MODÈLE LINÉAIRE GÉNÉRAL ET LE DOGME DE LA NORMALITÉ : UN OUTIL

Guy Mélard

Faculté Solvay Brussels School of Economics and Management,
ECARES, Université libre de Bruxelles,
CP114/4 Avenue Franklin Roosevelt 50, B-1050 Bruxelles,
& ITSE sprl, Bruxelles

Belgique, gmelard@ulb.ac.be

Plan de l'exposé

- ⇨Introduction = lien avec ma contribution à Angers
- ⇨Condition d'application de la régression (simple)
- ⇨Apports de la théorie
- ⇨Evolutions de la théorie
- ⇨Lien avec le nombre *n* d'observations
- ⇨Objet de l'outil
- ⇨Historique de l'outil
- ⇨Description de l'outil
- ⇨Démonstration
- ⇨Conclusions

Introduction

- ⇨Exposé lié à celui donné à Angers en 2012 (voir page suivante)
- ⇨Outil = programme de simulation mentionné dans la présentation
- ⇨Objectif : vérifier la nécessité pratique des conditions d'application théoriques
- ⇨En particulier, la supposition de normalité souvent mentionnée en statistique et très (*trop*?) souvent testée

Enseigner la régression multiple sans mathématiques. Est-ce possible et est-ce souhaitable ?

Guy Mélard

Faculté Solvay Brussels School of Economics and Management,
ECARES,
Université libre de Bruxelles,
CP114/4 Avenue Franklin Roosevelt 50, B-1050 Bruxelles,
Belgique, gmelard@ulb.ac.be

Contenu du chapitre de régression multiple

- ⇨contexte (avec illustrations)
- ⇨problématique de l'ajustement de données par une fonction du premier degré, à coefficients inconnus
- ⇨estimation des paramètres par la méthode des moindres carrés (sorties de logiciels, interprétation)
- ⇨problèmes numériques sous-jacents (quasi-colinéarité)
- ⇨mesure de la qualité de l'ajustement ($\hat{\sigma}^2$, R^2, $\overline{R}^2$ …)
- ⇨examen de la sensibilité des résultats vis-à-vis des données employées (échantillonnage)
- ⇨. . .

Contenu du chapitre de régression multiple

- ⇨approche d'inférence statistique, en particulier une introduction aux tests d'hypothèses (*t* de Student)
- ⇨conditions d'application et détection des conditions non satisfaites
- ⇨choix des variables (transformations, sélection)
- ⇨et choix des données (variables binaires)
- ⇨emploi de données temporelles
- ⇨prévision

L'hypothèse de normalité en statistique

⇨Base d'un certain nombre de tests paramétriques classiques :

- ✓Test de Student pour la moyenne (et test de comparaison de moyennes)
- ✓Test pour la variance (et test de comparaison associé)
- ✓Test d'analyse de variance (ANOVA) pour la comparaison de plusieurs moyennes et plans expérimentaux déduits (à effets fixes ou aléatoires)
- ✓Tests sur les coefficient de corrélation
- ✓Régression linéaire simple et multiple

Conditions d'application (régression simple)

⇨Variable explicative mesurée sans erreurs

⇨Relation linéaire

⇨Absence de colinéarité (ici avec la constante)

⇨Normalité des erreurs

⇨Donc absence de données aberrantes

⇨Homoscédasticité des erreurs

⇨Indépendance des erreurs

⇒ absence d'autocorrélation positive ou négative(*)

(*) Données chronologiques ou celles où l'ordre des données a un sens

Apports de la théorie

- ⇨Les conditions d'application sont nécessaires
- ⇨Dans une certaine mesure, certaines violations peuvent être traitées (p.ex. l'hétéroscédasticité)
- ⇨La littérature s'accorde sur la faible incidence de légers écarts par rapport aux conditions
- ⇨La démarche souvent recommandée est de **tester** la validité des conditions d'application
- ⇨Mais ces tests statistiques requièrent eux-mêmes des conditions d'application

Evolutions de la théorie

- ⇨ Nombreuses recommandations pour tendre vers la validité des conditions, par exemple transformation (logarithmique ou autre) de variable(s)
- ⇨ Quelques tentatives pour prendre en compte l'absence de validité des conditions d'application
  - ✓ Box & Watson (1962) Robustness to non-normality of regression tests : C_X = mesure de non-normalité des variables explicatives
  - ✓ Seber & Lee (2012) Linear regression analysis, Wiley
  - ✓ Glass et al. (1972) : ANOVA, ANCOVA
  - ✓ …

Résultats de Box-Watson (1962)

$\boldsymbol{y = 1'\beta_0 + X\beta + e}$, où $\mathbf{X}_{n\times p}\mathbf{1} = \mathbf{0}$

Si normalité : $R^2 \sim F_{p,n-p+1}$

Si distribution symétrique :

$R^2 \approx F_{\delta p,\ \delta(N-p+1)}$

avec $\delta^{-1} = 1 + C_X C_y / 2n$, $C_y = \gamma_2$

et $-2 < C_X < N$

Note. Ceci ne semble pas utilisé dans les logiciels statistiques en faveur de techniques plus modernes (permutation, robustesse, bootstrap, …)

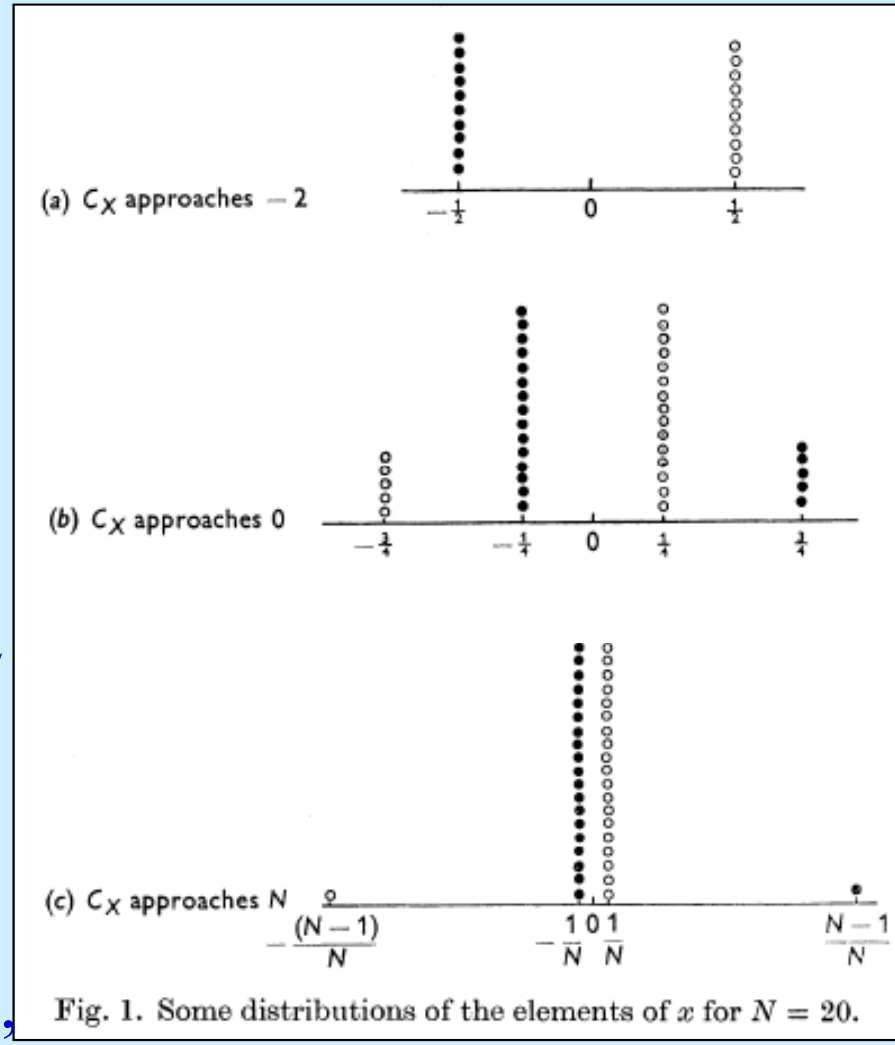


Fig. 1. Some distributions of the elements of x for $N = 20$.

Source : Box-Watson (1962)

SolvayBrusselsSchool of Economics and Management ULB

Lien avec le nombre n d'observations

- Typiquement on dit que la normalité est nécessaire pour des petits échantillons, pas pour de grands échantillons, « en vertu du théorème central limite »
- Très souvent, notamment pour la moyenne, on voit la condition $n \geq 30$ pour invoquer ce dernier
- Ceci ne repose sur RIEN : selon les cas (c'est-à-dire la distribution de la variable dans la population) on peut avoir $n \geq 10$ ou $n \geq 10000$
- Exemple de cette dernière situation : gain lors de jeux

Objet de l'outil

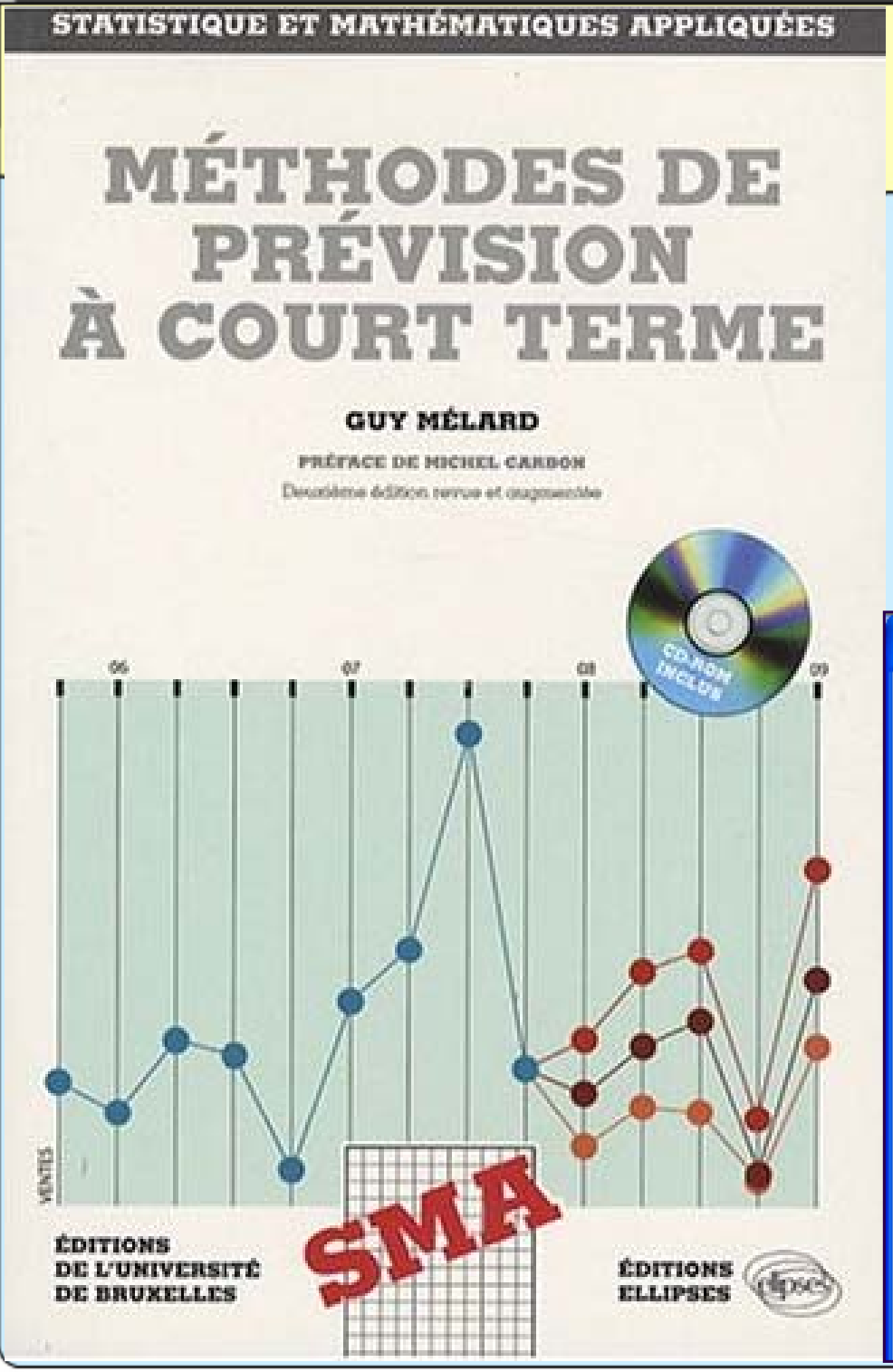
⇨Régression linéaire simple de y en x et en particulier le coefficient de régression b : $y = bx + e$ ($b_{vrai} = \beta$)

