MACROS MINITAB POUR LA RÉGRESSION LINÉAIRE

R. PALM*

1. INTRODUCTION

Malgré les développements continus des possibilités offertes par les logiciels statistiques, la mise au point d'un modèle de régression nécessite encore le plus souvent un grand nombre de manipulations pour la vérification de la validité du modèle.

L'Unité de Statistique et Informatique de la FUSAGx[1] propose aux utilisateurs du logiciel Minitab, des macros destinées à les aider dans la mise au point des modèles de régression. Ces macros peuvent être utilisées indépendamment, mais leur enchaînement dans deux macros générales permet d'obtenir, de manière automatique et dans un document de sortie condensé, une importante source d'informations utiles pour la validation des équations de régression.

Après quelques rappels théoriques généraux (paragraphe 2), nous examinerons successivement l'analyse des résidus (paragraphe 3) et les transformations de variables (paragraphe 4). Nous terminerons par un exemple (paragraphe 5) et par quelques conclusions (paragraphe 6).

L'objectif n'est pas d'étudier de façon exhaustive toutes les solutions proposées dans la littérature pour la validation des équations de régression. Des informations plus détaillées que celles présentées ici sont reprises dans une note technique [PALM et IEMMA, 2002/1]. Elles peuvent être trouvées également dans les livres relatifs à la régression, parmi lesquels nous citerons DRAPER et SMITH [1998], RYAN [1997] et WEISBERG [1982]. Nous nous limiterons à la présentation des notions qui interviennent directement dans les macros proposées. En particulier, différents choix ont dû être faits et ils ne seront pas justifiés dans cette note.

Les notices d'utilisation, les macros et la note technique mentionnée ci-dessus sont disponibles à l'adresse suivante :

`www.fsagx.ac.be/si/reglin/accueil.htm`

*Chargé de cours associé à la Faculté universitaire des Sciences agronomiques de Gembloux.
1. Faculté universitaire des Sciences agronomiques de Gembloux (Belgique).

2. RAPPELS THÉORIQUES

2.1. Modèle et conditions d'application

Soit y la variable à expliquer et $x_1, \ldots, x_p$ les p variables explicatives. Pour un individu donné, on considère le modèle suivant :

$$y_i = \beta_0 + \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + \ldots + \beta_p \, x_{ip} + \varepsilon_i \, .$$

Pour simplifier la présentation, on peut encore écrire ce modèle de la manière suivante :

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

avec :

$$\boldsymbol{x}_i = (1 \; x_{i1} \; \ldots \; x_{ip})$$

et

$$\boldsymbol{\beta}' = (\beta_0 \; \beta_1 \; \ldots \; \beta_p).$$

Pour l'ensemble des n individus d'un échantillon, on a :

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

Dans cette relation $\boldsymbol{y}$ est le vecteur des n réalisations de la variable à expliquer; $\boldsymbol{X}$ est la matrice de dimensions $n \times p'$ $(p' = p + 1)$ des valeurs observées des variables explicatives; $\boldsymbol{\beta}$ est le vecteur des p' paramètres et $\boldsymbol{\varepsilon}$ est le vecteur des n résidus théoriques et inconnus.

Les conditions d'application classiques relatives au modèle 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. Ces conditions correspondent aux conditions d'adéquation de la relation, de normalité, homoscédasticité et absence d'autocorrélation des résidus.

En plus des conditions relatives aux résidus, on considère que les variables explicatives sont connues sans erreur et que l'échantillon est un échantillon aléatoire et simple de n observations, conditionnellement aux vecteurs $\boldsymbol{x}_i$.

2.2. Estimation du modèle et résidus

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 :

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

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

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

le vecteur $\boldsymbol{e}$ étant le vecteur des résidus observés :

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X} \; \widehat{\boldsymbol{\beta}}.$$

Le vecteur $\widehat{\boldsymbol{y}}$ peut encore s'écrire :

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

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

Les éléments diagonaux de cette matrice sont égaux à :

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

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

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

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

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

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

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 :

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

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

$$r_i = e_i/\widehat{\sigma}\sqrt{(1 - h_{ii})}.$$

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

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

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

La plupart des éléments repris ci-dessus peuvent être obtenus par le logiciel Minitab et conservés dans la feuille de travail [X, 1994]. Le tableau 1 donne les sous-commandes de la commande REGRESS permettant d'obtenir ces éléments.

