NORMALITE DES RESIDUS

1. Introduction

Une grande partie de l'inférence statistique (test, intervalle de confiance, etc.) repose sur l'hypothèse de distribution normale $N(0,\sigma^2)$ du terme d'erreur. Vérifier cette hypothèse semble incontournable pour obtenir des résultats exacts.

Nous disposons des erreurs observées e_i, les résidus de la régression, pour évaluer les caractéristiques des erreurs théoriques ε_i.

Il semble néanmoins que les tests usuels restent valables, pour peu que l'on ait suffisamment d'observations (n>50). Nous pouvons vérifier l'hypothèse de normalité en utilisant des méthodes graphiques ou des tests. Nous citons la droite de Henry (Graphique Q-Q plot) pour la première approche et les tests de Jarque-Bera et Shapiro-Wilk

2. Graphique Q-Q plot

Il ne s'agit pas d'un test au sens statistique du terme. Le graphique Q-Q plot (quantile-quantile plot) est un graphique "nuage de points" qui vise à confronter les quantiles de la distribution empirique et les quantiles d'une

distribution théorique normale, de moyenne et d'écart type estimés sur les valeurs observées. Si la distribution est compatible avec la loi normale, les points forment une droite. Dans la littérature francophone, ce dispositif est appelé Droite de Henry.

a) Trier les résidus e_i de manière croissante, ce sont les quantiles observés.

b) Produire la fonction de répartition empirique, lissée en accord avec la loi normale F_i

c) Calculer les quantiles théoriques normalisées z_i en utilisant la fonction inverse de la loi normale centrée réduite.

d) En déduire les quantiles théoriques dé-normalisées $e_i^* = \hat{\sigma}_{e_i} z_i$. Si la distribution empirique cadre parfaitement avec la loi normale, les points (e_i, e_i^*) devraient être alignés sur la diagonale principale. Ici, $\hat{\sigma}_{e_i} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2}$

Bien souvent, on peut se contenter de ce diagnostic. Nous réagissons uniquement si l'écart avec la normalité est très marqué. Néanmoins, nous

pouvons consolider les conclusions en s'appuyant sur les tests de normalité. Nous citons quelques tests

3 Tests sur les résidus

3.1. Test de Jarque-Bera

Le coefficient de Skewness **(Skew)** ou coefficient d'asymétrie est une mesure de l'asymétrie de la distribution de la série autour de sa moyenne. Le Skewness est calculé de la manière suivante

$$Skew = \frac{\left[n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^3 \right]}{\left[n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right]^{3/2}} = \frac{\mu_3}{\sigma^3} \sim \mathrm{N}(0, \sqrt{6/n})$$

- Le coefficient de Kurtosis **(Kur)** ou coefficient d'aplatissement est une mesure de l'aplatissement de la distribution de la série. Le Kurtosis est calculé de la manière suivante

$$Kur = \frac{n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^4}{\left[n^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \right]^2} = \frac{\mu_4}{\mu_2^2} \sim \mathrm{N}(3, \sqrt{24/n})$$

Le Test de Jarque-Bera permet de savoir si la série est normalement distribuée. Il est basé sur la statistique JB qui est calculé par

$$JB = \frac{n-k}{6}\left[Skew^2 + \frac{(Kur-3)^2}{4}\right]$$

où k représente le nombre de coefficients estimés du modèle.

Sous l'hypothèse nulle de normalité, on a asymptotiquement $JB \sim \chi^2_{(2)}$

Règle de décision

Si $JB > \chi^2_{(2)}(\alpha)$ alors l'hypothèse nulle de normalité est rejetée au seuil de α

Si la p-value $< \alpha$ alors on rejette H_0 de normalité des résidus au seuil de α.

3.2. Test de Shapiro-Wilk

Très populaire, le test de Shapiro-Wilk est basé sur la statistique *W*. En comparaison des autres tests, il est particulièrement puissant pour les petits effectifs$(n \leq 50)$. La statistique du test s'écrit :

$$W = \frac{\left[\sum_{i=1}^{[n/2]} a_i \left(x_{(n-i+1)} - x_{(i)}\right)\right]^2}{\sum_i \left(x_{(i)} - \bar{x}\right)^2}$$

Où

_ $x(i)$ correspond à la série des données triées ;

_ $[n/2]$ est la partie entière du rapport $n/2$;

_ ai sont des constantes générées à partir de la moyenne et de la matrice de variance co-variance des quantiles d'un échantillon de taille n suivant la loi normale. Ces constantes sont fournies dans des tables spécifiques.

La statistique W peut donc être interprétée comme le coefficient de détermination (le carré du coefficient de corrélation) entre la série des quantiles générées à partir de la loi normale et les quantiles empiriques obtenues à partir des données.

Plus W est élevé, plus la compatibilité avec la loi normale est crédible.

La région critique, rejet de la normalité, s'écrit :

$R:C: : W < Wcrit$

Les valeurs seuils $Wcrit$ pour différents risques α et effectifs n sont lues dans la table de Shapiro-Wilk.

Remarque : l'implémentation dans le logiciel R a été évaluée (fonction shapiro.test(...).

