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

Guy Mélard[1]

[1] *Université libre de Bruxelles, ECARES CP114/4, Avenue Franklin Roosevelt 50, B-1050 Bruxelles, Belgique et ITSE sprl, Bruxelles, Belgique. gmelard@ulb.ac.be*

Résumé. En 2012, l'auteur [3] a présenté un exposé sous le titre "Enseigner la régression multiple sans mathématiques. Est-ce possible et est-ce souhaitable ?". Parmi les arguments figuraient une expérience sur les conditions d'application de la régression multiple qui mérite d'être mise en avant. Elle était déjà mentionnée dans [2, chapitre 7] et une vidéo est disponible sur le CD-Rom inclus. Cette vidéo est basée sur un script Matlab mais un des objectifs est de rendre la démarche plus accessible soit sous forme exécutable, soit sous forme de script R. Le script illustre la relative faible importance pratique de certaines des conditions d'application à l'aide d'une expérience de simulation de Monte-Carlo. On génère des données selon un modèle de régression linéaire simple dans les conditions théoriques du modèle linéaire général sauf qu'une des conditions d'application peut être violée : erreur sur la variable explicative, colinéarité (avec la constante), relation non linéaire, non normalité des erreurs (avec choix parmi quatre autres lois), présence de données aberrantes, hétéroscédasticité, autocorrélation positive des erreurs, autocorrélation négative des erreurs. Le but est de montrer que les conditions d'application ne sont pas toutes à mettre au même niveau et, en particulier, que la normalité des erreurs n'est pas nécessairement cruciale.

Mots-clés. Régression multiple, Modèle linéaire général, Conditions d'application.

Abstract. In 2012 the author [3] presented a talk on the subject "Teaching multiple regression without mathematics. Is it possible; is it desirable?". Among the arguments there was an experiment on the application conditions of multiple regression which is worth to be put in the light. It was already mentioned in [2, chapter 7] and a video is available on the included CD-Rom. That video is based on a Matlab script but one of the objectives is to make the approach more available, either under the form of an executable program, or under the form of a R script. Indeed, what is illustrated is the relatively small practical importance of some of the application conditions with the help of a Monte-Carlo simulation experiment. We generate data according to a simple linear regression model in the theoretical conditions of the general linear model except that one of the condition can be violated: error on the explanatory variable, multicollinearity (with the intercept), nonlinear relation, non normality of the errors (with a choice among four other laws), presence of outliers, heteroskedasticity, positive error autocorrelation, negative error autocorrelation. The aim is to show that the application conditions are not all to be put at the same level, and, in particular, that the normality of the errors in not necessarily crucial.

Keywords. Multiple regression, General linear model, Application conditions.

1 Introduction

Un grand nombre de méthodes statistiques ne sont valables, tout au moins pour des petits échantillons, que sous de conditions assez restrictives concernant le modèle sous-jacent. Il en est ainsi pour la régression linéaire simple ou multiple. On suppose que le modèle liant les variables explicatives à la variable dépendante est linéaire, avec des erreurs additives indépendantes, de moyenne nulle et de variance constante. De plus, les variables explicatives sont supposées mesurées

sans erreurs et être sans relation linéaire entre elles. On a évidemment raison d'insister sur ces conditions d'application qui permettent d'établir la distribution des estimateurs, au sens des moindres carrés, des coefficients de régression et notamment la distribution des statistiques de Student pour les coefficients de régression. Souvent, les lecteurs inattentifs interprètent d'ailleurs ces conditions en imposant que la distribution des observations de la variable dépendante soit normale, ce qui n'est pas du tout correct. Beaucoup d'auteurs insistent sur la vérification de ces conditions d'application, en particulier de la normalité des erreurs, suggérant ainsi d'appliquer des tests de normalité sur les résidus. Pourtant [1] a discuté cette supposition de normalité de manière théorique et a conclu en une certaine robustesse vis à vis de la non normalité, bien que cela dépende des valeurs prises par les variables explicatives. De même les économètres testent l'homoscédasticité et l'indépendance des erreurs (avec le test de Durbin-Watson, qui ne prend en compte que l'autocorrélation d'ordre 1). Entendons-nous bien, il est raisonnable de se demander si le modèle peut s'appliquer aux données mais tout tester n'est pas nécessairement une solution car les tests employés reposent aussi sur des conditions d'application. Il est plus important de se poser la question d'utilité du modèle. C'est un peu ce que nous avions préconisé dans [3].

