3.2.3) Test de Fisher partiel:

En (3.2.2) on a testé si TOUS LES $\beta_1, \beta_2, \ldots, \beta_p$ étaient nuls.

Dans cette section, on teste simultanément si certains β_i parmi $(\beta_1, \beta_2, \ldots, \beta_p)$ sont nuls.

NOTE: Pas β_0!!!

On teste donc: H_0: Un modèle "réduit" (noté M_0) dont certains $\beta_i = 0$ parmi $(\beta_1, \ldots, \beta_p)$ est acceptable

H_1: On doit utiliser le modèle "complet" (noté M_1) avec les p variables.

On utilise la statistique de Fisher partielle:

$$F^* = \frac{[SSE(M_0) - SSE(M_1)] / [d.\ell(SSE(M_0)) - d.\ell(SSE(M_1))]}{SSE(M_1) / d.\ell(SSE(M_1))}$$

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

$$F^* > F_\alpha\left(d\ell(SSE(M_0)) - d.\ell(SSE(M_1)) \;;\; d.\ell(SSE(M_1)) \right)$$

* Remarque: Si le modèle réduit de H_0 ne consiste qu'à $\beta_i = 0$ (un seul paramètre!); alors on aura que $F^* = t^2$

⇒ Dans ce cas SEULEMENT, le test de Fisher partiel est équivalent au test de Student.

Exemple: Soit le modèle de régression multiple suivant:

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 X_{t3} + \beta_4 X_{t4} + \varepsilon_t$$

Tester: $H_0: \beta_2 = \beta_3 = 0$

$H_1: \beta_2 \neq 0$ ou $\beta_3 \neq 0$ ou $(\beta_2 \neq 0$ et $\beta_3 \neq 0)$

Étape #1: Obtenir le tableau ANOVA pour le modèle sous H_0 ($= M_0 =$ modèle simplifié $= Y_t = \beta_0 + \beta_1 X_{t1} + \beta_4 X_{t4} + \varepsilon_t$)

... extraire $SSE(M_0)$ et $d.l.(SSE(M_0)) = n-3$ de ce tableau.

Étape #2: Obtenir le tableau ANOVA pour le modèle sous H_1 ($= M_1 =$ modèle complet $= Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \beta_3 X_{t3} + \beta_4 X_{t4} + \varepsilon_t$)

... extraire $SSE(M_1)$ et $d.l.(SSE(M_1)) = n-5$

Étape #3: Calculer

$$F^* = \frac{\left(SSE(M_0) - SSE(M_1)\right) / \left[(n-3)-(n-5)\right]}{SSE(M_1)/(n-5)}$$

↑ Δ du # de coef. à estimer!!

... puis rejeter H_0 au niveau $100\times(1-\alpha)\%$ si

$$F^* > F_\alpha(2, n-5)$$

3.3) Sélection d'un modèle optimal:

Lorsque l'on dispose de plusieurs variables explicatives $(X_1, X_2, \ldots, X_p)$, un modèle optimal est tel que:

1) Pouvoir prédictif MAXIMAL

2) Avec un nombre de variables MINIMAL

En régression, il existe plusieurs techniques (ou algorithmes) pour obtenir un modèle optimal

3.3.1) Technique #1: Essai de tous les modèles:

La stratégie la plus simple consiste à examiner tous les modèles possibles (2^p combinaisons)

On choisi le modèle ayant le plus grand R^2_{adj}, où

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

$$= 1 - (1-R^2) \times \left(\frac{n-1}{n-p-1}\right)$$

(*) Contrairement au R^2 normal, le R^2_{adj} pénalise par l'ajout de variables dans le modèle.

exemple: Si on dispose de X_1, X_2 et X_3 ; on ajuste les $2^3=8$ modèles possibles:

1) $Y = \beta_0 + \varepsilon$
2) $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
3) $Y = \beta_0 + \beta_1 X_2 + \varepsilon$
4) $Y = \beta_0 + \beta_1 X_3 + \varepsilon$
5) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$
6) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \varepsilon$
7) $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \varepsilon$
8) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

On fait le calcul du $R^2_{adj.}$ pour chaque modèle et on choisit le modèle avec le plus grand $R^2_{adj.}$

Remarque:

En pratique, cette méthode n'est pas "efficiente", car le temps d'exécution devient énorme lorsque p augmente:

p	2^p	
1	2	
2	4	
3	8	
⋮		
10 (~~+p+1~~)	1024	
⋮		
25	33554432	
⋮		
100	1.26×10^{30}	...ouf!