Exercice d'application :

Nous reprenons les données de l'exercice 1 du chapitre corrélation.

L'équation de régression : $y_i = a_0 + a_2 x_{i,1} + a_2 x_{i,2} + \varepsilon_i$

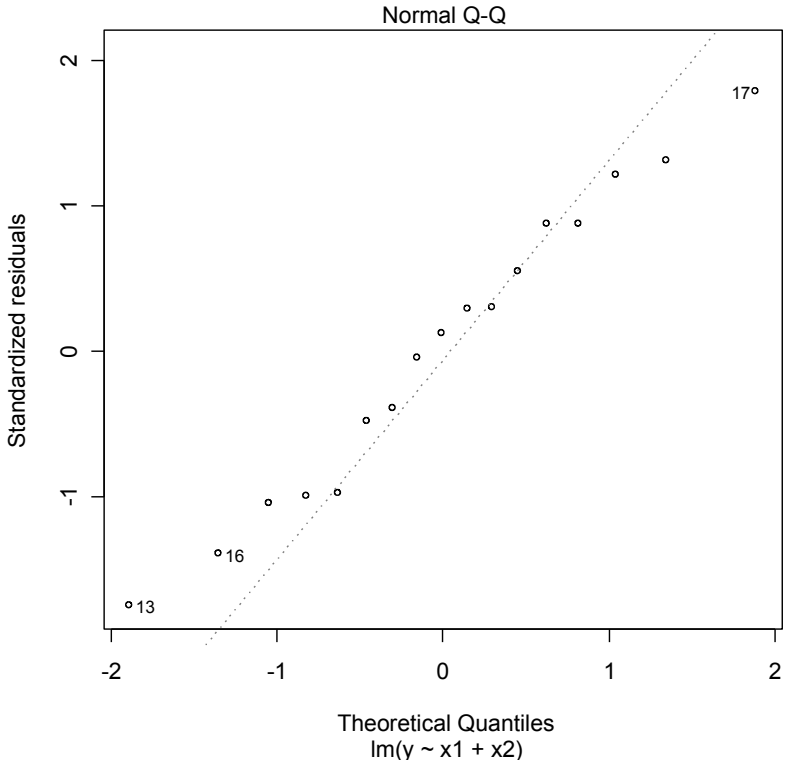
Où y est la consommation en textile, x_1 le prix du textile et x_2 le revenu par habitant.

```
y=c(99.2,99.0,100,111.6,122.2,117.6,121.1,136,154.2,153.6,158.5,140.6,
136.2,168,154.3,149,165.5)
x1=c(101,100.1,100,90.6,86.5,89.7,90.6,82.8,70.1,65.4,61.3,62.5,63.6,52
.6,59.7,59.5,61.3)
```

x2=c(96.7,98.1,100,104.9,104.9,109.5,110.8,112.3,109.3,105.3,101.7,95.4,96.4,97.6,102.4,101.6,103.8)

1- Utiliser le graphique Q-Q plot pour vérifier l'hypothèse de normalité
2- Tester la normalité des résidus en utilisant les tests de Jarque-Bera et de Shapiro-Wilk

1- Graphique Q-Q plot des résidus



MULTICOLINEARITE

1. Introduction

Le problème qui préoccupe l'économiste est de trouver des variables explicatives qui maximisent leur coefficient de corrélation avec la variable à expliquer tout en étant les moins corrélées entre elles

Nous présentons la notion de la corrélation partielle qui permet de déterminer l'apport relatif de chaque variable explicative à l'explication de la série endogène

2. Corrélation partielle

Le coefficient de corrélation partielle mesure la liaison entre deux variables lorsque l'influence d'une ou des autres variables est retirée

- Exemples :

$r_{yx_1x_2}$: - coefficient de corrélation partielle du premier ordre

- mesure la variance de y expliquée par la variable x_1 seule

(L'influence de x_2 étant retirée)

$r_{yx_1x_2x_3}$: - coefficient de corrélation partielle du deuxième ordre

- mesure la variance de y expliquée par la variable x_1

Seule (l'influence de x_2 et x_3 étant retirée)

- **Calcul du coefficient de corrélation partiel**

Soit

$$y_j = a_0 + a_1x_{1j} + a_2x_{2j} + ... + a_kx_{kj} + \varepsilon_j, \qquad j = 1,..., n$$

Deux méthodes

a) $r_{yx_j(autres\ variables)} = r(e_1, e_2)$

$r(e_1, e_2)$ est le coefficient de corrélation simple des résidus e_1 et e_2 des régressions de y sur x_{-j} et de x_j sur x_{-j} respectivement

(x_{-j} sont les variables explicatives autres que x_j)

b) $r_{yx_j(autres\ variables)} = \dfrac{t_i^2}{t_i^2 + (n-k-1)}$

$$t_i = \frac{\hat{a}_i}{s(\hat{a}_i)}$$

Cette relation n'est vérifiée que pour un coefficient de corrélation partiel d'ordre $k-1$

