La régression linéaire

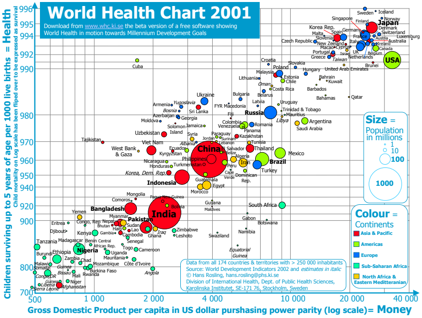
Stéphane Canu
stephane.canu@litislab.eu

M8 - Principes du traitement de l'information

March 26, 2019

Plan

Observation d'un couple de variables



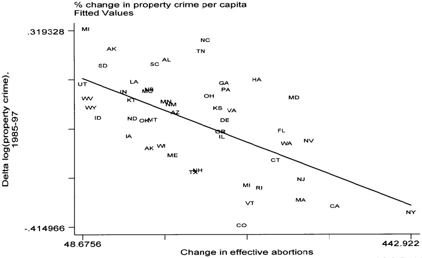
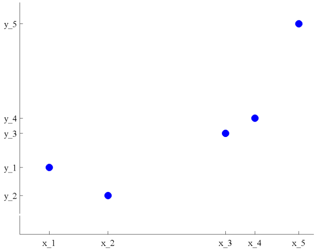


Figure 4b: Changes in Property Crime and Abortion Rates, 1985-1997

Figure : Exemples de données qui semblent linéairement liées
(http://www.math.yorku.ca/SCS/Gallery/, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=174508)

Le problème de la régression linéaire

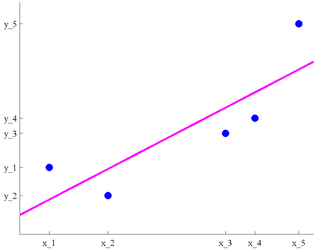


x	1	2	4	4,5	5,25
y	-1	-2	0,5	1	4

$\mathbf{x}$	x_1	x_2	x_3	x_4	x_5
$\mathbf{y}$	y_1	y_2	y_3	y_4	y_5

- n couples d'observations
  - x **: la variable explicative** :
    quantité d'additif, années, température
  - y **la variable à expliquer** :
    quantité de produit, nombre de ventes, concentration en ozone

Le problème de la régression linéaire



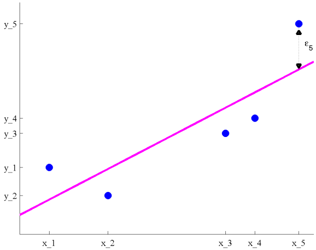
x	1	2	4	4,5	5,25
y	-1	-2	0,5	1	4

x	x_1	x_2	x_3	x_4	x_5
y	y_1	y_2	y_3	y_4	y_5

- n couples d'observations
  - x **: la variable explicative** :
    quantité d'additif, années, température
  - y **la variable à expliquer** :
    quantité de produit, nombre de ventes, concentration en ozone
- Trouver la droite qui représente « au mieux » la relation linéaire

$$y = ax + b$$

Le problème de la régression linéaire



x	1	2	4	4,5	5,25
y	-1	-2	0,5	1	4

x	x_1	x_2	x_3	x_4	x_5
y	y_1	y_2	y_3	y_4	y_5

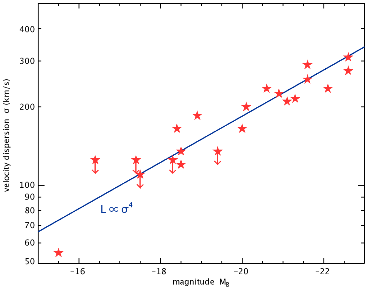
- *n* couples d'observations
  - x **: la variable explicative** :
    quantité d'additif, années, température
  - y **la variable à expliquer** :
    quantité de produit, nombre de ventes, concentration en ozone
- Trouver la droite qui représente « au mieux » la relation linéaire

$$y = ax + b$$

- difficulté : les observations ne sont pas exactement sur la droite : il y a du bruit

$$y_i = ax_i + b + \varepsilon_i$$

Le modèle linéaire



Donnnées (x_i, y_i) inconnues : (a, b, ε_i)

