Exemple: Soit les $n=5$ observations suivantes du couple (X_t, Y_t):

t	X_t	Y_t	...calculs: X_t^2	$X_t Y_t$
1	2	2	4	4
2	3	5	9	15
3	6	3	36	18
4	9	6	81	54
5	12	5	144	60
Totaux:	32	21	274	151

$$\Rightarrow \quad \hat{\beta}_1 = \frac{\sum_{t=1}^{n} X_t Y_t - n\bar{X}\bar{Y}}{\sum_{t=1}^{n} X_t^2 - n\bar{X}^2} = \frac{151 - (5)(21/5)(32/5)}{274 - (5)(32/5)^2} = \frac{83}{346} \approx 0.2399$$

$$\Rightarrow \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{21}{5} - \left(\frac{83}{346}\right) \times \left(\frac{32}{5}\right) \approx 2.6647$$

$\Rightarrow$ Modèle de régression: $Y_t = 2.6647 + 0.2399\, X_t + \varepsilon_t$

$\Rightarrow$ Prévisions et résidus:

t	$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t$	$\hat{\varepsilon}_t$
1	3.1445	-1.1445
2	3.3844	1.6156
3	4.1041	-1.1041
4	4.8238	1.1762
5	5.5435	-0.5435

$$\sum_{t=1}^{5} \varepsilon_t \approx -0.0003 \approx 0 !$$

Logiciel R:

```
x <- c(2,3,6,9,12)
y <- c(2,5,3,6,5)
```
... entrer les données

```
reg <- lm(y ~ x)
```
... estimation des paramètres

```
summary(reg)
```
... résumé de l'estimation

```
fitted(reg)
```
... valeurs de $\hat{Y}_t$

```
residuals(reg)
```
... résidus $\hat{\varepsilon}_t$

Logiciel SAS:

```
data donnees_exemple;
input x y;
cards;
2 2
3 5
6 3
9 6
12 5
run;
```
... entrer les données

```
proc reg data=donnees_exemple;
model y=x;
run;
```
... estimation des paramètres.

2.2.2) Caractéristiques du terme d'erreur:

Rappelons que :

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

$$\Rightarrow Y_t = E(Y_t | X_t) + \varepsilon_t$$

On a que:

1) $\mathbb{E}(\varepsilon_t) = 0$ ... par définition pour que $\mathbb{E}(Y_t) = \mathbb{E}(Y_t | X_t)$

2) $Var(\varepsilon_t) = \sigma^2$ ... constante par hypothèse

3) $Cov(\varepsilon_t, \varepsilon_s) = 0$, $t \neq s$ ... pas de corrélation par hypothèse

2.3) Propriétés de l'estimateur des moindres carrés (EMC):

2.3.1) Propriété 1: Estimateur sans biais:

- $$\mathbb{E}(\hat{\beta}_1) = \mathbb{E}\left(\frac{\sum_{t=1}^{n} (X_t - \bar{X})(Y_t - \bar{Y})}{\sum_{t=1}^{n} (X_t - \bar{X})^2} \right)$$

$$= \frac{\sum_{t=1}^{n} (X_t - \bar{X}) \mathbb{E}(Y_t - \bar{Y})}{\sum_{t=1}^{n} (X_t - \bar{X})^2}$$

$$= \frac{\sum_{t=1}^{n} (X_t - \bar{X})(E(Y_t) - E(\bar{Y}))}{\sum (X_t - \bar{X})^2}$$

Puisque le point $(\overline{X},\overline{Y})$ est toujours sur la droite de régression:

$$\mathbb{E}(\hat{\beta}_1) = \frac{\sum_{t=1}^{n}(X_t-\overline{X})\times(\beta_0+\beta_1 X_t-\beta_0-\beta_1\overline{X})}{\sum_{t=1}^{n}(X_t-\overline{X})^2}$$

$$= \frac{\sum_{t=1}^{n}(X_t-\overline{X})\,\beta_1\,(X_t-\overline{X})}{\sum_{t=1}^{n}(X_t-\overline{X})^2}$$