3. Relation entre coefficients de corrélation simple, partielle et multiple

- Cas d'un modèle de régression simple

$R^2_{y,x} = r^2_{yx}$ ou $1 - R^2_{y,x} = 1 - r^2_{yx}$

- Cas d'un modèle de régression multiple

$k = 2$

$1 - R^2_{yx_1x_2} = (1 - r^2_{yx_1})(1 - r^2_{yx_2x_1})$

$k = 3$

$1 - R^2_{yx_1x_2x_3} = (1 - r^2_{yx_1})(1 - r^2_{yx_2x_1})(1 - r^2_{yx_3x_1x_2})$

4. Multicolinéarité : Conséquences et détection

Dans la pratique, lorsque l'économiste modélise des phénomènes des phénomènes économiques, les séries explicatives sont toujours plus ou moins liées entres elles.

a) Conséquences:

- augmentation de la variance estimée de certains coefficients lorsque la colinéarité entre les variables explicatives augmente

- instabilité des estimateurs des coefficients: des faibles fluctuations concernant les données entraînent des fortes variations des valeurs estimées

- multicolinéarité parfaite : estimation des coefficients est impossible

b) Test de multicolinéarité:

Soit le modèle

$$y_j = a_0 + a_1 x_{1j} + a_2 x_{2j} + ... + a_k x_{kj} + \varepsilon_j, \qquad j = 1,..., n$$

Test de Klein

Ce test est basé sur la comparaison du coefficient de détermination R^2 et les coefficients de corrélation simple $r^2_{x_i x_j}$ entre les variables explicatives x_i et x_j pour $i \neq j$

Si $r^2_{x_i x_j} < R^2$ il y a présomption de multicolinéarité

Test de Farrar et Glauber

Les hypothèses de ce test

$$\begin{cases} H_0 : D = 1 \quad séries \quad \text{exp}\, licatives \quad orthogonal \ es \ (\text{cov}(x_i, x_j) = 0) \\ H_1 : D < 1 \quad séries \quad \text{exp}\, licatives \quad dépendante \ s \end{cases}$$

La règle de décision

$$d(\chi_c^2) = \begin{cases} a_0 : \chi_c^2 < \chi_K^2(\alpha) \\ a_1 : \chi_c^2 \ \square \ \chi_K^2(\alpha) \end{cases}$$

$$\chi_c^2 = -\left[n - 1 - \frac{2K+5}{6} \right] Ln(D) \ ; \qquad K = \frac{1}{2} k(k+1)$$

$$D = \begin{bmatrix} 1 & r_{x_1x_2} & r_{x_1x3} & . & r_{x_1xk} \\ r_{x_2x_1} & 1 & r_{x_2x_3} & . & r_{x_2x_k} \\ . & . & . & . & . \\ . & . & . & . & . \\ r_{x_kx_1} & r_{x_kx_2} & 0 & r_{x_kx_{k-1}} & 1 \end{bmatrix}$$

c) Solutions à la multicolinéarité

La méthode efficace consiste, lors de la spécification de modèle, à éliminer les variables explicatives susceptibles de représenter les mêmes phénomènes (corrélés entre elles). Pour ceci, l'économètre est souvent confronté au choix de plusieurs variables explicatives

Le problème devient : on dispose de k variables explicatives pour expliquer y : comment choisir convenablement un groupe de r variables ($r \leq k$) parmi les k variables ? Ce choix doit répondre à deux objectifs contradictoires :

- r doit être petit pour que le modèle soit facilement interprétable
- r doit être assez grand pour que l'ajustement de y soit correct

Critères du R^2 et du R^2 corrigé

Le critère de maximisation du $R^2 = 1 - \frac{SCR}{SCT}$ présente l'inconvénient de ne pas arbitrer entre la perte de degrés de liberté du modèle et l'ajustement qui en résulte. Il faut pénaliser un choix de r trop grand. Une façon de procéder est la suivante : Utiliser le coefficient de détermination corrigé

$$\bar{R}^2 = 1 - \frac{SCR\ /(n-r-1)}{SCT\ /(n-1)}$$

$\bar{R}^2 = 1 - \frac{n-1}{n-r-1}(1-R^2)$

Pour un modèle à r variables explicatives et retenir le modèle qui maximise $\bar{R}^2$. Ce critère favorise les modèles comportant un grand nombre de variables.

Critères AIC et BIC

On préfère utiliser les critères de Akaike (AIC) ou de Schwarz (BIC) afin de comparer des modèles. Le modèle qui maximise la fonction : --

$AIC(R_r)$ Akaike (Akaike Information Criterion) , $AIC(R_r) = Ln(\frac{SCR_r}{n}) + \frac{2r}{n}$,

R_r : Régression de y en r variables explicatives

- ou $BIC(R_r)$ de Schwarz (Schwarz Criterion ou Bayesian Information Criterion),

$BIC(R_r) = Ln(\frac{SCR_r}{n}) + \frac{rLn(n)}{n}$

Il existe plusieurs méthodes qui nous permettent de retenir le meilleur modèle,composé des variables, qui sont :

