

NOTES DE STATISTIQUE ET D'INFORMATIQUE

2002/1

CONDITIONS D'APPLICATION ET
TRANSFORMATIONS DE VARIABLES
EN RÉGRESSION LINÉAIRE

R. PALM et A.F. IEMMA

Faculté universitaire des
Sciences agronomiques

Centre de Recherches
agronomiques

GEMBLOUX

(Belgique)

CONDITIONS D'APPLICATION ET TRANSFORMATIONS DE VARIABLES EN RÉGRESSION LINÉAIRE

R. PALM* et A.F. IEMMA†

RÉSUMÉ

Cette note décrit d'abord des méthodes permettant de vérifier les conditions d'application de la régression (adéquation du modèle, normalité, homoscedasticité et indépendance des résidus) et présente ensuite les transformations de variables pouvant être utilisées en cas de non-respect de ces conditions d'application.

SUMMARY

This note first reviews methods used for checking the assumptions in linear regression (lack of fit, normality, homoscedasticity and independence of the residuals) and then describes variables transformations as remedial action when the assumptions are not fulfilled.

1. INTRODUCTION

Les méthodes classiques d'inférence statistique en régression multiple supposent qu'un ensemble de conditions d'application concernant le modèle et les données sont remplies. Le non-respect de ces conditions peut mettre en question les résultats de l'inférence statistique, suite à la modification des distributions d'échantillonnage des estimateurs qui peut rendre incorrect le calcul des intervalles de confiance ou le résultat des tests d'hypothèses.

Les conséquences varient évidemment avec l'importance de l'écart par rapport aux conditions théoriques. En pratique, il est donc nécessaire de vérifier si les conditions d'application sont remplies, au moins de façon approximative, avant de réaliser l'inférence.

*Chargé de cours associé à la Faculté universitaire des Sciences agronomiques de Gembloux.

†Senior bio-statisticien du Département de Recherche et Développement de Glaxo-SmithKline Biologicals, Rixensart) et Maître de conférences à la Faculté universitaire des Sciences agronomiques de Gembloux durant l'année académique 1990-1991.

Après un rappel de ces conditions d'application, nous passerons en revue diverses méthodes permettant de mettre en évidence le non-respect de celles-ci, successivement du point de vue de l'adéquation de la relation, de la normalité, de l'homoscédasticité et de l'indépendance des résidus (paragraphe 2).

Nous examinerons ensuite comment on peut tenter de résoudre le problème du non-respect des conditions par des transformations de variables, qui portent soit sur la variable à expliquer, soit sur les variables explicatives, soit à la fois sur la variable à expliquer et sur les variables explicatives (paragraphe 3).

Nous analyserons alors un exemple concret (paragraphe 4), avant de tirer quelques conclusions (paragraphe 5).

2. VÉRIFICATION DES CONDITIONS D'APPLICATION

2.1. Modèle et conditions d'application

On considère le modèle théorique suivant :

$$y = \mathbf{x}\boldsymbol{\beta} + \varepsilon,$$

dans lequel y est la variable dépendante, \mathbf{x} est le vecteur relatif aux variables explicatives, $\boldsymbol{\beta}$ est le vecteur des coefficients de régression théoriques et ε est le résidu. Pour tenir compte d'un éventuel terme indépendant, il suffit de considérer qu'une variable explicative est constante et égale à l'unité. Si p désigne le nombre de variables explicatives, à l'exclusion de l'éventuelle variable artificielle ajoutée pour prendre en compte le terme indépendant, les vecteurs \mathbf{x} et $\boldsymbol{\beta}$ sont respectivement de dimensions $1 \times p'$ et $p' \times 1$, p' étant égal à $p + 1$ si le modèle possède un terme indépendant et $p' = p$ sinon. Le résidu ε est une variable aléatoire, de même que y . Par contre, \mathbf{x} est un vecteur non aléatoire, connu sans erreur, et $\boldsymbol{\beta}$ est un vecteur de paramètres fixés mais inconnus.

Appliqué à un individu donné, noté i , le modèle s'écrit :

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i.$$

Pour les méthodes classiques d'inférence, les conditions d'application peuvent être résumées de la manière suivante : les résidus ε_i sont des réalisations indépendantes de variables aléatoires normales, de moyenne nulle et de variance constante et égale à σ^2 . Ces conditions correspondent aux conditions d'adéquation de la relation, de normalité, homoscédasticité et absence d'autocorrélation des résidus.

Le modèle de régression présenté ci-dessus est le modèle tout à fait classique. D'autres situations peuvent être considérées. En particulier, on pourrait s'intéresser au cas où les \mathbf{x}_i sont aléatoires ou sujets à erreurs. Ces situations particulières ne seront pas envisagées ici.

Soit un échantillon aléatoire et simple, conditionnellement à \mathbf{x} , de n individus provenant de la population décrite par le modèle ci-dessus. Pour ces n individus,

on a le modèle théorique :

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

\mathbf{y} étant le vecteur des n réalisations de la variable à expliquer, \mathbf{X} étant la matrice, de dimensions $n \times p'$, des valeurs observées des variables explicatives, $\boldsymbol{\beta}$ le vecteur des p' paramètres et $\boldsymbol{\varepsilon}$ le vecteur des n résidus théoriques et inconnus, de matrice de variances et covariances égale à $\sigma^2 \mathbf{I}$, \mathbf{I} étant la matrice identité, de dimensions $n \times n$.

Sous les conditions énoncées ci-dessus, un estimateur non biaisé et de variance minimum du vecteur $\boldsymbol{\beta}$ est donné par la méthode des moindres carrés :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y},$$

pour autant que la matrice $\mathbf{X}' \mathbf{X}$ soit non singulière. Une estimation non biaisée de la variance résiduelle, $\hat{\sigma}^2$ est donnée par :

$$\hat{\sigma}^2 = \mathbf{e}' \mathbf{e} / (n - p'),$$

le vecteur \mathbf{e} étant le vecteur des résidus observés :

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X} \hat{\boldsymbol{\beta}}.$$

Le vecteur $\hat{\mathbf{y}}$ peut encore s'écrire :

$$\hat{\mathbf{y}} = \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} = \mathbf{H} \mathbf{y},$$

\mathbf{H} étant la matrice de projection, de dimensions $n \times n$.

Les éléments diagonaux de cette matrice \mathbf{H} sont égaux à :

$$h_{ii} = \mathbf{x}_i (\mathbf{X}' \mathbf{X})^{-1} \mathbf{x}_i',$$

\mathbf{x}_i étant la $i^{\text{ème}}$ ligne de la matrice \mathbf{X} .

La matrice de variances et covariances du vecteur $\hat{\boldsymbol{\beta}}$ est donnée par :

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 (\mathbf{X}' \mathbf{X})^{-1},$$

et la matrice de variances et covariances des résidus observés est égale à :

$$\hat{\mathbf{V}}(\mathbf{e}) = \hat{\sigma}^2 (\mathbf{I} - \mathbf{H}).$$

On constate que, contrairement aux résidus théoriques ε_i , les résidus estimés e_i n'ont pas une variance constante et sont corrélés. En particulier, le résidu relatif à l'individu i a pour variance estimée :

$$\hat{v}(e_i) = \hat{\sigma}^2 (1 - h_{ii}).$$

Pour éliminer cette inégalité des variances des résidus observés, on définit des résidus standardisés, de variance constante et égale à l'unité :

$$r_i = e_i / \hat{\sigma} \sqrt{1 - h_{ii}}.$$

Les résidus peuvent encore être standardisés d'une manière légèrement différente :

$$t_i = e_i / \hat{\sigma}_{(i)} \sqrt{(1 - h_{ii})},$$

$\hat{\sigma}_{(i)}$ étant l'écart-type résiduel estimé à partir de l'équation de régression calculée après l'élimination de l'observation i .

Les différents résidus définis ci-dessus sont donnés, éventuellement en option, par les logiciels SAS et Minitab, notamment [SAS, 1989; X, 2000]. Ils interviennent dans la définition des paramètres permettant de quantifier l'influence sur la régression de chacun des individus de l'échantillon. Ces paramètres ont été décrits dans une note antérieure [PALM, 1988]. Des informations détaillées à ce sujet sont données par COOK et WEISBERG [1982].

Ces résidus sont utilisés également dans les paragraphes suivants pour la vérification des conditions d'application. A ce propos, il faut remarquer qu'on se trouve dans une situation un peu particulière, dans la mesure où on souhaite vérifier des conditions d'application relatives aux ε_i , alors qu'on ne dispose pas de valeurs de ces ε_i . Ainsi, pour vérifier la normalité d'une population, par exemple, on dispose habituellement d'un échantillon aléatoire et simple d'observations prélevées dans cette population. Dans le cas de la régression, par contre, on n'a pas un tel échantillon, car les ε_i sont non observables. On a uniquement n résidus observés e_i dont les propriétés sont différentes de celles des ε_i . En particulier, nous venons de signaler qu'ils sont corrélés et de variances inégales, même si les ε_i sont indépendants et de variance constante.

2.2. Adéquation de la relation

Dans le cas de la régression simple, l'adéquation du modèle peut être appréciée graphiquement sur le diagramme de dispersion de la variable à expliquer en fonction de la variable explicative. Dans le cas de modèles plus complexes (régression polynomiale et régression multiple), on peut établir un diagramme de dispersion des résidus observés (e_i , r_i ou t_i) en fonction d'une variable explicative particulière ou en fonction des valeurs estimées de y . Ce diagramme peut être complété par la surimposition des valeurs lissées par des moyennes mobiles par exemple. Ces valeurs lissées doivent fluctuer autour de zéro si le modèle est adéquat. Il faut noter qu'on porte en abscisse les valeurs estimées de y et non les valeurs observées, car ces dernières sont corrélées aux résidus e_i , la corrélation r_{ye} étant d'autant plus grande que le coefficient de détermination multiple R^2 est faible. On a, en effet, la relation suivante :

$$r_{ye} = \sqrt{1 - R^2}.$$

D'autres représentations sont encore proposées dans la littérature afin de vérifier la linéarité de l'effet d'une variable particulière quand on dispose de plusieurs variables explicatives. Ainsi, pour visualiser l'effet d'une variable explicative x_j sur la variable y après l'élimination de l'effet des $p - 1$ autres variables explicatives, on peut calculer, d'une part la régression de y en fonction de ces $p - 1$ variables explicatives et, d'autre part la régression de x_j en fonction de ces

mêmes $p-1$ variables explicatives. Les résidus de la première équation sont alors mis en graphique en fonction des résidus de la seconde équation. Le diagramme de dispersion ainsi obtenu, appelé *diagramme des résidus partiels*¹ donne les informations relatives à la nature de la liaison de y et de x_j , après l'élimination de l'effet des $p-1$ autres variables explicatives. La pente de la droite de régression passant par l'origine qui peut être calculée à partir de ce nuage de points est égale au coefficient de régression partielle de la variable x_j dans le modèle à p variables et la corrélation des deux séries de résidus est le coefficient de corrélation partielle de y et x_j , après l'élimination des $p-1$ autres variables explicatives. Un tel graphique s'interprète globalement comme un diagramme de dispersion en régression simple.

Une version différente est obtenue en portant en abscisse les valeurs de x_j et en ordonnée les valeurs $[LARSEN \text{ et } MCCLEARY, 1972]$:

$$e_i + x_{ij} \hat{\beta}_j \quad \text{avec} \quad e_i = y_i - \mathbf{x}_i \hat{\beta}.$$

Ce graphique est dénommé *graphique de la composante plus résidu*². On notera que les valeurs portées en ordonnée ne sont ni les résidus de y en l'absence de la variable explicative x_j , comme dans le graphique précédent, ni les résidus de y sur l'ensemble des variables, comme les e_i . La pente de la droite de régression passant pas l'origine relative à ce nuage de points est aussi égale au coefficient de régression partielle $\hat{\beta}_j$, mais la variance de ce coefficient de régression simple est plus faible que la variance de $\hat{\beta}_j$, la différence entre les deux variances étant d'autant plus importante que la variable x_j est fortement liée aux autres variables explicatives [RYAN, 1997]. Les résidus par rapport à la droite de régression dans ce second graphique sont également identiques à ceux du graphique précédent, mais l'allure générale des deux graphiques est différente.