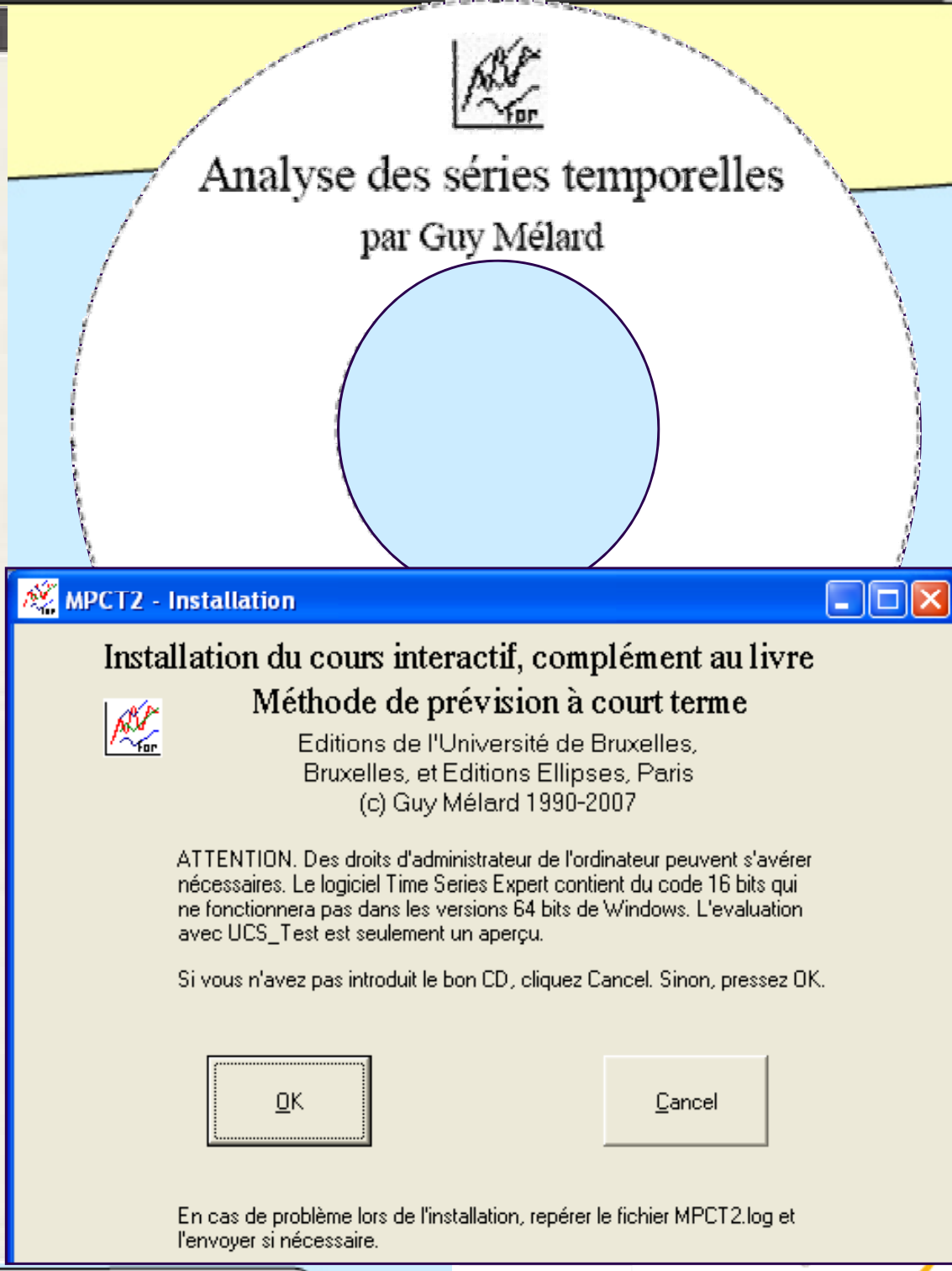
⇨Programme de simulation qui étudie l'effet de la violation d'une des conditions d'application théoriques sur les résultats de la régression, essentiellement

- ✓la distribution de la statistique de Student de b (Student à $n - 2$ degrés de liberté ?) et …
- ✓la distribution des résidus qui détermine la distribution de valeurs futures de y pour x donné

Historique de l'outil

- Mon premier programme Matlab (26/11/1996) basé sur un séminaire donné en 1979
- Utilisé comme illustration dans mes cours de statistique informatique (1996-2010), informatique (1997-2004), méthodes de prévision et techniques quantitatives de gestion (2001-2015)
- Programme et vidéo de démonstration présents sur le CD-ROM de M (2007), voir page suivante
- Tests sur octave (version anglaise seulement) et réalisation d'une version Matlab exécutable (2014)
- Conversion en une fonction R (2015) à la demande d'un arbitre


STATISTIQUE ET MATHÉMATIQUES APPLIQUÉES
MÉTHODES DE PRÉVISION À COURT TERME
GUY MÉLARD
PRÉFACE DE MICHEL CARDON
Deuxième édition revue et augmentée
CD-ROM INCLUS
06
07
08
09
VENTES
SMA
ÉDITIONS DE L'UNIVERSITÉ DE BRUXELLES
ÉDITIONS ELLIPSES


Analyse des séries temporelles
par Guy Mélard
MPCT2 - Installation
Installation du cours interactif, complément au livre
Méthode de prévision à court terme
Editions de l'Université de Bruxelles,
Bruxelles, et Editions Ellipses, Paris
(c) Guy Mélard 1990-2007
ATTENTION. Des droits d'administrateur de l'ordinateur peuvent s'avérer nécessaires. Le logiciel Time Series Expert contient du code 16 bits qui ne fonctionnera pas dans les versions 64 bits de Windows. L'evaluation avec UCS_Test est seulement un aperçu.
Si vous n'avez pas introduit le bon CD, cliquez Cancel. Sinon, pressez OK.
OK
Cancel
En cas de problème lors de l'installation, repérer le fichier MPCT2.log et l'envoyer si nécessaire.

Importance pratique de certaines des conditions d'application

UNE ANALYSE DE MONTE CARLO POUR LA REGRESSION LINEAIRE

Objet

Quelle est l'influence de la loi des erreurs sur la distribution de la statistique de Student pour le coefficient de régression d'une régression simple estimée par la méthode des moindres carrés où la variable explicative (x) représente le temps. On envisage 9 cas avec chaque fois une des conditions non satisfaite:

1. aucune
2. variable explicative avec erreurs
3. colinéarité
4. relation non linéaire
5. non normalité (4 lois autres de même dispersion (mesurée par l'intervalle interquartile égal dans chaque cas à 1,348) que la normale: uniforme, Laplace, Student, Student à 2 degrés de liberté, Cauchy)
6. données aberrantes
7. hétéroscédasticité
8. autocorrélation positive des erreurs
9. autocorrélation négative des erreurs

Méthode On a généré $R = 5000$ simulations d'une régression linéaire simple avec $\beta = 1$ sur $n = 100$ observations. Soit $h = 10$. On représente chaque fois six graphiques:

(1) les données de la première simulation avec relation vraie et la droite d'ajustement;

(2) les résidus de la première simulation en fonction des valeurs ajustées par le modèle estimé, avec les intervalles de confiance à 95% basés sur la loi normale;

(3) les résidus de la première simulation en fonction du temps, + intervalles de confiance (idem);

(4) la fonction de distribution empirique de l'erreur pour l'observation n° 110, comparée avec la loi normale;

(5) la fonction de distribution empirique de la statistique de Student, comparée à la loi de Student à $n - 2$ degrés de liberté (*pour examiner la validité du test de régression*);

(6) la fonction de distribution empirique du résidu au temps n° 110, comparée à la loi normale (*pour examiner la validité des intervalles de prévision*).

Importance pratique de certaines des conditions d'application

Quelle est l'influence de la loi des erreurs sur la distribution de la statistique de Student pour le coefficient de régression d'une régression simple estimée par la méthode des moindres carrés où la variable explicative (x) représente le temps (1 à n). On envisage 9 cas avec chaque fois une des conditions non satisfaite:

1. aucune
2. variable explicative avec erreurs
3. colinéarité
4. relation non linéaire
5. non normalité (4 lois autres de même dispersion (mesurée par l'intervalle interquartile égal dans chaque cas à 1,348) que la normale: uniforme, Laplace, Student, Student à 2 degrés de liberté, Cauchy)
6. données aberrantes
7. hétéroscédasticité
8. autocorrélation positive des erreurs
9. autocorrélation négative des erreurs

Choix des paramètres du programme

⇨ Autres distributions statistiques que la normale :

- ✓ loi uniforme ($k = 2$)
- ✓ loi de Laplace ou double exponentielle ($k = 3$)
- ✓ loi de Student à 2 degrés de liberté ($k = 4$)
- ✓ loi de Cauchy ($k = 5$)

} $C = 5$

⇨ Choix des paramètres au lieu de $X = (1\ 2\ \ldots\ n)$' :

- ✓ ervar $= n^{1/2}$ — $x_i \leftarrow x_i + \text{ervar}*Nalea_i$ — $C = 2$
- ✓ eps $= 10^{-10}$ — $x_i \leftarrow 1 + x_i*\text{eps}$ — $C = 3$
- ✓ quadr $= 20/n^2$ — ajout de quadr$*(x_i - n/2)^2$ — $C = 4$
- ✓ pertu $= 2/n$ — la proportion de données aberrantes
- ✓ ampli $= 10$ — l'amplitude des perturbations

} $C = 6$

si $Ualea_i <$ pertu alors $e_i \leftarrow e_i + \text{ampli}$

- ✓ heter $= 2/n$ — $e_i \leftarrow e_i\,(0{,}5 + \text{heter}*x_i)$ — $C = 7$
- ✓ autoc $= \pm 2.5/(n^{1/2})$ — $e_i \leftarrow e_i + \text{autoc}*e_{i-1}$ — $C = 8, 9$

Importance pratique de certaines des conditions d'application

Méthode On a généré $R = 5000$ simulations d'une régression linéaire simple avec $\beta = 1$ sur $n = 100$ observations. Soit $h = 10$.