$$= \beta_1\,\frac{\sum_{t=1}^{n}(X_t-\overline{X})^2}{\sum_{t=1}^{n}(X_t-\overline{X})^2}$$

$$\Rightarrow \boxed{\mathbb{E}(\hat{\beta}_1)=\beta_1}$$

Par conséquent,

$$\mathbb{E}(\hat{\beta}_0) = \mathbb{E}(\overline{Y}-\hat{\beta}_1\overline{X}) = \mathbb{E}(\overline{Y}) - \overline{X}\,\mathbb{E}(\hat{\beta}_1)$$

$$= \beta_0+\beta_1\overline{X}-\overline{X}\beta_1$$

$$= \beta_0$$

$$\Rightarrow \boxed{\mathbb{E}(\hat{\beta}_0)=\beta_0}$$

2.3.2) Propriété 2 : Variances et covariances:

$$\bullet\ Var(\hat{\beta}_1) = Var\left(\frac{\sum_{t=1}^{n}(X_t-\overline{X})(Y_t-\overline{Y})}{\sum_{t=1}^{n}(X_t-\overline{X})^2} \right)$$

$$= \frac{Var\left(\sum_{t=1}^{n}(X_t-\overline{X})Y_t - \sum_{t=1}^{n}(X_t-\overline{X}) \times \overline{Y} \right)}{\left[\sum_{t=1}^{n}(X_t-\overline{X})^2\right]^2}$$

$$= \frac{Var\left(\sum_{t=1}^{n}(X_t-\overline{X})Y_t \right) + Var\left(\overline{Y} \sum_{t=1}^{n}(X_t-\overline{X}) \right)}{\left[\sum_{t=1}^{n}(X_t-\overline{X})^2\right]^2}$$

$$= \frac{\sum_{t=1}^{n}(X_t-\overline{X})^2 Var(Y_t) + Var\left(\overline{Y} \times [n\overline{X} - n\overline{X}] \right)}{\left[\sum_{t=1}^{n}(X_t-\overline{X})^2\right]^2}$$

$$= \frac{\sum_{t=1}^{n}(X_t-\overline{X})^2 Var(\beta_0 + \beta_1 X_t + \varepsilon_t) + 0}{\left[\sum_{t=1}^{n}(X_t-\overline{X})^2\right]^2}$$

$$= \frac{\sum_{t=1}^{n}(X_t-\overline{X})^2 Var(\varepsilon_t)}{\left[\sum_{t=1}^{n}(X_t-\overline{X})^2\right]^2}$$

$$= \frac{\sum_{t=1}^{n} (X_t - \overline{X})^2 \sigma^2}{\left[\sum_{t=1}^{n} (X_t - \overline{X})^2\right]^2}$$

$$\Rightarrow \boxed{Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{t=1}^{n} (X_t - \overline{X})^2}}$$

* IMP

$$\bullet\ Var(\hat{\beta}_0) = Var(\overline{Y} - \hat{\beta}_1 \overline{X})$$

$$= Var(\overline{Y}) + Var(\hat{\beta}_1 \overline{X}) - 2\,Cov(\overline{Y}, \hat{\beta}_1 \overline{X})$$

$$= Var\left(\frac{\sum_{t=1}^{n} Y_t}{n}\right) + \overline{X}^2 Var(\hat{\beta}_1) - 2\overline{X} \times Cov(\overline{Y}, \hat{\beta}_1)$$

$$= \frac{n \times Var(Y)}{n^2} + \overline{X}^2 \times \left(\frac{\sigma^2}{\sum_{t=1}^{n} (X_t - \overline{X})^2}\right) - 2\overline{X}\,Cov(\overline{Y}, \hat{\beta}_1)^{*}$$

$$* \quad Cov(\overline{Y}, \hat{\beta}_1) = Cov\left(\frac{\sum_{t=1}^{n} Y_t}{n}, \frac{\sum_{s=1}^{n} (X_s - \overline{X})(Y_s - \overline{Y})}{\sum_{s=1}^{n} (X_s - \overline{X})^2}\right)$$