Definition (Le Modèle linéaire)

Le modèle linéaire pose la relation suivante entre la variable explicative x et la variable à expliquer y avec les paramètres inconnus (a, b, ε)

$$y = ax + b \;\; + \varepsilon \qquad \text{avec} \qquad \mathbf{a} = (a, b) \in \mathbb{R}^2$$

Observations = modèle + bruit $\qquad \mathbb{E}(\varepsilon) = 0$ et $\mathbb{E}(\varepsilon^2) = \sigma^2$

Vocabulaire

- variables explicatives . . . . $\mathbf{x} \in \mathbb{R}^p$
- variable à expliquer . . . . $y \in \mathbb{R}$
- erreur ou bruit . . . . ε
- paramètres scalaires . . . . $a, b \in \mathbb{R}$,
- paramètres vectoriels . . . . $\mathbf{a} \in \mathbb{R}^p$
- modèle . . . . $y = f(\mathbf{x}, \mathbf{a}) + \varepsilon$
- estimation . . . . $\widehat{a}$
- prédiction . . . . $\widehat{y} = f(\mathbf{x}, \widehat{a})$
- variables aléatoires . . . . ε et donc y et donc $\widehat{a}$

$$y = \underbrace{ax + b}_{f(x,\mathbf{a})} + \varepsilon$$

Un exemple : l'étalonnage d'un capteur

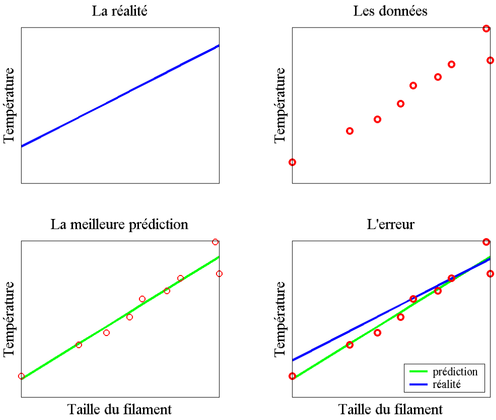


Figure : Les différentes phases de la régression

Plan

Le critère des moindres carrés

$$\min_{a,b} J(a,b)$$

$$J(a,b) = \sum_{i=1}^{n} \big(\underbrace{y_i - ax_i - b}_{\color{red}{\varepsilon_i}}\big)^2$$

Ce problème peut s'interpréter comme la recherche de la droite d'équation $ax + b$ passant « au mieux » (au sens des moindres carrés) parmi le nuage des observations (x_i, y_i) $i = 1, n$.

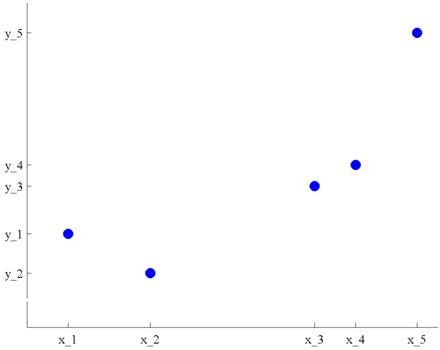


Figure : Les résidus de la régression

Le critère des moindres carrés

$$\min_{a,b} J(a,b)$$

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

Ce problème peut s'interpréter comme la recherche de la droite d'équation $ax + b$ passant « au mieux » (au sens des moindres carrés) parmi le nuage des observations (x_i, y_i) $i = 1, n$.

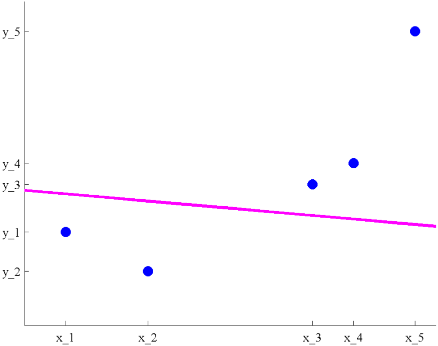


Figure : Les résidus de la régression

Le critère des moindres carrés

$$\min_{a,b} J(a, b)$$

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

Ce problème peut s'interpréter comme la recherche de la droite d'équation $ax + b$ passant « au mieux » (au sens des moindres carrés) parmi le nuage des observations (x_i, y_i) $i = 1, n$.