3.3.2) Technique #2 : Élimination régressive (backward elimination):

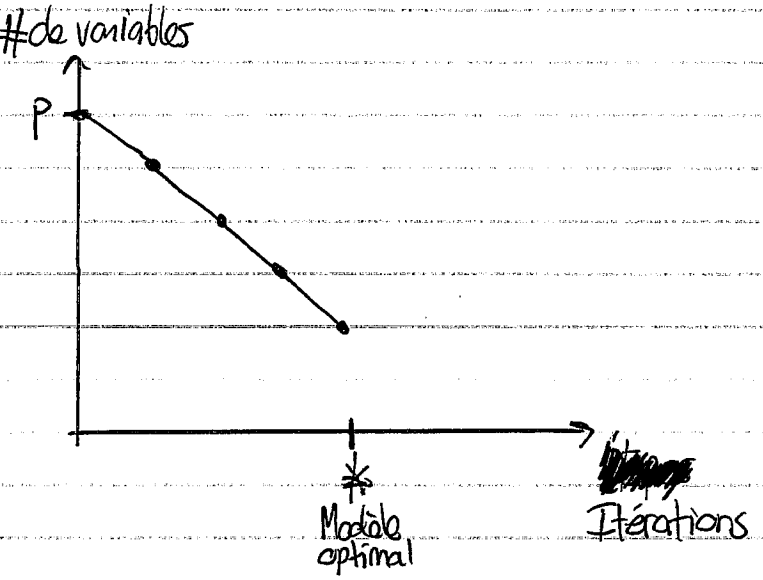
Étape 1 : Débuter avec toutes les variables disponibles dans le modèle

Étape 2 : Chercher LA variable qui génère la plus faible augmentation de SSE lorsqu'exclue du modèle ...la pire variable

Étape 3 : Utiliser les tests F partiels (3.2.3) pour tester si il est possible d'exclure la variable de l'étape 2.

Étape 4 : Continuer les étapes 2 et 3 jusqu'à ce qu'il n'y ait plus de variables à éliminer selon les tests F partiels

Illustration:



Remarque:

Le principal inconvénient de cette technique est qu'une variable éliminée ne peut jamais être être réintégrée.

3.3.3) Technique #3: Sélection progressive (forward selection):

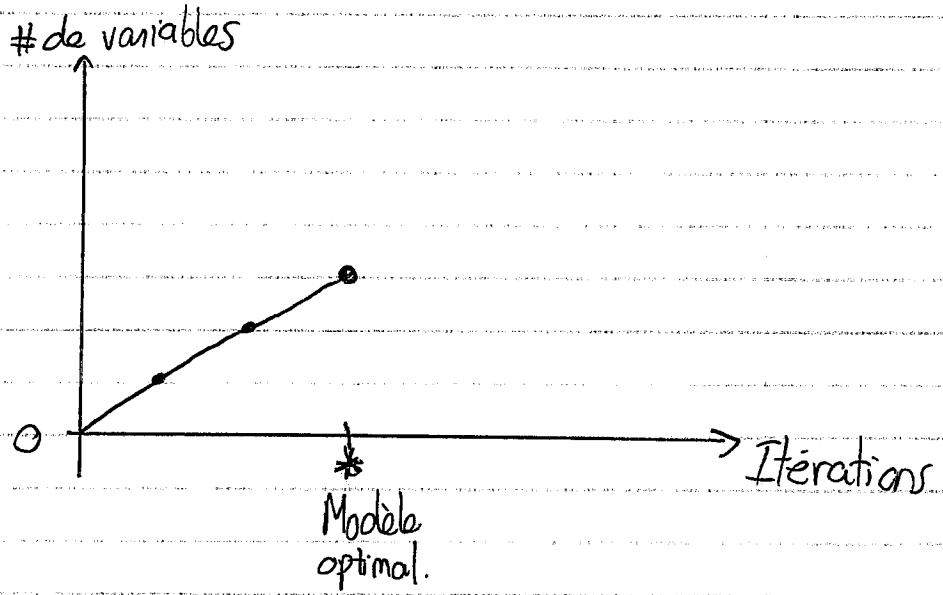
Étape 1: Débuter avec le modèle $Y = \beta_0 + \varepsilon$

Étape 2: Chercher LA variable qui génère la plus grande diminution du SSE lorsqu'incluse dans le modèle
... la meilleure variable

Étape 3: Utiliser les tests F partiels (3.2.3) pour tester si il est possible d'inclure la variable de l'étape 2.

