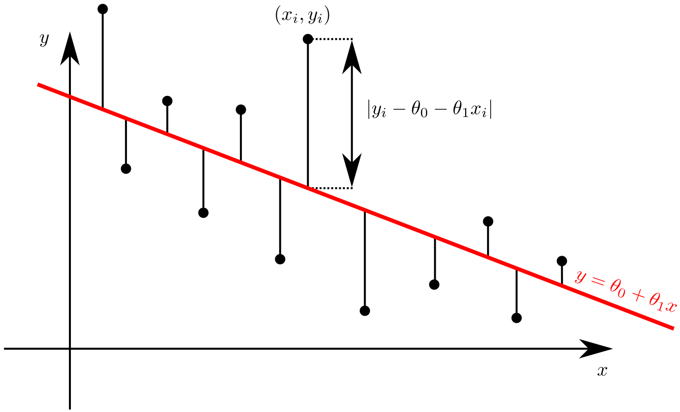
Objectif

Estimer θ_0 et θ_1 par des quantités $\hat{\theta}_0$ et $\hat{\theta}_1$ dépendant des observations (y_i, x_i) pour $i = 1, \ldots, n$

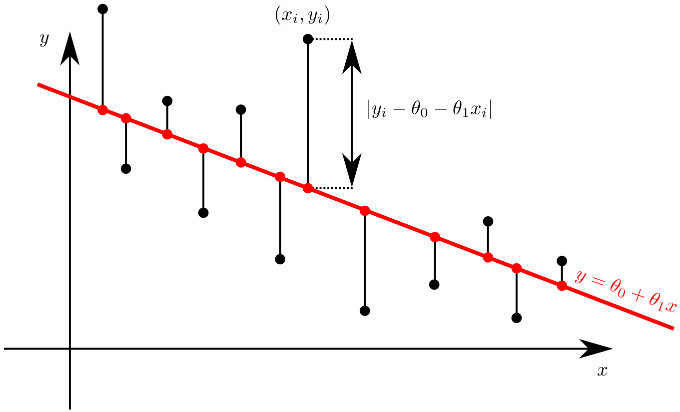
Estimateur des moindres carrés : visualisation

Statistique : Modèle de régression linéaire : formulation, résolution, analyse de performance

Joseph Salmon



Estimateur des moindres carrés : visualisation



Estimateur des moindres carrés : formalisation

$$\boxed{\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, \hat{\theta}_1) \in \mathop{\arg\min}_{(\theta_0,\theta_1)\in\mathbb{R}^2} \sum_{i=1}^{n} |y_i - \theta_0 - \theta_1 x_i|^2}$$

On cherche donc à minimiser une fonction de deux variables :

$$f(\theta_0, \theta_1) = f(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \theta_0 - \theta_1 x_i)^2$$

Solution : $\begin{cases} \hat{\theta}_0 = \bar{y}_n - \hat{\theta}_1 \bar{x}_n \\ \hat{\theta}_1 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \end{cases}$

Rem: la formule est vraie ssi $\mathbf{x} = (x_1, \ldots, x_n)^\top$ est non constant

Statistique : Modèle de régression linéaire : formulation, résolution, analyse de performance

Définitions

Prédicteur

On appelle **prédicteur** une fonction qui à une nouvelle valeur de la variable explicative x_{n+1} propose une estimation de la variable à expliquer

$$\mathrm{pred}(x_{n+1}) = \hat{\theta}_0 + \hat{\theta}_1 x_{n+1}$$

Rem: Souvent on note $\hat{y}_{n+1} = \mathrm{pred}(x_{n+1})$ s'il n'y pas d'ambiguité

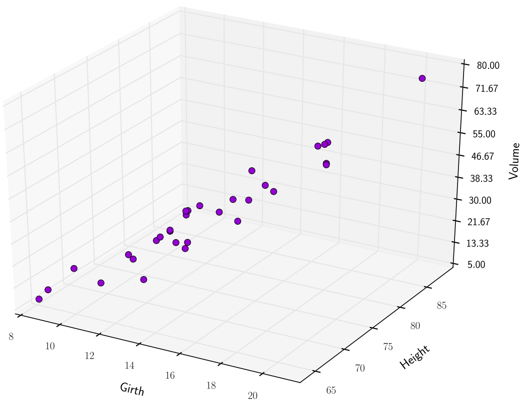
Résidus

On appelle **résidu** d'un prédicteur la différence entre la valeur observée et la valeur du prédicteur prise pour une valeur de la variable explicative observée :

$$r_i = y_i - \mathrm{pred}(x_i) = y_i - \hat{y}_i = y_i - (\hat{\theta}_0 + \hat{\theta}_1 x_i)$$