On représente chaque fois six graphiques:

(1) les données de la première simulation avec relation vraie et la droite d'ajustement;

(2) les résidus de la première simulation en fonction des valeurs ajustées par le modèle estimé, avec les intervalles de confiance à 95% basés sur la loi normale;

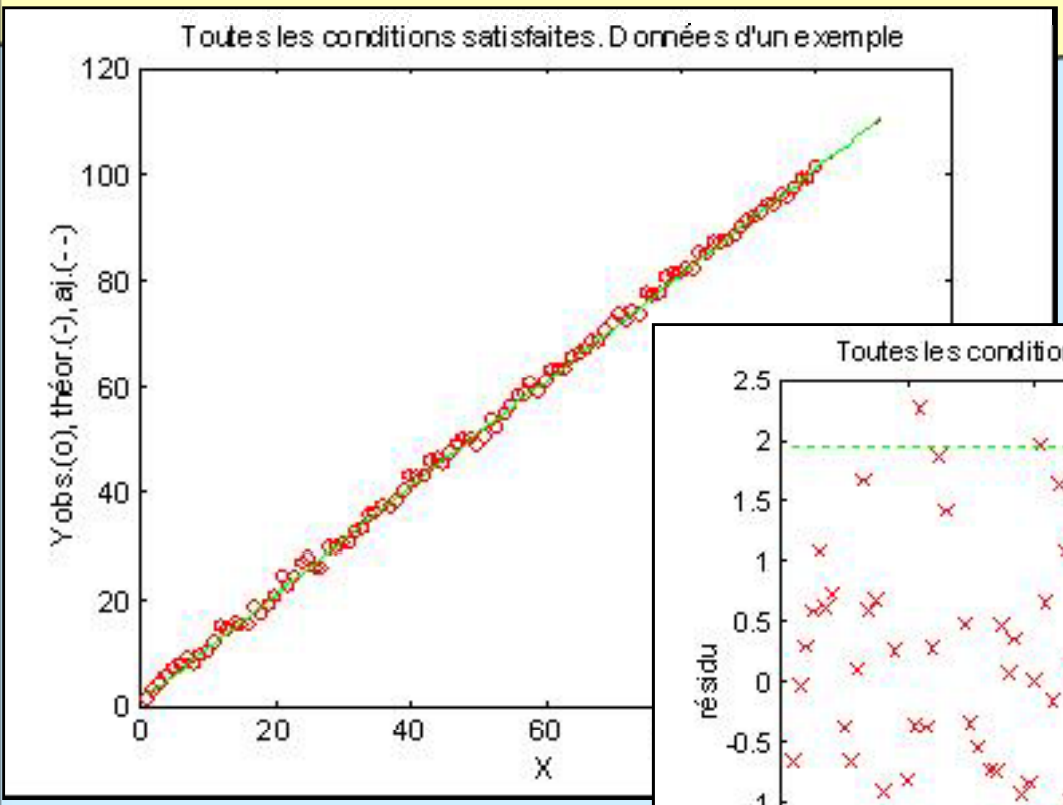
(3) les résidus de la première simulation en fonction du temps, + intervalles de confiance (idem);

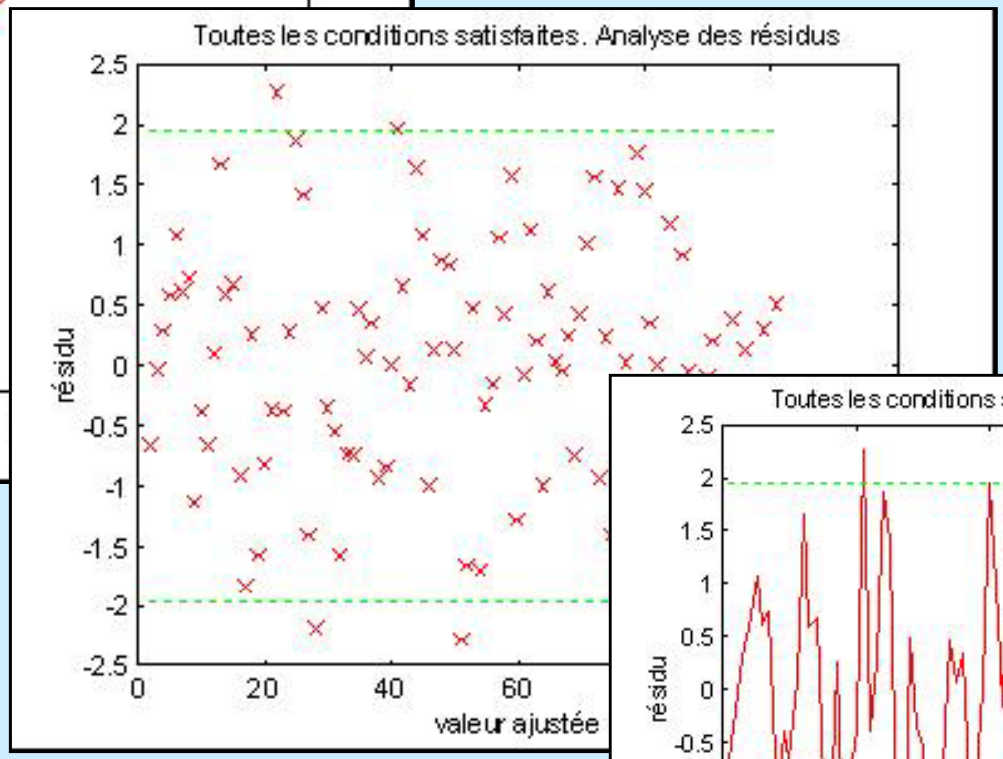
(4) la fonction de distribution empirique de l'erreur pour l'observation $n + h$, comparée avec la loi normale;

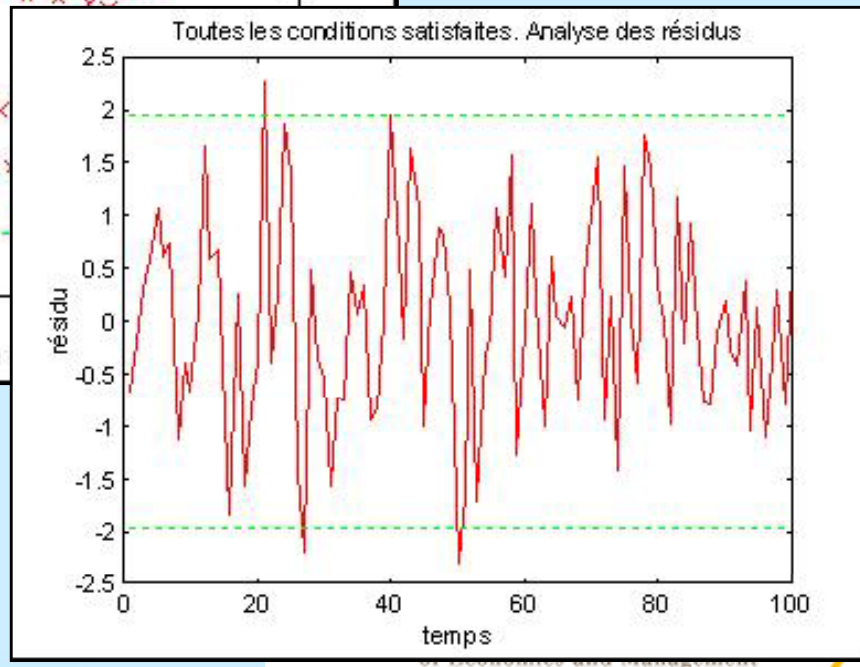
(5) la fonction de distribution empirique de la statistique de Student, comparée à la loi de Student à $n - 2$ degrés de liberté (*pour examiner la validité du test de régression*);

(6) la fonction de distribution empirique du résidu au temps $n + h$, comparée à la loi normale (*pour examiner la validité des intervalles de prévision*).