La nécessité d'une éventuelle transformation de x_j est mieux mise en évidence dans le cas du second graphique, car l'abscisse est la variable explicative elle-même. Par contre, le premier graphique met mieux en évidence les données influentes [CHATTERJEE et PRICE, 1991; WEISBERG, 1985].

Des extensions de ces types de graphiques et des informations concernant les propriétés des différentes représentations graphiques sont données, notamment, par BERK [1998], BERK et BOOTH [1995], COOK [1993, 1994, 1996], COOK et WEISBERG [1994, 1997].

Lorsqu'on dispose de plusieurs observations y_{ij} pour un même vecteur \mathbf{x}_i , on peut effectuer le test classique de linéarité, en décomposant la somme des carrés des écarts résiduelle obtenue à l'issue de la régression en deux composantes : la variation résiduelle pure et la composante liée à la non-linéarité.

Ce test nécessite des répétitions afin de déterminer l'erreur pure. Il est donc en général limité au cas où les données proviennent d'une expérience planifiée. En l'absence de répétitions, on peut regrouper des observations pour lesquelles les valeurs de \mathbf{x} sont à peu près identiques et utiliser les variations entre les réponses dans les groupes pour calculer l'erreur pure et en déduire l'erreur liée

1. En anglais : *partial residual plot, added variable plot*.

2. En anglais : *component plus residual plot, partial residual plot*.

à l'inadéquation du modèle. Il ne s'agit cependant pas d'un test rigoureux, les résultats dépendant de la manière dont les observations ont été regroupées.

La non-adéquation d'un modèle peut également être appréciée par la comparaison de ce modèle à des modèles alternatifs. Ainsi, dans le cas du modèle linéaire simple, la relation :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

peut être comparée aux deux relations suivantes :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x^2 \quad \text{et} \quad \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2.$$

Si le coefficient de détermination est plus grand pour le deuxième ou le troisième modèle ou si le coefficient β_2 du troisième modèle est significatif, on peut admettre l'existence d'une relation non linéaire entre y et x . Cette approche est utilisée pour juger de la pertinence d'une éventuelle transformation proposée comme remède à la non-linéarité (paragraphe 3.3).

Le logiciel Minitab propose une combinaison de tests, applicables en l'absence de répétitions et ayant pour objectif d'identifier la ou les variables responsables de l'inadéquation du modèle. Considérons d'abord le cas de la régression linéaire simple, pour laquelle deux tests sont réalisés. Pour le premier test, on considère les variables instrumentales z_0 et z_1 , telles que :

$$z_{i0} = 0 \quad \text{et} \quad z_{i1} = 0 \quad \text{si} \quad x_i \leq \bar{x}$$

et

$$z_{i0} = 1 \quad \text{et} \quad z_{i1} = x_i \quad \text{si} \quad x_i > \bar{x}.$$

On calcule les sommes de carrés des écarts résiduelles $SCE_r(1)$ et $SCE_r(2)$ associées aux deux modèles suivants :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{1}$$

et

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\alpha}_0 z_0 + \hat{\alpha}_1 z_1 \tag{2}.$$

La différence entre les deux sommes des carrés d'écarts :

$$SCE_r(1) - SCE_r(2) = R(\alpha_0, \alpha_1 | \beta_0, \beta_1)$$

correspond à la réduction de la somme des carrés des écarts liée à l'ajout des deux variables instrumentales dans le modèle initial et possède deux degrés de liberté. Cette composante est alors comparée à la composante résiduelle du modèle complet, $SCE_r(2)$, par l'intermédiaire du test F . Concrètement, cela revient à comparer une équation de régression simple unique (modèle (1)) à deux équations de régression simple, l'une étant ajustée aux données telles que $x_i \leq \bar{x}$ et l'autre aux données telles que $x_i > \bar{x}$.

Le deuxième test repose également sur la comparaison de deux modèles. Le premier modèle est le modèle (1) ajusté à l'ensemble des n données et le second modèle est le modèle (1) ajusté aux m données les plus centrales, pour lesquelles $h_{ii} \leq 2, 2/n$. La différence entre les deux sommes des carrés des écarts résiduelles,

qui est une mesure de l'inadéquation du modèle, possède $n - m$ degrés de liberté. On réalise alors le test F , en prenant comme base de comparaison la somme des carrés des écarts résiduelle relative au modèle obtenu sur les données les plus centrales.

Lorsqu'on dispose de p variables explicatives, le premier test décrit ci-dessus est appliqué en considérant successivement chacune des variables explicatives. Pour la variable x_j , on définit :

$$z_{ik} = 0 \quad (k = 0, \dots, p) \quad \text{si} \quad x_{ij} \leq \bar{x}_j$$

$$z_{i0} = 1 \quad \text{et} \quad z_{ik} = x_{ik} \quad (k = 1, \dots, p) \quad \text{si} \quad x_{ij} > \bar{x}_j,$$

et on considère le modèle complet :

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_i + \dots + \hat{\beta}_p x_p + \hat{\alpha}_0 z_0 + \hat{\alpha}_1 z_1 + \dots + \hat{\alpha}_p z_p,$$

qui correspond à l'ajustement de deux hyperplans indépendants, l'un pour les observations telles que $x_{ij} \leq \bar{x}_j$ et l'autre pour les observations telles que $x_{ij} > \bar{x}_j$. La sommes des carrés des écarts liée à l'ensemble des variables instrumentales est divisée en deux parties :

$$R(\alpha_0, \dots, \alpha_p \mid \beta_0, \dots, \beta_p) = R(\alpha_0, \alpha_j \mid \beta_0, \dots, \beta_p)$$

$$+ R(\alpha_1, \dots, \alpha_{j-1}, \alpha_{j+1}, \dots, \alpha_p \mid \beta_0, \dots, \beta_p, \alpha_0, \alpha_j).$$

La première somme des carrés des écarts, avec deux degrés de liberté, représente la courbure en x_j et le second terme, à $p - 1$ degrés de liberté, représente l'interaction entre x_j et les autres variables explicatives. Ces deux sources de variation sont testées en prenant la variation résiduelle du modèle complet comme base de comparaison.

Quant au deuxième test présenté pour la régression simple, il peut être généralisé sans difficulté au cas multivarié, les observations considérées comme centrales étant celles pour lesquelles $h_{ii} < (1, 1) (p + 1)/n$.

La procédure implique donc au total la réalisation de deux tests pour la régression simple et de $2p + 1$ tests pour la régression multiple à p variables explicatives. Pour maintenir le niveau de signification à α pour l'ensemble des tests, chaque test est réalisé en utilisant un niveau de signification α' égal à $\alpha/2$, pour la régression simple ou à $\alpha/(2p + 1)$, pour la régression multiple. Minitab fixe α à 0,1 et donne les probabilités ajustées associées aux différents tests, c'est-à-dire les probabilités multipliées par 2 ou par $(2p + 1)$. Enfin, les tests ne sont pas réalisés si le modèle initial ne présente pas d'ordonnée à l'origine. Les auteurs signalent que la procédure est expérimentale et doit être utilisée avec prudence [X, 1994].

2.3. Normalité des résidus

Comme signalé au paragraphe 2.1, la normalité des résidus théoriques, ε_i , est vérifiée à partir des résidus observés, e_i , qui n'ont pas les mêmes propriétés que les ε_i . En effet, les e_i sont corrélés et de variances inégales, même si les ε_i sont de variance constante.

L'inégalité des variances peut facilement être éliminée en considérant les résidus standardisés, r_i , comme le suggèrent DRAPER et SMITH [1998]. WEISBERG [1985] estime cependant qu'il n'est pas évident que l'utilisation de r_i ou de t_i soit préférable à l'utilisation de e_i . Pour les tests basés sur la corrélation des résidus et des scores normaux qui sera présenté ci-dessous, PFAFFENBERGER et DIELMAN [1991] ont montré que les e_i et les r_i donnent des résultats comparables mais que l'utilisation des t_i doit être déconseillée.

En plus de la corrélation et de l'inégalité des variances des e_i , se pose le problème connu sous l'appellation de supernormalité des résidus qui fait que les e_i peuvent avoir une distribution proche de la normale, même lorsque les ε_i ne sont pas normaux. Cette supernormalité s'explique si on exprime les résidus e_i en fonction des ε_i , par la relation suivante [WEISBERG, 1985] :

$$e_i = \varepsilon_i - \sum_{j=1}^n h_{ij} \varepsilon_j.$$

Le résidu e_i est donc égal à ε_i dont on soustrait une somme pondérée de tous les ε_j . Si le nombre de degrés de liberté de la variance résiduelle est faible et si certains éléments h_{ij} sont grands, le deuxième terme du membre de droite peut être plus important que le terme ε_i dans la détermination de la distribution des e_i . Et, en vertu du théorème central limite, cette somme aura une distribution approximativement normale, même lorsque les ε_i sont non normaux. Par conséquent, les tests de normalité, appliqués aux résidus observés ne peuvent qu'être approximatifs, du moins si l'effectif est faible [WEISBERG, 1985]. Pour des échantillons de grande taille, le terme ε_i domine et on peut s'attendre à ce que les tests appliqués aux e_i donnent des résultats comparables aux tests qui seraient appliqués directement aux ε_i , si ceux-ci étaient observables.

Pour vérifier la normalité des e_i , l'utilisateur dispose de méthodes graphiques et de tests statistiques.

Parmi les méthodes graphiques, on peut citer la comparaison de l'histogramme normé et de la fonction de densité de probabilité normale, la comparaison du polygone de fréquences relatives cumulées et de la fonction de répartition de la variable normale, l'établissement du diagramme de dispersion des proportions observées et des probabilités théoriques³ ou du diagramme de dispersion des quantiles observés et des quantiles théoriques⁴, appelé encore diagramme des scores normaux⁵. Des informations concernant ces graphiques sont données, notamment dans DAGNELIE [1998], SAS [1995] et SIEVERS [1986].

3. En anglais : *p-p plot, probability-probability plot.*

4. En anglais : *q-q plot, quantile-quantile plot.*

5. En anglais : *normal probability plot.*

Pour tenir compte du phénomène de supernormalité, ATKINSON [1981] propose de déterminer les limites de confiance du diagramme des scores normaux par simulation. La méthode consiste à simuler 19 vecteurs de n nombres aléatoires provenant d'une distribution normale réduite. Soit \mathbf{u}_k ($k = 1, \dots, 19$), ces vecteurs. On calcule ensuite les régressions de \mathbf{u}_k en fonction de \mathbf{X} et on enregistre les vecteurs des résidus, notés \mathbf{v}_k . Les éléments de chacun de ces vecteurs sont classés par ordre croissant. Parmi les 19 valeurs situées en rang 1 après ce classement, on retient la plus petite et la plus grande uniquement. On procède de la même manière pour les valeurs situées aux rangs 2, 3, \dots , n . Les valeurs maximum et minimum ainsi retenues sont des estimations des pourcentiles 5 et 95 de la distribution d'échantillonnage du $i^{\text{ème}}$ rang des résidus standardisés. Ce sont donc les limites de confiance inférieure et supérieure, au niveau $\alpha = 0, 10$, des statistiques d'ordre. Elles sont portées sur le graphique donnant les résidus standardisés en fonction des scores normaux et, si l'hypothèse de normalité est vérifiée, environ 90 % des résidus devraient se trouver dans les limites ainsi simulées. Un nombre de simulations supérieur à 19 ou le choix d'autres pourcentiles peut évidemment être envisagé. On pourrait par exemple simuler 100 échantillons et retenir le pourcentile 5 et 95, ou 2,5 et 97,5 de la distribution d'échantillonnage simulée des résidus de rang 1, de rang 2, etc. Obtenue à partir d'un plus grand nombre de simulations, l'enveloppe serait moins liée aux variations aléatoires.