- les plus corrélés avec la variable à expliquer
- les moins corrélées entre elles

a) Toutes les régressions possibles

Nous estimons les $(2^k - 1)$ régressions possibles et nous retenons le modèle dont le critère AIC ou SC est minimum et tous les variables explicatives sont significatives

b) L'élimination progressive (Backward Elimination)

Si la première équation peut être estimée alors nous éliminons les variables explicatives qui ne sont pas significatives et nous réestimons l'équation après chaque élimination jusqu'à l'obtention un modèle avec des variables dont les t de Student sont au dessus du seuil critique

c) La sélection progressive (Forward Régression)

- Sélectionner la variable x_i qui maximise $r^2_{yx_i}$
- Retenir la variable x_j qui maximise $r^2_{yx_jx_i}$ pour $j \neq i$
- S'arrêter lorsque les t de Student des variables explicatives sont inférieurs au seuil critique

d) La régression pas à pas (Stepwise Regression)

Nous reprenons la procédure précédente sauf qu'à chaque étape nous éliminons du modèle les variables explicatives dont les t de Student sont au dessus du seuil critique

e) La régression par étage (Stagewise Regression)

- Sélectionner la variable x_i qui maximise $r^2_{yx_i}$
- Calcul du résidu $e_1 = y - \hat{a}_0 - \hat{a}_1 x_i$
- Retenir la variable x_j qui maximise $r^2_{e_1x_j}$

- Calcul du résidu $e_2 = y - a_0 - a_1 x_1 - a_3 x_3$
- Retenir la variable x_l qui maximise $r^2_{e_2 x_l}$
- S'arrêter lorsque les coefficients de corrélation ne sont plus significativement différents de 0

Application.

Exercice

Un économiste cherche à expliquer la variable y à l'aide de quatre variables explicatives x_1, x_2, x_3, x_4. Il désire auparavant tester une éventuelle multicolinéarité entre ces quatre séries.

y=c(8.4,9.6,10.4,11.4,12.2,14.2,15.8,17.9,19.3,20.8)

x1=c(82.9,88,99.9,105.3,117.7,131,148.2,161.8,174.2,184.7)

x2=c(17.1,21.3,25.1,29.0,34.0,40.0,44.0,49.0,51.0,53.0)

x3=c(92,93,96,94,100,101,105,112,112,112)

x4=c(94,96,97,97,100 ,101,104,109,111,111)

1) appliquer le test de Klein
2) Effectuer le test de Farrar-Glauber
3) En utilisant l'élimination progressive et la régression pas à pas, sélectionner la ou les variables explicatives candidates celles dont le pouvoir explicatif est le plus important.

AUTOCORRELATION DES ERREURS

1 . Introduction

$$y_j = a_0 + a_1 x_{1j} + a_2 x_{2j} + ... + a_k x_{kj} + \varepsilon_j, \qquad j = 1,..., n$$

$$Y = Xa + \varepsilon$$

Parmi les hypothèses

$$V(\varepsilon_j) = \sigma^2$$

$$E(\varepsilon_j \varepsilon_{j'}) = 0 \quad : \quad j \neq j'$$

Nous supposons que $E(\varepsilon_j \varepsilon_{j'}) \neq 0$ pour $j \neq j'$

- La matrice des variances et covariances de l'estimateur de a n'est plus diagonale

Dans ce cas, l'estimateur de a

- est sans biais

- n'est pas à variance minimale

L'autocorrélation des erreurs : essentiellement dans les modèles en série temporelle

Les causes de l'autocorrelation

- Absence d'une variable explicative importante

- Une mauvaise spécification du modèle

Notre étude : trois étapes

- Comment détecter l' autocorrélation des erreurs **?**

- Quelles sont les conséquences ?

- Comment corriger ?

2 . Détection de l'aucorrélation des erreurs

La détection d'une éventuelle dépendance des erreurs s'effectue à

partir de l'analyse des résidus (seuls sont connus)