3. ANALYSE DES RÉSIDUS

3.1. Données anormales, exceptionnelles ou influentes

Au paragraphe 2.1, nous avons signalé que les résidus théoriques ε_i doivent être distribués selon une loi normale. Nous examinerons, au paragraphe 3.2 comment cette condition peut être vérifiée en pratique.

Tableau 1. Sous-commandes de la commande REGRESS de Minitab permettant d'enregistrer dans la feuille de travail les éléments repris dans la première colonne.

Eléments enregistrés	Sous-commandes
$\widehat{\boldsymbol{\beta}}$	COEFFICIENTS
$\boldsymbol{e}'\,\boldsymbol{e}$	MSE
e_i	FITS
r_i	SRESIDUALS
t_i	TRESIDUALS
h_{ii}	HI
D_i	COOKD

Il peut arriver cependant qu'une ou plusieurs observations présentent des résidus trop importants, en valeur absolue, alors que l'ensemble des résidus est compatible avec l'hypothèse de normalité. La mise en évidence de ces résidus anormaux peut se faire à partir des résidus standardisés t_i qui, sous les conditions d'application énoncées au paragraphe 2.1, possèdent une distribution t de STUDENT à $n - p' - 1$ degrés de liberté, p' étant le nombre de paramètres du modèle. On considère donc comme anormal tout résidu pour lequel :

$$|t_i| > t_{1-\alpha/2n}\,,$$

le niveau de signification du test, pour un résidu donné, étant fixé à α/n, car le test est appliqué n fois.

La macro NORMAL permet d'identifier ces individus.

On notera que le test ne peut être réalisé de manière rigoureuse que si, dans l'ensemble, les résidus sont normaux. On notera aussi que le niveau du test devrait être fixé à α, si l'utilisateur s'interroge *a priori*, sur la base de sa connaissance du problème, à propos du caractère éventuellement anormal d'une observation particulière.

Le caractère exceptionnel d'une observation peut également être lié au vecteur $\boldsymbol{x}$. Une mesure de l'éloignement, dans l'espace des variables explicatives, de l'observation caractérisée par le vecteur $\boldsymbol{x}_i$ par rapport au vecteur $\bar{\boldsymbol{x}}$ est donnée par la quantité h_{ii}, qui est d'autant plus grande que le vecteur $\boldsymbol{x}_i$ est éloigné de $\bar{\boldsymbol{x}}$. Plusieurs auteurs considèrent que les observations pour lesquelles h_{ii} est supérieur à $2p'/n$ ou à $3p'/n$ méritent une attention particulière. Ces limites sont cependant discutées par HOAGLIN et KEMPTHORNE [1986]. Minitab utilise, comme valeur seuil, la plus petite des deux valeurs suivantes : $3p'/n$ ou 0,99.

La caractéristique h_{ii} traduit le caractère exceptionnel d'une observation du

fait du vecteur $\boldsymbol{x}_i$, indépendamment de la valeur y_i. Elle mesure l'influence potentielle d'une observation sur les résultats de la régression. D'autres mesures sont proposées dans la littérature pour quantifier l'influence réelle d'une observation. Parmi celles-ci nous retenons la distance de COOK :

$$D_i = \frac{r_i^2}{p'} \frac{h_{ii}}{1 - h_{ii}} .$$

WEISBERG [1985] considère qu'une attention particulière doit être accordée aux observations pour lesquelles la distance de COOK est supérieure à l'unité. Une autre limite fréquemment admise est le pourcentile 50 de la distribution F de FISHER-SNEDECOR à p' et $n - p'$ degrés de liberté.

3.2. Test de normalité de RYAN et JOINER et enveloppe d'ATKINSON

Le test de normalité de RYAN et JOINER [1976] est basé sur la corrélation des résidus et des scores normaux correspondants. Ce test, réalisé par la macro NORMPLOT de Minitab, est équivalent au test de SHAPIRO et WILK [1965], considéré comme l'un des plus performants [ROYSTON, 1982; 1988].

Lorsque l'effectif de l'échantillon est faible, le test de normalité des résidus ε_i à partir des e_i ne peut qu'être approximatif car les propriétés des ε_i ne sont pas identiques aux propriétés des e_i. En particulier, les résidus e_i ne sont pas indépendants et ont la propriété de supernormalité, qui fait que les e_i peuvent avoir une distribution proche de la normale, même lorsque les ε_i ne sont pas normaux.