Vers des modèles avec multi-variés

Volume d'arbres en fonction de leur hauteur / circonférence



Dataset *trees* : `http://vincentarelbundock.github.io/Rdatasets/csv/datasets/trees.csv`

Statistique : Modèle de régression linéaire : formulation, résolution, analyse de performance

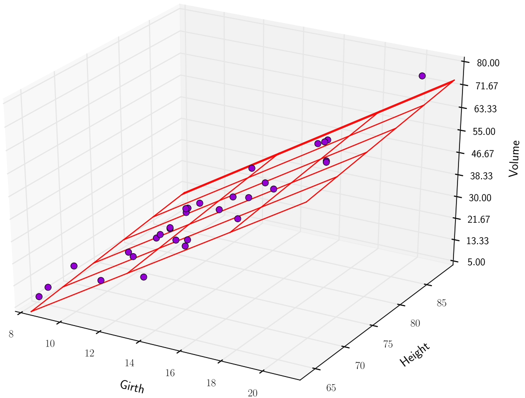
Introduction régression

Moindres carrés uni-dimensionnels

Moindres carrés multi-dimensionnels

Vers des modèles avec multi-variés

Volume d'arbres en fonction de leur hauteur / circonférence



Dataset *trees* : `http://vincentarelbundock.github.io/Rdatasets/csv/datasets/trees.csv`

Modélisation

On dispose de p variables explicatives

Modèle en dimension p

$$y_i = \theta_0^* + \sum_{j=1}^{p} \theta_j^* x_{i,j} + \varepsilon_i$$

$$\varepsilon_i \overset{i.i.d}{\sim} \varepsilon, \text{ pour } i = 1, \dots, n$$

$$\mathbb{E}(\varepsilon) = 0$$

De manière équivalente :

$$\begin{cases} y_1 & = \theta_0^* + \sum_{j=1}^{p} \theta_j^* x_{1,j} + \varepsilon_1 \\ & \vdots \\ y_n & = \theta_0^* + \sum_{j=1}^{p} \theta_j^* x_{n,j} + \varepsilon_n \end{cases}$$

Lejeune (2010) concernant le modèle linéaire (notamment)

Dimension p

Modéle matriciel

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} \theta_0^* \\ \theta_1^* \\ \vdots \\ \theta_p^* \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} X = \begin{pmatrix} 1 & x_{1,1} & \dots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \dots & x_{n,p} \end{pmatrix} \boldsymbol{\theta}^* = \begin{pmatrix} \theta_0^* \\ \theta_1^* \\ \vdots \\ \theta_p^* \end{pmatrix} \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$\boxed{\mathbf{y} = X\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}}$ ou $y_i = \langle X_{i,:}, \boldsymbol{\theta}^* \rangle + \varepsilon_i$ pour $i = 1, \dots, n$

Rem: $\boldsymbol{\theta}^*$ est le vrai paramètre du modèle que l'on veut retrouver.
Rem: On note aussi parfois $X = (\mathbf{1}_n, \mathbf{x}_1, \dots, \mathbf{x}_p)$

Estimateur des moindres carrés

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta} \in \mathbb{R}^{p+1}}{\arg\min} \left(\|\mathbf{y} - X\boldsymbol{\theta}\|_2^2\right) = \sum_{i=1}^{n} \left[y_i - \left(\theta_0 + \sum_{j=1}^{p} \theta_j x_{i,j} \right) \right]^2$$

Équations normales

La CNO nous assure que le minimiseur $\hat{\boldsymbol{\theta}}$ satisfait l'équation :

$$\boxed{(X^\top X)\hat{\boldsymbol{\theta}} = X^\top \mathbf{y}}$$

Rem: résidus orthogonaux aux variables $X^\top(X\hat{\boldsymbol{\theta}} - \mathbf{y}) = 0$

Formule fermée

Si la matrice X est de plein rang (*i.e.*, si $X^\top X$ inversible)

$$\hat{\boldsymbol{\theta}} = (X^\top X)^{-1} X^\top \mathbf{y}$$

Références I

M. Lejeune.
Statistiques, la théorie et ses applications.
Springer, 2010.

W. McKinney.
Python for Data Analysis : Data Wrangling with Pandas, NumPy, and IPython.
O'Reilly Media, 2012.