2.5.2) Background 2 : Degrés de liberté :

Le nombre de "degrés de liberté" (=d.l.) d'une "somme de carrés" (=SS...) est :

→ Le nombre de composants "indépendants" dans la somme

ou

→ Le nombre minimal de fonctions de $Y_1, \ldots, Y_n$ qu'il faut connaître par obtenir la somme

ou

→ Pour SST et SSE seulement :

$$d.l = \begin{pmatrix} \text{Nombre de termes} \\ \text{dans la somme} \end{pmatrix} - \begin{pmatrix} \text{Nombre de paramètres} \\ \text{estimés dans cette somme} \end{pmatrix}$$

Ainsi :

- $SST = \sum_{t=1}^{n} (Y_t - \bar{Y})$: n termes $- \left(1 \text{ param. estimé} : \bar{Y}\right) = \boxed{(n-1)\ d.l.}$

- $SSE = \sum_{t=1}^{n} (Y_t - \hat{Y}_t)^2 = \sum_{t=1}^{n} (Y_t - \hat{\beta}_0 - \hat{\beta}_1 X_t)^2$: n termes $- \begin{pmatrix} 2 \text{ param estimés :} \\ \hat{\beta}_0 \text{ et } \hat{\beta}_1 \end{pmatrix} = \boxed{(n-2)\ d.l.}$

- $SSR = \sum_{t=1}^{n} (\hat{Y}_t - \bar{Y})^2 = \sum_{t=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 X_t - \hat{\beta}_0 - \hat{\beta}_1 \bar{X})^2 = \underbrace{\hat{\beta}_1^2}_{f(Y_1, \ldots, Y_n)} \times \underbrace{\sum_{t=1}^{n} (X_t - \bar{X})^2}_{f(X_1, \ldots, X_n)}$

∴ une seule fonction des $Y_1, \ldots, Y_n$ doit être connue pour obtenir SSR $\Rightarrow$ $\boxed{1\ d.l}$

Remarque:

- On sait que:

$$SST = SSE + SSR$$

- On note aussi que

$$d.\ell(SST) = d.\ell(SSE) + d.\ell(SSR)$$
$$(n-1) = (n-2) + (1)$$

*On aurait donc pu retrouver $d.\ell(SSR) = d.\ell(SST) - d.\ell(SSE)$!

2.5.3) Tableau d'analyse de la variance (=ANOVA):

Source de la variance	Somme des carrés (SS)	Degrés de liberté (d.ℓ)	Carrés moyens (MS)	Ratio de Fisher (F)
Régression	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
Erreur (=résidus)	SSE	n-2	$MSE = \frac{SSE}{n-2}$	
Total	SST	n-1		

* Ce type de tableau est utilisé dans tous les logiciels de régression pour évaluer la qualité du modèle!

Exemple: On reprend l'exemple de la section (2.2.1)

t	X_t	Y_t	$\hat{Y}_t = \hat{\beta}_0 + \hat{\beta}_1 X_t$	$\hat{\varepsilon}_t$
1	2	2	3.1445	-1.1445
2	3	5	3.3844	1.6156
3	6	3	4.1041	-1.1041
4	9	6	4.8238	1.1762
5	12	5	5.5435	-0.5435
Totaux	32	21		

$$\Downarrow$$

$$\bar{Y} = 21/5$$

- $SSE = \sum_{t=1}^{5} (Y_t - \hat{Y}_t)^2 = \sum_{t=1}^{\hat{}} \varepsilon_t^2 = 6.8179$

- $SSR = \sum_{t=1}^{5} (\hat{Y}_t - \bar{Y})^2 = 3.9821$

- $SST = SSE + SSR = 6.8179 + 3.9821 = 10.8000$

ANOVA:

Source	SS	d.l	MS	F
Régression	3.9821	1	3.9821	1.7522
Erreur	6.8179	3	2.2726	
Total	10.8000	4		

R^2:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = \frac{3.9821}{10.8000} = 36.98:$$

Seulement 36.98 de la variabilité des Y_t est expliquée par la variabilité des X_t

$\Rightarrow$ Régression pas très utile.