Dans le cours en auto-apprentissage inclus dans [2], nous avions proposé un exercice qui consistait à vérifier l'effet des conditions d'application sur la régression linéaire simple. Dans la plupart des cas, on utilisait une variable explicative unique, prenant les valeurs 1, ..., n, où la valeur par défaut de n était 10. On procédait par simulation en générant un premier jeu de données que l'on traitait et qu'on montrait sous différents angles : diagramme de dispersion de la variable dépendante en fonction de la variable explicative et de même pour les valeurs prédites de la variable dépendante obtenues par la méthode des moindres carrés. Ensuite on générait un certain nombre N, par défaut $N = 1000$, et on examinait la distribution des erreurs, la distribution de la statistique de Student du coefficient de régression de la variable x et la distribution des erreurs de prévision pour $x = n + h$ (par défaut $h = 3$). Ceci était fait soit dans le respect des conditions d'application, soit en supposant qu'une seule d'entre elles n'était pas remplie, et ceci évidemment de manière bien précise.

Pratiquement, il s'agissait d'une vidéo montrant l'exécution d'un script sous Matlab. Le script lui-même était fourni mais n'était utilisable que par les possesseurs de Matlab. Le script utilisait un menu qui offrait le choix du traitement automatique ou manuel (permettant de changer n, N et h), puis le choix entre toutes les conditions d'application ou la violation d'une seule condition parmi les dix. La vidéo était réalisée pour le traitement automatique (donc avec $n = 10$, $N = 1000$ et $h = 3$).

2 L'outil et son utilisation

Comme indiqué ci-dessus, la vidéo n'illustre que le mode automatique donc avec peu de données et peu de simulations et l'outil est un script pour Matlab appelé `condreg.m` qui nécessite donc que l'utilisateur dispose de Matlab. En réalité, après quelques modifications, le script fonctionne dans octave (http://www.gnu.org/software/octave/), tout au moins la version anglaise `condrege` (car les caractères accentués des titres des graphiques produisent des erreurs dans la version française).

Nous avons donc investigué la possibilité de s'affranchir de ces contraintes. Une première manière a consisté à générer un programme exécutable à partir de la boîte à outil Compiler de Matlab. Nous avons ainsi généré un exécutable pour Windows appelé `condreg.exe`, voir Figure 1. En employant une version Linux de Matlab, nous avons de même pu produire un exécutable pour Linux. Compte tenu de la grande popularité de R, nous avons ensuite entrepris de traduire le script Matlab en R (même si les figures jointes viennent encore de la version pour Matlab).

Les figures 2 à 13 montrent les résultats de l'exécution de conreg quand $n = 100$, $N = 5000$ et $h = 10$ et ceci dans deux cas :

- le cas 1 où les données ont été générées quand toutes les conditions sont satisfaites ;

- le cas 5 où les données ont été générées quand toutes les conditions sont satisfaites sauf que les erreurs sont distribuées selon une loi de Cauchy au lieu d'une loi normale, mais de même étendue interquartile.

On voit successivement d'abord les trois premiers graphiques qui portent sur une première simulation avec les points représentés en rouge alors que les situations attendues sont montrées en vert :

- (figures 2 et 8) un diagramme de dispersion avec la variable dépendante en fonction de la variable explicative et de même pour les valeurs prédites de la variable dépendante obtenues par la méthode des moindres carrés : les points (en rouge) sont bien distribués autour de la droite de régression (en vert) avec toutefois des écarts importants dans le cas d'erreurs distribuées avec une loi de Cauchy ;
- (figures 3 et 9) un diagramme des résidus (en rouge) en fonction de la variable explicative; si la plupart des résidus sont dans l'intervalle $\pm 1{,}96\,\hat{v}$ (en vert) dans le cas normal, où $\hat{v}$ est l'écart-type résiduel estimé, il n'en est rien dans le cas de la loi de Cauchy ;
- (figures 4 et 10) un diagramme des résidus (en rouge) en fonction de l'indice de l'observation, c'est-à-dire x lui-même dans ces deux cas; ces graphiques sont ici identiques aux précédents sauf que les points sont reliés par des segments de droite ; dans le cas d'erreurs sur la variable ou d'erreur de spécification avec un modèle quadratique en x, ces diagrammes seraient différents des précédents.