Étape 4: Continuer les étapes 2 et 3 jusqu'à ce qu'il n'y ait plus de variables à inclure selon les tests F partiels

Illustration:



Remarque:

↳ principal inconvénient de cette technique est qu'une variable incluse ne peut jamais être éliminée par la suite.

3.3.4) Technique #4: Régression pas à pas (stepwise regression):

*Combinaison de backward et forward.

Étape 1: Débuter avec le modèle $Y = \beta_0 + \varepsilon$

Étape 2: Chercher LA meilleure variable (qui génère la plus grande diminution du SSE si incluse)

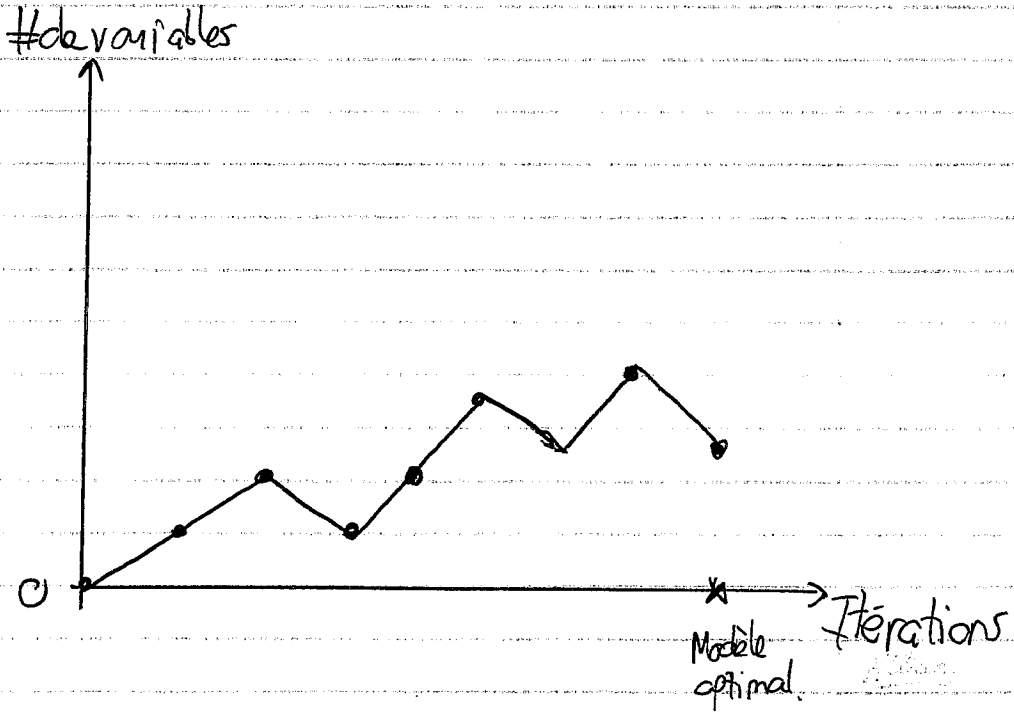
Étape 3: Utiliser les tests F partiels pour tenter d'inclure la variable de l'étape 2

Étape 4: Chercher LA pire variable (qui génère la plus faible diminution du SSE si exclue)

Étape 5: Utiliser les tests F partiels pour tenter d'exclure la variable de l'étape 4.

Étape 6: Continuer les étapes 2 à 5 jusqu'à ce que l'algorithme élimine la variable qui vient d'entrer.

Illustration:



Exemple: $n=20$ observations

Variables du modèle	SSE	SSR	SST	R^2_{adj}
$\emptyset$	10	0	10	0%
X_1	5	5	10	47,2%
X_2	9	1	10	5%
X_3	8	2	10	15.5%
X_1, X_2	4	6	10	55.3%
X_1, X_3	3.9	6.1	10	56.4% *
X_2, X_3	8.5	1.5	10	5%
X_1, X_2, X_3	3.8	6.2	10	54.9%

Technique 1: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \varepsilon$

Technique 2: • Modèle initial = $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

• Pire variable : X_2

$\Rightarrow$ H_0: Modèle avec X_1 et X_3 $\Rightarrow$ $F = \dfrac{(3.9 - 3.8)/1}{3.8/16} = 0.4211$

