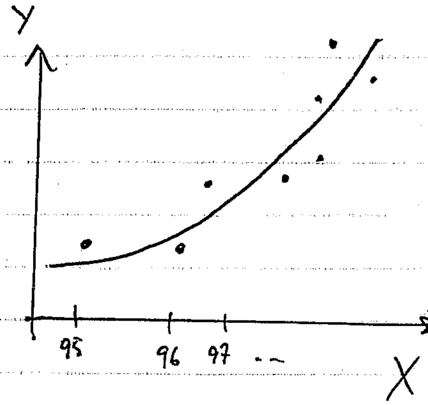
3) Régression exponentielle:

$$Y = \beta_0 \times e^{\beta_1 X} * \varepsilon$$

Y → Sévérité des sinistres

X → temps

ε → Erreur aléatoire



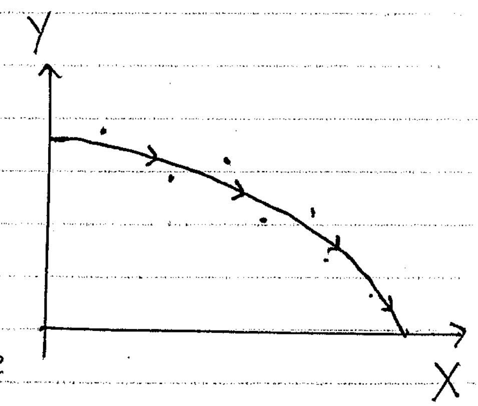
NOTE: $\ln(Y) = \ln(\beta_0) + \beta_1 X + \ln(\varepsilon)$

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

... peut se voir comme une régression linéaire simple

NOTE: Ce modèle est aussi appelé:
- Régression multiplicative
- Régression log-linéaire

4) Régression quadratique:



$$Y = \beta_0 + \beta_1 \times X + \beta_2 \times X^2 + \varepsilon$$

Y → Hauteur d'un objet lancé

X, X^2 → Temps

ε → Erreur aléatoire || Vent etc...

Note: En posant $X_1 = X$ et $X_2 = X^2$, ce modèle peut se voir comme:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

... une régression multiple !!!

* Dans ce cours, nous allons nous concentrer sur les <u>modèles linéai</u>

→ Plus simples

→ Plusieurs modèles peuvent se ramener à un modèle linéai
(ex: reg. multiplicative, reg. quadratique)

→ Constituent souvent une très bonne approximation de la réalité qui peut être très complexe (ex: assurance!!!)

→ Se généralisent facilement
(ex: GLM: generalized linear models)

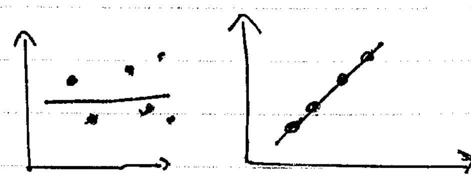
<u>Le problème</u>: Trouver les paramètres $\beta_0, \beta_1, \dots, \beta_p$ tels que

$$\varepsilon = Y - f(X_1, \dots, X_p ; \beta_0, \beta_1, \dots, \beta_p) \text{ est "petite"}$$

<u>Quelle mesure d'erreur utiliser?</u>

1) <u>Erreur totale?</u> $\displaystyle\sum_{t=1}^{n} \varepsilon_t = \sum_{t=1}^{n} \left(Y_t - (\beta_0 + \beta_1 X_t) \right)$

→ Facile à mettre à 0 ↳ <u>ex</u>:



→ Pas fiable !

2) <u>Erreur absolue</u>?: $\displaystyle\sum_{t=1}^{n} |\varepsilon_t| = \sum_{t=1}^{n} \left| Y_t - (\beta_0 + \beta_1 X_t) \right|$

→ Très robuste

→ Compliqué mathématiquement (minimiser $\sum |\varepsilon_t|$ implique de dériver cette fonction!!!)

3) Erreur quadratique ?: $\sum_{t=1}^{n} \varepsilon_t^2 = \sum_{t=1}^{n} \left[Y_t - (\beta_0 + \beta_1 X_t) \right]^2$

→ Plus simple mathématiquement

→ Donne beaucoup de poids aux grandes erreurs

* Cette mesure est celle que nous allons étudier *

2.2) Le modèle de régression linéaire simple:

On dispose de n paires d'observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$

Exemple: X_t : # d'années de scolarité de l'actuaire t

Y_t : Salaire de l'actuaire t

Idée:

- Pour $X_t = 0$; on a $Y_t = \beta_0$ (ex: β_0 = Salaire d'un stagiaire!)
  « en moyenne »

- Pour chaque année additionnelle de scolarité, le salaire augmente de β_1 unitées « en moyenne »