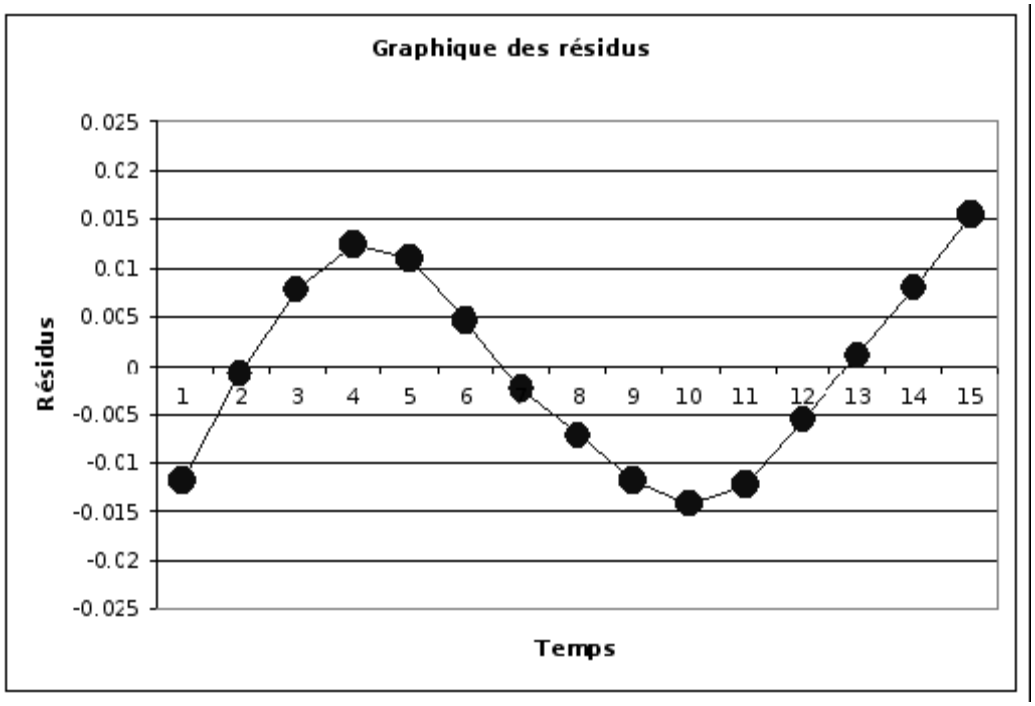
a) Analyse graphique des résidus

L'analyse graphique des résidus permet le plus souvent de détecter un processus de reproduction des erreurs

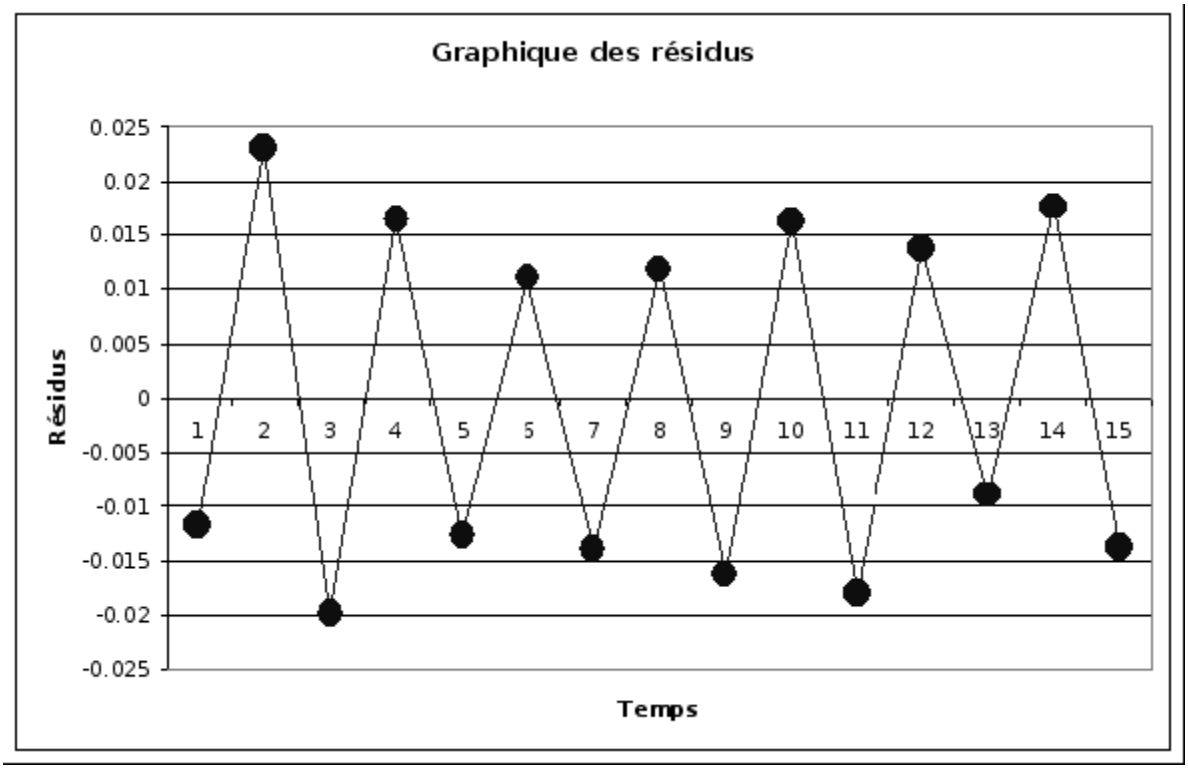
Pour ceci : visualiser les points $(\hat{y}_j, e_j)$

Autocorrélation positive : les résidus sont pendant plusieurs

périodes successives soit positifs, soit négatifs



Autocorrélation négative: les résidus sont alternées



Dans la majeure partie : cette autocorrélation est difficile à détecter

b) Test de Durbin Watson

- permet de détecter une autocorrélation d'ordre 1

$$\varepsilon_j = \rho\varepsilon_{j-1} + \nu_t \quad \text{avec} \quad \nu_t \sim \mathrm{N}(0,\sigma)$$

Test $\begin{cases} H_0 : \rho = 0 \quad absence\ d'autocorrélation \\ H_1 : \rho \neq 0 \quad présence\ d'autocorrélation \end{cases}$