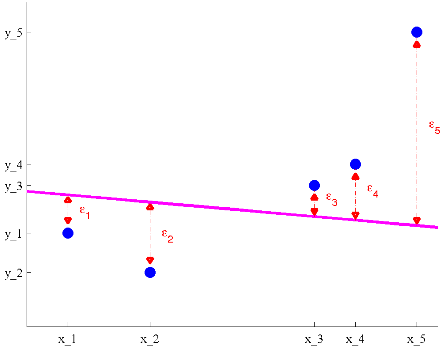


Figure : Les résidus de la régression

Le critère des moindres carrés

$$\min_{a,b} J(a,b)$$

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

Ce problème peut s'interpréter comme la recherche de la droite d'équation $ax + b$ passant « au mieux » (au sens des moindres carrés) parmi le nuage des observations (x_i, y_i) $i = 1, n$.

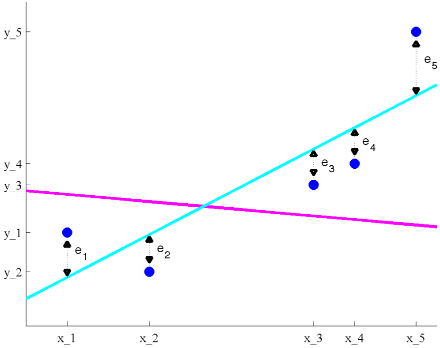


Figure : Les résidus de la régression

Le problème des moindres carrés

Les données dont nous disposons peuvent aussi être vue comme un système de n équations à $2+n$ inconnues (a, b et ε). Ce système s'écrit de la manière suivante :

$$\begin{cases} ax_1 + b + \varepsilon_1 & = y_1 \\ \quad ... & \quad ... \\ ax_i + b + \varepsilon_i & = y_i \\ \quad ... & \quad ... \\ ax_n + b + \varepsilon_n & = y_n \end{cases}$$

on recherche a et b qui minimisent simultanément tous les $\varepsilon_i, i = 1, n$

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

Calcul du gradient

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

$$\left\{\begin{array}{rcll}
\dfrac{\partial J(a,b)}{\partial a} & = & \displaystyle\sum_{i=1}^{n}(ax_i + b - y_i)\,x_i & = \displaystyle\sum_{i=1}^{n}\left(ax_i^2 + bx_i - y_i x_i\right) \\
 & & & = \displaystyle\sum_{i=1}^{n} ax_i^2 + \sum_{i=1}^{n} bx_i - \sum_{i=1}^{n} y_i x_i \\
 & & & = a\displaystyle\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i x_i \\
\dfrac{\partial J(a,b)}{\partial b} & = & \displaystyle\sum_{i=1}^{n}(ax_i + b - y_i) & = \displaystyle\sum_{i=1}^{n} ax_i + \sum_{i=1}^{n} b - \sum_{i=1}^{n} y_i \\
 & & & = a\displaystyle\sum_{i=1}^{n} x_i + bn - \sum_{i=1}^{n} y_i
\end{array}\right.$$

Conditions d'optimalité

$(\widehat{a}, \widehat{b})$ est solution du problème $\min\limits_{a,b} J(a,b) \quad \Leftrightarrow \quad \begin{cases} \dfrac{\partial J(\widehat{a}, \widehat{b})}{\partial a} = 0 \\ \dfrac{\partial J(\widehat{a}, \widehat{b})}{\partial b} = 0 \end{cases}$

$$\begin{cases} \dfrac{\partial J(\widehat{a}, \widehat{b})}{\partial a} = 0 \\ \dfrac{\partial J(\widehat{a}, \widehat{b})}{\partial b} = 0 \end{cases} \quad \Leftrightarrow \quad \begin{array}{llll} \widehat{a} \displaystyle\sum_{i=1}^n x_i^2 + \widehat{b} \sum_{i=1}^n x_i & = \displaystyle\sum_{i=1}^n y_i x_i & \times n & (1) \\ \widehat{a} \displaystyle\sum_{i=1}^n x_i + \widehat{b}\, n & = \displaystyle\sum_{i=1}^n y_i & \times \displaystyle\sum_{i=1}^n x_i & (2) \end{array}$$

$$\widehat{a} \left(n \sum_{i=1}^n x_i^2 - \Big(\sum_{i=1}^n x_i\Big)^2 \right) = n \sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i$$