- Ainsi, « en moyenne » on a :

$$\mathbb{E}(Y_t | X_t) = \beta_0 + \beta_1 X_t$$

Habituellement, la relation n'est pas exacte dans la réalité. La "différence" (ou l'erreur) est notée ε_t et est assumée aléatoire :

$$\varepsilon_t = Y_t - \mathbb{E}(Y_t | X_t) = Y_t - (\beta_0 + \beta_1 X_t)$$

En réorganisant ; on trouve le modèle suivant :

$$Y_t = \underbrace{\beta_0}_{\substack{\text{Ordonnée à l'origine} \\ \| \\ \text{Salaire si } X_t = 0}} + \underbrace{\beta_1}_{\substack{\text{Pente} \\ \| \\ \text{Taux de croissance} \\ \text{du salaire avec } X_t}} \underbrace{X_t}_{\substack{\text{\# années} \\ \text{de scolarité}}} + \underbrace{\varepsilon_t}_{\substack{\text{Erreur aléatoire} \\ \| \\ \text{résidu}}}$$

avec Y_t : salaire

<u>But</u> : Trouver les paramètres β_0 et β_1 de façon à minimiser l'erreur ε_t

<u>Note</u> : Si ε_t est minimal, cela veut dire que $Y_t \approx \beta_0 + \beta_1 X_t$

$\Rightarrow$ Donc la droite de régression $(\beta_0 + \beta_1 X_t)$ est une bonne approximation de Y_t.

2.2.1) <u>Coefficients de régression :</u>

Les paramètres (ou coefficients) β_0 et β_1 sont déterminés en minimisant l'erreur quadratique (méthode des moindres carrés)

$$S(\beta_0, \beta_1) = \sum_{t=1}^{n} \varepsilon_t^2$$

$$= \sum_{t=1}^{n} \left(Y_t - (\beta_0 + \beta_1 X_t)\right)^2$$

$$= \sum_{t=1}^{n} \left(Y_t - \beta_0 - \beta_1 X_t\right)^2$$

Note: on met un "^" pour représenter l'estimation ... les paramètres "optimaux"

Minimisation:

① $\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_0} = 0 \Rightarrow -2 \sum_{t=1}^{n} (Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t) = 0$

$$\Rightarrow \boxed{\sum_{t=1}^{n} Y_t - n \times \hat{\beta}_0 - \hat{\beta}_1 \sum_{t=1}^{n} X_t = 0}$$

② $\frac{\partial S(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_1} = 0 \Rightarrow -2 \sum_{t=1}^{n} (Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t) \times X_t = 0$

$$\Rightarrow \boxed{\sum_{t=1}^{n} X_t Y_t - \hat{\beta}_0 \sum_{t=1}^{n} X_t - \hat{\beta}_1 \sum_{t=1}^{n} X_t^2 = 0}$$

Deux équations (① et ②) et deux inconnus ($\hat{\beta}_0$ et $\hat{\beta}_1$):

De ①: $\hat{\beta}_0 = \left(\frac{\sum_{t=1}^{n} Y_t}{n} \right) - \hat{\beta}_1 \left(\frac{\sum_{t=1}^{n} X_t}{n} \right)$

IMP: $\Rightarrow \boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}}$

De ②: $\hat{\beta}_1 = \frac{\sum_{t=1}^{n} X_t Y_t - \hat{\beta}_0 \sum_{t=1}^{n} X_t}{\sum_{t=1}^{n} X_t^2}$

On met ① dans ② $\hat{\beta}_1 = \frac{\sum_{t=1}^{n} X_t Y_t - (\bar{Y} - \hat{\beta}_1 \bar{X}) \sum_{t=1}^{n} X_t}{\sum_{t=1}^{n} X_t^2}$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum_{t=1}^{n} X_t Y_t - (\bar{Y} - \hat{\beta}_1 \bar{X}) \times n \bar{X}}{\sum_{t=1}^{n} X_t^2}$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum_{t=1}^{n} X_t Y_t - n \bar{X}\bar{Y} + \hat{\beta}_1 \times n \times \bar{X}^2}{\sum_{t=1}^{n} X_t^2}$$

On isole $\hat{\beta}_1$:

IMP :

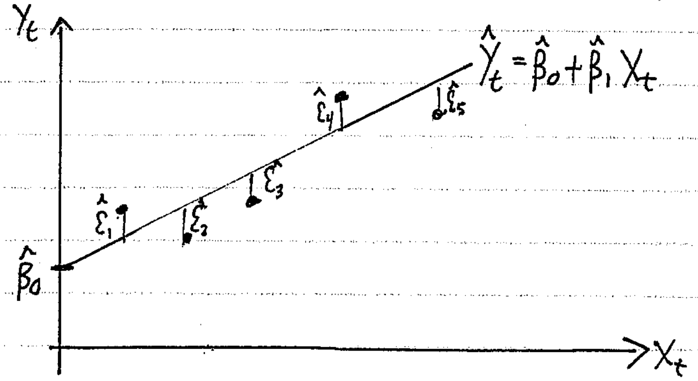
$$\boxed{\hat{\beta}_1 = \frac{\sum_{t=1}^{n} X_t Y_t - n \bar{X}\bar{Y}}{\sum_{t=1}^{n} X_t^2 - n \bar{X}^2}}$$