Aucune condition non vérifiée. Figures 1 à 3



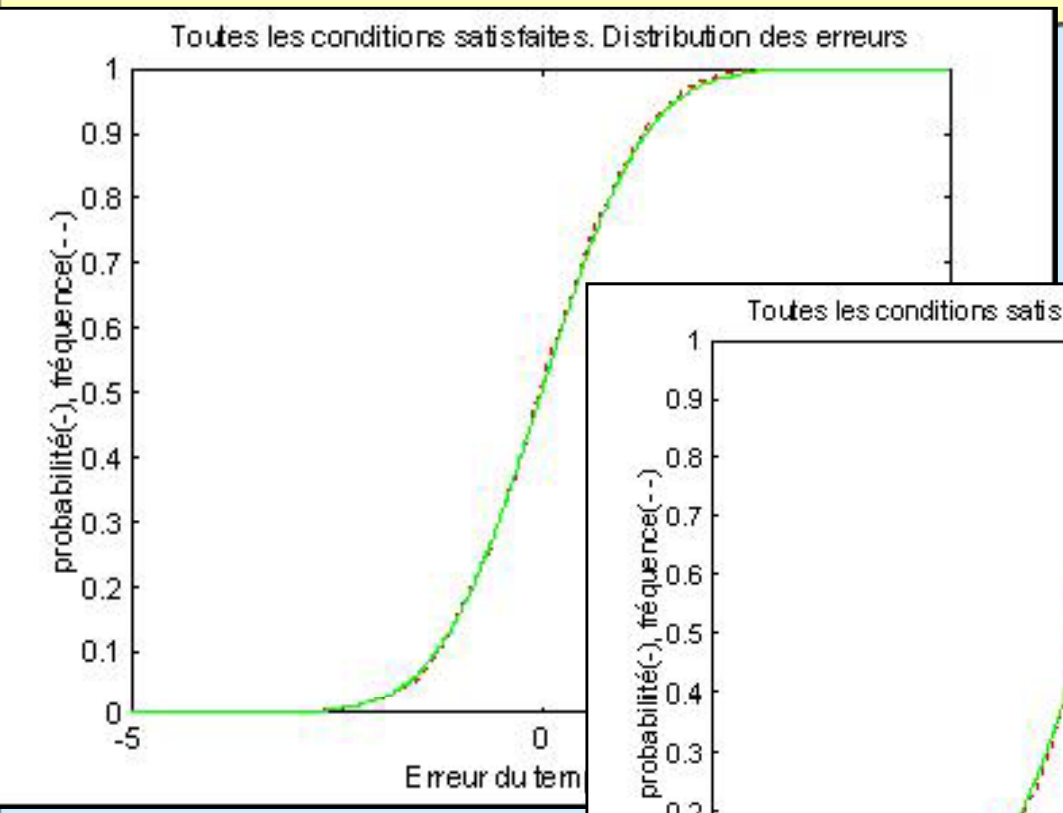


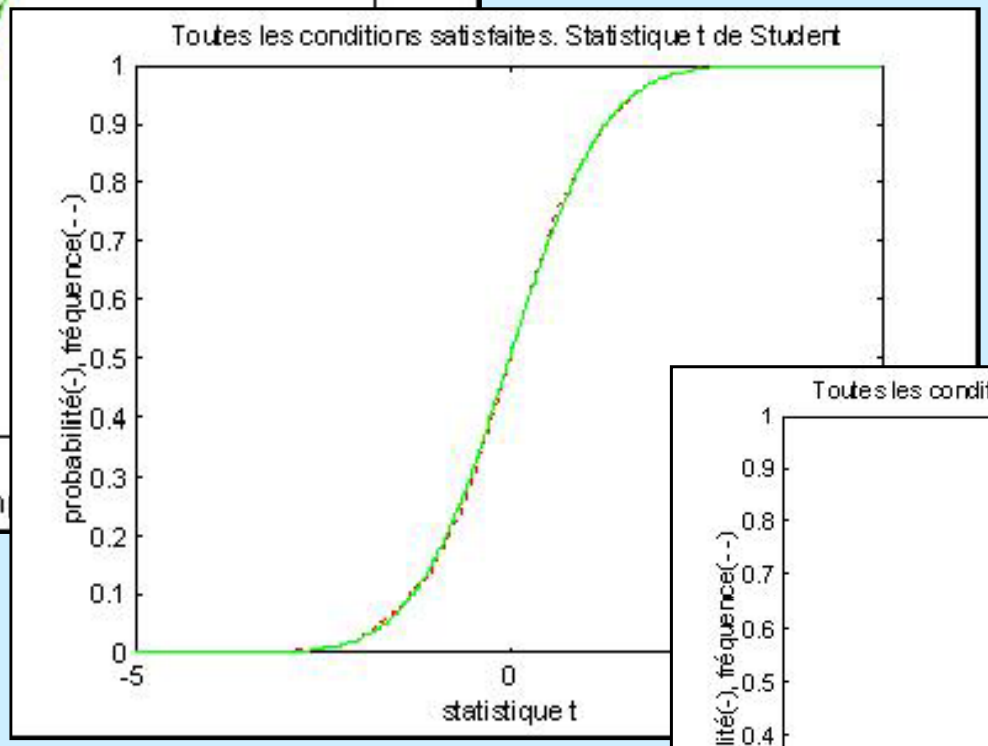


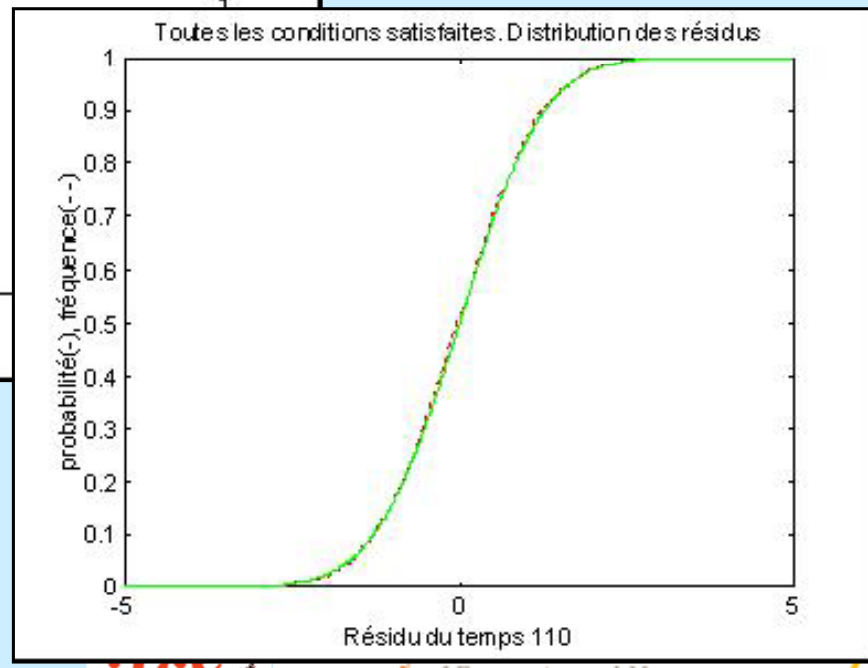
Pourquoi ces graphiques?

N.B. Une séquence vidéo ($R = 1000$, $n = 10$, $h = 3$) et un document ($R = 5000$, $n = 100$, $h = 10$) dans le CD de Mélard (2007) . $\beta = 1$