$$= \frac{1}{n} \times \frac{1}{\sum_{s=1}^{n} (X_s - \overline{X})^2} \times Cov\left(\sum_{t=1}^{n} Y_t, \sum_{s=1}^{n} (X_s - \overline{X}) Y_s - \overline{Y} \underbrace{\sum_{s=1}^{n} (X_s - \overline{X})}_{=0\ !!!}\right)$$

$$= \frac{1}{n} \times \frac{1}{\sum_{s=1}^{n} (X_s - \overline{X})^2} \sum_{t=1}^{n} \sum_{s=1}^{n} Cov\left(Y_t, (X_s - \overline{X}) Y_s\right)$$

$$= \frac{1}{n} \times \frac{1}{\sum_{s=1}^{n} (X_s - \overline{X})^2} \sum_{t=1}^{n} \sum_{s=1}^{n} (X_s - \overline{X}) \, Cov(\varepsilon_t, \varepsilon_s)$$

$$= \frac{1}{n} \times \frac{1}{\sum_{s=1}^{n} (X_s - \overline{X})^2} \left(\sum_{\substack{t=1 \\ : t \neq s}}^{n} \sum_{s=1}^{n} (X_s - \overline{X}) \times 0 + \sum_{\substack{t=1 \\ : t = s}}^{n} \sum_{s=1}^{n} (X_s - \overline{X}) \times \sigma^2 \right)$$

$$= \frac{1}{n} \times \frac{1}{\sum_{s=1}^{n} (X_s - \overline{X})^2} \times \left(\sigma^2 \sum_{t=1}^{n} (X_s - \overline{X}) \right)$$

$$= \frac{1}{n} \times \frac{1}{\sum_{s=1}^{n} (X_s - \overline{X})^2} \times \sigma^2 \left(n\overline{X} - n\overline{X} \right)$$

$$= 0 \ !!!$$

Donc;

$$\boxed{Var(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\overline{X}^2 \sigma^2}{\sum_{t=1}^{n} (X_t - \overline{X})^2}}$$

⊛ IMP

Finalement,

$$\boxed{\begin{aligned} Cov(\hat{\beta}_0, \hat{\beta}_1) &= Cov(\overline{Y} - \hat{\beta}_1 \overline{X}, \hat{\beta}_1) \\ &= Cov(\overline{Y}, \hat{\beta}_1) - \overline{X} \, Var(\hat{\beta}_1) \\ &= \underset{\substack{\uparrow \\ \text{voir } ⊛}}{0} - \overline{X} \times \frac{\sigma^2}{\sum_{t=1}^{n} (X_t - \overline{X})^2} \end{aligned}}$$

IMP

2.3.3) Propriété 3 : Optimalité :

Le théorème de Gauss–Markov établit que l'estimateur des moindres carrés est l'estimateur non biaisé à variance minimale

Idée de la preuve :

(1) Considérer l'estimateur $\oplus^* = \sum_{t=1}^{n} C_t \times Y_t$

(2) Minimiser $Var(\oplus^*)$ sous la contrainte que $E(\oplus^*) = \beta$; où

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

2.4) Régression passant par l'origine :

Dans certaines situations, il est possible que l'on souhaite forcer la droite de régression à passer par l'origine

ex : Y_t : Consommation d'essence en L d'une voiture
X_t : # de Km parcouru

... $X_t = 0 \Rightarrow Y_t = 0$!!!

Dans ce cas, on peut postuler le modèle suivant :

$$Y_t = \beta X_t + \varepsilon_t$$

On peut montrer que $\boxed{\hat{\beta} = \dfrac{\sum_{t=1}^{n} X_t Y_t}{\sum_{t=1}^{n} X_t^2}}$ ... même procédure qu'en (2.2.1)

2.5) Analyse de la variance:

Un tableau d'analyse de la variance permet d'évaluer la qualité de l'ajustement du modèle aux observations

