RéMuS

Manuel Théorique

par

Luc Perreault

Pierre Bruneau

Louis Mathier

Bernard Bobée

Hugues Perron

INRS-Eau | **Hydro-Québec**

Janvier 1993

TABLE DES MATIÈRES

CHAPITRE I

INTRODUCTION

1.1 Problématique

Un des outils priviligiés par Hydro-Québec pour gérer efficacement les ressources hydriques est la simulation de production énergétique aux différents sites du réseau. La simulation à un site donné utilise les débits mensuels moyens calculés à partir des débits enregistrés quotidiennement. Or, certains débits mensuels moyens peuvent être manquants. En effet, Hydro-Québec ne peut calculer cette donnée si plusieurs mesures de débits quotidiens sont manquantes ou si elles sont jugées erronées. Pour certains sites, les débits moyens mensuels peuvent même être manquants pendant plusieurs mois consécutifs. C'est pourquoi il importe de reconstituer ces valeurs manquantes afin que les modèles de simulation de production énergétique donnent des prévisions fiables.

Pour reconstituer les débits mensuels manquants à un site donné, Hydro-Québec utilise actuellement le logiciel ***REMUL*** qui est une adaptation du programme ***HEC-4*** développé aux États-Unis par le Corps des Ingénieurs de l'Armée des États-Unis (Beard, 1971). Le logiciel ***REMUL*** permet, à l'aide de régressions multiples effectuées sur le logarithme des données, d'estimer les données manquantes à un site sur une base mensuelle à partir des données de p sites voisins. Ainsi, l'ensemble des données sur lesquelles le modèle est ajusté comprend le logarithme des débits moyens d'un mois donné à $p+1$ sites mesurés pour n années concommittantes. Suite à l'ajustement du modèle de régression, l'équation générale utilisée par ***REMUL*** pour reconstituer le débit moyen manquant d'une année donnée à l'aide des p mesures disponibles aux autres sites, peut s'exprimer de la façon suivante :

$$\ln q_i = \hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}_k \ln Q_{ik} + \sqrt{1-R^2}\, u_i \qquad (1.1)$$

où:

- $\ln q_i$ est l'estimation du logarithme du débit mensuel manquant de l'année i au site considéré;
- $\ln Q_{ik}$ est le logarithme du débit mensuel moyen mesuré au site voisin k pour l'année i ;
- $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ sont les paramètres estimés par la régression multiple;
- R^2 est le coefficient de détermination de la régression multiple;
- u_i est un nombre aléatoire provenant d'une loi normale centrée-réduite.

Le débit moyen mensuel reconstitué q_i de l'année i fourni par ***REMUL*** est enfin obtenu par la transformation inverse, c'est-à-dire :

$$q_i = \exp\left\{\hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}_k \ln Q_{ik} + \sqrt{1-R^2}\, u_i\right\} \qquad (1.2)$$

Toutefois, le logiciel ***REMUL*** comporte plusieurs lacunes importantes reliées aux hypothèses théoriques qui doivent être vérifiées lors de l'utilisation des modèles régressifs. Entre autres,

- aucune procédure d'analyse des résidus n'est disponible pour valider le modèle;
- aucun test statistique n'est effectué pour examiner si les paramètres sont significativement différents de zéro;
- par défaut, on transforme les débits mensuels moyens à l'aide du logarithme. On suppose donc a priori, sans vérification, que les données proviennent d'une loi lognormale;
- aucune correction ou méthode alternative n'est disponible pour tenir compte d'un éventuel problème de multicollinéarité, c'est-à-dire d'une possible corrélation entre les données des sites utilisés comme variables explicatives dans le modèle;

- l'équation (1.2) ne permet pas de conserver la structure de corrélation qui pourrait exister entre différents sites à reconstituer.