Comme le signale RYAN [1997], l'enveloppe simulée doit être considérée comme un outil permettant une meilleure appréciation d'une éventuelle non-normalité, plutôt que comme un test statistique. En particulier, lorsque l'hypothèse de normalité est vraie, la probabilité qu'un résidu se situe hors des limites peut être bien plus grande que 10 %, surtout si n est grand. Inversement, une dissymétrie même importante de la distribution des résidus, qui se traduira par une courbure dans le diagramme des scores normaux, n'implique pas nécessairement que les points se trouvent en dehors de l'enveloppe simulée.

Parmi les tests classiques de normalité, on peut citer le test de SHAPIRO et WILK [1965], considéré comme l'un des plus performants [ROYSTON, 1982, 1988] et le test de RYAN et JOINER [1976] qui est basé sur la corrélation des résidus et des scores normaux et qui est équivalent au test précédent [X, 1994].

L'utilisation conjointe d'une méthode graphique et d'un test statistique permet de vérifier visuellement si le caractère non normal de la distribution des résidus est lié ou non à la présence d'une donnée ou d'un petit nombre de données aberrantes, au sein d'un ensemble de données présentant globalement une distribution normale. Dans de telles situations, il peut être utile d'éliminer ces données avant la réalisation du test de normalité.

La décision d'éliminer ces données peut éventuellement être prise à l'issue d'un test statistique de détection de résidus aberrants. Un tel test peut se faire à partir des résidus standardisés t_i qui, dans les conditions énoncées au paragraphe 2.1, possèdent une distribution de STUDENT à $n - p' - 1$ degrés de liberté, p' étant le nombre de paramètres du modèle. Le niveau de signification du test sera égal à α si l'utilisateur suspecte *a priori*, c'est-à-dire avant d'avoir examiné les t_i , une donnée particulière d'être anormale, par exemple sur la base de sa

connaissance du problème. En l'absence de cette identification *a priori*, ce qui est le plus souvent le cas en pratique, lorsqu'on s'intéresse aux résidus t_i les plus importantes en valeur absolue, le niveau de signification sera fixé à α/n , le test étant en fait appliqué n fois. Des tables donnant les valeurs de $t_{1-\alpha/2n}$ en fonction du nombre d'observations et du nombre de variables sont données par WEISBERG [1985]. Le test de détection de données aberrantes ne peut cependant être réalisé, de manière rigoureuse, que si, dans l'ensemble les résidus sont normaux.

2.4. Homoscédasticité

Nous considérons d'abord le cas d'une seule variable explicative et nous verrons ensuite la généralisation au cas de plusieurs variables explicatives.

L'inégalité des variances conditionnelles peut être visualisée par le diagramme de régression de y en x ou des résidus de la régression en fonction de x . Dans ce dernier cas, il est d'ailleurs préférable de considérer les résidus standardisés, r_i , plutôt que les résidus non standardisés car, comme nous l'avons signalé au paragraphe 2.1, ces derniers ne sont pas de variance constante, même sous l'hypothèse d'homoscédasticité des ε_i .

Pour mieux mettre en évidence l'inégalité des variances conditionnelles, on peut remplacer les résidus par une fonction de ces résidus. Parmi les transformations proposées, on peut citer la valeur absolue des résidus, le carré des résidus, la racine cubique du carré des résidus. Une discussion de ces diverses solutions est donnée dans CARROLL et RUPPERT [1988]. Ces derniers proposent la représentation du logarithme des résidus absolus en fonction du logarithme de x . Ce type de graphique donne, en effet, des informations concernant la nature de la relation liant l'écart-type conditionnel à la valeur de x , car il s'apparente à un graphique du logarithme de la variance conditionnelle en fonction du logarithme de la variable explicative. Dans l'interprétation de ce graphique, il faut cependant veiller à ne pas accorder une importance exagérée aux quelques valeurs négatives très grandes liées aux résidus absolus pratiquement nuls.

CARROLL et RUPPERT [1988] insistent sur l'intérêt qu'il y a à éliminer le signe des résidus, afin de doubler la densité des points permettant ainsi de mieux apprécier visuellement la variabilité des résidus. Ils signalent également qu'une difficulté d'interprétation résulte de l'inégale densité de points le long de l'axe des x . En effet, en l'absence d'hétéroscédasticité, une forte densité de points sur une portion de l'axe des x induit, en moyenne, une plus grande amplitude des résidus qu'une faible densité.

Il faut noter aussi qu'une inadéquation du modèle, conduisant à des résidus absolus importants pour des valeurs estimées extrêmes peut parfois suggérer, à tort, une inégalité des variances conditionnelles. Il en va de même pour la présence éventuelle de résidus absolus importants, liés à des données aberrantes. Dans ces conditions, il peut être utile d'examiner l'allure générale du graphique, abstraction faite de ces quelques résidus particuliers.

Le diagramme de dispersion d'une fonction des résidus absolus peut être complété par la superposition d'une relation donnant l'évolution moyenne des

résidus absolus ou d'une fonction de ces résidus. Un modèle simple, comme la droite de régression peut convenir dans certains cas. Un coefficient de régression positif indique alors que la dispersion des résidus augmente avec x , alors qu'un coefficient négatif indique une diminution de la dispersion avec x . L'ajustement d'un modèle rigide comme la droite de régression peut être remplacé par une forme de lissage, par moyennes mobiles par exemple.

On peut aussi tenter de synthétiser le diagramme de dispersion par le calcul du coefficient de corrélation. Un coefficient positif peut être l'indication d'une variabilité croissante avec x , alors qu'un coefficient négatif est l'indice d'une variabilité décroissante avec x . Le test de signification d'un tel coefficient n'est cependant qu'approximatif, car les résidus dépendent des paramètres estimés. CARROLL et RUPPERT [1988] proposent de remplacer le coefficient de corrélation classique par le coefficient de corrélation de rang de SPEARMAN qui est moins sensible à la présence de données extrêmes.

Outre le test approximatif du coefficient de corrélation de rang, plusieurs tests statistiques peuvent être réalisés. Parmi ceux-ci, le test de BARTLETT ou de HARTLEY [DAGNELIE, 1998], applicables lorsque plusieurs observations sont caractérisées par une même valeur de x , le test de GOLDFELD et QUANDT [1965] et le test de BREUSCH et PAGAN [1979] ont été discutés dans une note antérieure PALM, 1994]. Nous nous limitons à la présentation du dernier test cité, qui a aussi été proposé par COOK et WEISBERG [1983].

Le test consiste à vérifier la nullité de λ dans le modèle de variance conditionnelle suivant :

$$\sigma_{y|x_i}^2 = \sigma^2 [\exp(\lambda x_i)].$$

A partir des résidus de la régression, e_i , on définit une nouvelle variable e'_i :

$$e'_i = e_i^2 / \tilde{\sigma}^2 \quad \text{avec} \quad \tilde{\sigma}^2 = \sum_{i=1}^n e_i^2 / n.$$

On calcule alors la somme des carrés des écarts, $SCE_{rég}$, liée à la régression de e'_i en fonction de x_i , ainsi que la quantité :

$$\chi_{obs}^2 = SCE_{rég} / 2,$$

qui suit une distribution χ^2 à un degré de liberté lorsque l'hypothèse de nullité de λ est vérifiée.

Les approches graphiques et numériques décrites ci-dessus peuvent être généralisées au cas de plusieurs variables explicatives. Pour les représentations graphiques et le calcul du coefficient de corrélation de rang, on choisit *a priori* l'une ou l'autre variable explicative dont on pense qu'elle pourrait être pertinente pour expliquer la non-constance des variances conditionnelles. On peut également remplacer la variable explicative par les valeurs estimées de la variable à expliquer. Pour le test de BREUSCH et PAGAN [1979] ou de COOK et WEISBERG [1983], la régression de e' en fonction de x peut être remplacée par la régression de e' en fonction de \hat{y} , ou par une régression multiple en fonction de toutes ou d'un sous-ensemble de variables explicatives. Dans ce cas, et sous

l'hypothèse d'homoscédasticité, la quantité χ_{obs}^2 suit une distribution χ^2 à k degrés de liberté, k étant le nombre de paramètres de l'équation, à l'exclusion du terme indépendant.

Ainsi donc, la mise en évidence de l'hétéroscédasticité peut se faire par des représentations graphiques ou par le calcul de paramètres, certains de ceux-ci pouvant faire l'objet d'un test statistique. Les représentations graphiques donnent une meilleure idée de la nature de l'hétéroscédasticité en vue d'une éventuelle modélisation ultérieure. Elles permettent également de mieux apprécier l'influence éventuelle de quelques observations particulières sur cette hétéroscédasticité. Par contre, les paramètres résument davantage l'information par la quantification de l'importance de l'hétéroscédasticité et permettent de ce fait de comparer des situations différentes, liées par exemple à des transformations de variables. Ils donnent cependant moins d'information concernant la nature de l'hétéroscédasticité et peuvent être fortement liés à quelques observations particulières.

2.5. Indépendance des résidus

L'objectif est de vérifier si les résidus ε_i et ε_j , relatifs à deux observations i et j quelconques, sont indépendants.

Sous l'hypothèse de normalité, l'indépendance est assurée dès qu'il y a non-corrélation. En pratique, on vérifie la non-corrélation des résidus ε_i et ε_j à partir des résidus observés e_i et e_j . De plus, on suppose que s'il existe une forme de corrélation, celle-ci est liée à un facteur particulier tel que le temps ou l'espace.

Lorsque les données sont ordonnées dans le temps ou dans l'espace et que les intervalles séparant deux données successives sont constants, la liaison entre les résidus successifs est mesurée par les coefficients d'autocorrélation, ρ_s , avec $s = |i - j|$ et $\rho_0 = 1$. On considère donc que la corrélation entre deux résidus i et j ne dépend que du nombre d'observations séparant i et j . Ainsi, la corrélation entre ε_2 et ε_1 est la même que la corrélation entre ε_3 et ε_2 , entre ε_4 et ε_3 , etc. Elle est égale à ρ_1 . La corrélation entre ε_3 et ε_1 est la même que la corrélation entre ε_4 et ε_2 , etc. Elle est égale à ρ_2 . Et ainsi de suite pour ρ_3 , ρ_4 , etc.

La notion d'autocorrélation est une notion fondamentale dans l'étude des séries chronologiques. Des informations à ce sujet sont données dans une précédente note [PALM, 1987].

L'indépendance des résidus peut être vérifiée par le test de DURBIN et WATSON. De manière stricte, ce test de vérifie l'hypothèse nulle :

$$H_0 : \rho_s = 0$$

contre l'alternative :

$$H_1 : \rho_s = \rho_1^s,$$

qui correspond au cas où les résidus successifs proviendraient d'un modèle autorégressif d'ordre 1, s'écrivant :

$$\varepsilon_i = \rho_1 \varepsilon_{i-1} + \varepsilon_i^*,$$

dans lequel les résidus ε_i^* sont des réalisations indépendantes d'une variable normale de moyenne nulle et d'écart-type σ .

Mais en pratique, le test est souvent utilisé pour tester l'hypothèse nulle :

$$H_0 : \rho_1 = 0$$

contre l'alternative :

$$H_1 : \rho_1 \neq 0,$$

sans référence à un modèle spécifique particulier d'autocorrélation pour l'hypothèse alternative. Il peut en résulter une certaine perte de puissance [DRAPER et SMITH, 1998].

DURBIN et WATSON proposent le calcul de la quantité :

$$d = \sum_{i=2}^n (e_i - e_{i-1})^2 \Big/ \sum_{i=1}^n e_i^2,$$

qui est approximativement égale à :

$$d = \left(2 \sum_{i=2}^n e_i^2 - 2 \sum_{i=2}^n e_i e_{i-1} \right) \Big/ \sum_{i=1}^n e_i^2 \simeq 2(1 - \hat{\rho}_1).$$

La statistique de DURBIN et WATSON varie donc entre 0, lorsque $\hat{\rho}_1 = 1$, et 4, lorsque $\hat{\rho}_1 = -1$.

Sous l'hypothèse nulle, l'espérance mathématique de la statistique vaut 2 et la distribution est symétrique. Les tables proposées par DRAPER et SMITH [1998] donnent des couples de valeurs critiques, notées d_L et d_U pour des niveaux de probabilité $\alpha = 0,01, 0,025$ et $0,05$, et ce, en fonction du nombre d'observations n et du nombre de variables explicatives p .