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 $\boldsymbol{u}_k$ $(k = 1, \ldots, 19)$, ces vecteurs. On calcule ensuite les régressions de $\boldsymbol{u}_k$ en fonction de $\boldsymbol{X}$ et on enregistre les vecteurs des résidus, notés $\boldsymbol{v}_k$. 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 $2, 3, \ldots, n$. Les maximums et minimums ainsi retenus sont des estimations des pourcentiles 5 et 95 de la distribution d'échantillonnage du $i^{ème}$ 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.

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.

Les enveloppes d'ATKINSON peuvent être simulées par la macro ENVATK, le nombre d'échantillons simulés et le degré de confiance étant fixés par l'utilisateur.

3.3. Test d'homoscédasticité de BREUSCH et PAGAN

L'égalité des variances conditionnelles peut être vérifiée par le test de BREUSCH et PAGAN [1979], également connu sous le nom de test de COOK et WESISBERG [1983].

Pour une seule variable explicative, le test consiste à vérifier la nullité de λ dans le modèle de variance conditionnelle suivant :

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

A partir des résidus de la régression, e_i, on définit une nouvelle variable $e_i^{'}$:

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

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

$$\chi^2_{obs} = \text{SCE}_{rég}/2,$$

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

Dans le cas de plusieurs variables explicatives, la régression de e' en fonction de x peut être remplacée par la régression de e' en fonction de $\widehat{y}$, ou par une régression multiple en fonction de toutes ou d'un sous-ensemble de variables explicatives. Sous l'hypothèse d'homoscédasticité, la quantité χ^2_{obs} suit alors une distribution χ^2 à k degrés de liberté, k étant le nombre de paramètres de l'équation, à l'exclusion du terme indépendant.

Le test de BREUSCH et PAGAN est réalisé par la macro BREUPAG.

3.4. La macro RESANA

L'enchaînement des macros citées dans les paragraphes 3.1 à 3.3 est réalisé par la macro RESANA qui :

- calcule la régression,
- identifie les données anormales, exceptionnelles et influentes,
- établit le graphique des scores normaux,

– réalise les tests de RYAN et JOINER et de BREUSCH et PAGAN.

Lorsque le nombre de données est inférieur à 30, le graphique des scores normaux est complété par l'enveloppe simulée d'ATKINSON.

La macro refait une seconde fois l'ensemble des calculs, après suppression des données anormales, exceptionnelles ou influentes, c'est-à-dire celles pour lesquelles une des trois conditions suivantes au moins est réalisée :

$|t_i| > t_{1-\alpha/2n}$ à $n - p' - 1$ degrés de liberté,

$h_{ii} > k\,p'/n$,

$D_i > F_{0,5}$ à p' et $n - p$ degrés de liberté.

Par défaut, la valeur α est fixée à 0,05 et la valeur k à 3, mais ces valeurs peuvent être modifiées par l'utilisateur. Différentes options permettent également de supprimer une partie des informations.

4. TRANSFORMATION DE VARIABLES

4.1. Transformation de la variable à expliquer

Une transformation de la variable à expliquer affecte l'adéquation de la relation, la normalité et l'homoscédasticité des résidus. En l'absence d'une base théorique permettant de sélectionner une transformation particulière susceptible d'améliorer les conditions d'application, on peut retenir la transformation de BOX et COX, dont une des formulations est la suivante :

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

$\bar{x}_g$ étant la moyenne géométrique de y.

La valeur λ est définie à partir des observations elles-mêmes. 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 formule donnée ci-dessus et on détermine l'équation de régression :

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

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

$$L(\lambda) = -\frac{n}{2}\log_e \mathrm{SCE}_r(\lambda).$$

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, $\widehat{\lambda}$, 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.

La macro BOXCOXLL calcule le logarithme de la vraisemblance pour des valeurs de $\widehat{\lambda}$ variant dans un intervalle fixé par l'utilisateur.

Une autre méthode de détermination de λ a été proposée par ATKINSON [1981]. Elle nécessite le calcul d'une variable explicative nouvelle, g, fonction de y :

$$g = y\,[\log_e(y/\bar{y}_g) - 1] + \log_e \bar{y}_g + 1$$

et le calcul de l'équation de régression multiple prenant en compte cette variable explicative supplémentaire :

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

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

$$\widehat{\lambda} = 1 - \widehat{\phi}\,.$$

La macro BOXCOXAT détermine la valeur de $\widehat{\lambda}$ selon la méthode décrite ci-dessus.