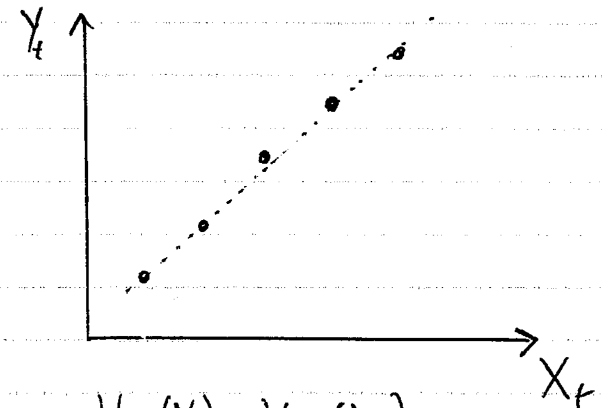
Idée:

(I) Si on décide de modéliser Y_t sans la régression (= analyse statistique des risques actuariels...), alors Y est vu comme une v.a avec une certaine variance $= Var(Y)$

(II) En utilisant la régression pour modéliser Y_t en fonction de X_t une partie de $Var(Y_t)$ est "expliquée" par $Var(X_t)$, alors que l'autre partie reste "inexpliquée"

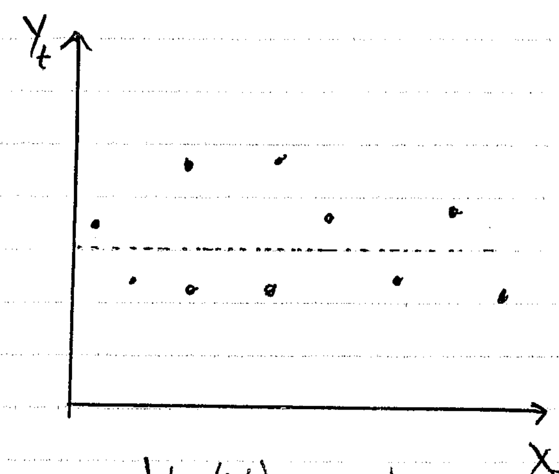
(III) L'utilité de la régression = Proportion de $Var(Y_t)$ expliquée par $Var(X_t)$

exemple:



$Var(Y_t) = Var(X_t)$

$\Rightarrow Var(Y_t) = 100\% \times Var(X_t)$: rég. utile!



$Var(Y_t) = \underbrace{0\% \, Var(X_t)}_{\text{expliquée}} + \underbrace{100\% \, Var(\varepsilon_t)}_{\text{inexpliquée}}$

$\Rightarrow$ rég. inutile $\Rightarrow$ utiliser stat. de base

2.5.1) Background 1 : Somme des carrés :

La variance totale dans les Y_t est proportionnelle à :

$$SST = \sum_{t=1}^{n} (Y_t - \bar{Y})^2$$

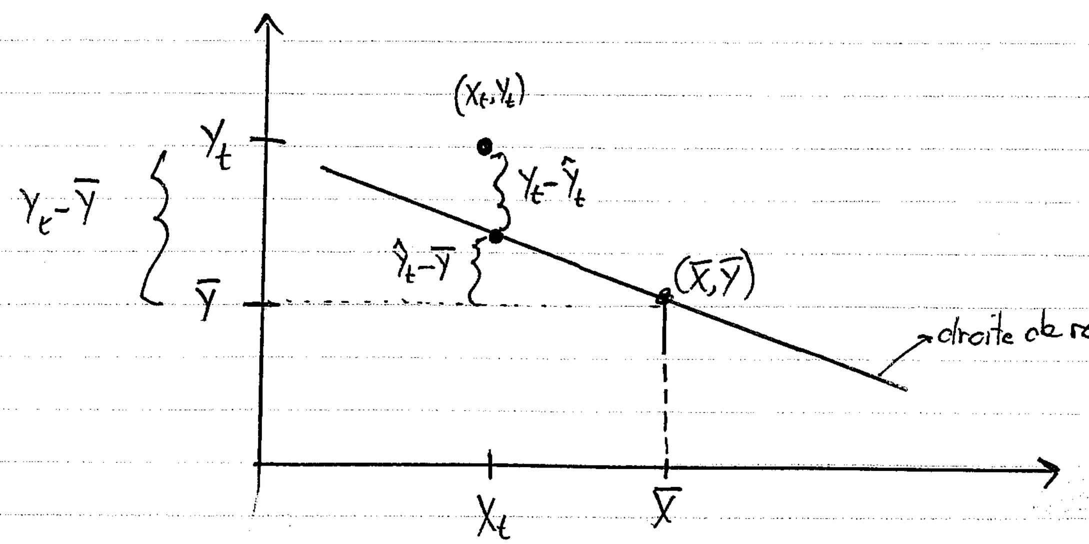
Décomposition :