Deux équations linéaires à deux inconnues

$$\widehat{a} = \frac{n \displaystyle\sum_{i=1}^n y_i x_i - \sum_{i=1}^n y_i \sum_{i=1}^n x_i}{n \displaystyle\sum_{i=1}^n x_i^2 - \Big(\sum_{i=1}^n x_i\Big)^2} \qquad \widehat{b} = \frac{\displaystyle\sum_{i=1}^n y_i - \widehat{a} \sum_{i=1}^n x_i}{n}$$

Réécriture de la solution

si l'on note $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ et $V_x = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$ la variance des x_i

$$\begin{aligned}
\widehat{a} &= \frac{n\sum_{i=1}^{n} y_i x_i - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \frac{1}{n}\sum_{i=1}^{n} y_i \; \frac{1}{n}\sum_{i=1}^{n} x_i}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \overline{y}\,\overline{x}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\mathrm{cov}(\mathbf{x}, \mathbf{y})}{V_x}
\end{aligned}$$

$$\begin{aligned}
\widehat{b} &= \frac{\sum_{i=1}^{n} y_i - \widehat{a}\sum_{i=1}^{n} x_i}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i - \widehat{a}\,\frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \overline{y} - \widehat{a}\,\overline{x}
\end{aligned}$$

Le théorème des moindres carrés

Théorème des moindres carrées

Soient $(x_i, y_i), i = 1, n$ un ensemble de couples d'observations.

La solution du problème de minimisation de la somme des carrées des erreurs

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

est donnée par $\widehat{a}$ et $\widehat{b}$ définis par :

$$\widehat{a} = \frac{\displaystyle\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\displaystyle\sum_{i=1}^{n} (x_i - \overline{x})^2} = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{V_x} \qquad \text{et} \qquad \widehat{b} = \overline{y} - \widehat{a}\,\overline{x}$$

$\widehat{a}$ et $\widehat{b}$ que l'on appelle aussi les estimateurs au sens des moindres carrés.

Le calcul pratique

Pratiquement, il est pratique de calculer les quantités suivante :

$$SX = \sum_{i=1}^{n} x_i \quad \overline{x} = \frac{SX}{n} \quad SXX = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$SY = \sum_{i=1}^{n} y_i \quad \overline{y} = \frac{SY}{n} \quad SYY = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$SXY = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}$$

et en déduire :

$$\widehat{a} = \frac{SXY}{SXX} \qquad \text{et} \qquad \widehat{b} = \overline{y} - \widehat{a}\,\overline{x}$$

Le tableau des moindres carrés

variables explicatives x	variables à expliquer y	prédiction $y_p = \widehat{a}x_i + \widehat{b}$	résidus $\varepsilon_i = y - y_p$
0.1000	2.3196	2.0681	0.2514
0.3000	2.1000	2.3061	-0.2061
0.7000	2.5836	2.7820	-0.1984
0.8000	2.7000	2.9010	-0.2010
0.9500	3.4335	3.0795	0.3540

```
sx2 = sum((x-mean(x)).^2), sy2 = sum((y-mean(y)).^2), sxy=((x-mean(x))'*(y-mean(y))
sx2 =    0.5080                     sy2 =    1.0299                          sxy =    0.6044

a = sxy/sx2 =    1.1898
b = mean(y) - a*mean(x)  =    1.9492

yp = a*x+b
e = y - yp
```

Le poids des individus et le choc des variables

Une première remarque permet d'évaluer l'influence de chacune des observation dans le calcul de coefficient directeur $\widehat{a}$. En effet on a :

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

$w_i = \dfrac{x_i - \overline{x}}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$ mesure le poids de l'observation x_i dans le calcul de $\widehat{a}$.