En pratique, la valeur de λ obtenue par le maximum de vraisemblance ou par la méthode d'ATKINSON est généralement arrondie, de manière à retrouver une transformation plus classique (transformation racine carrée, logarithmique, inverse, etc.). D'autre part, lorsque la valeur retenue pour λ est différente de 0 ou de 1, on utilise, par la suite, la transformation y^λ, la transformation plus compliquée, $w(\lambda)$, donnée ci-dessus ne présentant aucun avantage supplémentaire, une fois la valeur de λ fixée. Si $\widehat{\lambda} = 0$, on utilise indifféremment le logarithme naturel ou le logarithme décimal et on néglige la constante $\bar{y}_g$ de la formule initiale.

Enfin, 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.

4.2. Transformation des variables explicatives

Comme pour la variable à expliquer, la transformation de BOX et COX peut être utilisée pour transformer une variable explicative particulère x_j. Sous sa forme simplifiée, cette transformation s'écrit :

$$z_j = \begin{cases} x_j^\alpha & \text{si } \alpha \neq 0 \\ \log_e x_j & \text{si } \alpha = 0 \end{cases}$$

En pratique, il faut décider si une transformation est utile et, dans l'affirmative, il faut fixer la valeur de α.

Supposons qu'on s'intéresse à la transformation de la première variable explicative, x_1. On considère, d'une part, le modèle de régression à p variables suivant :

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon \qquad (1)$$

et, d'autre part, le modèle à $p+1$ variables :

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \gamma x_1 \log_e x_1 \varepsilon \qquad (2)$$

Le test classique de signification appliqué au coefficient γ du second modèle permet de juger de l'intérêt de la transformation de x_1 et, lorsque la transformation s'avère utile, une estimation de α est donnée par :

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

$\widehat{\gamma}$ étant obtenu par le modèle (2) et $\widehat{\beta}_1$ étant la valeur estimée lors de l'ajustement du modèle (1).

Ce résultat provient d'un développement en série de Taylor et 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 $\alpha_2^{'}$, 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 le coefficient γ est non significatif.

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] analyse, par contre, simultanément la transformation de toutes les variables en calculant, à chaque itération, les p variables $x_j' \log_e x_j'$ $(j = 1, \ldots, p)$.

La macro BOXTID1 procède à l'étude séquentielle des variables explicatives, alors que la macro BOXCOXP réalise simultanément la transformation des p variables.

4.3. La macro BOCOTI

Pour un ensemble de valeurs λ fixées par l'utilisateur, la macro BOCOTI réalise la transformation de Box et Cox de la variable à expliquer et, en option, la transformation de Box et Tidwell des p variables explicatives, considérées simultanément. Les différents paramètres synthétiques suivants sont calculés :

– le logarithme de la vraisemblance;

– le coefficient de détermination, calculé sur la variable transformée :

$$\mathrm{R}^2(w) = 1 - \sum_{i=1}^{n}(w_i - \widehat{w}_i^2) \Big/ \sum_{i=1}^{n}(w_i - \bar{w})^2 \,;$$

– le coefficient $\mathrm{R}^2(y)$ obtenu par la relation :

$$\mathrm{R}^2(y) = 1 - \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \Big/ \sum_{i=1}^{n}(y_i - \bar{y})^2 \,;$$

– la valeur χ^2_{obs} du test de Breusch et Pagan effectué sur les données transformées en considérant que la variance conditionnelle est fonction des valeurs estimées de la variable à expliquer après transformation éventuelle, $\widehat{w}_i$;

– la corrélation entre les résidus et les scores normaux calculés pour la variable transformée, w.

L'utilisation de cette macro permet de choisir la valeur λ de la transformation de Box et Cox, en tenant compte, non seulement de la vraisemblance, mais aussi de l'adéquation de la relation, de la normalité et de l'homoscédasticité des résidus après transformation, le choix résultant, en pratique, le plus souvent d'une forme de compromis.