Aucune condition non vérifiée. Figures 4 à 6



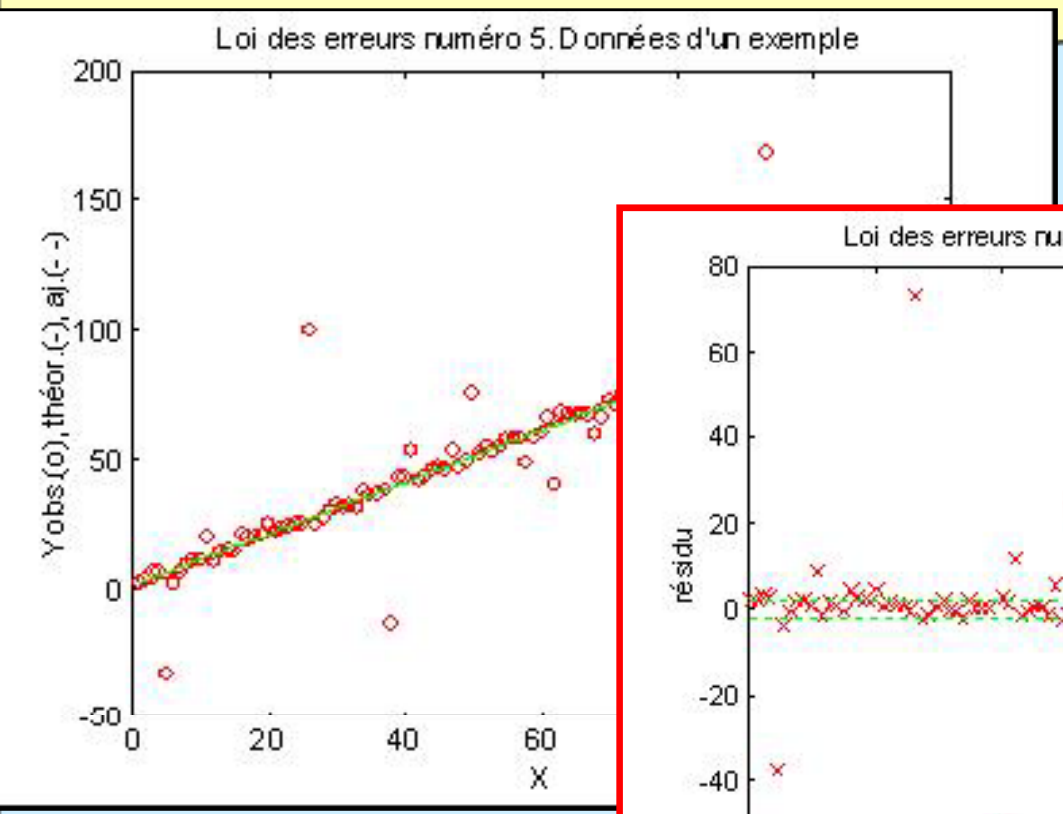


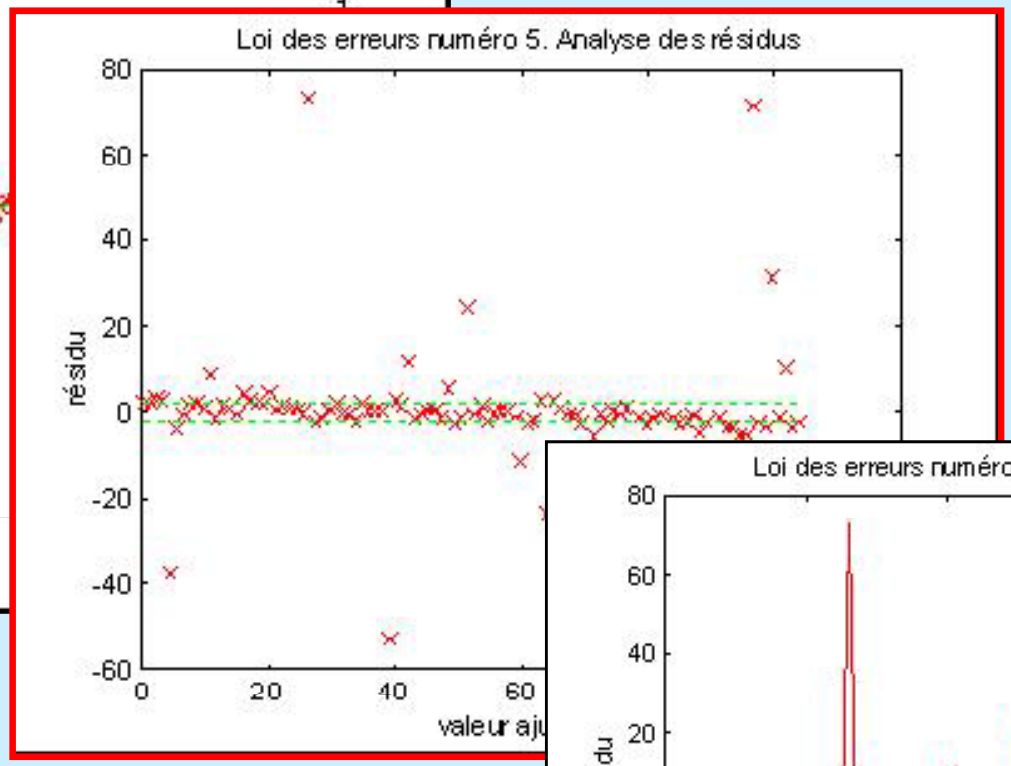


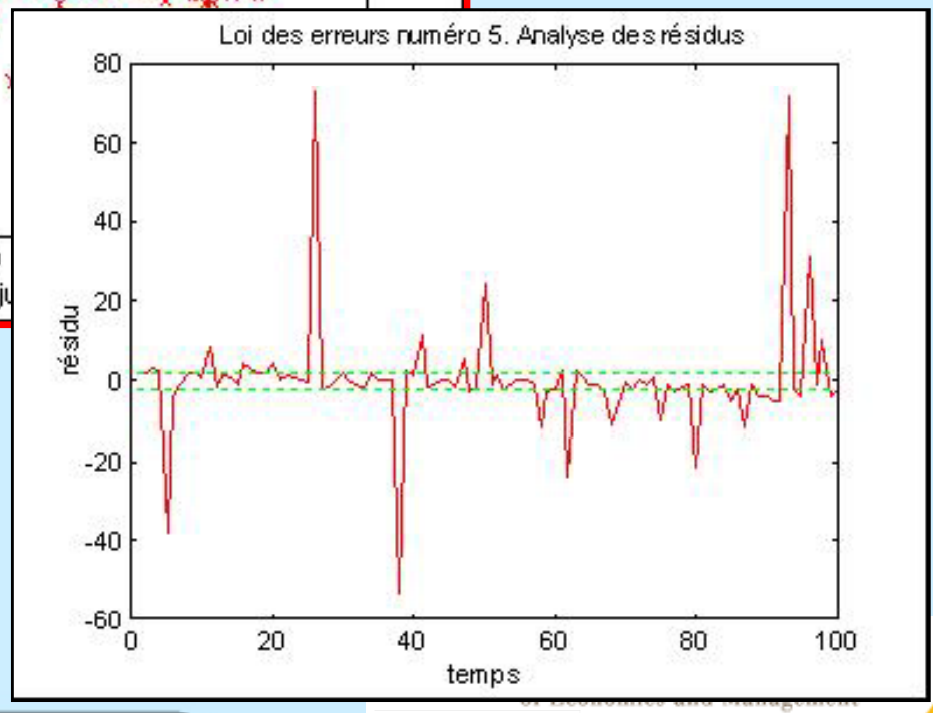
Pourquoi ces graphiques?

Conclusion : tout va bien

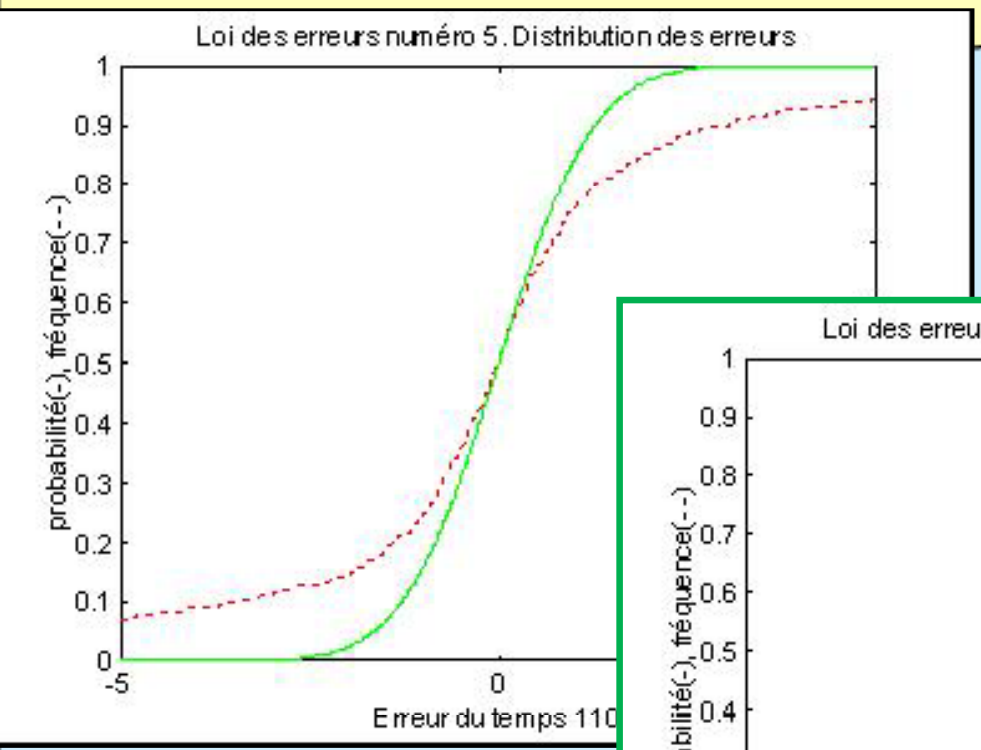
Erreurs avec loi de Cauchy. Figures 1 à 3