Le lien de $\widehat{a}$ avec la corrélation est évident puisque l'on a :

$$\widehat{a} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{s_x^2}$$

Il peut aussi s'écrire en fonction de la corrélation :

$$\widehat{a} = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{s_x s_x}\frac{s_y}{s_y} = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{s_x s_y}\frac{s_y}{s_x} = \text{cor}(\mathbf{x}, \mathbf{y})\frac{s_y}{s_x}$$

L'équation de la droite de régression s'écrit aussi :

$$y = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{s_x^2}(x - \overline{x}) + \overline{y}$$

car

$$\begin{aligned} y &= \widehat{a}x + \widehat{b} \\ &= \widehat{a}x + \overline{y} - \widehat{a}\,\overline{x} \\ &= \widehat{a}(x - \overline{x}) + \overline{y} \quad = \frac{\text{cov}(\mathbf{x},\mathbf{y})}{s_{\mathbf{x}}^{2}}(x - \overline{x}) + \overline{y} \end{aligned}$$

La droite de régression passe par le point $(\overline{x}, \overline{y})$, le centre de gravité du nuage de points car $\widehat{a\overline{x}} + \widehat{b} = \overline{y}$.

Prédiction et erreur de prédiction

Pour un $\mathbf{x}$ donné, $y_x = f(\mathbf{x})$ est une v.a.

$$y_x = \widehat{a}\mathbf{x} + \widehat{b} + \varepsilon = \overline{y} + \widehat{a}(\mathbf{x} - \overline{x}) + \varepsilon$$

on calcule son espérance : $\mathbb{E}(y_x) = a\mathbf{x} + b$
que l'on estime par :

$$\widehat{\mathbb{E}(y_x)} = \widehat{a}\mathbf{x} + \widehat{b} = \overline{y} + \widehat{a}(\mathbf{x} - \overline{x})$$

Il est aussi possible d'associer à cette prévision une mesure d'incertitude :

$$V(y_x) = V(\overline{y}) + V(\widehat{a})(\mathbf{x} - \overline{x})^2 + \sigma^2$$

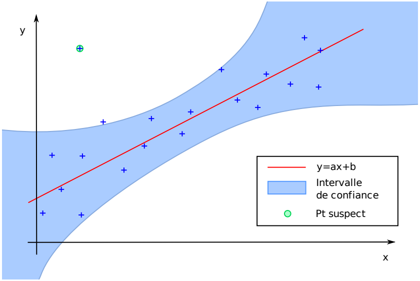


Figure : prédiction avec la régression

Théorème : pour un x donné il est très probable que (avec une probabilité α)

$$y_x \in \left\{ \overline{y} + \widehat{a}(\mathbf{x} - \overline{x}) \pm t_{\frac{\alpha}{2}}\, \widehat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(\mathbf{x} - \overline{x})^2}{\sum_{i=1}^{n}(\mathbf{x}_i - \overline{x})^2}} \right\}$$

Plus on s'éloigne de la moyenne des x, plus l'incertitude grandit

Démonstration

$$
\begin{aligned}
V(\widehat{a}) &= V\left(\frac{\sum_{i=1}^{n}(y_i-\overline{y})(x_i-\overline{x})}{\sum_{i=1}^{n}(\mathbf{x}_i-\overline{x})^2}\right)\\
&= \left(\tfrac{1}{\sum_{i=1}^{n}(\mathbf{x}_i-\overline{x})^2}\right)^2 V\left(\sum_{i=1}^{n}\left(a(x_i-\overline{x})+\varepsilon\right)(x_i-\overline{x})\right)\\
&= \left(\tfrac{1}{\sum_{i=1}^{n}(\mathbf{x}_i-\overline{x})^2}\right)^2 V\left(\sum_{i=1}^{n}\varepsilon\,(x_i-\overline{x})\right)\\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(\mathbf{x}_i-\overline{x})^2}
\end{aligned}
$$