Les coefficients $\mathrm{R}^2(w)$ et $\mathrm{R}^2(y)$ mesurent des choses différentes et doivent être interprétés de façon adéquate. Le coefficient $\mathrm{R}^2(w)$ mesure la part de la variance de la variable w, c'est-à-dire de la variable y après transformation, qui est expliquée par les variables explicatives. Les valeurs ne peuvent donc être comparées que pour une même valeur de λ. Dans ce cas, la comparaison de $\mathrm{R}^2(w)$ avant et après transformation des x_j permet de juger de l'intérêt de la transformation des variables explicatives.

Les coefficients $\mathrm{R}^2(y)$ peuvent, par contre, être comparés pour les différentes valeurs de λ. On notera qu'ils sont négatifs lorsque la somme des carrés des résidus $y_i - \widehat{y}_i$ est supérieure à la somme des carrés des écarts $y_i - \bar{y}$. Les valeurs négatives de ce paramètre correspondent donc à de très mauvaises relations.

5. APPLICATION

5.1. Données

Pour illustrer l'utilisation des macros présentées dans les paragraphes précédents, nous reprenons l'exemple traité dans Palm et Iemma [2002/1], relatif à la construction d'un modèle de croissance de la taille des exploitations laitières d'une région du Brésil, à partir d'observations réalisées sur 162 exploitations [Moro, 1995].

La variable à expliquer est le taux de croissance (TAUX), défini comme le rapport entre le nombre de vaches en 1992 et le nombre de vaches en 1986. Les trois variables explicatives sont :

- le nombre annuel moyen de visites effectuées par les techniciens, augmenté d'une unité pour éviter les valeurs nulles (VULG);
- le nombre de vaches laitières en 1986 (UGB86);
- la valeur des actifs, en milliers de dollars (VACTIFS).

Nous allons tout d'abord utiliser la macro BOCOTI, afin de vérifier l'intérêt de transformations de variables. Ensuite, pour le modèle retenu, nous procéderons à une analyse plus approfondie des résultats avec la macro RESANA.

5.2. Résultats de la macro BOCOTI

La figure 1 reprend les résultats de l'exécution de cette macro. Pour réduire la taille de la figure, le pas de variation de λ a volontairement été fixé à 0,5. On a vérifié que l'utilisation d'un pas plus petit ne change pas les conclusions. Le nombre maximum d'itérations pour la recherche de la transformation de Box et Tidwell a été fixé à 5.

En l'absence de transformations de variables explicatives, on constate que la valeur de λ qui maximise la vraisemblance est égale à zéro, suggérant l'intérêt de la transformation logarithmique. Pour $\lambda = 0$, la valeur χ^2_{obs} relative au test de Breusch et Pagan est nulle et la corrélation entre les résidus et les scores normaux, r_{ns}, est égale à 0,997, valeurs correspondant respectivement au minimum et au maximum observé. La transformation logarithmique améliore donc très nettement les conditions d'application puisque, pour $\lambda = 1$ on a $\chi^2_{obs} = 21,7$ et $r_{ns} = 0,961$.

On constate aussi que la transformation de la première variable permet d'améliorer légèrement l'adéquation de la relation. En effet, pour $\lambda = 0$, $\mathrm{R}^2(w)$ est égal à 0,46 en l'absence de transformation de la variable VULG et de 0,54 après transformation de celle-ci. De même, $\mathrm{R}^2(y)$ passe de 0,31 à 0,42 après transformation de la variable VULG. La valeur $\alpha = -0,125$ obtenue pour $\lambda = 0$ suggère, ici aussi, la transformation logarithmique de la variable VULG.

5.3. Utilisation de la macro RESANA

Nous considérons donc le modèle suivant :

$$\log_e \mathrm{TAUX} = \widehat{\beta}_0 + \widehat{\beta}_1 \log_e \mathrm{VULG} + \widehat{\beta}_2 \,\mathrm{UGB86} + \widehat{\beta}_3 \,\mathrm{VACTIFS} + e\,.$$

La figure 2 reprend une partie des résultats obtenus par la macro RESANA.

On constate que, pour ce modèle, le test de non linéarité, réalisé par la commande expérimentale XLOF de Minitab, ne détecte pas de non-linéarité.