Concrètement, les tests se réalisent de la manière suivante. Pour le test unilatéral, $H_0 : \rho_1 = 0$ contre l'alternative $H_1 : \rho_1 > 0$,

- si $d < d_L$: on rejette H_0 au niveau α ;
- si $d > d_U$: on ne rejette pas H_0 au niveau α ;
- si $d_L \leq d \leq d_U$: le test ne permet pas de conclure au niveau α .

Pour le test unilatéral, $H_0 : \rho_1 = 0$ contre l'alternative $H_1 : \rho_1 < 0$,

- si $4 - d < d_L$: on rejette H_0 au niveau α ;
- si $4 - d > d_U$: on ne rejette pas H_0 au niveau α ;
- si $d_L \leq 4 - d \leq d_U$: le test ne permet pas de conclure au niveau α .

Enfin, pour le test bilatéral, $H_0 : \rho_1 = 0$ contre l'alternative $H_1 : \rho_1 \neq 0$,

- si $d < d_L$ ou $4 - d < d_L$: on rejette H_0 au niveau 2α ;
- si $d > d_U$ ou $4 - d > d_U$: on ne rejette pas H_0 au niveau 2α ;
- si $d_L \leq d < d_U$ ou $d_L < 4 - d < d_U$: le test ne permet pas de conclure au niveau 2α .

La présence de valeurs de d ne conduisant pas à une décision claire peut être embarrassante en pratique. DRAPER et SMITH [1998] proposent de considérer cette situation comme un signal d'alarme ou de rejeter l'hypothèse nulle, en admettant que le risque de première espèce est légèrement supérieur à la valeur nominale.

La détection de l'autocorrélation par le test de DURBIN et WATSON ne concerne que l'autocorrélation entre données adjacentes (autocorrélation de rang 1). Si des autocorrélations existent entre des résidus séparés de plus d'une unité de temps ou d'espace, alors que l'autocorrélation de rang 1 est faible, elles pourront passer inaperçues par ce test.

Les programmes utilisés pour l'analyse des séries chronologiques fournissent des estimations de ces autocorrélations, qui, sous l'hypothèse nulle $H_0 : \rho_s = 0$, ont une distribution d'échantillonnage asymptotiquement normale, de moyenne nulle et d'erreur-standard approximativement égale à :

$$\frac{1}{\sqrt{n}} \sqrt{\frac{n-s}{n+2}} \simeq \frac{1}{\sqrt{n}}.$$

On peut donc, en première approximation et pour autant que n soit suffisamment grand, considérer comme différents de zéro, au niveau $\alpha = 0,05$, les coefficients d'autocorrélation qui, en valeur absolue, dépassent $2/\sqrt{n}$.

Il s'agit d'un test rapide et très approximatif, qui ne tient pas compte du nombre de variables explicatives dans le modèle. A titre de comparaison, pour $n = 30$, on considérerait comme significatifs tous les coefficients d'autocorrélation supérieurs, en valeur absolue, à $2/\sqrt{30}$, soit 0,365. Compte tenu de la relation existant entre $\hat{\rho}_1$ et la statistique d de DURBIN et WATSON, une autocorrélation au rang 1 égale à 0,365 correspond à la valeur :

$$d \simeq 2(1 - 0,365) \simeq 1,27.$$

Les valeurs d_L et d_U lues dans la table relative à $\alpha = 0,025$ sont de 1,25 et 1,38 pour $p = 1$ et de 0,98 et 1,73 pour $p = 5$. On constate que la valeur approximative $d = 1,27$ se trouve dans la zone de non-conclusion du test de DURBIN et WATSON.

D'autre part, le risque de première espèce global lié à la réalisation de la séquence de tests sur les coefficients d'autocorrélation successifs est largement plus élevé que le risque associé à un test unique. Ainsi, dans le cas d'un niveau α par test de 0,05 et en supposant qu'on teste la signification des 20 premiers

coefficients d'autocorrélation, on peut s'attendre à obtenir un coefficient significatif en moyenne par application de la séquence de tests, lorsque l'hypothèse nulle est vraie. Plus que le test lui-même, c'est l'importance et le rang des autocorrélations les plus grandes qui présente de l'intérêt. Ainsi, une autocorrélation significative au rang quatre pour des données trimestrielles sera davantage de nature à remettre en question le modèle qu'une autocorrélation de même importance, mais située au rang six par exemple. L'autocorrélation de rang quatre peut, en effet, indiquer la présence d'un phénomène saisonnier qu'il y aurait lieu de prendre en considération dans le modèle.

Un test global de la signification des k premiers coefficients d'autocorrélation a été proposé par BOX et PIERCE [1970] et ensuite amélioré par LJUNG et BOX [1978]. Ce test contrôle le risque α mais le non-rejet de l'hypothèse nulle ne signifie pas nécessairement que les résidus sont non corrélés car un coefficient d'autocorrélation significatif peut être camouflé, dans la statistique χ_{obs}^2 qui est calculée pour ce test, si les autres coefficients sont très proches de zéro. Inversement, le rejet de l'hypothèse nulle donne peu d'indications à l'utilisateur, qui devra de toute manière examiner les coefficients individuels.

Par ailleurs, si les données ne sont pas ordonnées dans le temps ou dans l'espace ou ne correspondent pas à des observations faites à des intervalles de temps ou d'espaces égaux, le test de DURBIN et WATSON et les autres tests de signification des coefficients d'autocorrélation ne sont plus applicables, de façon stricte. Ils peuvent cependant être utilisés à titre indicatif, lorsque les données sont ordonnées selon un critère extérieur, comme par exemple l'ordre alphabétique, la taille des individus, etc. Une autocorrélation trop différente de zéro devrait attirer l'attention de l'utilisateur et l'encourager à examiner sérieusement ses données en vue d'améliorer le modèle, par exemple par la prise en compte d'autres variables explicatives.

3. TRANSFORMATIONS DE VARIABLES

3.1. Types et effets des transformations

L'objectif des transformations de variables est de modifier le modèle initial de manière à obtenir un nouveau modèle pour lequel les conditions d'application sont vérifiées, ou du moins pour lequel le non-respect de ces conditions ne constitue plus un problème majeur. Dans certains cas, la transformation permet également de simplifier la relation.

Les transformations peuvent porter uniquement sur la variable à expliquer ou bien uniquement sur les variables explicatives ou, au contraire, simultanément sur la variable à expliquer et sur les variables explicatives.

La transformation de y modifie à la fois la relation fonctionnelle et la distribution des résidus du point de vue de la normalité et de l'hétéroscédasticité.

A l'inverse, la transformation d'une ou de plusieurs variables explicatives vise essentiellement à améliorer l'adéquation du modèle du point de vue de la

relation fonctionnelle, c'est-à-dire à vérifier la condition de nullité de l'espérance mathématique des résidus théoriques. Cette transformation n'a pas d'effet sur la normalité, l'homoscédasticité et l'indépendance des résidus théoriques. Toutefois, la transformation entraîne une modification des résidus observés, qui sont utilisés pour la vérification de ces conditions d'application.

Enfin, la transformation de la variable à expliquer et des variables explicatives combine les effets des deux transformations (effets sur la relation, sur la normalité et sur l'hétéroscédasticité).

Quelle que soit la nature de la transformation réalisée, il sera toujours utile, à l'issue de la transformation, de vérifier l'effet de celle-ci sur l'adéquation de la relation retenue, sur la normalité et l'homoscédasticité des résidus. Dans la pratique, un compromis devra fréquemment être accepté, la transformation jugée la plus adéquate pour un critère donné, par exemple la normalité, ne sera pas automatiquement la plus adéquate pour un autre critère, par exemple l'homoscédasticité.

3.2. Transformation de la variable à expliquer

Nous avons signalé, au paragraphe 3.1, qu'une transformation de la variable à expliquer affecte l'adéquation de la relation, la normalité et l'homoscédasticité des résidus.

Dans certaines situations, cette transformation peut être choisie sur une base théorique, en particulier lorsqu'elle vise à stabiliser les variances conditionnelles. Ainsi, la transformation racine carrée permet de stabiliser les variances chaque fois qu'il y a proportionnalité entre les moyennes conditionnelles et les variances conditionnelles, comme dans le cas des distributions de POISSON. La transformation arcsinus racine carrée, appelée aussi transformation angulaire, $\sin^{-1} \sqrt{y}$, s'applique aux variables binomiales. La transformation logarithmique s'utilise lorsque le coefficient de variation est constant. Cette transformation mérite d'être envisagée chaque fois que le rapport du maximum au minimum de y est très grand, supérieur à 1.000 par exemple. La transformation $1/y$ est souvent appliquée quand la variable à expliquer est le temps d'attente d'un événement. Des informations plus complètes concernant les diverses transformations ci-dessus sont données, entre autres, par CHATTERJEE et PRICE [1991], DAGNELIE [1998], DRAPER et SMITH [1998] et WEISBERG [1985].

Lorsque la variable y comporte des valeurs nulles, on remplace habituellement y par $y + 1$ dans les différentes transformations. En présence de valeurs négatives, l'ajout à y d'une constante assez petite avant la transformation permet de résoudre le problème de façon satisfaisante, pour autant que les valeurs de y ne soient pas toutes faibles [WEISBERG, 1985].

On notera aussi que les transformations ci-dessus, sélectionnées dans le but de stabiliser les variances conditionnelles, ont également souvent un effet bénéfique sur la normalité des résidus.

Les transformations racine carrée, logarithmique et inverse, mentionnées ci-dessus, sont des cas particuliers de la transformation puissance de BOX et COX :

$$w(\lambda) = \begin{cases} y^\lambda & \text{si } \lambda \neq 0 \\ \log_e y & \text{si } \lambda = 0, \end{cases}$$

obtenus en donnant à λ respectivement la valeur de 0,5, 0 et -1.

En l'absence d'une base théorique permettant de sélectionner une transformation adéquate, on peut retenir la transformation de BOX et COX et déterminer la valeur de λ à partir des observations elles-mêmes.

Cette transformation peut s'écrire sous une forme un peu plus compliquée :

$$w(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda \bar{y}_g^{\lambda-1} & \text{si } \lambda \neq 0 \\ \bar{y}_g \log_e y & \text{si } \lambda = 0, \end{cases}$$

\bar{y}_g étant la moyenne géométrique de y . Cette seconde présentation a l'avantage d'être continue en fonction de λ et de permettre la comparaison directe des sommes de carrés d'écartes résiduelles obtenues après ajustement du modèle de régression pour différentes valeurs de λ .

Une première méthode de détermination de λ est la méthode du maximum de vraisemblance. Pour une valeur fixée de λ , on calcule les w_i par la seconde formule donnée ci-dessus et on détermine l'équation de régression :

$$\mathbf{w} = \mathbf{X} \hat{\boldsymbol{\beta}} + \mathbf{e},$$

les vecteurs \mathbf{w} , $\hat{\boldsymbol{\beta}}$ et \mathbf{e} dépendant de λ . A partir des résidus, on calcule la somme des carrés des écartes résiduelle $SCE_r(\lambda)$, elle aussi fonction de λ , et on en déduit le logarithme de la vraisemblance par la relation :

$$L(\lambda) = -\frac{n}{2} \log_e SCE_r(\lambda).$$

La fonction de vraisemblance peut ainsi être calculée pour une série de valeurs λ comprises entre -2 et 2 , par exemple. L'estimation du maximum de vraisemblance, $\hat{\lambda}$, correspond à la valeur pour laquelle la fonction de vraisemblance est maximum. Si la valeur de λ qui maximise la vraisemblance se situe hors de l'intervalle $(-2, 2)$, on peut mettre en doute l'utilité de la méthode pour les données en question.

On peut encore calculer, de manière approximative, l'intervalle de confiance de λ . Cet intervalle comprend toutes les valeurs de λ telles que :

$$L(\lambda) > L(\hat{\lambda}) - \frac{1}{2} \chi_{1-\alpha}^2,$$

$\chi_{1-\alpha}^2$ étant le pourcentile $1 - \alpha$ de la distribution χ^2 à un degré de liberté. Les limites, λ_1 et λ_2 , sont déterminées à partir du graphique donnant la vraisemblance en fonction de λ ou par interpolations linéaires à partir des couples $L(\lambda)$ et λ [DRAPER et SMITH, 1998].