Remarques:

1) On note $\hat{\varepsilon}_t$ les « résidus » générés par le modèle estimé :

$$\hat{\varepsilon}_t = Y_t - \hat{Y}_t$$

$$\Rightarrow \varepsilon_t = Y_t - (\hat{\beta}_0 - \hat{\beta}_1 X_t) \quad ; \text{pour } t = 1, \dots, n$$

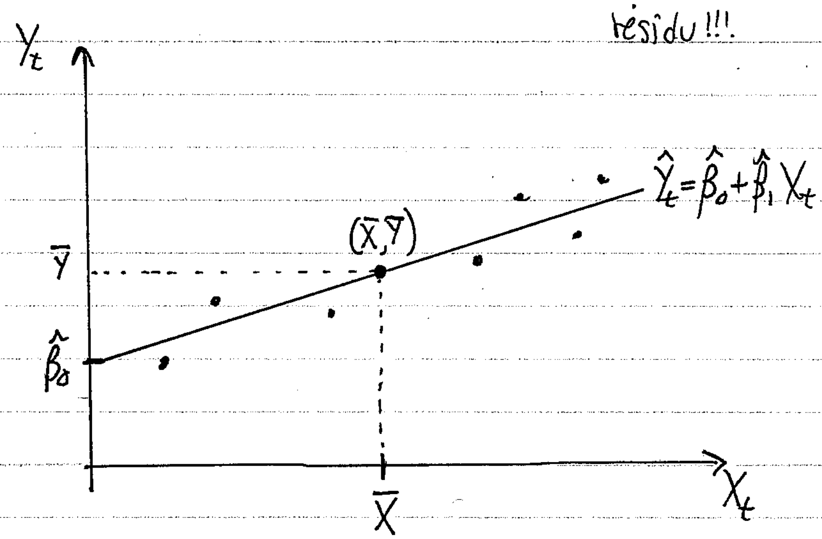


2) Le "centre de gravité" (ou centre de masse) des données $(\bar{X}, \bar{Y})$ se trouve sur la droite de régression

Preuve:

De ①: $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

$$\Rightarrow \bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} + \underset{\substack{\equiv \\ \uparrow \\ \text{aucun} \\ \text{résidu !!!}}}{0}$$



3) On note : $\rightarrow S_{XX} = \sum_{t=1}^{n} (X_t - \bar{X})^2 = \sum_{t=1}^{n} (X_t^2 - 2X_t\bar{X} + \bar{X}^2)$

$$= \sum_{t=1}^{n} X_t^2 - 2\bar{X}\sum_{t=1}^{n} X_t + n\bar{X}^2$$

$$= \sum_{t=1}^{n} X_t^2 - 2\bar{X} \times n\bar{X} + n\bar{X}^2$$

$$= \left(\sum_{t=1}^{n} X_t^2\right) - n \times \bar{X}^2$$

$$\rightarrow S_{XY} = \sum_{t=1}^{n} (X_t - \bar{X}) \times (Y_t - \bar{Y}) = \overset{\text{m̂ démo.}}{\ldots} = \left(\sum_{t=1}^{n} X_t Y_t\right) - n \times \bar{X}\bar{Y}$$

4) Sous la remarque (3) ; on a que :

$$\boxed{\hat{\beta}_1 = \frac{S_{XY}}{S_{XX}}} \quad \text{et} \quad \boxed{\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}}$$

5) La somme des résidus de tout modèle de régression linéaire est nulle :

Preuve : $\sum_{t=1}^{n} \hat{\varepsilon}_t = \sum_{t=1}^{n} \left(Y_t - (\hat{\beta}_0 + \hat{\beta}_1 X_t)\right)$

$$\overset{\text{De } (1)}{=} \sum_{t=1}^{n} \left(Y_t - (\overline{Y} - \hat{\beta}_1 \overline{X}) - \hat{\beta}_1 X_t\right)$$

$$= \sum_{t=1}^{n} Y_t - \sum_{t=1}^{n} \overline{Y} + \hat{\beta}_1 \sum_{t=1}^{n} \overline{X} - \hat{\beta}_1 \sum_{t=1}^{n} X_t$$

$$= n \times \overline{Y} - n \times \overline{Y} + \hat{\beta}_1 \times n \times \overline{X} - \hat{\beta}_1 \times n \times \overline{X}$$

$$= 0 \ !!!$$

↑↑↑↑

Rendu ici

07-09-2010