L'utilisation de ce logiciel ne permet donc pas de vérifier la validité et l'adéquation du modèle. Ces lacunes peuvent amener l'usager à faire une reconstitution et une interprétation erronées des résultats. C'est pourquoi, dans le cadre d'un projet de partenariat financé par Hydro-Québec et subventionné par le CRSNG, l'INRS-Eau a eu le mandat de développer le logiciel ***RéMuS***. Ce nouveau logiciel, beaucoup plus souple que ***REMUL***, permet à l'usager d'effectuer quatre types de régressions (multiple, ridge, pas à pas et multivariée) répondant aux différentes contraintes pratiques que rencontrent les hydrologues d'Hydro-Québec. De plus, pour tous ces modèles régressifs, l'usager peut effectuer des tests graphiques et statistiques sur les données et les paramètres, ainsi qu'une analyse complète des résidus. L'hydrologue peut donc avoir une idée précise de la qualité des données reconstituées et des limites du modèle qu'il a choisi. Enfin, ***RéMuS*** donne les débits reconstitués accompagnés de statistiques et de tests qui permettent d'apprécier la qualité des résultats.

1.2 Le logiciel *RéMuS*

Le logiciel ***RéMuS*** est composé de quatre modules indépendants comme le montre la Figure 1.1. Le module *Configuration* contient les options permettant de signaler au logiciel les périphériques, les répertoires et les couleurs d'écran désirés par l'usager. Le module *Format* permet à l'usager de spécifier la structure du fichier source où se trouvent les données originales. Le module *Initialisation* contient des options permettant à l'usager de sélectionner 21 variables, c'est-à-dire 21 sites (incluant le site où l'on désire effectuer la reconstitution), qui peuvent être utilisées lors de la modélisation. Enfin, le module *Modélisation* permet d'effectuer les quatre types de régression linéaire disponibles, un ensemble de tests statistiques et graphiques, et la reconstitution de données. Le guide de l'usager (Perron, 1993a) et le guide du programmeur (Perron, 1993b) donnent respectivement les détails techniques et informatiques concernant la structure et le fonctionnement du logiciel ***RéMuS***.

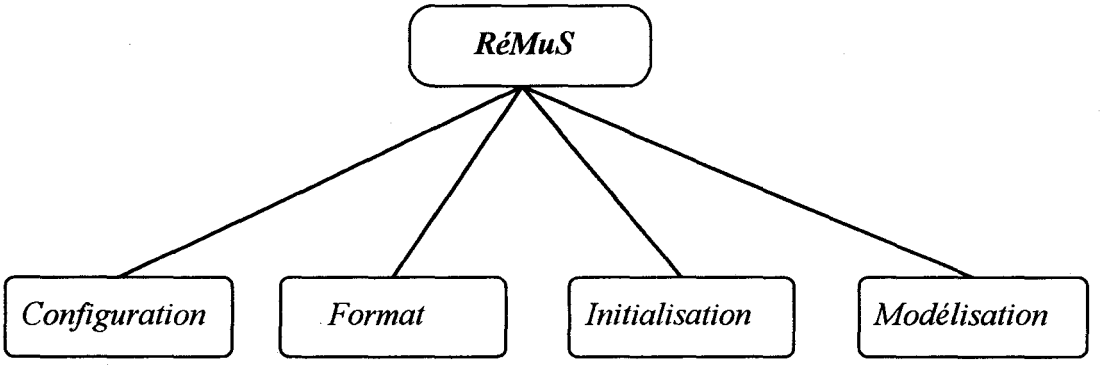


Figure 1.1. Structure du logiciel ***RéMuS***

Dans le présent rapport, les différentes méthodes statistiques disponibles dans le module *Modélisation* sont décrites. Les quatre types de régressions linéaires disponibles dans ***RéMuS*** sont: la régression multiple, la régression ridge, la régression pas à pas et la régression multivariée. Ces méthodes sont accompagnées d'options permettant à l'usager d'effectuer des tests statistiques et graphiques, des transformations de variables, une étude complète des résidus et enfin, une reconstitution des données manquantes.