H_1: Modèle avec X_1, X_2 et X_3

$\Rightarrow$ $F < F_{5\%}(1,16) = 4.49$

$\Rightarrow$ On accepte H_0

$\Rightarrow$ On exclut X_2

• Pire variable : X_1

$\Rightarrow$ H_0: Modèle avec X_3

H_1: Modèle avec X_1 et X_3 $\Rightarrow$ $F = \dfrac{(5 - 3.9)/1}{3.9/17} = 4.79 > 4.45 = F_{5\%}(1,17)$

$\Rightarrow$ On rejette H_0 $\Rightarrow$ X_3 n'est pas exclut / Modèle $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \varepsilon$

3.4) Régression avec variables indicatrices:

Permettent de traiter des variables explicatives catégoriques (=non-numériques) dans les modèles.

ex: • Couleur des yeux (bleu, brun, vert, autres)
• Type de véhicule (sport, autre)
• Emploi (ACT, ETU, RTR, GOU, autres)
⋮

Par inclure une variable catégorique ayant r valeurs possibles; on doit créer (r-1) variables indicatrices

ex: -Couleur des yeux: $X_{t,1} = 1\{Couleur_t = Bleu\}$
$X_{t,2} = 1\{Couleur_t = Brun\}$
$X_{t,3} = 1\{Couleur_t = Vert\}$

-Type de véhicule: $X_{t,4} = 1\{Type_t = Sport\}$

-Emploi: $X_{t,5} = 1\{Emploi_t = ACT\}$
$X_{t,6} = 1\{Emploi_t = ETU\}$
$X_{t,7} = 1\{Emploi_t = RTR\}$
$X_{t,8} = 1\{Emploi_t = GOU\}$

; où $1\{A\} = \begin{cases} 1 & \text{, si A vrai} \\ 0 & \text{, sinon.} \end{cases}$

Modèle de régression:

$$Y_t = \beta_0 + \beta_1 X_{t,1} + \beta_2 X_{t,2} + \beta_3 X_{t,3} + \beta_4 X_{t,4} + \beta_5 X_{t,5} + \beta_6 X_{t,6} + \beta_7 X_{t,7} + \beta_8 X_{t,8} + \varepsilon_t$$

... cela revient encore une fois à bien définir la matrice shéma (X)

exemple: $n = 5$ observations

Y_t	$Couleur_t$	$Type_t$	$Emploi_t$
70	Bleu	Autres	ETU
75	Brun	Sport	GOU
50	Vert	Autres	Autres
55	Autres	Autres	Autres
85	Brun	Sport	ACT

On peut utiliser le modèle de régression multiple suivant:

$$Y = X\beta + \varepsilon$$

... en posant

$$Y = \begin{bmatrix} 70 \\ 75 \\ 50 \\ 55 \\ 85 \end{bmatrix}$$

... et

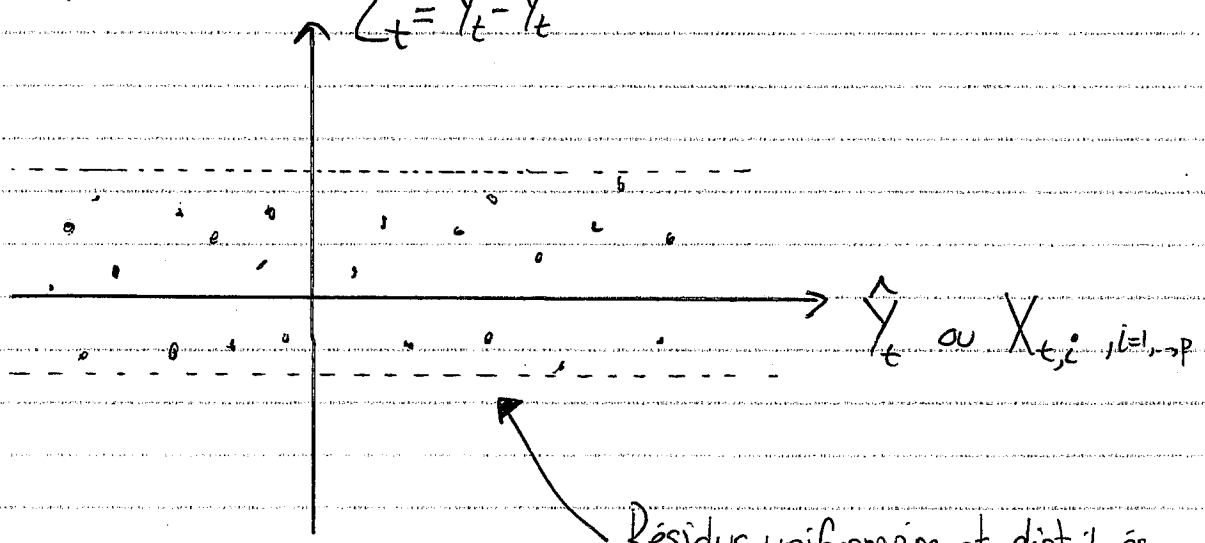
$$X = \begin{array}{c} \begin{matrix} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