On voit ensuite les trois graphiques suivants qui portent sur l'ensemble des simulations représentées en rouge alors que les situations espérées sont montrées en vert :

- (figures 5 et 11) la fonction de distribution empirique des erreurs pour l'observation $n + h$ (en rouge) comparée à la fonction de la distribution normale de moyenne 0 et d'écart-type 1 (en vert) ; il y a évidemment accord pour la figure 5 mais pas du tout pour la figure 11 où la loi de Cauchy est employée ;
- (figures 6 et 12) la fonction de distribution empirique de la statistique de Student associée au coefficient de régression pour x (en rouge), comparée à la fonction de distribution d'une loi de Student à $n - 2$ degrés de liberté (en vert) ; on remarque qu'il y a accord dans les deux cas, bien que dans le second cas une distribution de Student ne soit pas du tout justifiée car une des suppositions fondamentales, la normalité des erreurs, n'est pas vérifiée ; il semble donc que les tests et intervalles de confiance pour le coefficient de régression ne seront pas aussi faux qu'on pourrait croire ;
- (figures 7 et 13) la fonction de distribution empirique des erreurs de prévisions basées sur le modèle de régression estimé par la méthode des moindres carrés et l'écart-type des résidus estimés (en rouge), comparée à la fonction de la distribution normale de moyenne 0 et d'écart-type approprié (en vert) ; il y a un assez large accord pour la figure 7 qui aurait encore été meilleur si on avait employé la loi exacte, une loi de Student à $n - 2$ degrés de liberté, mais pas du tout pour la figure 13 où l'on devine que les intervalles de prévision à 90 % basés sur un distribution normale auront en fait une probabilité de couverture bien différente.

Nous avons pris cet exemple du cas 5 qui explique bien le titre de l'exposé. Nous aurions pu prendre le cas de la présence de données aberrantes ou le cas d'autocorrélation positive des erreurs (illustré toutefois dans la figure 1) qui montrent mieux la nécessité des conditions d'application. Comme les économètres le savent bien, l'autocorrélation positive des erreurs implique que les tests sur le coefficient de régression sont extrêmement biaisés au point d'être inutilisables.

3 Extensions et conclusions

L'outil est évidemment utile dans le contexte qui a été décrit. Il faut tenir compte dans les

conclusions de [1] qui indiquent une certaine robustesse vis à vis de la normalité mais en précisant que le degré de robustesse est dépendant de la distribution des valeurs prises par les variables explicatives. L'avantage de notre outil est de permettre de visualiser ces propriétés. Pour cela on peut changer la variable explicative. On peut adapter l'outil à l'inférence sur la moyenne par le test de Student, à la comparaison de deux moyennes par le test de Student, à l'analyse de variance à un facteur à effets fixes, à la régression multiple. Actuellement seules des distributions symétriques sont considérées pour les erreurs mais cela peut être modifié. Bien sûr, comme une seule des suppositions est violée à la fois, cela n'implique pas que les conclusions restent valables quand plusieurs des conditions ne sont pas remplies simultanément. Ceci pourrait être effectué.

Bibliographie

[1] Box, G.E.P. et Watson, G.S. (1962) Robustness to non-normality of regression tests, *Biometrika*, 49, 93–106.

[2] Mélard, G. (2007), *Méthodes de prévision à court terme*, 2e édition, Editions de l'Université de Bruxelles, Bruxelles et Ellipses Edition Marketing à Paris (avec CD-ROM).

[3] Mélard, G. (2012), Enseigner la régression multiple sans mathématiques. Est-ce possible et est-ce souhaitable ?", Communication au CFIES'2012, Société française de Statistique, Angers 12-14 septembre. http://homepages.ulb.ac.be/~gmelard/rech/Angers2012.pdf