La statistique utilisée : $DW = \dfrac{\sum_{j=2}^{n} (e_j - e_{j-1})^2}{\sum_{j=1}^{n} e_j^{\,2}}$ (ou d)

Où e_j sont les résidus

$$DW = \frac{\sum_{j=2}^{n} e_j^{\,2} + \sum_{j=2}^{n} e_{j-1}^{\,2} - 2\sum_{j=2}^{n} e_j e_{j-1}}{\sum_{j=1}^{n} e_j^{\,2}}$$

Pour n grand : $DW = 2(1-r)$ (1)

où $r = \dfrac{\sum_{j=2}^{n} e_j e_{j-1}}{\sum_{j=1}^{n} e_j{}^2}$

r est le coefficient de corrélation de la régression par les MCO de e_t en e_{t-1}

D'après (1)

- DW varie entre 0 et 4
- $DW < 2 \Rightarrow Autocorrél \quad ation \quad positive$
- $DW > 2 \Rightarrow Autocorrél \quad ation \quad négative$
- $DW = 2 \Rightarrow Pas \quad d'autocorrél \quad ation$

En raison de la dépendance de DW avec la matrice X des valeurs critiques de DW ne peuvent être calculées pour tous les cas possibles Durbin et Watson ont tabulé les valeurs critiques de DW, d_L et d_U, au seuil de 5% en fonction de la taille de l'échantillon et le nombre de variables explicatives

Selon la valeur de DW empirique nous pouvons conclure

$DW < d_L \Rightarrow rejeter \ H_0 \ Autocorrél \ ation \ positive$

$DW > 4 - d_L \Rightarrow rejeter \ H_0 \ Autocorrél \ ation \ négative$

$d_U < DW < 4 - d_U \Rightarrow accepter \ H_0$

$d_L < DW < d_U \ ou \ 4 - d_U < DW < 4 - d_L \Rightarrow \ zone \ de \ doute$

Conditions d'utilisation :

- Présence d'un terme constant dans le modèle
- la variable y ne figure parmi les variables explicatives
- n > 15

3. Estimation en cas d'autocorrélation des erreurs

Si nous retenons l'hypothèse d'une autocorrélation des erreurs d'ordre 1, le modèle linéaire s'écrit

$$y_j = a_0 + a_1 x_{1j} + a_2 x_{2j} + \ldots + a_k x_{kj} + \varepsilon_j, \qquad j = 1,\ldots, n$$

$$Y = Xa + \varepsilon$$

$$\varepsilon_j = \rho \varepsilon_{j-1} + \nu_t \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

avec $\nu_t \sim \mathrm{N}(0, \sigma)$; $|\rho| < 1$

$$E(\nu_j \nu_{j'}) = 0 \quad : \quad j \neq j'$$

En utilisant (2)

$$\varepsilon_j = \nu_j + \rho \nu_{j-1} + \rho^2 \nu_{j-2} + \ldots + \rho^h \nu_{j-h} + \ldots$$

Ce processus tend vers 0 puisque $|\rho| < 1$

$$E(\varepsilon_j) = 0 \; ; \quad V(\varepsilon_j) = \frac{\sigma^2}{1-\rho^2}$$

En utilisant les propriétés de ν_j

$$E(\varepsilon_j \varepsilon_{j-1}) = \rho\sigma^2$$

$$E(\varepsilon_j \varepsilon_{j-2}) = \rho^2\sigma^2$$

$$E(\varepsilon_j \varepsilon_{j-i}) = \rho^i\sigma^2$$

La matrice des variances et covariances de l'erreur

$$\Omega = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & . & \rho^{n-1} \\ \rho & 1 & . & \rho^{n-2} \\ . & . & . & . \\ \rho^{n-1} & \rho^{n-2} & . & 1 \end{bmatrix}$$

$$\Omega^{-1} = \frac{1-\rho^2}{\sigma^2}\begin{bmatrix} 1 & -\rho & 0 & . & 0 \\ -\rho & 1+\rho^2 & -\rho & . & . \\ 0 & -\rho & . & . & 0 \\ . & . & . & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & -\rho & 1 \end{bmatrix}$$

L'estimateur des Moindres Carrés Généralisés (MCG) est

$$\hat{a} = \left(X^t \Omega^{-1} X\right)^{-1} X^t \Omega^{-1} Y$$

Dans le cas où ρ est inconnu

- estimer ρ par régression de e_j sur e_{j-1}

$$\hat{\rho} = \frac{\sum_{j=2}^{n} e_j e_{j-1}}{\sum_{j=1}^{n} e_j^{\,2}} \quad \textbf{ou} \quad \hat{\rho} = 1 - \frac{D}{2}$$

- estimer les coefficients a_i par régression sur les quasi-différences

$$y_j - \rho y_{j-1} = b_0 + a_1(x_{1j} - \rho x_{1j-1}) + \ldots + a_k(x_{kj} - \rho x_{kj-1}) + v_j$$

Les paramètres estimés par MCO sont alors

$$\hat{a}_1, \hat{a}_2, \ldots \hat{a}_k \quad \text{et} \quad \hat{a}_0 = \frac{\hat{b}_0}{1 - \hat{\rho}}$$

Ceci nous donne l'estimateur des moindres carrés quasi-généralisés MCQG .