Le chapitre II traite des principaux fondements théoriques de la régression linéaire multiple. On y présente le modèle, l'estimation des paramètres ainsi que les tests effectués sur ces derniers (tests de Student et de Fisher). Une section de ce chapitre est aussi consacrée à l'étude des résidus du modèle (tests graphiques et statistiques pour vérifier les hypothèses de base de la régression). Enfin, nous présentons l'équation générale permettant de reconstituer les données et les tests disponibles dans ***RéMuS*** pour vérifier si la moyenne et la variance des mesures originales sont conservées (tests de Wilcoxon et de Levene). Dans le chapitre III, nous abordons le problème de la multicollinéarité et de ses effets sur les résultats de la régression multiple. Nous présentons ensuite la régression ridge qui s'avère une alternative au modèle classique pour pallier ce problème. Le chapitre IV est consacré à la régression pas à pas (stepwise regression) qui permet une sélection automatique des variables explicatives les plus pertinentes. La régression pas à pas est un outil intéressant d'exploration des données lorsque l'usager doit effectuer un grand nombre de régressions ou lorsque le nombre de variables explicatives est important. Le dernier chapitre traite de l'utilisation de la régression multivariée dans une perspective d'estimation régionale. On y aborde le problème de la conservation de la structure de corrélation qui existe entre plusieurs sites d'une même région (corrélations spatiales) pour lesquels on veut reconstituer des données. On montre que le modèle de régression multivariée permet de conserver cette structure lors de la reconstitution simultanée des débits à plusieurs sites. Ce modèle est décrit en détail dans le chapitre IV.

CHAPITRE II

REGRESSION MULTIPLE

2.1 Introduction

Dans bien des domaines, la régression linéaire multiple est l'une des méthodes statistiques les plus utilisées pour étudier la relation entre une variable dépendante et plusieurs variables indépendantes (ou variables explicatives). Le problème qui nous concerne ici va au-delà de l'étude de l'association entre différentes variables (débits moyens mensuels à différents sites)[1]. En effet, nous désirons non seulement établir la relation entre les variables à l'étude (variable dépendante et plusieurs variables explicatives), mais aussi utiliser cette relation à des fins de reconstitution.

En théorie statistique, la régression multiple a fait l'objet de nombreux articles scientifiques et d'ouvrages très complets. C'est pourquoi, nous ne donnons pas ici une présentation exhaustive des techniques reliées à la régression multiple mais plutôt une brève description des méthodes implantées dans le logiciel ***RéMuS***. Pour plus de détails sur le sujet, nous invitons le lecteur à consulter Neter *et al.* (1985) et Draper et Smith (1966).

Dans ce chapitre, nous présentons le modèle général de régression linéaire multiple ainsi que les hypothèses sous-jacentes à ce type d'analyse. Nous examinons les aspects théoriques concernant l'estimation des paramètres et les techniques statistiques permettant de faire de l'inférence sur ces derniers. Enfin, nous traitons du problème de l'adéquation du modèle aux données observées ainsi que du problème de la reconstitution des données manquantes.

---

[1]Nous entendons par variables, tout au long du présent rapport, les débits mensuels moyens aux différents sites ou tout autre type de mesures qui peuvent faire l'objet d'une analyse de régression multiple.

2.2 Le modèle de régression linéaire multiple

Considérons une variable dépendante Y et p variables explicatives supposées indépendantes $X_1, X_2, \ldots, X_p$. Alors, le modèle général de régression multiple est de la forme :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon \tag{2.1}$$

où ε est une variable aléatoire distribuée selon une loi normale.

On suppose que les réalisations des variables explicatives $X_1, X_2, \ldots, X_p$, que l'on note $x_1, x_2, \ldots, x_p$, sont connues exactement. On note aussi la réalisation de la variable dépendante Y par y. Ainsi, pour une série particulière de l'ensemble des variables (une réalisation donnée), le modèle s'écrit :

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i \tag{2.2}$$

ou encore

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i \tag{2.3}$$

où :

- l'indice $i = 1, 2, \ldots, n$ réfère à la série particulière de l'ensemble des $p + 1$ variables, par exemple les observations d'une année i donnée;
- n désigne la taille d'échantillon. Dans ***RéMuS***, n est par exemple le nombre d'années pour lesquelles toutes les stations considérées dans le modèle possèdent une mesure. On appelle alors n le nombre de données concomitantes.
- $\beta_0, \beta_1, \ldots, \beta_p$ sont les paramètres de la régression multiple;
- ε_i, $i = 1, 2, \ldots, n$, sont des variables aléatoires indépendantes et identiquement distribuées selon une loi normale centrée de variance σ^2.