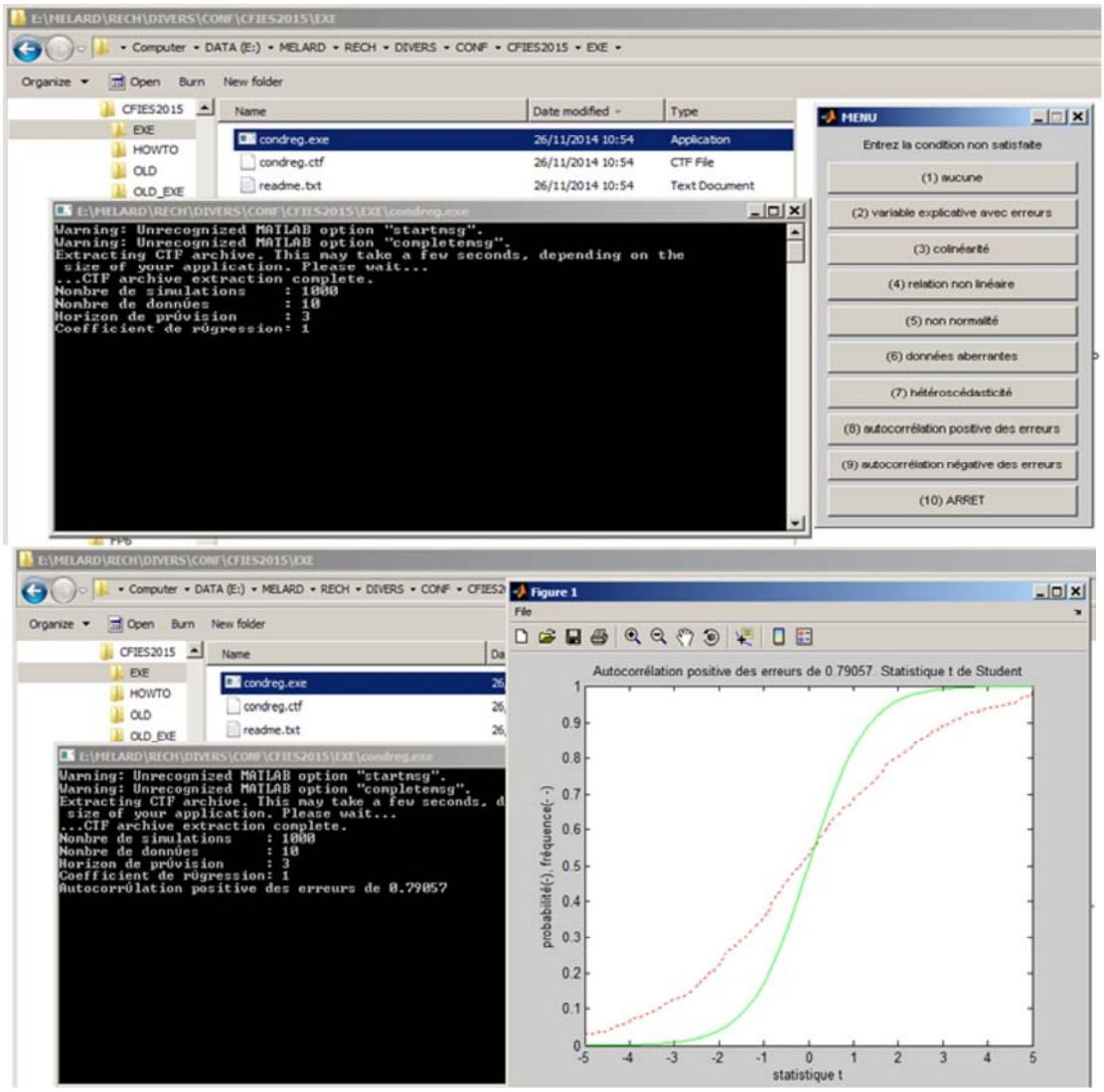


FIGURE 1 – Exécution du programme exécutable `condreg.exe`

Figure 2

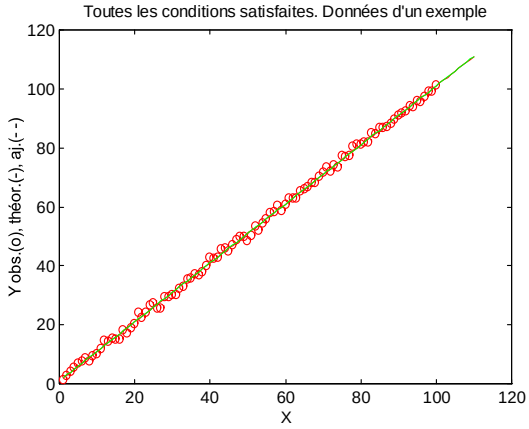


Figure 3

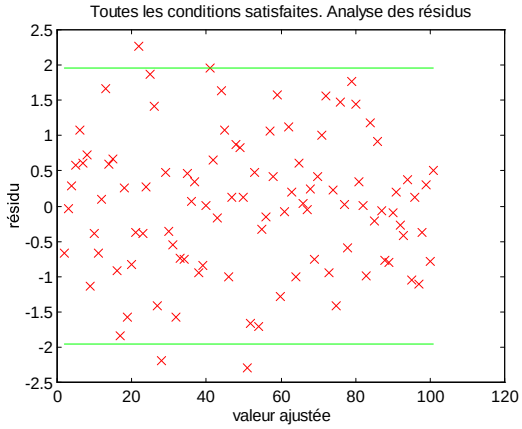