Logiciel R:

anova(reg)

Logiciel SAS:

proc reg génère automatiquement le tableau ANOVA

Logiciel Excel:

Voir Macro complémentaires/Utilitaire d'analyse/Régression Linéaire

2.6) Intervalles de confiance (I.C.) et tests d'hypothèses:

Contexte: On poursuit l'objectif des sections (2.3) et (2.5), soit de valider la qualité du modèle de regression.

Remarque importante: Jusqu'à maintenant, nous n'avons fait aucune hypothèse quant à la distribution des v.a. ε_t.

(*) Il sera par contre nécessaire d'assigner une loi à ε_t pour les I.C. et les tests.

2.6.1) Distribution des variables aléatoires:

On suppose que:

$$\varepsilon_t \overset{i.i.d}{\sim} N(0, \sigma^2)$$

Conséquences:

1) $$\left(Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t\right) \sim N\left(\beta_0 + \beta_1 X_t \,,\, \sigma^2\right)$$

2) Puisque $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des fonctions linéaires de $Y_1, \ldots, Y_n$:

- $\hat{\beta}_0 \sim N(\beta_0 \,;\, Var(\hat{\beta}_0))$
- $\hat{\beta}_1 \sim N(\beta_1 \,;\, Var(\hat{\beta}_1))$

Voir section (2.3.2) pour les variances!

3) Un estimateur sans biais pour σ^2 est:

$$\boxed{\hat{\sigma}^2 = S^2 = MSE = \frac{SSE}{d.l(SSE)} = \frac{\sum_{t=1}^{n}(Y_t - \hat{Y}_t)^2}{n-2} = \frac{\sum_{t=1}^{n}\hat{\varepsilon}_t^2}{n-2}.}$$

4) On peut montrer que :

$$\boxed{\left(\frac{SSE}{\sigma^2}\right) \sim \chi^2(n-2)}$$

2.6.2) Intervalle de confiance pour β_1: (...et non pour $\hat{\beta}_1$!!!)

Puisque $\hat{\beta}_1 \sim N(\beta_1, Var(\hat{\beta}_1))$

On a que $\left(\frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var(\hat{\beta}_1)}}\right) \sim N(0,1)$

IMPORTANT: Si: le σ^2 est estimé par S^2 dans la formule de $Var(\hat{\beta}_1)$, c-à-d.:

$$\widehat{Var}(\hat{\beta}_1) = \frac{S^2}{\sum_{t=1}^{n}(X_t - \bar{X})^2}$$

Alors:

$$\left(\frac{\hat{\beta} - \beta_1}{\sqrt{\widehat{Var}(\hat{\beta}_1)}}\right) \sim t(n-2)$$

Un intervalle de confiance au niveau $100 \times (1-\alpha)\%$ pour β_1 est donc:

$$\hat{\beta}_1 \pm t_{\frac{\alpha}{2}}(n-2) \times \sqrt{\hat{Var}(\hat{\beta}_1)}$$

$$\Rightarrow \boxed{\hat{\beta}_1 \pm t_{\frac{\alpha}{2}}(n-2) \times \sqrt{\frac{S^2}{\sum_{t=1}^{n}(X_t-\bar{X})^2}}}$$

2.6.3) Intervalle de confiance pour β_0:

De manière similaire, un intervalle de confiance au niveau $100 \times (1-\alpha)\%$ pour β_0 est :

$$\hat{\beta}_0 \pm t_{\frac{\alpha}{2}}(n-2) \times \sqrt{\hat{Var}(\hat{\beta}_0)}$$

$$\Rightarrow \boxed{\hat{\beta}_0 \pm t_{\frac{\alpha}{2}}(n-2) \times \sqrt{\frac{S^2}{n} + \frac{S^2\bar{X}^2}{\sum_{t=1}^{n}(X_t-\bar{X})^2}}}$$

2.6.4) Tests d'hypothèses sur les paramètres:

Principales questions auxquelles on aimerait répondre:

(1) L'ordonnée à l'origine (β_0) est-elle significativement différente de 0?