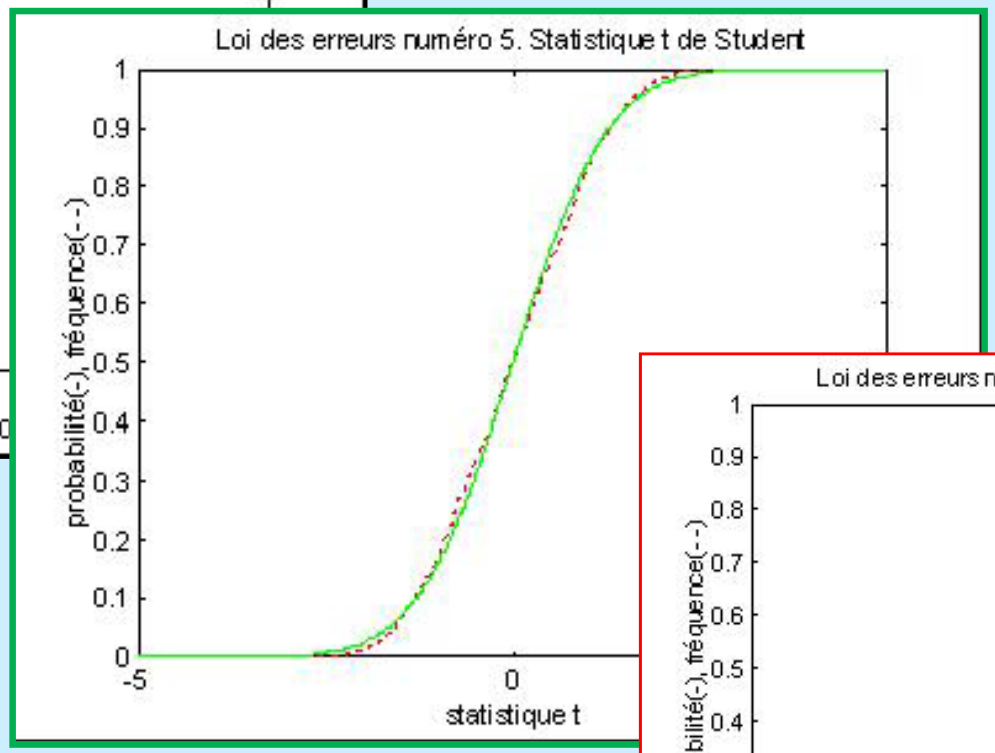


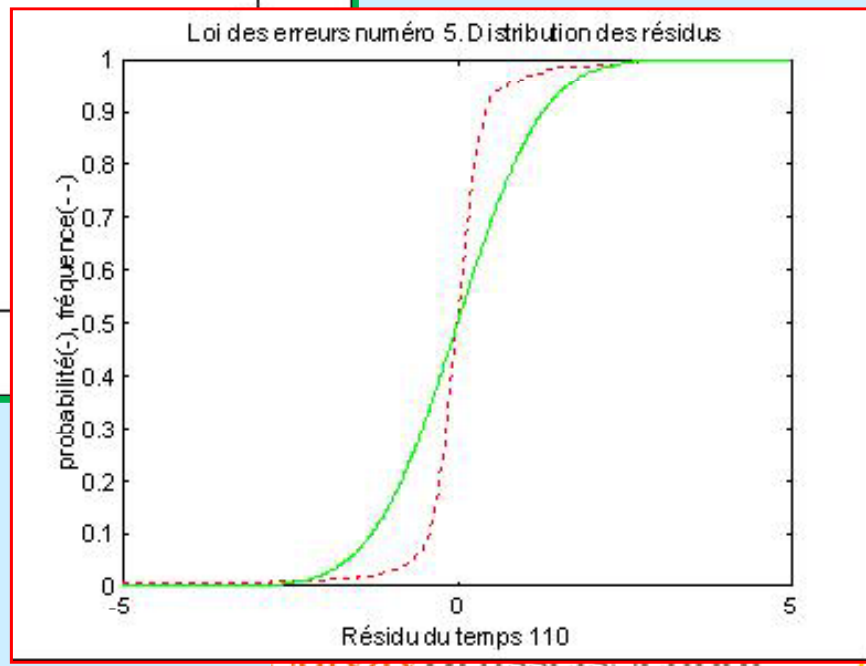




Erreurs avec loi de Cauchy. Figures 4 à 6

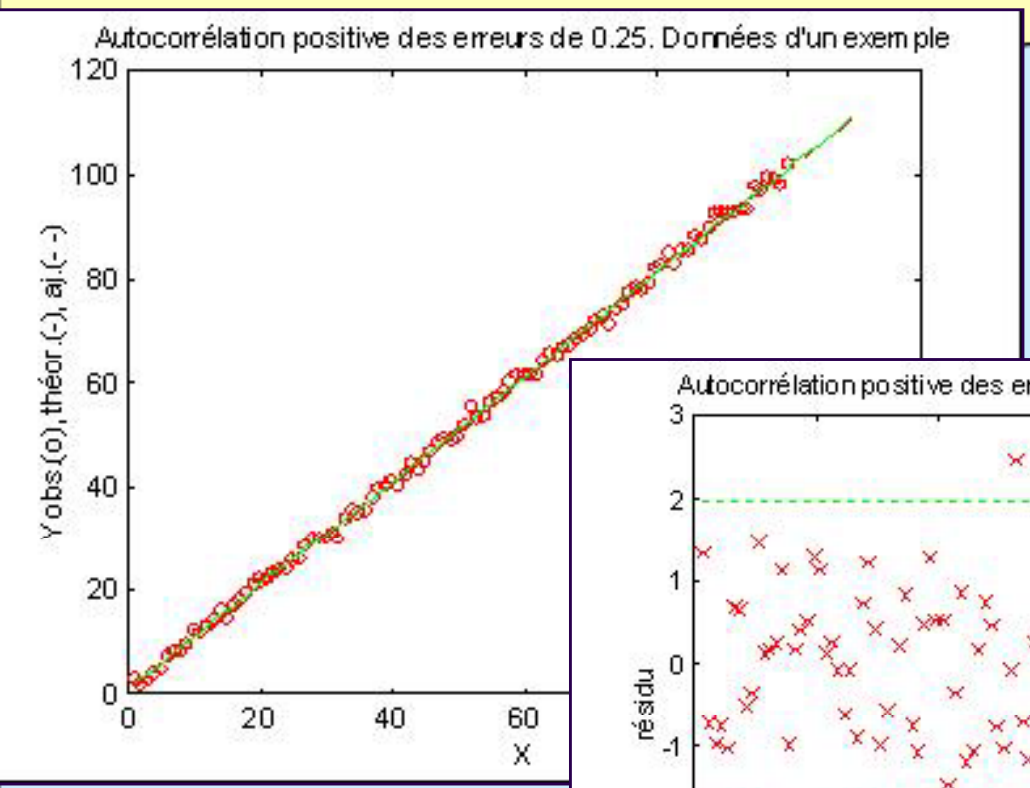


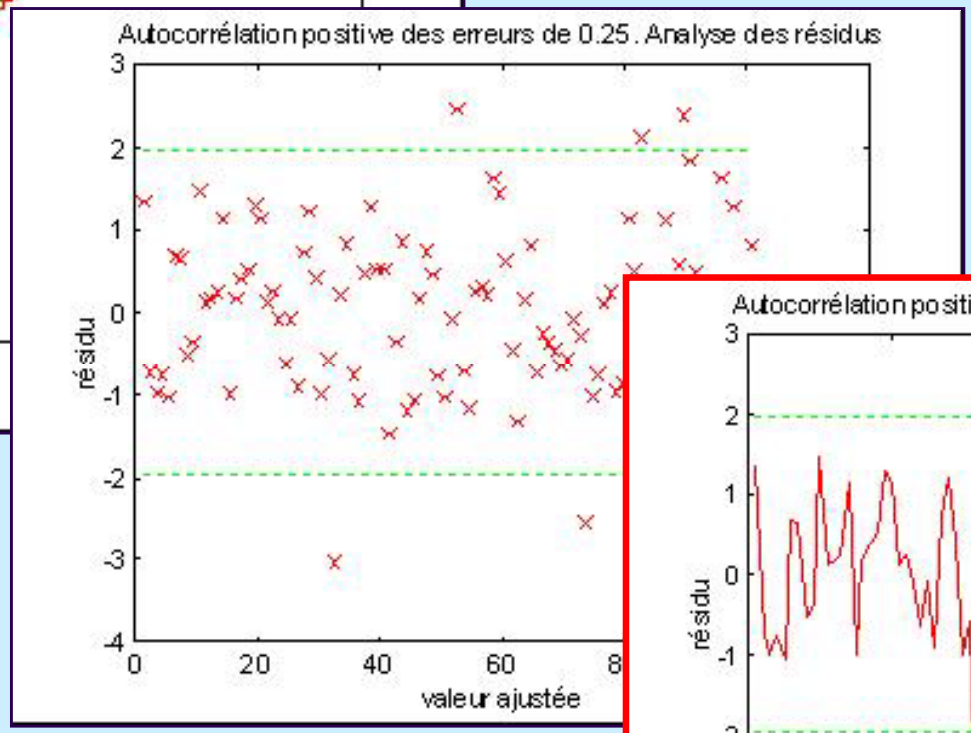


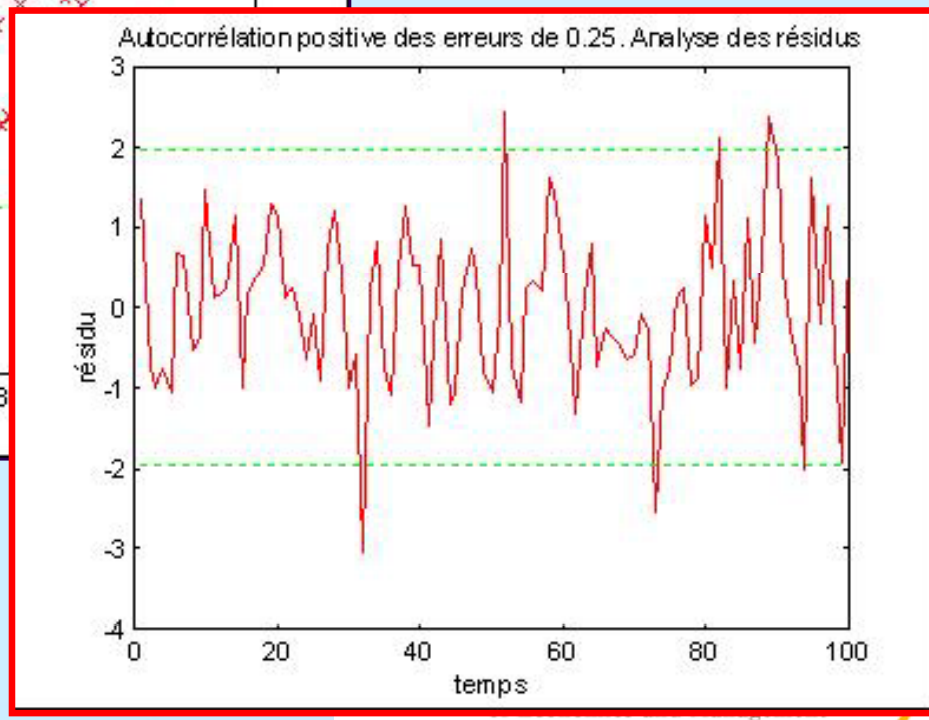


Conclusion : tests de Student Ok
Mais intervalles de prévision trop étroits

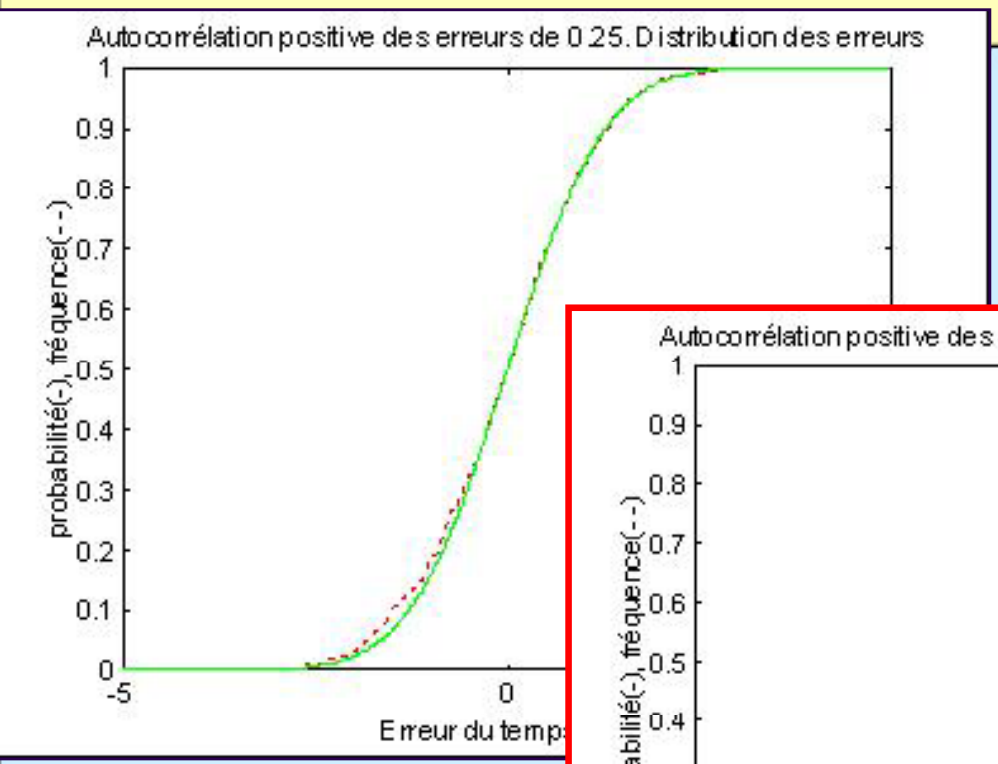
Autocorrélation de +0,25. Figures 1 à 3

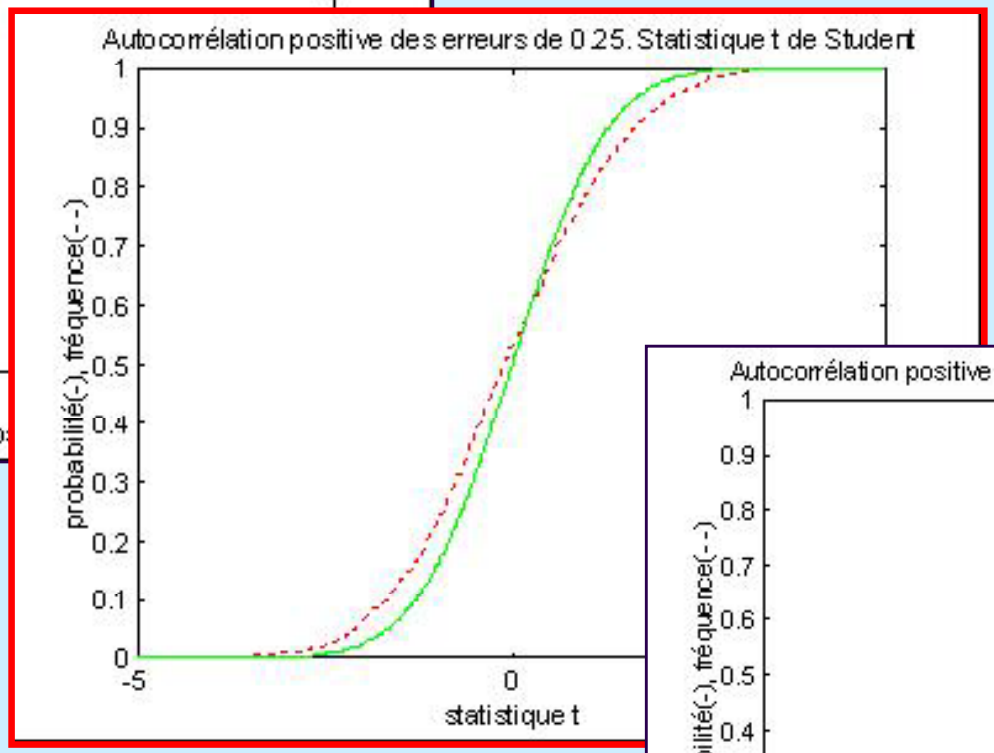


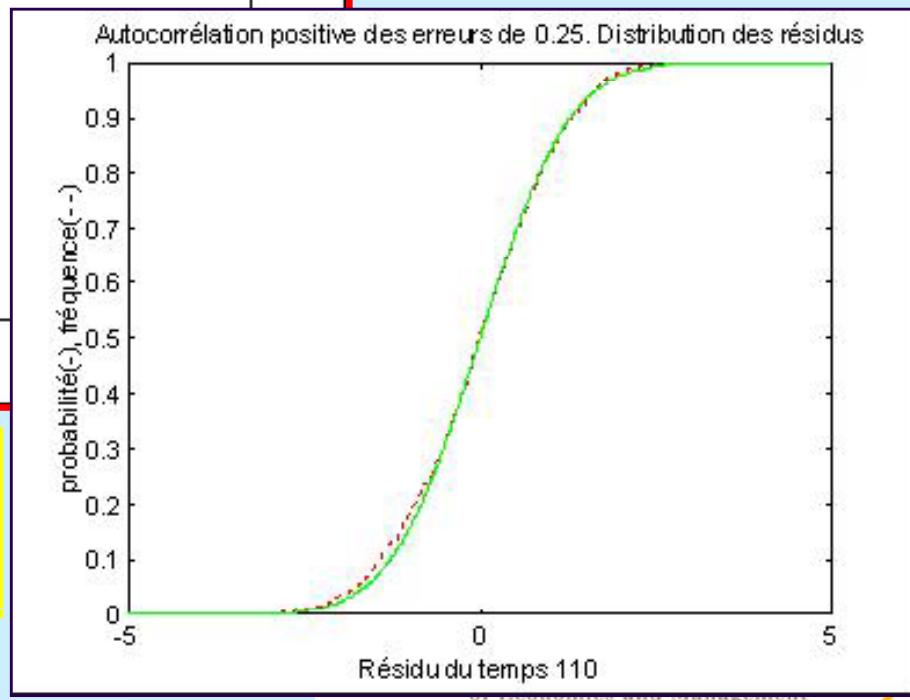