Le modèle (2.1) est dit *multiple* puisqu'il fait intervenir plus d'une variable explicative, et *linéaire* parce que celles-ci apparaissent dans le modèle à la puissance 1. Trois hypothèses de base concernant les termes ε_i sont associées au modèle de régression multiple:

1. *Normalité des erreurs* : ε_i est une variable aléatoire distribuée selon une loi normale centrée de variance σ^2;

2. *Absence de corrélation des erreurs* : les termes ε_i et ε_j relatifs à deux observations i et j n'ont aucune corrélation entre eux, $Cov\{\varepsilon_i, \varepsilon_j\} = 0, \quad i \neq j$;

3. *Homoscédasticité* : la variance des ε_i est constante quelles que soient les valeurs des variables explicatives $x_{i1}, x_{i2}, \ldots, x_{ip}$, $Var\{\varepsilon_i\} = \sigma^2, \quad \forall i$.

Comme on le verra plus loin, toute l'inférence que l'on peut faire suite à une analyse de régression multiple, en particulier l'interprétation des tests effectués sur les paramètres, repose sur ces hypothèses. L'analyse des résidus (Section 2.5), après estimation des paramètres du modèle, permettra de les vérifier.

La spécification de ces hypothèses nous permet maintenant de caractériser la variable dépendante Y . On remarque que Y est la somme de deux composantes : le terme constant $\beta_0 + \sum_{k=1}^{p} \beta_k X_k$ et le terme aléatoire ε (éq. 2.1). Ainsi, Y est une variable aléatoire provenant d'une loi de probabilité, étant donnée la réalisation d'une série particulière i des variables $X_1, X_2, \ldots, X_p$, de moyenne

$$E\{Y | X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_p = x_{ip}\} = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} \tag{2.4}$$

et de variance

$$Var\{Y | X_1 = x_{i1}, X_2 = x_{i2}, \ldots, X_p = x_{ip}\} = \sigma^2 \tag{2.5}$$

On note que la variance de Y et celle de ε sont identique pour une réalisation i fixée de $X_1, X_2, \ldots, X_p$. La fonction de régression (éq. 2.4) relie donc la moyenne de la distribution conditionnelle de Y étant donnés $x_1, x_2, \ldots, x_p$ aux valeurs spécifiques $x_{i1}, x_{i2}, \ldots, x_{ip}$, et la réalisation y_i est supérieure ou inférieure d'une quantité ε_i à la valeur correspondante de la fonction de régression (éq. 2.2 et 2.4). Lorsqu'une seule variable explicative est utilisée dans le modèle de régression (régression simple), cette fonction est une droite. Les paramètres β_0 et β_1 sont alors respectivement l'ordonnée à l'origine et la pente de cette droite. Si le modèle possède

deux variables explicatives, la fonction de régression est un plan. Enfin, si plus de deux variables sont utilisées, la fonction est un hyperplan.

2.3 Estimation des paramètres

La régression multiple nécessite de nombreux calculs et le système des équations normales à résoudre pour déterminer les paramètres devient rapidement lourd à mesure que le nombre de variables explicatives augmente. Ainsi, un modèle linéaire multiple avec p variables explicatives présentera une système de $p+1$ équations à $p+1$ inconnues que l'on doit résoudre. La résolution en écriture matricielle permet d'estimer les paramètres simultanément et plus facilement par inversion de matrice, si la matrice inverse existe. Nous présentons donc ici la régression multiple sous forme matricielle.