Figure 4

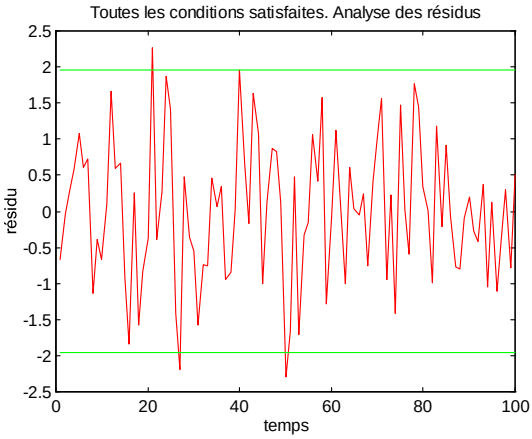


Figure 5

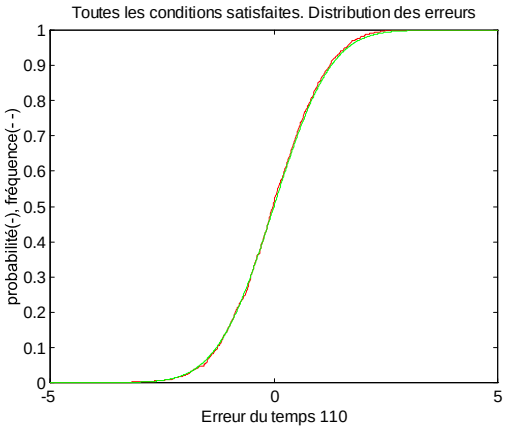


Figure 6

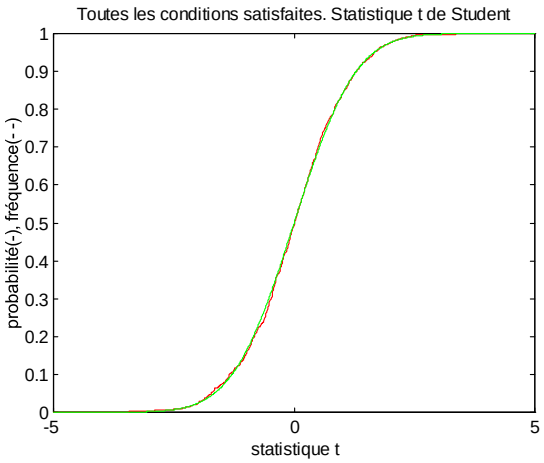
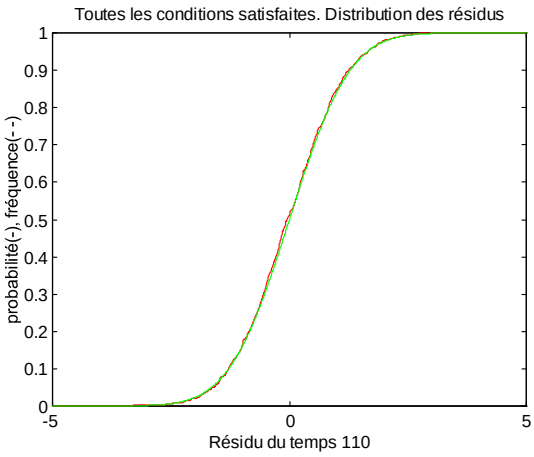