Autocorrélation de +0,25. Figures 4 à 6







Conclusion : tests de Student très biaisés
Mais intervalles de prévision Ok

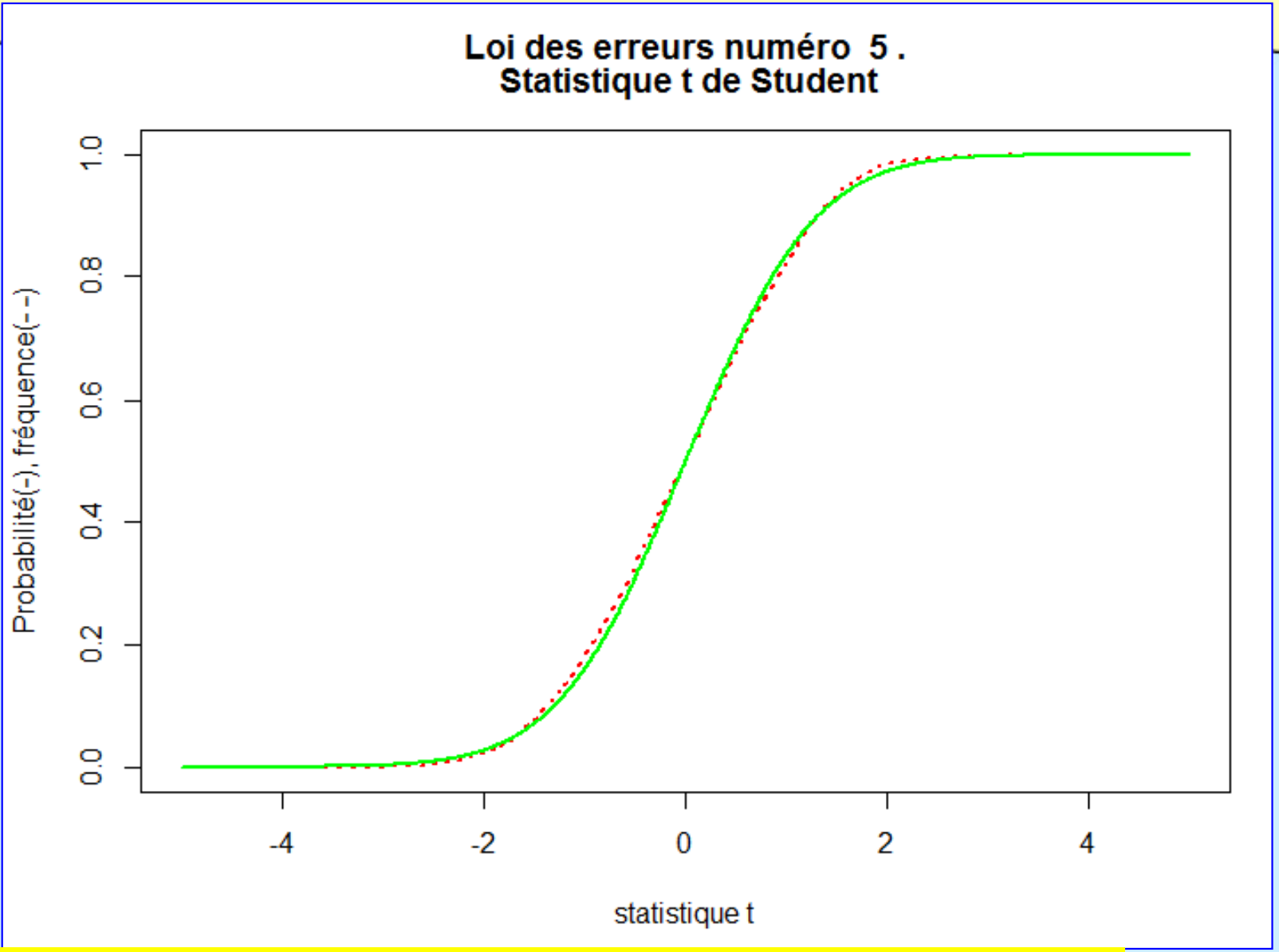
Développement de l'outil sous R (1)

⇨Sous la forme d'une fonction

```
condreg <- function(C=1, k=1, n=10, H=3, R=1000, beta=1)
```

- ✓C : conditions satisfaites (1=toutes, 2= toutes sauf ...)
- ✓k : loi des erreurs (1 sauf pour C = 5)
- ✓n : nombre d'observations
- ✓h : horizon de prévision
- ✓R : nombre de réplications
- ✓beta : vrai coefficient de régression β