D'où

$$
\begin{aligned}
V(y_x) &= V(\overline{y}) + V(\widehat{a})(\mathbf{x}-\overline{x})^2+\sigma^2\\
&= \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^{n}(\mathbf{x}_i-\overline{x})^2}(\mathbf{x}-\overline{x})^2+\sigma^2\\
&= \sigma^2\left(1+\frac{1}{n}+\frac{(\mathbf{x}-\overline{x})^2}{\sum_{i=1}^{n}(\mathbf{x}_i-\overline{x})^2}\right)
\end{aligned}
$$

Puis on estime σ^2 par la moyenne pondérée des carrés des résidus

$$
\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i-(\widehat{a}x_i+\widehat{b})\right)^2
$$

Codage de la régression en matab et en python

```
sx2 = sum((x-mean(x)).^2),
sy2 = sum((y-mean(y)).^2),
sxy=((x-mean(x))'*(y-mean(y)))

a = sxy/sx2
b = mean(y) - a*mean(x)

yp = a*x+b
e = y - yp

n = length(e);
sigma2_hat = 1/(n-2)* e'*e;

xt = 0:0.01:1;
yt = a*xt+b;
t = icdf('t',0.025,n-2);
interval = t*sqrt(sigma2_hat*
     (1+1/n+(xt-mean(x)).^2/sx2));

plot(xt,yt);
hold on
plot(xt,yt+interval,'--')
plot(xt,yt-interval,'--')
```

```
sx2 = np.sum((x-np.mean(x))**2)
sy2 = np.sum((y-np.mean(y))**2)
sxy=((x-np.mean(x)).T @ (y-np.mean(y)))

a = sxy/sx2
b = np.mean(y) - a*np.mean(x)

yp = a*x+b;
e = y - yp

n = len(e);
sigma2_hat = 1/(n-2)* e.T@e

xp =  np.arange(0,1,.01)
yp = a*xp+b
tv = t.ppf(0.025,n-2)
interval = tv*np.sqrt(sigma2_hat*
     (1+1/n+(xp-np.mean(x))**2/sx2));

plt.plot([xp, xp],[yp-interval, yp+interval]
plt.plot([xm,xM],[ym,yM],'gv-', linewidth=2)
plt.plot(x,y,'b+')
plt.show()
```

ACP pour $p = 2$ et régression linéaire

Régression linéaire :

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

ACP :

$$\min_{a,b} \sum_{i=1}^{n} \left\| (x_i, y_i) - \text{ proj sur } \Delta_{(a,b)} \right\|^2$$

où $\Delta_{(a,b)}$ est la droite d'équation $y = ax + b$.

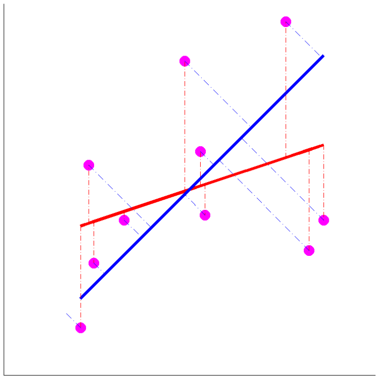
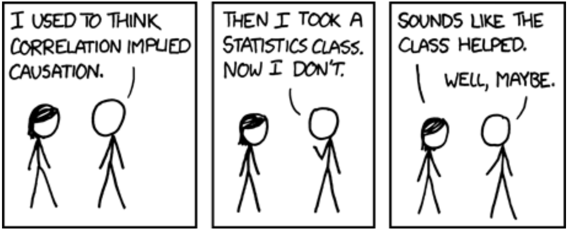


Figure : Régression linéaire et ACP (régression orthogonale)

Conclusion



- observation = modèle + bruit
  - modèle linéaire
  - certains modèles non linéaires peuvent être linéarisés
- Moindres carré
  - facile à calculer
- reste à faire
  - passer au multi dimensionnel (plusieurs variables explicatives)
  - valider le modèle : diagnostique de la régression

Repéres bibliographiques

- http://papers.ssrn.com/sol3/papers.cfm?abstract_id=174508