Pour exprimer le modèle de régression multiple à p variables explicatives (éq. 2.2) sous forme matricielle, définissons les matrices suivantes :

$$\underset{n\times 1}{\mathbf{Y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \underset{n\times(p+1)}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

$$\underset{(p+1)\times 1}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \qquad \underset{n\times 1}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Le vecteur colonne $\mathbf{Y}$ contient les n valeurs observées de la variable dépendante (par exemple, les débits mensuels moyens du site à reconstituer mesurés sur n années), la matrice $\mathbf{X}$ les valeurs correspondantes des p variables explicatives (par exemple, les débits mensuels moyens aux sites voisins mesurés sur n années) et une colonne de 1. Le vecteur $\boldsymbol{\beta}$ contient les $p+1$ paramètres de la régression, et enfin le vecteur $\boldsymbol{\varepsilon}$ les n termes d'erreur aléatoire. Donc, sous forme matricielle, le modèle de régression linéaire multiple (2.2) s'exprime de la façon suivante :

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times(p+1)}{\mathbf{X}}\,\underset{(p+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}} \tag{2.6}$$

On peut aussi transposer les résultats et les hypothèses présentés à la section précédente en écriture matricielle. Ainsi, on a :

$$E\{\boldsymbol{\varepsilon}\} = \begin{bmatrix} E\{\varepsilon_1\} \\ E\{\varepsilon_2\} \\ \vdots \\ E\{\varepsilon_n\} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

$$Var\{\boldsymbol{\varepsilon}\} = \begin{bmatrix} Var\{\varepsilon_1\} & \cdots & \cdots & Cov\{\varepsilon_1, \varepsilon_n\} \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ Cov\{\varepsilon_n, \varepsilon_1\} & \cdots & \cdots & Var\{\varepsilon_n\} \end{bmatrix} = \begin{bmatrix} \sigma^2 & \cdots & \cdots & 0 \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}_n$$

où $\mathbf{I}_n$ est la matrice identité d'ordre n (matrice dont les termes sur la diagonale principale sont égaux à 1 et dont tous les autres termes sont nuls). On déduit alors que le vecteur aléatoire $\mathbf{Y}$ possède une moyenne et une matrice de variances-covariances données respectivement par :

$$E\{\mathbf{Y}\} = \mathbf{X}\boldsymbol{\beta} \quad \text{et} \quad Var\{\mathbf{Y}\} = \sigma^2 \mathbf{I}_n$$

Ainsi, en pratique, on connait la matrice $\mathbf{X}$ des observations des variables explicatives et la matrice $\mathbf{Y}$ des observations de la variable dépendante. Pour déterminer la fonction de régression, il suffit alors d'estimer les paramètres $\beta_0, \beta_1, \ldots, \beta_p$. Pour obtenir de bons estimateurs, on emploie la méthode des moindres carrés qui consiste à minimiser la somme des carrés des résidus e_i définis comme suit :

$$e_i = y_i - \left(\hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}_k x_{ik} \right) \tag{2.7}$$

$\hat{\beta}_0$ étant l'estimateur de β_0 et $\hat{\beta}_k$ l'estimateur de β_k. La fonction à minimiser, sous forme matricielle, s'exprime alors de la façon suivante :

$$Q = (\mathbf{Y} - \mathbf{Xb})'(\mathbf{Y} - \mathbf{Xb}) = \mathbf{Y'Y} - \mathbf{bX'Y} \tag{2.8}$$

où $\mathbf{b} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p)$, $\mathbf{X}'$ désignant la transposée d'une matrice $\mathbf{X}$. On peut montrer (Draper et Smith, 1966) que cette procédure nous amène à résoudre un système de p équations à p inconnues pour obtenir le vecteur des paramètres estimés $\mathbf{b}$. Ce système d'équations s'écrit :

$$(\mathbf{X'X})\mathbf{b} = \mathbf{X'Y} \tag{2.9}$$

et on déduit :

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y} \tag{2.10}$$

où $\mathbf{X}'$ est la matrice transposée de $\mathbf{X}$ et $(\mathbf{X'X})^{-1}$ est la matrice inverse de $(\mathbf{X'X})$.