n = 30. Erreurs avec loi de Cauchy. Figure 5

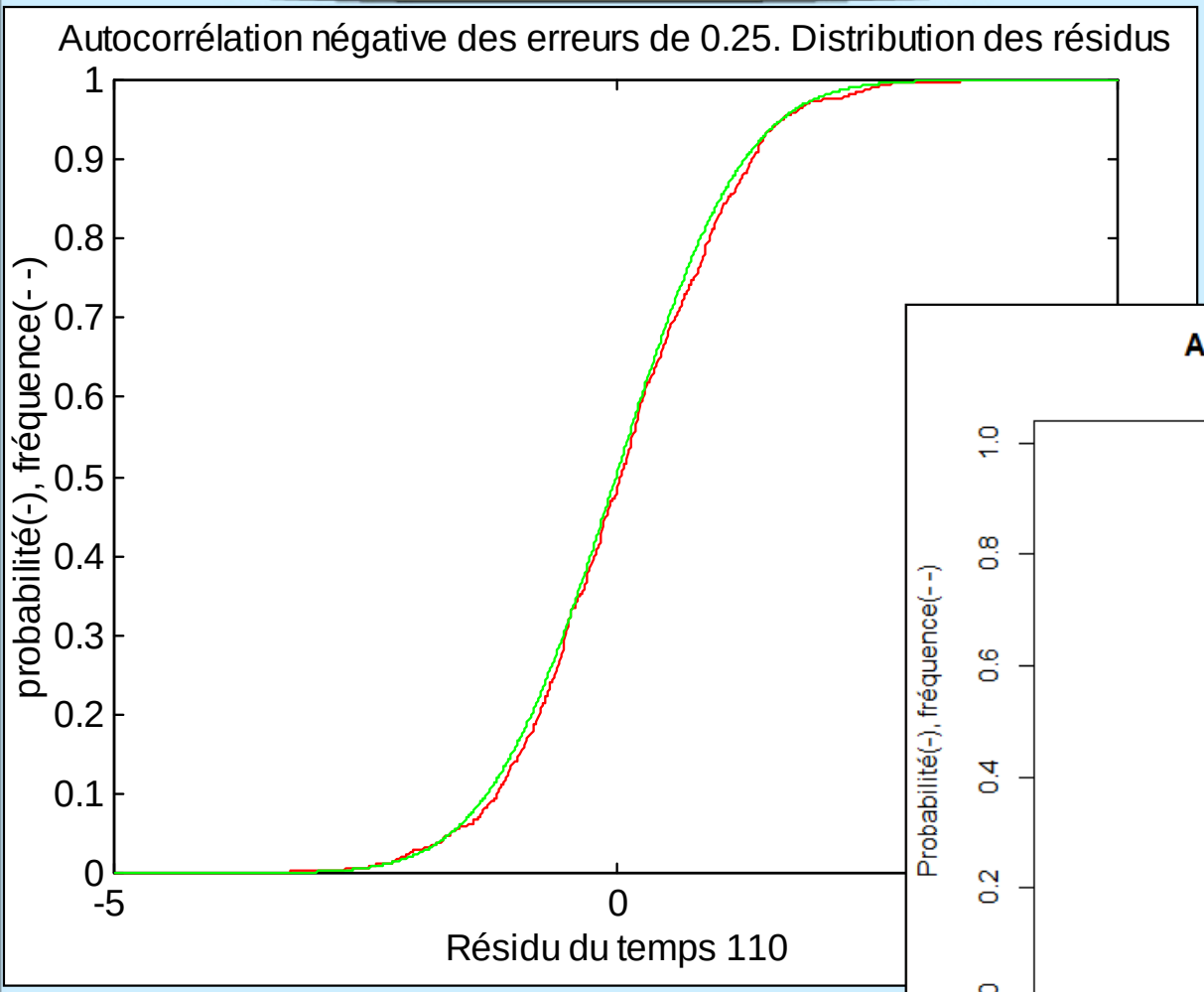


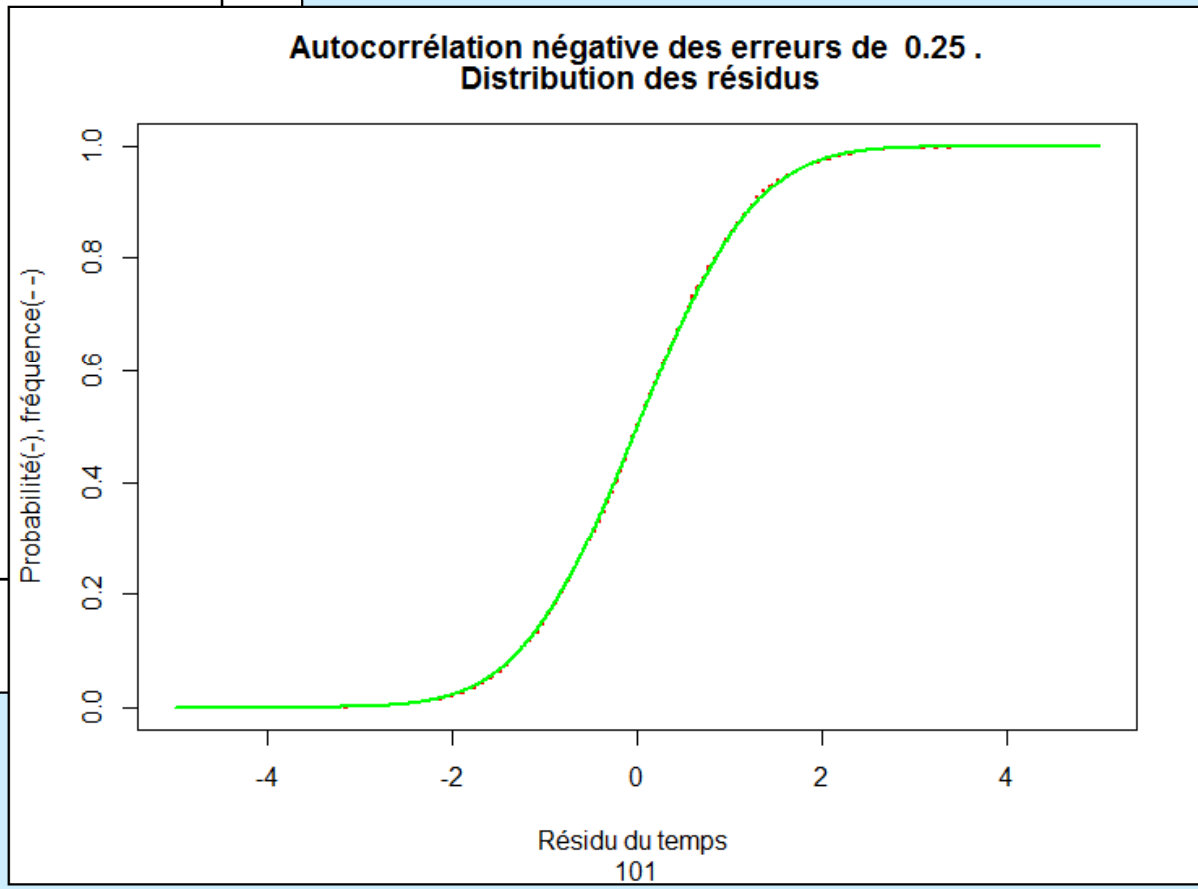

```
> source("condreg.R")
> condreg(C=5, k=5, n=30, H=3, R=5000)
```

Développement de l'outil sous R (2)

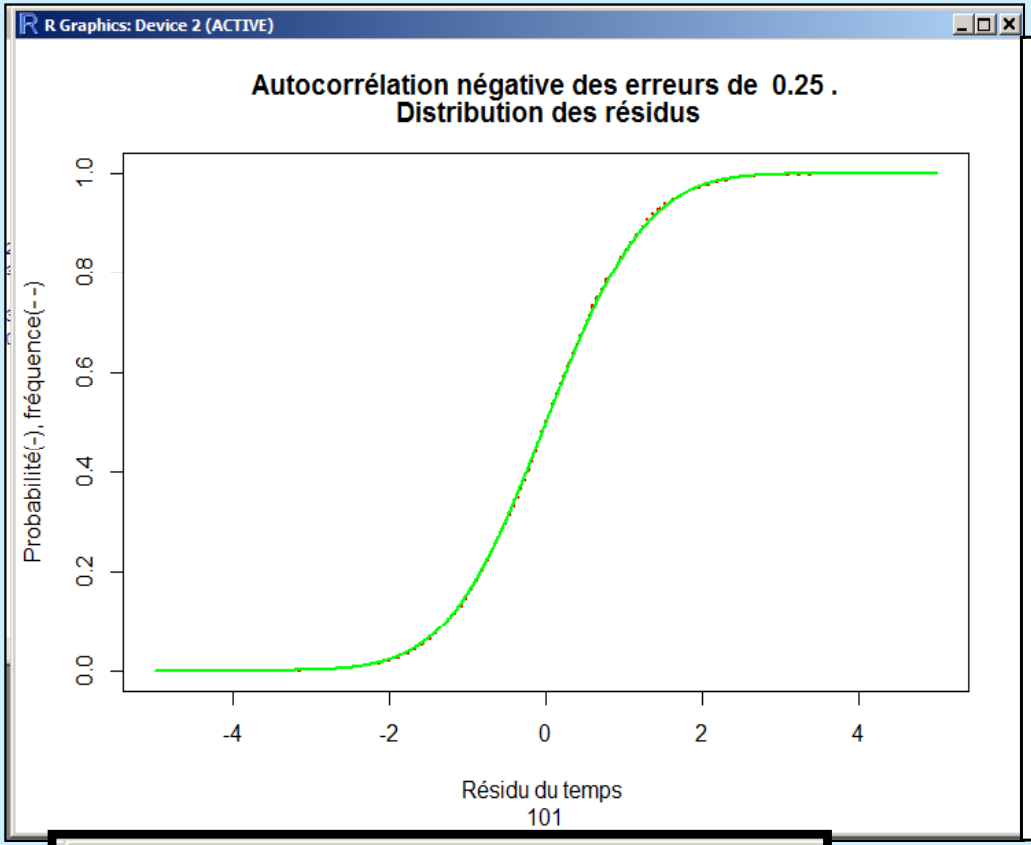
- ⇨Traduction du programme Matlab condreg.m
- ⇨Ma première fonction R, condreg.R, mais plus laborieux que mon premier programme Matlab
- ⇨Parties délicates :
  - ✓Remplacement du script avec menus par une fonction avec arguments par défaut
  - ✓"cov" est différent
  - ✓Noms de fonctions à changer
  - ✓Graphiques (titres, traits, couleurs, plusieurs tracés)
  - ✓Générations de fichiers graphiques (postscript p.ex.)

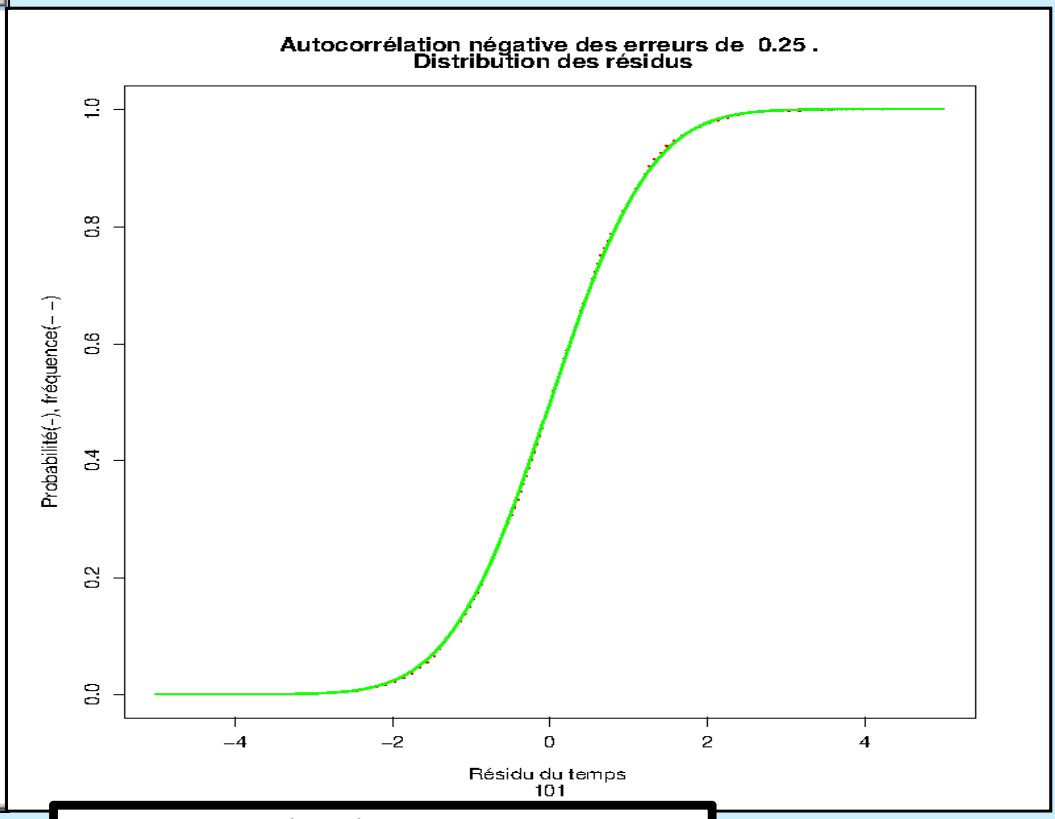
Comparaison Matlab-R

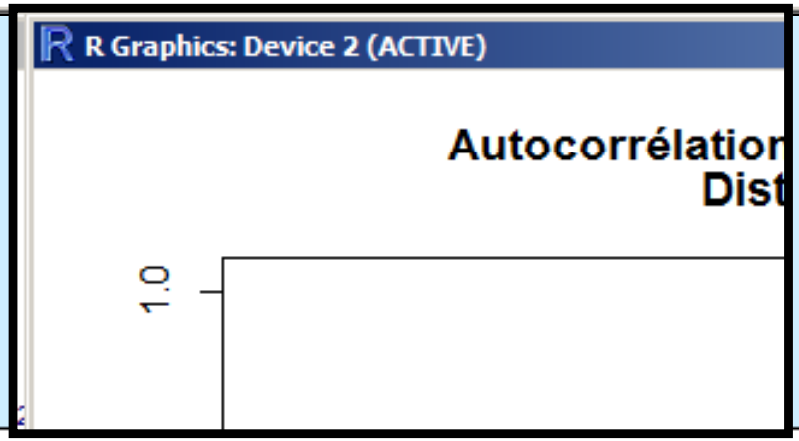


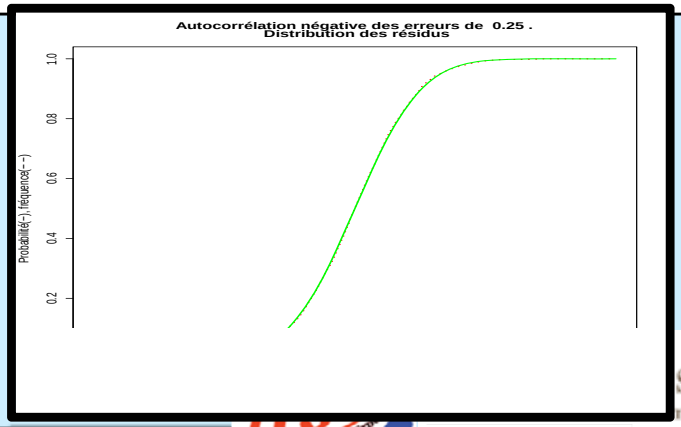


Comparaison sortie graphique R-fichier EPS









Conclusions

⇨Outil utile pour se rendre compte de l'absence d'une des conditions d'application

⇨En particulier la normalité peut être peu essentielle!

⇨On peut modifier l'outil :

- ✓inférence sur la moyenne par le test de Student,
- ✓comparaison de deux moyennes par le test de Student
- ✓analyse de variance à un facteur à effets fixes
- ✓régression multiple
- ✓distributions non symétriques
- ✓absence simultanée de plusieurs des conditions

Merci de votre attention

N’hésitez pas à faire part de vos suggestions

gmelard@ulb.ac.be

Outil:
http://homepages.ulb.ac.be/~gmelard/rech/condreg.zip