Remarque : Il existe une autre méthode (Méthode de Cohrane-Orcutt) basée sur une estimation itérative des coefficients de régression et du paramètre ρ (Voir le livre de Regis Bourbonnais « économétrie »)

4. Le test des séquences (Test de Wald-Wolfowitz) : De façon plus générale, si on veut tester

H_0 : $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ sont des v.a. indépendantes identiquement distribuées

Soit n_+ (resp n_-) le nombre de résidus positifs (resp le nombre de résidus négatifs) dans la série des résidus et $n = n_+ + n_-$.

Une séquence de la suite des signes est une suite maximale de signes successifs constants et R est le nombre de séquences

Par exemple pour la suite de $n = 18$:

++----+-++++-+++--

n_+=10 ; n_-=8 ; $R = 8$

Sous H_0 , la loi de R dépend uniquement de n_+et n_- et elle est donnée par des tables statistiques pour $n_+ \leq 20$et $n_- \leq 20$

Si R est faible il y'a association et si R est élevé la liaison est négative

Les tables statistiques donnent $r_-(\alpha)$resp$(r_+(\alpha))$ le plus grand k vérifiant $P(R \leq k|H_0) \leq \alpha$ resp (le plus petit k vérifiant $P(R \geq k|H_0) \leq \alpha$

Pour n grand R est approximativement normale de moyenne $\mu_R = \frac{2n_+ n_-}{n} + 1$et de variance $\sigma_R^2 = \frac{(\mu_R - 1)(\mu_R - 2)}{n-1}$

Nous rejetons H_0 si $|Z| > z_{\alpha/2}$

Où $Z = \frac{R - \mu_R}{\sigma_R}$ $\sim N(0,1)$ et $P[Z > z_{\alpha/2}] = \alpha/2$

5. Applications

Exercice 1

L'objectif est de prédire la consommation de textile à partir du revenu par tête des personnes et du prix. Nous disposons d'observations sur 17 années de 1923.

y=c(99.2,99.0,100,111.6,122.2,117.6,121.1,136,154.2,153.6,158.5,140.6,136.2,168,154.3,149,165.5)
x1=c(101,100.1,100,90.6,86.5,89.7,90.6,82.8,70.1,65.4,61.3,62.5,63.6,52.6,59.7,59.5,61.3)
x2=c(96.7,98.1,100,104.9,104.9,109.5,110.8,112.3,109.3,105.3,101.7,95.4,96.4,97.6,102.4,101.6,103.8)

L'équation de régression : $y_i = a_0 + a_2 x_{i,1} + a_2 x_{i,2} + \varepsilon_i$

Où y est la consommation en textile, x_1 le prix du textile et x_2 le revenu par habitant. Nous voulons tester l'indépendance des erreurs :

a) Effectuer le test de Durbin-Watson
b) Effectuer le test des séquences

a) Test de Durbin-Watson

- A l'aide du logiciel SPSS nous obtenons
- Coefficients:

$\hat{a}_0 = 13{,}70$, $\hat{a}_1 = -1.3$ et $\hat{a}_2 = 1.0$

- $DW = 2.019$
- Pour un test bilatéral à 10% , les valeurs critiques données par la table de Durbi-Watson pour $n = 17$ et $k = 2$: $d_L = 1.02$ et $d_U = 1.54$
- Nous constatons que $d_U < DW < 4 - d_U$

Accepter H_0

Remarque : Nous pouvons obtenir la valeur DW en utilisant R

```
L=lm(y~x1+x2)
n=17
e=residuals(L)
e1=e[1 :n-1]
e2=e[2 :n]
DW=sum((e2-e1)^2)/sum(e^2); DW
```

b) Test des séquences

n_+=9 ; n_-=8 ;

La région de rejet bilatérale au niveau 5% donnée par la table du test des séquences : $W = \{R \leq 5\} \cup \{R \geq 14\}$

Observant R=8 , nous acceptons H_0 et donc nous retenons la non corrélation . Nous concluons que les résidus sont indépendants, ils sont générés par un processus aléatoire.

Exercice 2

On dispose des données relatives à la consommation et au revenu disponible des ménages dans un pays donné de 1960 à 1984

```
C=c(176.5,186.9,199.4,211.7,222.1,230.6,242.3,254.2,266.1,283.5,293.8
,312.6,330.1,346.2,363.6,369.0,389.3,401.8,415.9,430.6,440.7,455.1,462
.2,466.6,468.5)
```

Re=c(208.2,218.7,239.0,251.5,262.8,275.2,287.4,302.4,315.6,330.8,352.9,375.6,397.0,418.6,440.3,453.2,465.8,481.9,504.0,513.7,518.0,539.3,548.1,544.9,542.5)

On suppose que la relation qui lie la consommation au revenu est linéaire et stable, ce qui correspond au modèle économétrique suivant