3.5) Analyse qualitative des résidus:

Même si les tests t et F sont concluants, le modèle choisi peut ne pas être adéquat.

L'analyse qualitative (ou graphique) des résidus $\varepsilon_t = Y_t - \hat{Y}_t$ est la principale façon de valider un modèle sélectionné

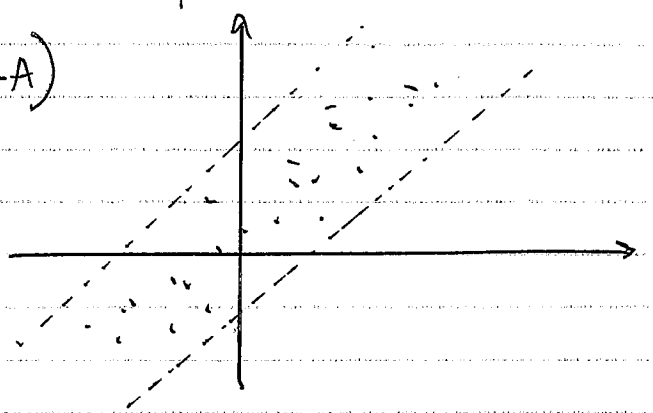
Graphique #1: (à l'examen!)



Résidus uniformément distribués autour de l'axe des "x" ⇒ Idéal
⇒ Pas de problème

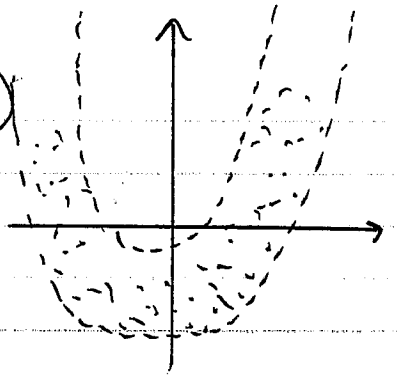
Problèmes possibles:

1-A)



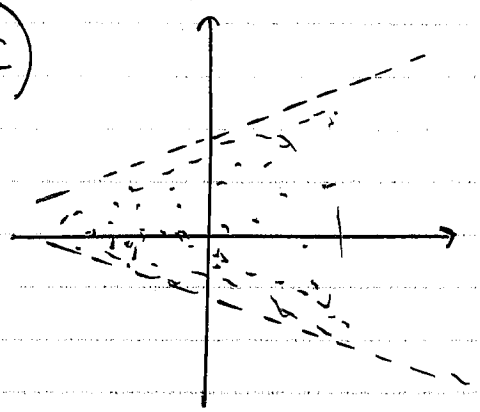
Il manque probablement un terme linéaire dans X.

1-B)



Il manque probablement un terme (une variable) quadratique dans X

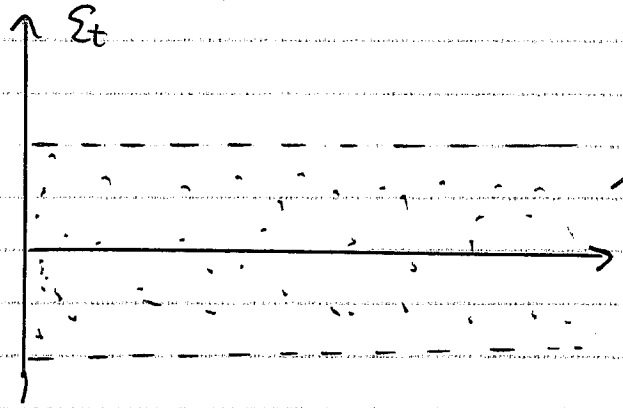
1-C)