En pratique, on arrondira souvent la valeur de $\hat{\lambda}$, surtout si la valeur arrondie se trouve dans l'intervalle de confiance. Ainsi, si les limites de confiance sont par exemple 0,4 et 0,7 on prendra $\lambda = 0,5$, qui correspond à la transformation racine carrée.

Une autre méthode de détermination de λ a été proposée par ATKINSON [1981]. Elle est décrite par WEISBERG [1985] notamment et permet d'obtenir rapidement une estimation à partir du développement en série de TAYLOR au point $\lambda = 1$ de la transformation puissance.

On calcule d'abord la variable g , fonction de y :

$$g_i = y_i [\log_e(y_i/\bar{y}_g) - 1] + \log_e \bar{y}_g + 1.$$

Ensuite, on détermine l'équation de régression suivante :

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \phi \mathbf{g} + \boldsymbol{\varepsilon},$$

Si le coefficient ϕ est non significatif, on considère qu'il n'y a pas lieu d'effectuer de transformation de y . Si, au contraire, le coefficient est significatif, on estime λ par la relation :

$$\hat{\lambda} = 1 - \hat{\phi},$$

la valeur étant le plus souvent arrondie de manière à retrouver une transformation plus classique.

L'estimation de λ par la méthode du maximum de vraisemblance et par la méthode d'ATKINSON est illustrée au paragraphe 4.2.

Comme pour les transformations mentionnées plus haut, la transformation de BOX et COX n'est applicable que si les y_i sont tous positifs. En présence de valeurs nulles ou négatives, on peut ajouter une constante à y avant d'employer la méthode. On dispose cependant de peu d'informations quant au choix de cette constante [WEISBERG, 1985].

D'autre part, la transformation de BOX et COX est sensible aux données influentes et des méthodes ont été proposées pour détecter de telles observations [KIM *et al.*, 1996].

En pratique, lorsque la valeur retenue pour λ est différente de 0 ou 1, on utilise, par la suite, la transformation y^λ , la formulation plus compliquée ne présentant plus aucun intérêt, une fois la valeur de λ fixée. Pour $\lambda = 0$, on utilise indifféremment le logarithme népérien ou le logarithme décimal et on néglige la constante \bar{y}_g présente dans la formule plus compliquée.

Enfin, il est utile de vérifier si les diverses conditions d'application sont acceptables après transformation car, comme nous l'avons déjà signalé, la transformation modifie la relation, la normalité et l'homoscédasticité des résidus.

3.3. Transformation de variables explicatives

La transformation de variables explicatives peut intervenir sous la forme de l'introduction dans le modèle de nouvelles variables, fonction d'une ou de plusieurs variables initiales. C'est l'approche utilisée dans la régression polynomiale.

Elle présente l'inconvénient de compliquer quelque peu le modèle et d'augmenter le nombre de paramètres à estimer. L'utilité des termes supplémentaires est généralement vérifiée par un test de signification classique.

Une autre approche consiste à remplacer une variable explicative donnée par une fonction de cette variable. Cette seconde approche n'est cependant valable que si y est une fonction monotone croissante ou décroissante de x , du moins lorsque x est une variable positive.

Parmi les transformations les plus courantes, on peut citer :

$$z = x^2, \quad z = \sqrt{x}, \quad z = \log x \quad \text{et} \quad z = 1/x,$$

qui sont des cas particuliers de la transformation puissance de BOX et COX, mentionnée au paragraphe précédent :

$$z = \begin{cases} x^\alpha & \text{si } \alpha \neq 0 \\ \log_e x & \text{si } \alpha = 0. \end{cases}$$

La question qui se pose en pratique est de décider si une transformation est nécessaire et, dans l'affirmative, de fixer la valeur de α et ce, pour chacune des variables explicatives.

Une solution à cette double question peut être obtenue par un développement en série de TAYLOR. Considérons en effet un modèle initial à p variables :

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon \quad (1)$$

et supposons qu'on s'intéresse à la transformation de la première variable explicative. Après transformation, le modèle s'écrit :

$$y = \beta_0 + \beta_1 x_1^\alpha + \dots + \beta_p x_p + \varepsilon \quad (2)$$

Le paramètre α pourrait être estimé en même temps que les paramètres $\beta_0, \beta_1, \dots, \beta_p$, par la méthode des moindres carrés non linéaires. Une solution approchée est obtenue en considérant le développement en série de TAYLOR de x_1^α , au point $\alpha = 1$ et limité aux deux premiers termes :

$$\begin{aligned} x_1^\alpha &\simeq x_1 + (\alpha - 1) \left(\frac{d x_1^\alpha}{d \alpha} \right)_{\alpha=1} \\ &\simeq x_1 + (\alpha - 1) x_1 \log_e x_1, \end{aligned}$$

En remplaçant x_1^α par son développement en série dans l'équation (2), on a :

$$\begin{aligned} y &\simeq \beta_0 + \beta_1 [x_1 + (\alpha - 1) x_1 \log_e x_1] + \dots + \beta_p x_p + \varepsilon \\ &\simeq \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \gamma x_1 \log_e x_1 + \varepsilon \quad (3) \end{aligned}$$

avec $\gamma = \beta_1(\alpha - 1)$.

Cette dernière équation de régression peut être ajustée par les moindres carrés ordinaires. Le test classique de signification appliqué au coefficient γ , permet de juger de l'intérêt de la transformation et une estimation de α est donnée par :

$$\hat{\alpha} = (\hat{\gamma}/\hat{\beta}_1) + 1,$$

$\hat{\gamma}$ étant obtenu par l'ajustement du modèle (3) mais $\hat{\beta}_1$ étant estimé à partir du modèle initial (1).

BOX et TIDWELL [1962] ont proposé une procédure itérative. A la première itération, la valeur de l'exposant est fixée à l'unité, soit $\alpha = 1$. Au terme des calculs ci-dessus, on obtient une nouvelle estimation de l'exposant, soit α_1 . On recommence alors les calculs en remplaçant, dans le modèle initial, la variable x_1 par x_1' :

$$x_1' = x_1^{\alpha_1}.$$

On retrouve une nouvelle valeur α_2' , qui donne lieu à une nouvelle transformation :

$$x_1' = (x_1^{\alpha_1})^{\alpha_2'} = x_1^{\alpha_2} \quad \text{avec} \quad \alpha_2 = \alpha_1 \alpha_2'.$$

Et ainsi de suite, pour les itérations suivantes, la procédure s'arrêtant lorsque l'exposant ne change pratiquement plus ou lorsque le coefficient γ est non significatif. La procédure est illustrée au paragraphe 4.2.

Dans le cas de la régression multiple, il peut être utile de transformer ainsi plusieurs variables explicatives. WEISBERG [1985] suggère une approche séquentielle, étudiant successivement chacune des variables explicatives. RYAN [1997] étudie, par contre, simultanément la transformation de toutes les variables en calculant, à chaque itération, les p variables $x_j' \log_e x_j'$ ($j = 1, \dots, p$).

L'utilisation de la transformation de BOX et TIDWELL peut poser des problèmes de colinéarité des variables, si le rapport entre le maximum et le minimum de la variable x_j est limité, car dans ce cas la corrélation entre x_j et $x_j \log_e x_j$ est toujours très élevée. Le problème est encore accentué si les variables explicatives présentent de fortes corrélations.

Indépendamment de la méthode de BOX et TIDWELL, il faut noter que le choix empirique d'une transformation sera toujours difficile lorsque le rapport du maximum au minimum d'une variable est inférieur à 10 car x_j et x_j^α sont toujours fortement corrélés, du moins pour des valeurs positives de α . Il sera par conséquent difficile de choisir entre x_j et $x_j^{0,5}$, par exemple.

La méthode développée par BOX et TIDWELL peut également conduire à des résultats absurdes, la valeur obtenue pour un exposant étant très largement en dehors du domaine attendu, qui est de -2 à 2. Ces résultats aberrants proviennent de la mauvaise estimation de γ_j ou de β_j .

3.4. Transformation de la variable à expliquer et des variables explicatives

Une première situation résulte de la combinaison des transformations examinées aux deux paragraphes précédents : on effectue une transformation de y

et une transformation d'une ou de plusieurs variables explicatives, les transformations réalisées sur les variables explicatives pouvant être différentes de celles réalisées sur la variable y . Ainsi, disposant de deux variables explicatives, x_1 et x_2 , on pourrait, par exemple, arriver à la conclusion qu'un modèle du type :

$$\log y = \beta_0 + \beta_1 \sqrt{x_1} + \beta_2 x_2 + \varepsilon$$

soit acceptable, à la fois du point de vue de la normalité et de l'homoscédasticité des résidus et de la forme de la relation.

Théoriquement, la distribution des ε est indépendante de la relation et la transformation de y devrait pouvoir être étudiée indépendamment des transformations des variables explicatives. Dans la pratique, il en est effectivement ainsi pour autant qu'on dispose de répétitions. Dans ce cas, en effet, les résidus purs peuvent être estimés à partir de y uniquement, conditionnellement aux x . Par contre, dans le cas plus général, la normalité et l'homoscédasticité des résidus ε_i sont appréciées à partir des e_i , qui eux dépendent de l'adéquation du modèle.

D'autre part, les transformations de variables explicatives nécessaires pour rendre le modèle linéaire dépendent de la transformation effectuée sur y .

Il en résulte que les deux types de transformation doivent être recherchés de façon simultanée. RYAN [1997] propose la démarche suivante. La fonction de vraisemblance est calculée pour différentes valeurs de λ comprises entre -2 et 2, en l'absence de transformation des x_j . Sur la base de cette vraisemblance ainsi que de critères de normalité et d'homoscédasticité, un intervalle plus réduit de valeurs acceptables de λ est retenu. Ensuite, les valeurs α_j de la transformation de COX et TIDWELL des variables explicatives, ainsi que les mesures de normalité et d'homoscédasticité sont calculées pour différentes valeurs de λ de cet intervalle réduit. Enfin, les valeurs de λ et α_j retenues sont celles qui conduisent au modèle offrant le meilleur compromis pour les critères de normalité, homoscédasticité et d'adéquation de la relation.

WEISBERG [1985] propose une solution quelque peu différente. D'abord, toutes les variables explicatives pour lesquelles le rapport du maximum sur le minimum est supérieur à 10 subissent une transformation logarithmique. Ensuite, la variable à expliquer est transformée après estimation de λ par la méthode du maximum de vraisemblance ou par la méthode d'ATKINSON. Enfin, la méthode de BOX et TIDWELL est appliquée séquentiellement aux variables explicatives ayant une valeur t_{obs} importante.

Dans certaines situations, la même transformation est appliquée aux deux membres de l'équation. L'exemple le plus courant est la double transformation appliquée en vue de linéariser un modèle non linéaire. C'est, par exemple, le cas pour la fonction puissance :

$$y = \beta_0 x^{\beta_1} \varepsilon,$$

qui, après double transformation logarithmique, s'écrit :

$$\log y = \log \beta_0 + \beta_1 \log x + \log \varepsilon,$$

ou pour le modèle :

$$y = x / (\beta_0 + \beta_1 x + \varepsilon)$$

qui, après transformation inverse, s'écrit :

$$1/y = \beta_0/x + \beta_1 + \varepsilon/x.$$

On constate effectivement que, après transformation, les modèles sont linéaires et les paramètres peuvent être estimés par les moindres carrés ordinaires. Toutefois, l'inférence statistique ne peut être réalisée, de manière correcte, que si les résidus transformés, $\log \varepsilon$ pour le premier modèle et ε/x pour le second modèle, remplissent les conditions d'application (normalité, homoscedasticité, indépendance, nullité de l'espérance mathématique).