$$(Y_t - \bar{Y}) = Y_t - \hat{Y}_t + \hat{Y}_t - \bar{Y}$$

$$\Rightarrow (Y_t - \bar{Y}) = (Y_t - \hat{Y}_t) + (\hat{Y}_t - \bar{Y})$$

$$\Rightarrow \underbrace{(Y_t - \bar{Y})}_{\substack{\text{Variation} \\ \text{totale} \\ \text{de } Y_t}} = \underbrace{(\hat{Y}_t - \bar{Y})}_{\substack{\text{Variation} \\ \text{de } \hat{Y}_t \\ = \\ \text{Variation} \\ \text{expliquée} \\ \text{par la} \\ \text{régression}}} + \underbrace{(Y_t - \hat{Y}_t)}_{\substack{\hat{\varepsilon}_t \\ = \\ \text{résidu} \\ = \\ \text{Variation inexpliquée} \\ \text{par la régression}}}$$

Illustration :



Par conséquent, on a que:

$$SST = \sum_{t=1}^{n} \left[(\hat{Y}_t - \bar{Y}) + (Y_t - \hat{Y}_t) \right]^2$$

$$= \underbrace{\sum_{t=1}^{n} (\hat{Y}_t - \bar{Y})^2}_{SSR} + \underbrace{\sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2}_{SSE} + \underbrace{2 \sum_{t=1}^{n} (\hat{Y}_t - \bar{Y})(Y_t - \hat{Y}_t)}_{\star}$$

SSR ↑ régression

SSE ↑ Erreur

$$\star = 2 \sum_{t=1}^{n} (\hat{Y}_t - \bar{Y})(Y_t - \hat{Y}_t)$$

$$= 2 \sum_{t=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 X_t - \hat{\beta}_0 - \hat{\beta}_1 \bar{X})(Y_t - \bar{Y} + \bar{Y} - \hat{Y}_t)$$

$$= 2 \sum_{t=1}^{n} \hat{\beta}_1 (X_t - \bar{X})(Y_t - \bar{Y} + \hat{\beta}_0 + \hat{\beta}_1 \bar{X} - \hat{\beta}_0 - \hat{\beta}_1 X_t)$$

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

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

$$= 2 \hat{\beta}_1 (S_{XY} - \hat{\beta}_1 S_{XX})$$

Section (2.2.1) remarque (4)

$$= 2 \hat{\beta}_1 \left(S_{XY} - \left(\frac{S_{XY}}{S_{XX}}\right) \times S_{XX} \right)$$

$$= 2 \hat{\beta}_1 (S_{XY} - S_{XY}) = 0 !$$

Ainsi :

$$\boxed{SST = \underbrace{SSR}_{\substack{\text{variation}\\ \text{expliquée}\\ \text{par la}\\ \text{régression}}} + \underbrace{SSE}_{\substack{\text{variation}\\ \text{inexpliquée,}\\ \text{ou résiduelle}}}}$$

<u>Intuitivement :</u>

- Dans un bon modèle (régression utile), on aimerait que
  $\Rightarrow SST \approx SSR$ ... $Var(Y_t) \approx Var(X_t)$
  ou
  $\Rightarrow SSE \approx 0$ ... variation résiduelle faible !

- On définit le <u>coefficient de détermination</u> :

$$\boxed{R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}}$$

: % de la variance dans Y_t expliquée par la régression

=

1 − (% de la variance dans Y_t inexpliquée par la régression)

$\Rightarrow R^2 \in [0, 1]$

$\Rightarrow R^2 = 100\% \Rightarrow$ régression utile ; $R^2 = 0\% \Rightarrow$ rég. inutile !

Rendu ici
14-09-2010