Le logiciel ***RéMuS*** utilise cette expression pour calculer les paramètres de la régression multiple. L'inversion de la matrice est effectuée par élimination gaussienne (Johnson et Riess, 1982). Une analyse complète de la variabilité et quelques tests sur les paramètres sont également présentés dans le logiciel. La section qui suit traite de ces statistiques permettant d'interpréter les résultats de la régression multiple.

2.4 Analyse de la variance et inférence sur les paramètres

Suite à l'estimation des paramètres de la régression, on peut calculer le vecteur $\hat{\mathbf{Y}}$ des valeurs prédites $\hat{y}_i$, et le vecteur correspondant $\mathbf{e}$ des écarts résiduels $e_i = y_i - \hat{y}_i$. Sous forme matricielle, les valeurs prédites de la variable dépendante ainsi que les résidus s'expriment respectivement comme suit :

$$\hat{\mathbf{Y}} = \mathbf{Xb} \quad \text{et} \quad \mathbf{e} = \mathbf{Y} - \mathbf{Xb}$$

A l'aide de ces valeurs il est possible d'étudier les différentes sources de variation associées à la variable aléatoire dépendante Y. Comme dans tout ensemble d'observations utilisé en analyse statistique, il y a nécessairement une variabilité associée à Y. En effet, si toutes les observations y_i étaient identiques, donc égales à la moyenne $\bar{y}$, il n'y aurait pas lieu de faire une étude statistique. La variabilité des y_i est traditionnellement mesurée en terme d'écart par rapport à la moyenne :

$$y_i - \bar{y} \tag{2.11}$$

La mesure de la variation totale, notée *SCT*, est la somme des carrés de ces écarts :

$$SCT = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \mathbf{Y'Y} - \frac{1}{n}\mathbf{Y'1\,1'Y} \qquad \text{(variabilité totale)} \tag{2.12}$$

où **1** est un vecteur colonne de dimension n dont tous les éléments sont égaux à 1. *SCT* est ici l'abréviation de *somme des carrés totale*. Si $SCT = 0$, toutes les observations sont égales. La variabilité des y_i est d'autant plus grande que SCT est grande. En fait, cette somme de carrés mesure l'incertitude à prédire la variable *Y* lorsqu'aucune variable explicative n'est considérée.

Toutefois, lorsqu'on utilise la régression multiple, c'est-à-dire que l'on tient compte de l'information contenue dans $X_1, X_2, \ldots, X_p$ au sujet de *Y*, la variation traduisant cette incertitude est celle des observations autour de la fonction de régression. La mesure globale de la variabilité des données avec le modèle de régression est la *somme des carrés des écarts SCE* :

$$SCE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \mathbf{Y'Y} - \mathbf{b'X'Y} \quad \text{(variabilité non expliquée par la régression)} \tag{2.13}$$

Si $SCE = 0$, toutes les observations tombent sur la fonction de régression (une droite si $p = 1$, un plan si $p = 2$ et un hyperplan si $p > 2$). Plus grande est la valeur de *SCE*, plus grande est la variation des observations y_i autour de la fonction de régression. Ainsi, cette somme des carrés mesure la variation des y_i, ou l'incertitude à prédire ces valeurs, lorsque l'information provenant des variables explicatives est prise en compte.

Maintenant, quelle mesure de variabilité pourrait quantifier la différence entre SCT et SCE, et ainsi traduire le gain obtenu par le modèle de régression multiple, c'est-à-dire en considérant $X_1, X_2, \ldots, X_p$, pour expliquer la variabilité totale des y_i ? On peut montrer (Neter *et al.*, 1985) que cette différence est la *somme des carrés de la régression SCR* définie comme suit :

$$SCR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \mathbf{b'X'Y} - \frac{1}{n}\mathbf{Y'1\,1'Y} \quad \text{(variabilité expliquée par la régression)} \tag{2.14}$$

Ainsi, $SCT = SCR + SCE$ représente le partionnement de la somme des carrés totale qui est égale à la variabilité expliquée par la régression plus la variabilité résiduelle. Les écarts utilisés dans *SCR* sont tout simplement les différences entre les valeurs prédites par la fonction de régression et

leur moyenne (notons que la moyenne des valeurs prédites est égale à la moyenne des observations de la variable dépendante). Si la fonction de régression est horizontale (aucune relation entre la variable dépendante et les variables explicatives : paramètres nuls), c'est-à-dire $\hat{y}_i - \bar{y} = 0$ pour tout i, alors $SCR = 0$. Sinon, SCR est positive. Plus la valeur de SCR est grande par rapport à SCT, plus le modèle de régression contribue à expliquer la variabilité totale des observations y_i.