La variance est probablement non constante

⇒ Violation de l'hyp. #2.
(--- $Var(\varepsilon_t) = \sigma^2$)

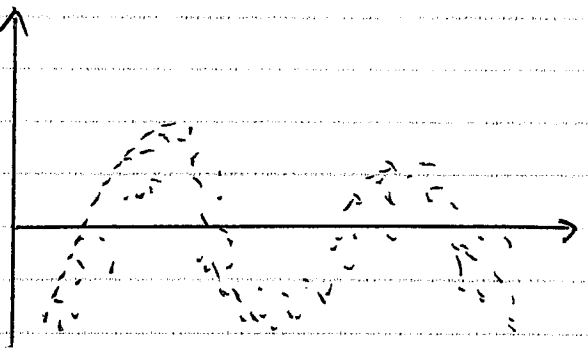
Graphique #2: (pas à l'examen, mais intéressant)



Résidus uniformément distribués autour de l'axe des "x"
⇒ Idéal
⇒ Pas de problèmes!

Problèmes possibles:

2-A)

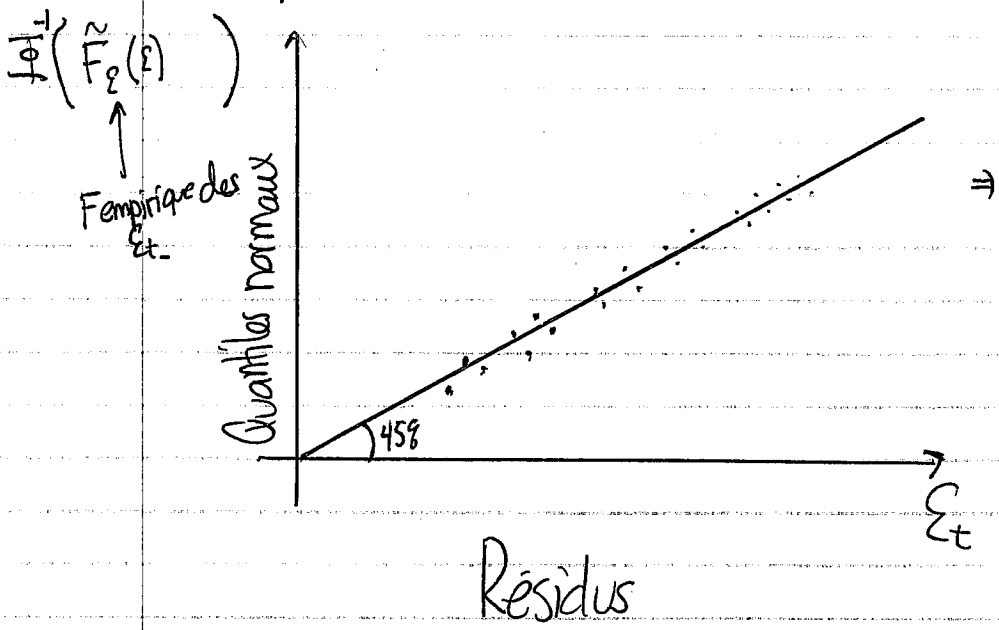


Les résidus sont probablement corrélés entre eux

⇒ Violation de l'hyp #3
(--- $Cov(\varepsilon_t, \varepsilon_s) = 0$, $t \neq s$)

Dans ce cas ⇒ Séries chrono. !!!

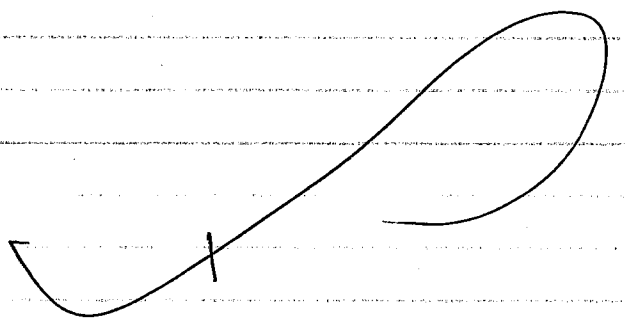
Graphique #3 : Quantiles normaux : (à l'examen!)



⇒ ≈ droite à 45° ⇒ Idéal

⇒ $\varepsilon_t \sim N(0,1)$

⇒ Hyp. #4 validé!



↑↑↑
Rendu ici
12-10-2010.