$C_t = a_0 + a_1 R_t + \varepsilon_t \quad , \quad t = 1,..., 25$

Nous proposons de déceler une éventuelle autocorrélation d'ordre 1 des erreurs, pour cela on demande :

a) d' estimer les coefficients du modèle ;
b) d'effectuer l'analyse graphique des résidus ;
c) d'effectuer le test des séquences
d) de calculer la statistique de Durbin et Watson et d'effectuer le test ;
e) d'en corriger dans le cas d'autocorrélation des erreurs : estimation de ρ, regression sur les quasi-différences

Solution

a) Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.10851	3.75488	-0.295	0.77
R	0.84262	0.00921	91.485	<2e-16 ***

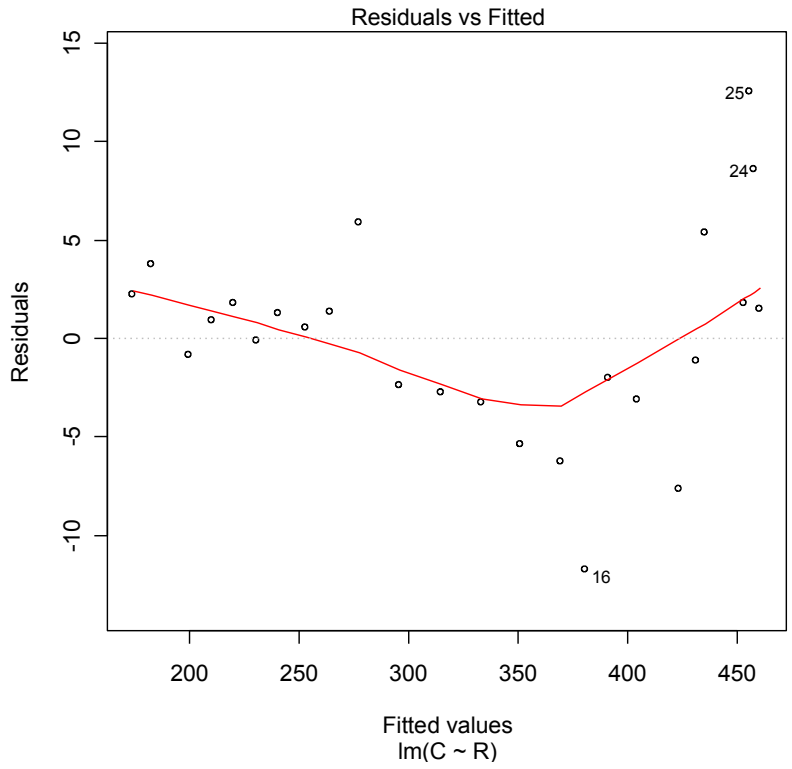
Residual standard error: 5.237 on 23 degrees of freedom

Multiple R-Squared: 0.9973, Adjusted R-squared: 0.9971

F-statistic: 8370 on 1 and 23 DF, p-value: < 2.2e-16

$\hat{a}_0 = -1.10$; $\hat{a}_1 = 0.84$

b)



c) $n_+ = 13; n_- = 12;$

La région de rejet est **:** $W = R \leq 8 \vee R \geq 19$

En observant R=7, le caractère iid des variables ε_i est rejeté

d) $DW = \dfrac{\sum_{j=2}^{n} (e_j - e_{j-1})^2}{\sum_{j=1}^{n} e_j^{\,2}}$

```
L=lm(C~Re)
n=25
e=residuals(L)
e1=e[1 :n-1]
e2=e[2 :n]
DW=sum((e2-e1)^2)/sum(e^2); DW
```

DW=0.695

Pour α=5% , les valeurs critiques données par la table de Durbi-Watson pour $n = 25$ et $k = 1$: $d_L = 1.29$ et $d_U = 1.45$

- Nous constatons que $DW < d_L \Rightarrow rejeter\ H_0\ : Autocorrél ation\ \ positive$

d) ρ est estimé par $\hat{\rho} = 1 - \frac{DW}{2} = 0.6525$

- estimer les coefficients a_0 et a_1 par régression sur les quasi-différences

$$C_j - \hat{\rho} C_{j-1} = b_0 + a_1 (R_j - \hat{\rho} R_{j-1}) + v_j \ , \ j = 2,..., n$$

Les paramètres estimés par MCO sont alors

```
n=25
C1=C[1 :n-1]
C2=C[2 :n]
Re1=Re[1 :n-1]
Re2=Re[2 :n]
rau=1-(DW/2)
CT=C2-(rau*C1)
ReT=Re2-(rau*Re1)
LT=lm(CT~ReT)
```

summary(LT)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.37689	3.53880	-0.389	0.701
ReT	0.85100	0.02321	36.660	<2e-16 ***

Residual standard error: 4.223 on 22 degrees of freedom
Multiple R-squared: 0.9839, Adjusted R-squared: 0.9832
F-statistic: 1344 on 1 and 22 DF, p-value: < 2.2e-16

a0=-1.376/(1-rau)

C'est-à-dire ; $\hat{b}_0 = -1.3$; $\hat{a}_1 = 0.85$

Ceci nous donne l'estimateur des moindres carrés quasi-généralisés MCQG .

$\hat{a}_0 = -3.95$ et $\hat{a}_1 = 0.85$