La Figure 2.1 illustre graphiquement les écarts utilisés dans chacune des sommes de carrés présentées ci-haut pour le cas particulier de la régression simple ($Y = b_0 + b_1 X$). Les Figures 2.1(a), 2.1(b) et 2.1(c) donnent respectivement les écarts considérés dans les somme des carrés totale, des résidus et de la régression, tandis que la Figure 2.1(d) illustre le partitionnement de *SCT*.

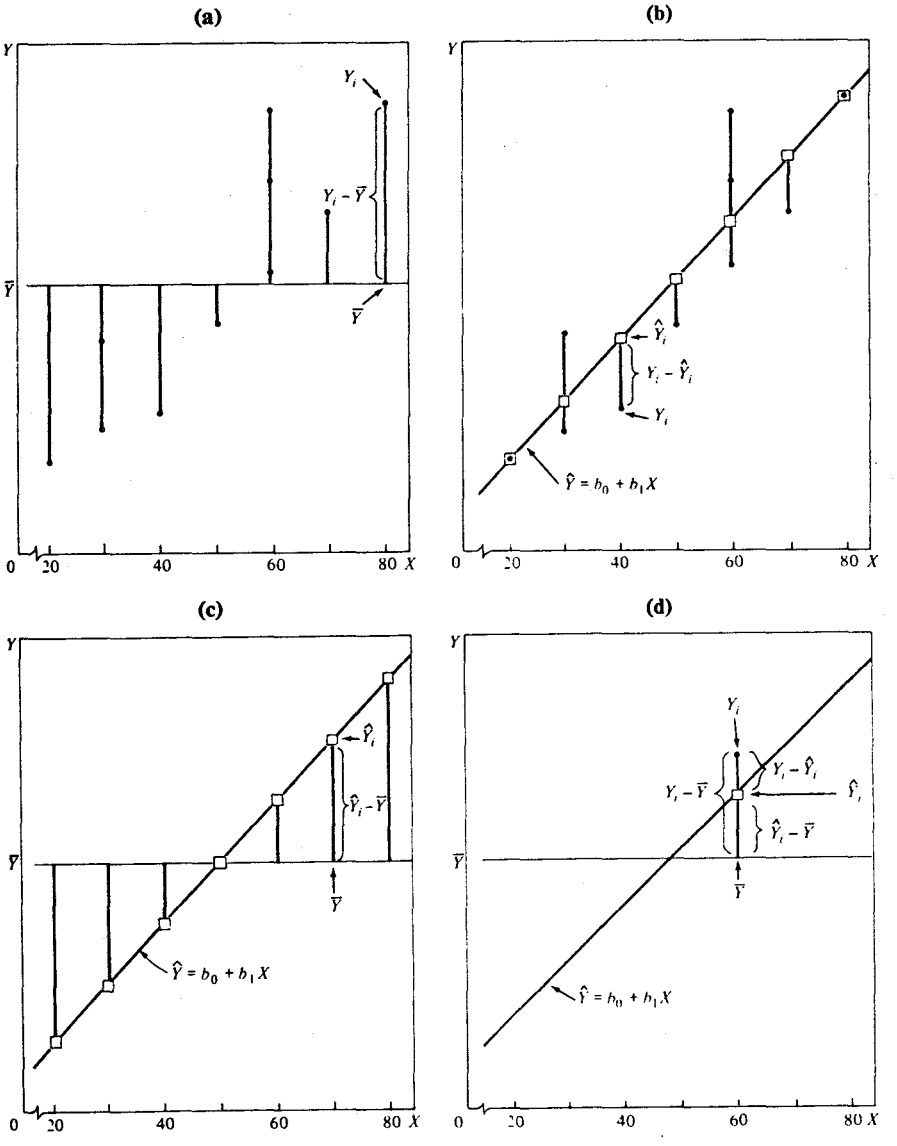


Figure 2.1. Partionnement des écarts $y_i - \bar{y}$ (cette figure est tirée de Neter et al., 1985).