LAMBDA	LOGLIK	R2W	R2Y	CHI2	RNS
-2.00	-492.3	0.3670	0.0276	266.0	0.8064
-1.50	-419.0	0.4256	0.0402	144.3	0.8863
-1.00	-364.9	0.4678	-1.1415	52.9	0.9529
-0.50	-333.1	0.4806	0.2098	9.6	0.9880
0.00	-323.4	0.4616	0.3126	0.0	0.9968
0.50	-333.5	0.4167	0.3516	7.0	0.9878
1.00	-360.2	0.3562	0.3544	21.7	0.9610
1.50	-400.5	0.2918	0.3384	39.6	0.9185
2.00	-451.7	0.2330	0.2791	57.6	0.8679

Transformation de BOX-TIDWELL

LAMBDA	LOGLIK	R2W	R2Y	CHI2	RNS
-2.00	-488.5	0.3966	0.0998	208.9	0.7681
-1.50	-411.5	0.4767	-2.1939	138.3	0.8729
-1.00	-354.0	0.5348	0.1953	57.6	0.9558
-0.50	-322.0	0.5468	0.4009	14.7	0.9869
0.00	-311.6	0.5351	0.4174	1.7	0.9961
0.50	-324.1	0.4804	0.4243	2.8	0.9815
1.00	-351.7	0.4203	0.4203	36.3	0.9687
1.50	-393.0	0.3541	0.3954	52.1	0.9263
2.00	-444.9	0.2948	0.3640	98.1	0.8867

LAMBDA ALPHA.1-ALPHA.p

Row	LAMBDA	ALPHA.1	ALPHA.2	ALPHA.3
1	-2.0	-1.02950	2.25339	-2.43972
2	-1.5	0.18164	2.05258	-1.27797
3	-1.0	0.16633	1.10988	-1.86813
4	-0.5	-0.18101	1.00000	-1.07552
5	0.0	-0.12539	1.00000	1.00000
6	0.5	-0.05548	1.00000	1.00000
7	1.0	0.40468	-0.19082	1.00000
8	1.5	0.38400	0.37211	1.00000
9	2.0	0.37894	-0.47544	1.00000

Figure 1. Résultats de la macro BOCOTI.

```
The regression equation is LTAUX = 0.129 + 0.245 LVULG - 0.00917
UGB86 - 0.00156 VACTIFS

Predictor        Coef     SE Coef          T        P
Constant      0.12891     0.07370       1.75    0.082
LVULG         0.24491     0.02607       9.39    0.000
UGB86       -0.009169    0.002082      -4.40    0.000
VACTIFS    -0.0015597   0.0007246      -2.15    0.033

S = 0.3803      R-Sq = 53.8%     R-Sq(adj) = 52.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     26.6399      8.8800     61.38    0.000
Residual Error   158     22.8566      0.1447
Total            161     49.4966

No evidence of lack of fit (P > 0.1)

   Number of nonmissing in ROUT = 4

   4 DONNEES TELLES QUE HI> 0.07

   0 DONNEES TELLES QUE D>0.84 (  4 ET 158 d.l)

DONNEES INFLUENTES OU EXCEPTIONNELLES (ID Y X.1-X.P R H D

Row IDOUT    YOUT  XOUT.1 XOUT.2 XOUT.3    ROUT     HOUT      DOUT

  1  124  0.00000 0.00000     30    220  1.36659  0.113876  0.060000
  2  153 -0.13926 0.00000     63    300  2.21084  0.145362  0.207837
  3  161 -0.23572 2.56495    127    340  2.03974  0.181426  0.230532
  4  162 -1.17118 0.00000    130    120  0.24214  0.264403  0.005269


PAS DE DONNEES TELLES QUE t > t théorique =3.690 (dl= 157)

Corrélation entre résidus et scores normaux et valeurs théoriques:
(après suppression des données anormales éventuelles) cor= 0.99650
cor10= 0.99309  cor05= 0.99167  cor01= 0.98841


TEST DE BREUSCH et PAGAN : chi2-Z =    1.2  Prob = 0.28077
```

Figure 2. Résultats de la macro RESANA.

On constate aussi que quatre observations présentent des valeurs de h_{ii} telle que :

$$h_{ii} > 3\,p'/n = (3)(4)/162 = 0,0754\,.$$

Il s'agit des observations 124, 153, 161 et 162. Le diagramme de dispersion de UGB86 et VACTIFS, non repris ici, montre clairement le caractère exceptionnel de ces quatre observations, qui ne sont cependant pas particulièrement influentes, puisque les valeurs D_i sont largement inférieures à $F_{0,5}$ à $p' = 4$ et $n - p' = 158$ degrés de liberté, qui vaut 0,84.