Les deux exemples ci-dessus constituent des cas particuliers de la méthode plus générale proposée par CARROLL et RUPPERT [1988]. Partant de la relation suivante :

$$y = f(x, \beta) + \varepsilon,$$

ils proposent de transformer y et $f(x, \beta)$ par la transformation de BOX et COX discutée précédemment, ou encore par d'autres transformations. Si la fonction $f(x, \beta)$ est linéaire, la fonction transformée sera cependant le plus souvent non linéaire et les paramètres devront être estimés par les moindres carrés non linéaires. L'utilisation de cette approche peut se justifier dans le cas où la forme fonctionnelle est adéquate mais où la normalité ou l'homoscedasticité n'est pas vérifiée. La double transformation permet de conserver l'adéquation de la relation, tout en modifiant la distribution des résidus.

La méthode des moindres carrés généralisés peut également être vue comme une transformation des deux membres de l'équation. En effet, considérons le modèle :

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{avec } \mathbf{V}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{V}.$$

Soit \mathbf{P} une matrice non singulière telle que :

$$\mathbf{P}' \mathbf{P} = \mathbf{P} \mathbf{P}' = \mathbf{P}^2 = \mathbf{V}.$$

En prémultipliant le modèle par \mathbf{P}^{-1} , on obtient [DRAPER et SMITH, 1998] :

$$\mathbf{P}^{-1} \mathbf{y} = \mathbf{P}^{-1} \mathbf{X} \boldsymbol{\beta} + \mathbf{P}^{-1} \boldsymbol{\varepsilon},$$

ou encore :

$$\mathbf{w} = \mathbf{z} \boldsymbol{\beta} + \boldsymbol{\delta},$$

les résidus $\boldsymbol{\delta}$ étant de matrice de variances et covariances égale à $\sigma^2 \mathbf{I}$.

En pratique, l'utilisation de la transformation nécessite la connaissance de \mathbf{P}^{-1} , ou du moins d'une estimation de cette matrice.

Dans le cas de la régression pondérée, \mathbf{V} est une matrice diagonale dont les éléments sont proportionnels aux variances des ε_i :

$$v_{ii} = k \sigma_{\varepsilon_i}^2,$$

et donc inversement proportionnels aux poids à attribuer aux observations. Prémultiplier par \mathbf{P}^{-1} revient donc à diviser y_i et x_i par $\sqrt{v_{ii}}$:

$$w_i = y_i / \sqrt{v_{ii}}, \quad z_i = x_i / \sqrt{v_{ii}} \quad \text{et} \quad \delta_i = \varepsilon_i / \sqrt{v_{ii}}.$$

Une présentation plus détaillée de la régression pondérée est donnée dans une autre note [PALM, 1994].

Dans le cas de données autocorrélées et dans les conditions décrites au paragraphe 2.5, la matrice \mathbf{V} est la matrice d'autocorrélation des résidus. Pour des données ordonnées dans le temps ou dans l'espace, l'élément v_{ij} de la matrice \mathbf{V} est égal à ρ_k avec $k = |i - j|$ et $\rho_0 = 1$. Sur le plan théorique, la procédure à l'avantage d'être très générale mais elle nécessite l'estimation des autocorrélations jusqu'au rang n , ce qui en pratique n'est généralement pas possible.

Une autre solution pour des données corrélées est la méthode de COCHRAN et ORCUTT, décrite notamment par CHATTERJEE et PRICE [1991]. Elle consiste à déterminer d'abord le modèle sans tenir compte de l'autocorrélation. Les résidus e_i sont alors utilisés pour estimer ρ_1 :

$$\hat{\rho}_1 = \frac{\sum_{i=2}^n e_i e_{i-1}}{\sum_{i=2}^n e_i^2}.$$

On transforme ensuite les variables y et x_j de la manière suivante :

$$w_i = y_i - \hat{\rho}_1 y_{i-1} \quad \text{et} \quad z_{ij} = x_{ij} - \hat{\rho}_1 x_{i-1,j}$$

et on ajuste le modèle :

$$w_i = \gamma_0 + \gamma_1 z_{i1} + \dots + \gamma_p z_{ip} + \delta_i.$$

Les paramètres γ_0 et γ_j sont liés aux paramètres du modèle initial par les relations suivantes :

$$\hat{\beta}_0 = \hat{\gamma}_0 / (1 - \hat{\rho}_1) \quad \text{et} \quad \hat{\beta}_j = \hat{\gamma}_j.$$

La procédure peut être utilisée de manière itérative. Si les résidus e'_i présentent des autocorrélations, on recalcule les résidus du modèle obtenu en tenant compte de l'autocorrélation estimée à l'étape précédente et, à partir de ces résidus, on détermine une nouvelle estimation de ρ_1 . La procédure s'arrête lorsque les estimations $\hat{\rho}_1$ ne se modifient plus d'une étape à l'autre. CHATTERJEE et PRICE [1991] estiment cependant que, si après la première itération, les résidus présentent encore une autocorrélation, l'utilisateur doit rechercher des solutions alternatives à la méthode de COCHRAN et ORCUTT. Au lieu de la procédure itérative, on peut aussi ajuster le modèle :

$$w_i = \gamma_0 + z_i \gamma + \delta_i$$

pour différentes valeurs de ρ_1 et retenir comme estimations les paramètres $\hat{\rho}_1$, $\hat{\gamma}_0$ et $\hat{\gamma}_j$ qui rendent minimum la somme des carrés des écarts résiduelle [CHATTERJEE et PRICE, 1991].

La méthode de COCHRAN et ORCUTT repose sur une structure des résidus du type autorégressif d'ordre 1, tout comme le test de DURBIN et WATSON décrit au paragraphe 2.5. Si des autocorrélations existent entre des résidus séparés de

plus d'une unité de temps ou d'espace, elles ne pourront pas être éliminées par la procédure ci-dessus. De même, si les données sont ordonnées selon un autre caractère que le temps ou l'espace, les techniques ci-dessus sont inappropriées.

Enfin, signalons encore que, dans certains cas, l'autocorrélation peut être liée à l'absence, dans le modèle, d'une variable particulière dont les observations successives sont corrélées. L'introduction de cette variable dans le modèle peut parfois complètement éliminer le phénomène d'autocorrélation. Dans une telle situation, il est préférable d'éliminer l'autocorrélation par l'ajout d'une variable plutôt que par une forme de modélisation de l'autocorrélation. Des exemples sont donnés par CHATTERJEE et PRICE [1991].

4. APPLICATION

4.1. Données et modèle initial

Un chercheur a réalisé des observations sur 162 exploitations laitières dans une région du Brésil, en vue de modéliser la croissance des exploitations de cette région [MORO, 1995].

La variable à expliquer est le taux de croissance, défini comme le rapport entre le nombre de vaches en 1992 et le nombre de vaches en 1986. Cette variable sera désignée par le nom TAUX dans les sorties du logiciel Minitab notamment, ou par le symbole y dans certaines formules, afin d'en simplifier l'écriture.

Parmi les variables explicatives disponibles, nous avons sélectionné les trois variables les plus corrélées, en corrélation simple, au taux de croissance: le nombre annuel moyen de visites effectuées par les techniciens (x_1 ou VULG), le nombre de vaches laitières en 1986 (x_2 ou UGB86) et la valeur des actifs, en milliers de dollars (x_3 ou VACTIFS). Compte tenu du rythme des visites (annuel, semestriel, trimestriel, ...), ce nombre correspond à une variable discontinue ne pouvant prendre qu'une dizaine de valeurs différentes, comprises entre 0 (pas de visites) et 48 (4 visites mensuelles). Pour permettre des transformations de la variable, ce nombre a été augmenté d'une unité.

Les variables UGB86 et VACTIFS sont assez corrélées ($r = 0,73$) et la variable VULG est non corrélée aux deux autres variables explicatives ($r \simeq 0,08$). Les corrélations simples des trois variables explicatives avec le taux de croissance sont relativement faibles et du même ordre de grandeur; elles sont comprises entre 0,38 et 0,44 en valeur absolue.

La figure 1 donne les diagrammes de dispersion de la variable à expliquer avec les trois variables explicatives. Ces graphiques montrent qu'on doit s'attendre à des problèmes d'inégalité de variances, de non linéarité, et peut-être aussi de non-normalité des résidus.

La régression multiple du taux de croissance en fonction des trois variables explicatives a été calculée et la figure 2 donne l'analyse des résidus, réalisée avec la procédure %RESPLOTS de Minitab.

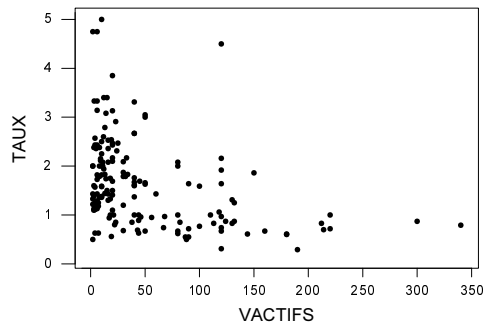
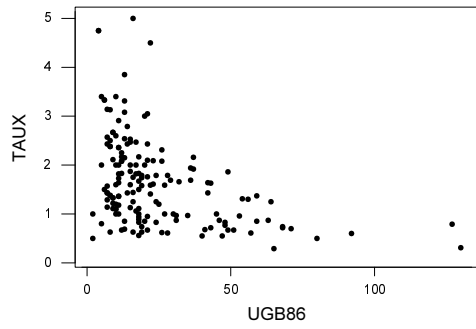
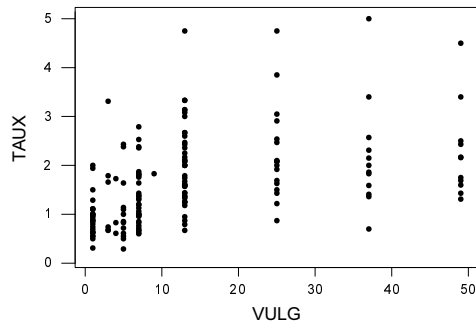


Figure 1. Diagramme de dispersion du taux d'accroissement en fonction des trois variables explicatives.

Résidus standardisés (modèle initial)

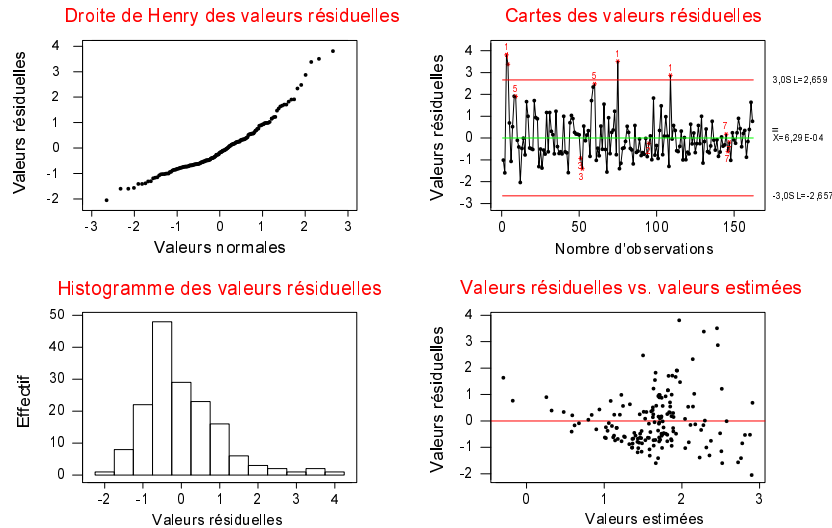


Figure 2. Examen des résidus du modèle initial par la commande %RESPLOTS de Minitab.

Le graphique des résidus standardisés en fonction des scores normaux présente une allure non linéaire et l'histogramme des résidus montre une dissymétrie gauche assez prononcée. Le coefficient de corrélation entre les résidus standardisés et les scores normaux est égal à 0,961. Cette valeur est inférieure à la valeur seuil du test de RYAN et JOINER [1976], égale à 0,992 pour un niveau de signification de 0,05, ce qui confirme bien le caractère non normal des résidus.

Le graphique des résidus en fonction des valeurs estimées montre que la variance des résidus augmente avec la valeur estimée. Le test de BREUSCH et PAGAN, réalisé en prenant les valeurs estimées des taux de croissance comme variable explicative dans le modèle de variance, donne la valeur $\chi_{obs}^2 = 21,7$. Cette valeur étant très nettement plus grande que $\chi_{0,95}^2 = 3,84$, on rejette indiscutablement l'hypothèse d'homoscédasticité.