<u>Sinon</u>: Considérer le modèle $Y_t = \beta_1 \times X_t + \varepsilon_t$

(2) La pente (β_1) est-elle significativement différente de 0?

<u>Sinon</u>: Considérer le modèle $Y_t = \beta_0 + \varepsilon_t$

$\Downarrow$

Analyse stat. des risques act

$\Downarrow$

Régression inutile!

Pour tester...

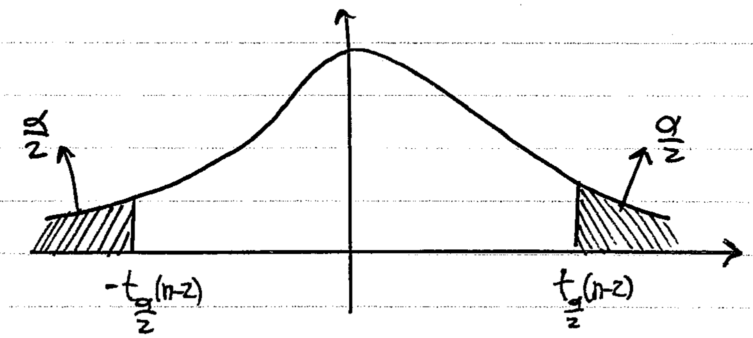
$$H_0 : \beta_0 = 0 \qquad\qquad H_0 : \beta_1 = 0$$
$$H_1 : \beta_0 \neq 0 \qquad\qquad H_1 : \beta_1 \neq 0$$

...on utilise la statistique...

$$t = \frac{\hat{\beta}_0 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} \qquad\qquad t = \frac{\hat{\beta}_1 - 0}{\sqrt{\widehat{Var}(\hat{\beta}_1)}}$$

...et on rejette H_0 au niveau de confiance $100 \times (1-\alpha)\%$ si :

$$|t| > t_{\frac{\alpha}{2}}(n-2)$$



* Prob. de "se tromper" en rejetant H_0 $= \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha$!!!

Remarque :

Hypothèse plus générale :

$$H_0 : \beta_0 = \beta_0^* \qquad\qquad H_0 : \beta_1 = \beta_1^*$$
$$H_1 : \beta_0 \neq \beta_0^* \qquad\qquad H_1 : \beta_1 \neq \beta_1^*$$

On utilise la statistique :

$$t = \frac{\hat{\beta}_0 - \beta_0^*}{\sqrt{\widehat{Var}(\hat{\beta}_0)}} \qquad\qquad t = \frac{\hat{\beta}_1 - \beta_1^*}{\sqrt{\widehat{Var}(\hat{\beta}_1)}}$$

... on rejette H_0 au niveau de confiance $100 \times (1-\alpha)\%$ si $|t| > t_{\frac{\alpha}{2}}(n-2)$.

exemple:

Dans une régression sur un ensemble de $n=14$ observations, on a obtenu:

$$\hat{Y}_t = 68,494 - 0.468\,X_t\,,$$

ainsi que

$$\hat{Var}(\hat{\beta}) = \hat{Var}\left(\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix}\right) = \begin{bmatrix}\hat{Var}(\hat{\beta}_0) & \hat{Cov}(\hat{\beta}_0,\hat{\beta}_1)\\ \hat{Cov}(\hat{\beta}_0,\hat{\beta}_1) & \hat{Var}(\hat{\beta}_1)\end{bmatrix}$$

$$= \begin{bmatrix}66,8511 & 1.2544\\ 1.2544 & 0.0237\end{bmatrix}$$

Question 1: Tester si β_0 est significativement différent de 0?