On constate aussi qu'aucun résidu ne doit être considéré comme aberrant, car les résidus standardisés, t_i, sont tous inférieurs à $t_{0,9998}$ à $n - p' - 1 = 157$ degrés de liberté.

Enfin, le test de RYAN et JOINER et le test de BREUSCH et PAGAN ne montrent aucun signe de non-normalité ou d'hétéroscédasticité.

Dans la mesure où les quatre données exceptionnelles repérées ci-dessus ne sont pas influentes et où rien ne nous permet de mettre en doute l'exactitude de ces données, il ne nous semble pas opportun de les éliminer. Pour cette raison, les résultats du calcul de la seconde régression, après l'élimination des quatre données en question, qui sont fournis par RESANA, ne sont pas repris dans la figure 2. Les résultats sont en fait très proches de ceux de la figure 2, le point le plus important étant sans doute l'identification de huit nouvelles observations considérées comme exceptionnelles du point de vue de h_{ii}. La détection d'un second ensemble de données exceptionnelles après suppression d'un premier ensemble s'explique par le caractère nettement dissymétrique des variables UGB86 et VACTIFS.

6. CONCLUSIONS

Les différentes macros proposées devraient permettre à l'utilisateur de vérifier, avec rapidité et facilité, les résultats d'un modèle de régression et de l'orienter, si nécessaire, vers des modèles alternatifs, par des transformations de variables.

Comme nous l'avons signalé dans l'introduction, les solutions qui sont proposées ne sont pas les seules solutions envisageables. Les choix qui ont été faits peuvent être discutés. Pour les tests de normalité et d'hétérogénéité notamment, de nombreuses solutions alternatives existent.

D'autre part, lorsque les conditions d'application ne sont pas remplies, le recours aux transformations de BOX et COX n'est pas l'unique solution. Ainsi, par exemple, en présence d'hétéroscédasticité, l'utilisation de la régression pondérée peut être envisagée, en particulier si la nature de la relation est adéquate et si les résidus sont normaux.

Bien qu'ils soient susceptibles d'orienter l'utilisateur dans le choix d'une transformation, les outils proposés ne doivent pas être utilisés de manière aveugle. En particulier, la transformation de BOX et TIDWELL des variables explicatives

ne conduit pas toujours à des résultats réalistes, c'est-à-dire à des coefficients α compris entre -2 et 2. C'est d'ailleurs la raison pour laquelle, dans les macros, les coefficients sont fixés à l'unité lorsque la valeur estimée dépasse 3, en valeur absolue.

Rappelons aussi que, de manière générale, les macros se basent sur un modèle linéaire avec ordonnée à l'origine. Pour un modèle passant par l'origine, diverses modifications devraient être apportées aux macros, notamment chaque fois que la commande REGRESS est utilisée ainsi que lors du calcul des degrés de liberté pour les valeurs théoriques relatives aux distributions t (test du caractère anormal d'un résidu) ou lors du calcul des valeurs seuils pour les h_{ii} et D_i, dans la recherche des données exceptionnelles et influentes.

7. BIBLIOGRAPHIE

Atkinson A.C. [1981]. Two graphical displays for outlying and influential observations in regression. *Biometrika* 68, 13-20.

Box G.E.P., Tidwell P.W. [1962]. Transformations of the independent variables. *Technometrics* 4, 531-550.

Breusch T.S., Pagan A.R. [1979]. A simple test for heteroscedasticity and random coefficient variation. *Econometrica* 47, 1287-1294.

Cook R.D., Weisberg S. [1983]. Diagnostics for heteroscedasticity in regression. *Biometrika* 70, 1-10.

Draper N.R., Smith H. [1998]. *Applied regression analysis.* New York, Wiley, 706 p.

Hoaglin D.G., Kempthorne P.J. [1986]. Comment. *in* Chatterjee, Hadi : influential observations, high leverage points and outliers in linear regression. *Stat. Sci.* 1, 408-412.

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., Iemma A.I. [2002]. Conditions d'application et transformations de variables en régression linéaire. *Notes Stat. Inform.* (Gembloux) 2002/1, 34 p.

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.

SHAPIRO S.S., WILK M.B. [1965]. An analysis of variance test for normality (complete samples). *Biometrika* 52, 591-611.

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.