Enfin, le test d'adéquation du modèle proposé par Minitab révèle l'inadéquation du modèle due, notamment, à la non-linéarité de la variable VULG.

4.2. Transformations de variables

Dans la mesure où la non-normalité et l'inégalité de variances sont très marquées, une transformation de la variable à expliquer semble tout à fait indiquée. Nous allons rechercher la valeur de λ de la transformation de BOX et COX, d'abord par la méthode du maximum de vraisemblance et ensuite par la méthode d'ATKINSON, afin d'illustrer, de manière concrète, les deux approches.

Dans la pratique, on se limite le plus souvent à une seule méthode.

La moyenne géométrique de la variable à expliquer est égale :

$$\bar{y}_g = \exp \left[\frac{1}{162} \sum_{i=1}^{162} \log_e y_i \right] = \exp(0,35573) = 1,4264.$$

Pour la méthode du maximum de vraisemblance, il faut calculer la variable w , pour une valeur donnée de λ . Considérons, à titre d'illustration, le cas où $y_i = 1,50$ et $\lambda = 2$. Pour cette observation, on a :

$$w_i = (y_i^\lambda - 1) / \lambda \bar{y}_g^{\lambda-1} = (1,50^2 - 1) / (2) (1,4264) = 0,4382.$$

Pour $\lambda = 2$, la régression de w_i en fonction des trois variables explicatives, VULG, UGB86 et VACTIFS, conduit à une somme des carrés des écarts résiduelle égale à 264,025. Le logarithme de la vraisemblance est par conséquent égal à :

$$L(2) = -\frac{n}{2} \log_e \text{SCE}_r = -\frac{162}{2} \log_e 264,025 = -451,7.$$

La vraisemblance a été calculée pour les différentes valeurs de λ comprises entre -2 et 2 , par pas de $0,05$. Le maximum de cette fonction se situe en $\lambda = 0,00$ et vaut $-323,4$. L'intervalle de confiance de λ correspond à l'ensemble des valeurs pour lesquelles :

$$L(\lambda) > L(0,00) - 0,5 \chi_{1-\alpha}^2 = -323,4 - (0,5)(3,84) = -325,3,$$

soit approximativement :

$$-0,20 < \lambda < 0,20.$$

En estimant λ par la méthode du maximum de vraisemblance, on conclut donc qu'il y a lieu d'effectuer une transformation logarithmique du taux de croissance.

Pour déterminer λ par la méthode d'ATKINSON, il faut calculer la variable g . Pour $y_i = 1,50$, par exemple, on trouve :

$$\begin{aligned} g_i &= y_i [\log_e (y_i / \bar{y}_g) - 1] + [\log_e \bar{y}_g + 1] \\ &= 1,50 [\log_e (1,50 / 1,4264) - 1] + [\log_e 1,4264 + 1] = -0,0694. \end{aligned}$$

Le calcul de la régression multiple de y fonction des trois variables explicatives et de g donne, comme coefficient de cette variable g :

$$\hat{\phi} = 1,332.$$

Le coefficient étant significativement différent de zéro ($t_{obs} = 14,1$), on conclut que λ est significativement différent de l'unité, la valeur estimée étant égale à :

$$\hat{\lambda} = 1 - \hat{\phi} = 1 - 1,333 = -0,333.$$

Pour cet exemple, les deux méthodes de détermination de λ donnent des résultats numériques légèrement différents : la méthode du maximum de vraisemblance indique clairement l'opportunité d'une transformation logarithmique et la méthode d'ATKINSON propose une transformation moins habituelle, mais pas fondamentalement en contradiction avec une transformation logarithmique. En conclusion, il semble bien qu'une transformation logarithmique puisse résoudre le problème de la non-normalité et de l'hétéroscédasticité des résidus.

Pour vérifier l'intérêt des transformations de variables explicatives, la méthode de BOX et TIDWELL a été appliquée. Les variables suivantes ont été calculées :

$$z_j = x_j \log_e x_j \quad (j = 1, \dots, 3)$$

et la régression multiple de $\log_e y$ en fonction des trois variables explicatives initiales x_1 , x_2 et x_3 et des variables construites z_1 , z_2 et z_3 a été calculée. Parmi ces dernières, seule la variable z_1 est significative ($t_{obs} = -5,6$). On conclut donc qu'il n'y a pas lieu de procéder à une transformation des variables x_2 et x_3 . Par contre, une transformation de x_1 peut être recommandée.

A partir des deux équations suivantes :

$$\log_e y = 0,4070 + 0,01741 x_1 - 0,009019 x_2 - 0,001729 x_3$$

et

$$\log_e y = 0,01725 + 0,1355 x_1 - 0,009198 x_2 - 0,001421 x_3 - 0,02988 x_1 (\log_e x_1),$$

on trouve :

$$\hat{\alpha} = \frac{\hat{\gamma}}{\hat{\beta}} + 1 = \frac{-0,02988}{0,01741} + 1 = -0,716.$$

Cette valeur est une première estimation de la puissance à laquelle il faudrait élever la variable x_1 . Nous la notons $\hat{\alpha}_1$.

A partir des deux équations suivantes :

$$\log_e y = 0,9236 - 0,8315 x_1^{-0,716} - 0,009433 x_2 - 0,001667 x_3$$

$$\text{et} \quad \log_e y = 1,4064 - 1,2489 x_1^{-0,716} - 0,009128 x_2 - 0,001477 x_3 \\ + 1,4978 x_1^{-0,716} \left(\log_e x_1^{-0,716} \right),$$

on obtient :

$$\hat{\alpha}'_2 = \frac{1,4978}{-0,8315} + 1 = -0,801$$

et

$$\hat{\alpha}_2 = \hat{\alpha}_1 \hat{\alpha}'_2 = (-0,716)(-0,801) = 0,574.$$

En continuant les calculs, on trouve, pour les itérations suivantes :

$$\hat{\alpha}_3 = -0,162, \quad \hat{\alpha}_4 = 0,0842, \quad \hat{\alpha}_5 = 0,0341, \quad \text{etc.}$$

A partir de la quatrième itération, les résultats sont relativement stables et le coefficient γ est non significatif. La valeur obtenue pour le paramètre α est donc proche de zéro et on conclut à l'intérêt de la transformation logarithmique de la variable VULG.

Résidus standardisés (modèle final)

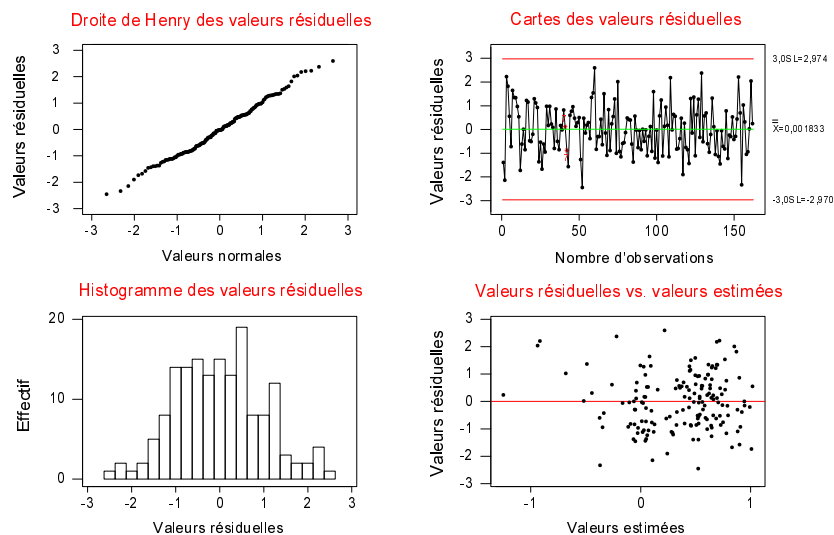


Figure 3. Examen des résidus du modèle final par la commande %RESPLOTS de Minitab.

4.3. Modèle final

Le modèle retenu s'écrit donc :

$$\log_e y = \hat{\beta}_0 + \hat{\beta}_1 \log_e x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

et les résultats de l'analyse des résidus par la commande %RESPLOTS de Minitab sont données dans la figure 3.

Du point de vue des conditions d'application, ce modèle est satisfaisant. La corrélation entre les scores normaux et les résidus standardisés est égale à 0,9965. La valeur seuil étant approximativement égale à 0,992, au niveau 0,05, on accepte l'hypothèse de normalité des résidus. Pour le test de BREUSCH et PAGAN, réalisé en prenant les valeurs estimées des logarithmes des taux comme variable explicative dans le modèle de variance, on a $\chi_{obs}^2 = 1,2$ et on accepte l'hypothèse d'homoscédasticité. Enfin, le test d'adéquation proposé par Minitab ne signale aucune trace d'inadéquation du modèle.

A titre de comparaison, le tableau 1 donne le coefficient de corrélation entre les scores normaux et les résidus standardisés, la valeur χ_{obs}^2 relative au test de BREUSCH et PAGAN, ainsi que le coefficient de détermination multiple pour le modèle initial, pour le modèle final ainsi que pour deux autres modèles. Pour le premier de ces deux modèles, aucune transformation des variables explicatives n'a été faite, alors que pour le second, au contraire, les trois variables explicatives ont été soumises à une transformation logarithmique. On constate qu'après

Tableau 1. Résultats du test de RYAN et JOINER (r_{sn}), du test de BREUSCH et PAGAN (χ_{obs}^2), du test d'adéquation du modèle proposé par Minitab et valeur R^2 pour différents modèles (NS : test non significatif au niveau 0,05; * : test significatif au niveau 0,05; ** : test significatif au niveau 0,01; *** : test significatif au niveau 0,001).

Variables à expliquer	Variables explicatives	r_{sn}	χ_{obs}^2	Adéquation	R^2
TAUX	VULG, UGB86, VACTIFS	0,961**	21,7***	***	34,4
\log_e TAUX	\log_e VULG, UGB86, VACTIFS	0,996 NS	1,2 NS	NS	52,9
\log_e TAUX	VULG, UGB86, VACTIFS	0,997 NS	0,0 NS	***	45,1
\log_e TAUX	\log_e VULG, \log_e UGB86, \log_e VACTIFS	0,995 NS	0,0 NS	*	51,0

transformation logarithmique de la variable à expliquer les résidus peuvent être considérés comme normaux et de variance constante, que les variables explicatives aient subi ou non une transformation logarithmique.

Les différences entre les trois derniers modèles se marquent sur la qualité des ajustements, mesurée par le coefficient R^2 ajusté : après transformation de la variable VULG uniquement, le modèle est un peu meilleur qu'après transformation des trois variables et nettement meilleur que le modèle sans transformation des variables explicatives. Il faut noter que la comparaison des valeurs R^2 peut se faire, de manière valable, uniquement si on compare les trois derniers modèles, pour lesquels la variable à expliquer est identique.

Les résultats des tests d'adéquation des modèles proposés par Minitab confirme la supériorité du modèle retenu sur les autres modèles.

Enfin, on a vérifié qu'aucun résidu n'est anormal et qu'aucune donnée n'est particulièrement influente, le plus grand résidu standardisé, t_i , est en effet égal à 2,60, alors que la valeur théorique vaut :

$$t_{1-\alpha/2n} = t_{0,9998} = 3,7,$$

et la plus grande distance de COOK, égale à 0,23, est largement inférieure à l'unité, valeur au-delà de laquelle une attention particulière doit être accordée aux observations [WEISBERG, 1985].

5. CONCLUSIONS

Dans les paragraphes précédents, de nombreux outils ont été décrits pour vérifier le respect des conditions d'application des modèles de régression. Ces outils sont graphiques ou numériques et dans le second cas, il s'agit du calcul de paramètres, éventuellement associés à des tests statistiques.

Les méthodes graphiques ont l'avantage de donner une vision plus globale du problème, par la mise en évidence de données particulières ou anormales, susceptibles d'influencer fortement les valeurs obtenues pour les paramètres. Par contre, les représentations graphiques se prêtent moins bien aux comparaisons de plusieurs modèles, par exemple pour apprécier l'amélioration du respect des conditions d'application à la suite de la transformation d'une variable. Les paramètres présentent l'avantage de la concision et les tests statistiques permettent la prise de décision sur une base plus objective, pour autant qu'ils soient suffisamment exacts.