Figure 7



FIGURES 2-7 – Cas 1. Toutes les conditions satisfaites

Figure 8

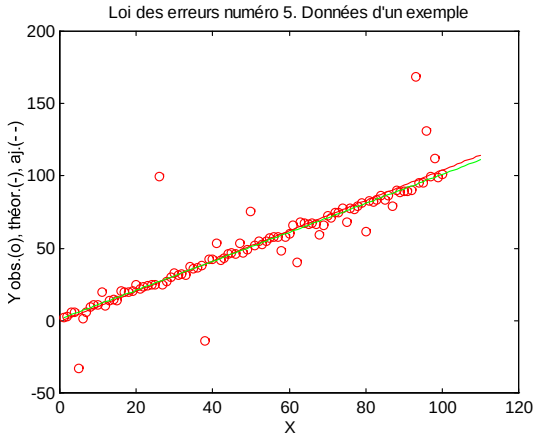


Figure 9

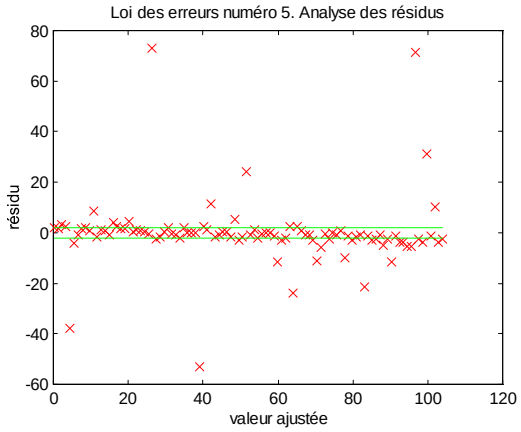


Figure 10

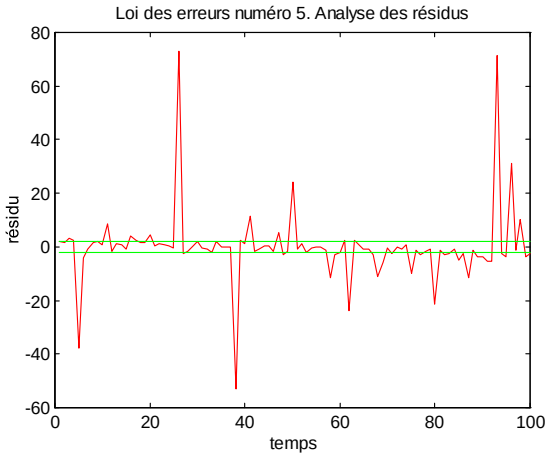


Figure 11

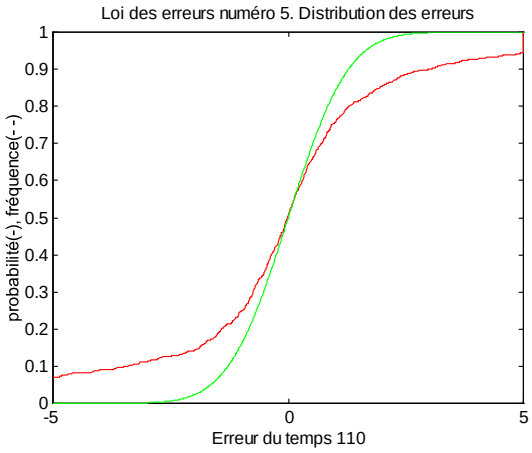


Figure 12

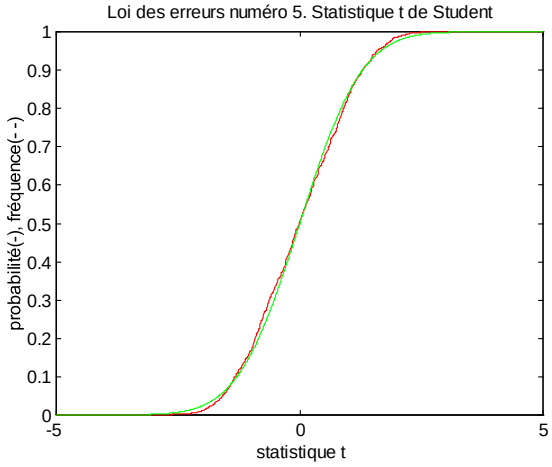
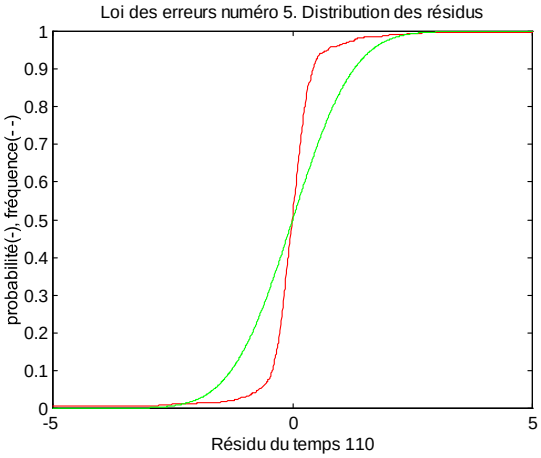


Figure 13



FIGURES 8-13 – Cas 5 : loi des erreurs numéro 5 (Cauchy)