$$H_0 : \beta_0 = 0 \quad (=\text{hyp. nulle})$$
$$H_1 : \beta_0 \neq 0$$

$$\Rightarrow \quad t = \frac{\hat{\beta}_0 - 0}{\sqrt{\hat{Var}(\hat{\beta}_0)}} = \frac{68.494 - 0}{\sqrt{66.8511}} = 8.38$$

Puisque $|8.38| > t_{\frac{0,05}{2}}(14-2) = 2.18$, on rejette H_0 au niveau de confiance 95%.

* Ordonnée à l'origine significative!

Question 2: Tester si β_1 est significativement différent de 0 ?
= tester si la régression est utile ?

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

$$\Rightarrow \quad t = \frac{-0.468 - 0}{\sqrt{0.0237}} = -3{,}040$$

Puisque $|-3.040| > t_{\frac{0.05}{2}}(14-2) = 2.18$, on rejette H_0 au niveau de confiance 95%

*Il y a 95% de chance que la régression soit utile !

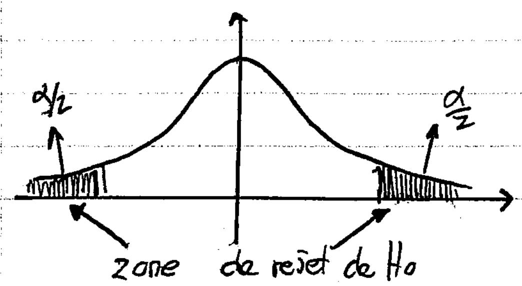
Question 3: Tester si β_1 est significativement négatif ?

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 < 0$$

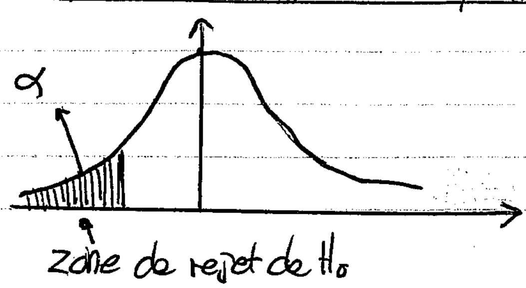
$$\Rightarrow \quad t = \frac{-0.468 - 0}{\sqrt{0.0237}} = -3.040 \qquad (\text{Même stat. } t\,!)$$

***Attention**: Ce test est unilatéral !!!

Test bilatéral (Questions 1 & 2):



Test unilatéral de la question 3:



Puisque $-3.040 < -t_{0.05}(14-2) = -1.78$, on rejette H_0 au niveau de confiance $100\times(1-\alpha)\% = 95\%$ $(\alpha = 0.05)$

↑↑↑ Rendu ici 21-09-2010

* La pente de la droite est significativement négative

Question 4: Obtenir un I.C. au niveau $100\times(1-0.05)\% = 95\%$ pour β_0?

$$\beta_0 \in \hat{\beta}_0 \pm t_{\frac{0.05}{2}}(14-2) \times \sqrt{\hat{Var}(\hat{\beta}_0)}$$

$$\Rightarrow \beta_0 \in 68.494 \pm 2.18 \times \sqrt{66.8511}$$

$$\Rightarrow \beta_0 \in \left]50.670 \,;\, 86.318\right[$$

*Retour à la question 1: Puisque cet I.C. ne comprend pas la valeur 0 ; on valide le test de la question 1!

Question 5: Obtenir un I.C au niveau 95% pour β_1?

$$\beta_1 \in -0.468 \pm 2.18 \times \sqrt{0.0237}$$

$$\Rightarrow \beta_1 \in \left]-0.804 \,,\, -0.132\right[$$

*Retour aux questions 2 et 3: L'I.C ne comprend pas le 0 et est strictement négatif ⇒ ok!