Les méthodes décrites pour la vérification des conditions d'application ne constituent d'ailleurs pas un relevé exhaustif des outils disponibles dans la littérature. Elles ont toutes leurs particularités. L'effort qu'un utilisateur est prêt à consacrer lors de l'analyse des résultats dépend sans doute des circonstances, mais le plus souvent il choisira un nombre limité de méthodes en fonction, notamment, de ses préférences et des facilités offertes par les logiciels dont il dispose. A ce sujet, il faut signaler aussi que la multiplication des tests sur un même ensemble de données augmente le risque global de première espèce.

Pour analyser l'exemple du paragraphe 4, nous avons utilisé le test basé sur les scores normaux pour la normalité, le test χ^2 de BREUSCH et PAGAN pour l'homoscédasticité et le test proposé par Minitab pour l'adéquation du modèle. Ces tests ont été complétés par l'examen des divers graphiques des résidus de la procédure %RESPLOT de Minitab. D'autres paramètres ont été calculés et plusieurs autres graphiques ont été établis. Les résultats obtenus, qui ne sont pas repris dans cette note, confirment les commentaires qui ont été faits.

A propos des transformations de variables, les méthodes décrites pour la recherche de λ (paragraphe 3.2) et α (paragraphe 3.4) constituent des méthodes objectives et générales de recherche des transformations. Comme le signalent DRAPER et SMITH [1998], le fait qu'une méthode générale d'analyse existe ne signifie pas qu'elle doit être utilisée dans tous les cas. Souvent une représentation graphique révèle clairement la nécessité d'une transformation, telle que la transformation logarithmique ou la transformation inverse. Ainsi, pour l'exemple du paragraphe 4, la transformation logarithmique est la transformation qu'un utilisateur quelque peu expérimenté aurait proposé sur la base des diagrammes de dispersion de la figure 1. L'estimation de λ ne fait que confirmer ce choix *a priori*. Pour les variables explicatives le choix *a priori* d'une transformation est sans doute moins évident.

Une certaine controverse existe également à propos des conséquences de l'estimation du paramètre λ sur l'inférence statistique. Certains auteurs estiment qu'il faut tenir compte du fait que λ a été estimé; d'autres, au contraire, consi-

dèrent que, une fois le paramètre λ estimé, l'inférence classique peut être poursuivie normalement. Des références bibliographiques à ce sujet sont données par WEISBERG [1985].

Dans le cas de transformations, la vérification des conditions d'application doit se faire sur les données transformées. Ainsi, pour une transformation logarithmique par exemple, les résidus qui doivent répondre aux conditions d'application, éventuellement après standardisation, sont les quantités :

$$e_i = \log y_i - \widehat{\log y_i},$$

$\widehat{\log y_i}$ étant l'estimation de $\log y_i$ donnée par le modèle. Par la transformation exponentielle, on peut évidemment obtenir \widehat{y}_i et calculer les résidus $(y_i - \widehat{y}_i)$. Ceux-ci ne feront cependant pas l'objet d'une étude particulière du point de vue de la normalité, de l'hétéroscédasticité, de la présence de données anormales, etc.

Par contre, on ne peut pas, en général, comparer directement les valeurs R^2 obtenues pour des ajustements faisant intervenir diverses transformations de la variable à expliquer. Une solution est de recalculer dans ce cas la valeur R^2 après retour dans l'espace initial, selon la formule proposée par KVALSETH [1985] :

$$R^2 = 1 - \left[\sum_{i=1}^n (y_i - \widehat{y}_i)^2 \right] / \left[\sum_{i=1}^n (y_i - \bar{y})^2 \right].$$

L'examen direct des valeurs de R^2 obtenues après transformation permet cependant de comparer des modèles pour lesquels la variable à expliquer est identique, comme nous l'avons fait lors de la comparaison des trois derniers modèles repris dans le tableau 1.

6. BIBLIOGRAPHIE

- ATKINSON A.C. [1981]. Two graphical displays for outlying and influential observations in regression. *Biometrika* 68, 13-20.
- BERK K.N. [1998]. Regression diagnostic plots in 3-D. *Technometrics* 40, 39-47.
- BERK K.N., BOOTH D.E. [1995]. Seeing a curve in multiple regression. *Technometrics* 37, 385-398.
- BOX G.E.P., TIDWELL P.W. [1962]. Transformations of the independent variables. *Technometrics* 4, 531-550.
- BOX G.E.P., PIERCE D.A. [1970]. Distribution of residual autocorrelations in autoregressive-integrated average time series models. *J. Amer. Stat. Assoc.* 65, 1509-1516.
- BREUSCH T.S., PAGAN A.R. [1979]. A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47, 1287-1294.
- CARROLL R.J., RUPPERT D. [1988]. *Transformations and weighting in regression*. New York, Chapman and Hall, 249 p.

- CHATTERJEE S., PRICE B. [1991]. *Regression analysis by example*. New York, Wiley, 298 p.
- COOK R.D. [1993]. Exploring partial residual plots. *Technometrics* 35, 351-362.
- COOK R.D. [1994]. On the interpretation of regression plots. *J. Amer. Stat. Assoc.* 89, 177-189.
- COOK R.D. [1996]. Added-variable plots and curvature in linear regression. *Technometrics* 38, 275-278.
- COOK R.D., WEISBERG S. [1982]. *Residuals and influence in regression*. New York, Chapman and Hall, 230 p.
- COOK R.D., WEISBERG S. [1983]. Diagnostics for heteroscedasticity in regression. *Biometrika* 70, 1-10.
- COOK R.D., WEISBERG S. [1994]. *An introduction to regression graphics*. New York, Wiley, 253 p.
- COOK R.D., WEISBERG S. [1997]. Graphics for assessing the adequacy of regression models. *J. Amer. Stat. Assoc.* 92, 490-499.
- DAGNELIE P. [1998]. *Statistique théorique et appliquée. Tome 2 : inférence statistique à une et à deux dimensions*. Bruxelles, De Boeck et Larcier, 659 p.
- DRAPER N.R., SMITH H. [1998]. *Applied regression analysis*. New York, Wiley, 706 p.
- GOLDFELD S.M., QUANDT R.E. [1965]. Some tests for homoscedasticity. *J. Amer. Stat. Assoc.* 60, 535-547.
- KIM C., STORER B.E., JEONG M. [1996]. A note on BOX-COX transformation diagnostics. *Technometrics* 38, 178-180.
- KVALSETH T.O. [1985]. Cautionary note about R^2 . *Amer. Stat.* 39, 279-285.
- LARSEN W.A., MCCLEARY S.J. [1972]. The use of partial residual plots in regression analysis. *Technometrics* 14, 781-790.
- LJUNG G.M., BOX G.E.P. [1978]. On a measure of lack of fit in time series models. *Biometrics* 65, 297-303.
- MORO S. [1995]. *Etude économétrique des variables internes qui influencent la croissance des entreprises laitières dans la Zona da Mata, Etat de Minas Gerais, Brésil* (thèse de doctorat). Gembloux, Faculté des Sciences agronomiques, 274 p.
- PALM R. [1987]. Etude des séries chronologiques par la méthode de BOX et JENKINS. *Notes Stat. Inform.* (Gembloux) 87/2, 40 p.
- PALM R. [1988]. Les critères de validation des équations de régression linéaire. *Notes Stat. Inform.* (Gembloux) 88/1, 27 p.
- PALM R. [1994]. La régression linéaire pondérée: principes et applications. *Notes Stat. Inform.* (Gembloux) 94/4, 20 p.
- PFÄFFENBERGER R.C., DIELMAN T.E. [1991]. Testing normality of regression disturbances. *Comput. Stat. Data Anal.* 11, 265-273.

- ROYSTON J.J. [1982]. An extension of SHAPIRO and WILK W test for normality to large samples. *Appl. Stat.* 31, 115-124.
- ROYSTON J.J. [1988]. SHAPIRO-WILK W statistics. In: KOTZ S., JOHNSON N.L. (edit.). *Encyclopedia of statistical sciences* (vol. 8). New York, Wiley, 430-431.
- RYAN T.P. [1997]. *Modern regression methods*. New York, Wiley, 515 p.
- RYAN T.A., JOINER B.L. [1976]. *Normal probability plots and tests for normality*. Pennsylvania State University, 12 p.
- SAS INSTITUTE INC. [1989]. *SAS/STAT. User's guide*, version 6. Fourth edition (2 volumes). Cary NC, SAS Institute Inc. 943 + 846 p.
- SAS INSTITUTE INC. [1995]. *SAS/QC^R Software: Usage and reference, Version 6* (vol. 1). Cary NC, SAS Institute Inc. 847 p.
- SHAPIRO S.S., WILK M.B. [1965]. An analysis of variance test for normality (complete samples). *Biometrika* 52, 591-611.
- SIEVERS G.L. [1986]. Probability plotting. In: KOTZ S., JOHNSON N.L. (edit.). *Encyclopedia of statistical sciences* (vol. 7). New York, Wiley, 232-237.
- WEISBERG S. [1985]. *Applied linear regression*. New York, Wiley, 324 p.
- X. [1994]. *Minitab reference manual, release 10 for windows*. PA State College, Minitab, 1047 p.
- X. [2000]. *Minitab user's guide: data analysis and quality tools, release 13 for windows*. PA State College, Minitab, 963 p.

La collection

NOTES DE STATISTIQUE ET D'INFORMATIQUE

réunit divers travaux (documents didactiques, notes techniques, rapports de recherche, publications, etc.) émanant des services de statistique et d'informatique de la Faculté des Sciences agronomiques et du Centre de Recherches agronomiques de Gembloux (Belgique).

Quelques titres récents:

PALM R. [1996]. Cartes de contrôle : les cartes de SHEWHART. *Notes Stat. Inform.* (Gembloux) 96/2, 41 p.

PALM [1996]. Cartes de contrôle : combinaison de résultats, données corrélées ou multivariées. *Notes Stat. Inform.* (Gembloux) 96/3, 37 p.

PRÉVOT H. [1997]. Introduction au système d'exploitation WINDOWS NT. *Notes Stat. Inform.* (Gembloux) 97/1, 35 p.

VAN BELLE L., CLAUSTRIAUX J.J. [1997]. Introduction à l'analyse des données par le logiciel Minitab sous Windows. *Notes Stat. Inform.* (Gembloux) 97/2, 22 p.

PRÉVOT H. [1998]. Les outils de base du réseau Internet. *Notes Stat. Inform.* (Gembloux) 98/1, 24 p.

PALM [1998]. L'analyse en composantes principales : principes et applications. *Notes Stat. Inform.* (Gembloux) 98/2, 31 p.

BROSTAUX Y. [1999]. Introduction au système d'exploitation Unix. *Notes Stat. Inform.* (Gembloux) 99/1, 15 p.

CLAUSTRIAUX J.J., IEMMA A.F. [1999]. A propos des qualificatifs complet, orthogonal et équilibré en analyse de la variance. *Notes Stat. Inform.* (Gembloux) 99/2, 14 p.

IEMMA A.F., CLAUSTRIAUX J.J. [1999]. Etude des hypothèses de l'analyse de la variance à deux critères de classification : approche par l'exemple. *Notes Stat. Inform.* (Gembloux) 99/3, 14 p.

PALM R. [1999]. L'analyse discriminante décisionnelle : principes et application. *Notes Stat. Inform.* (Gembloux) 99/4, 41 p.

PALM R. [1999]. Indices d'aptitude des procédés de production. *Notes Stat. Inform.* (Gembloux) 99/5, 26 p.

PALM R. [2000]. L'analyse de la variance multivariée et l'analyse canonique discriminante : principes et applications. *Notes Stat. Inform.* (Gembloux) 2000/1, 40 p.

Faculté universitaire des Sciences agronomiques
Avenue de la Faculté d'Agronomie 8
5030 GEMBLoux (Belgique)

D/2002/2371/1