Les différentes sources de variation (*SCT*, *SCE* et *SCR*) nous amènent de façon naturelle à définir une mesure permettant de quantifier l'effet de l'introduction des variables explicatives $X_1, X_2, \ldots, X_p$ dans le modèle sur la réduction de la variation totale de Y. Cette statistique est le coefficient de détermination multiple R^2 qui s'exprime de la façon suivante :

$$R^2 = \frac{SCT - SCE}{SCT} = \frac{SCR}{SCT} = 1 - \frac{SCE}{SCT} \tag{2.15}$$

Puisque $0 \leq SCE \leq SCT$, on a que $0 \leq R^2 \leq 1$. Ce coefficient peut être interprété comme le pourcentage de réduction de la variation totale obtenu en utilisant les variables explicatives $X_1, X_2, \ldots, X_p$ dans le modèle, ou simplement comme le pourcentage de variation totale expliquée par le modèle. Ainsi, plus R^2 est grand, plus l'apport des variables explicatives est important pour expliquer la variabilité totale de Y. Cette mesure permet de juger de l'adéquation du modèle. Toutefois, il est important de noter que R^2 augmente à mesure que l'on ajoute des variables explicatives au modèle, et ce même si ces variables ne sont pas significativement reliées à la variable dépendante. Cette mesure d'adéquation doit donc être utilisée avec prudence.

À partir des différentes sommes des carrés définies ci-haut, il est possible de construire un ensemble de tests statistiques vérifiant différentes hypothèses sur les paramètres du modèle. Dans ce qui suit, nous présentons les tests incorporés dans ***RéMuS*** qui permettent de faire de l'inférence sur les paramètres de la régression multiple.

Premièrement, pour examiner s'il existe globablement une relation linéaire entre la variable dépendante Y et les variables explicatives $X_1, X_2, \ldots, X_p$, c'est-à-dire tester les hypothèses :

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad \text{contre} \quad H_1 : \textit{au moins un } \beta_k \neq 0, k \neq 0$$

on utilise la statistique F :

$$F = \frac{n-p-1}{p} \frac{SCR}{SCE} \tag{2.16}$$

La règle de décision pour effectuer ce test à un niveau de signification donné α est la suivante :

- Si $F \leq F_{p,n-p-1}(1-\alpha)$, on accepte H_0 (pas de relation linéaire)
- Si $F > F_{p,n-p-1}(1-\alpha)$, on rejette H_0 (existence d'une relation linéaire)

où $F_{p,n-p-1}(1-\alpha)$ est le quantile de probabilité au non-dépassement 1 - α de la loi de Fisher à p et n - p - 1 degrés de liberté. Une table de ces quantiles est disponible dans plusieurs livres de statistique dont Neter *et al.* (1985). Elle est reproduite dans l'annexe A. En plus de fournir la valeur de la statistique F calculée, ***RéMuS*** donne la probabilité au dépassement P associée à cette valeur (P-value). Cette mesure correspond à la probabilité que la statistique du test soit supérieure à ce que l'on a calculé avec les données. Une grande valeur de cette probabilité supporte l'hypothèse H_0 alors qu'une faible valeur favorise le rejet de cette hypothèse. Il est important de noter que le rejet de l'hypothèse nulle H_0, qui signifie l'existence d'une relation linéaire, ne nous assure pas que les prédictions obtenues à partir du modèle seront satisfaisantes (voir Section 2.5, Adéquation du modèle et analyse des résidus).

Pour approfondir l'interprétation des résultats du modèle, ***RéMuS*** permet aussi d'effectuer un test sur chacun des paramètres β_k $(k = 0, 1, \ldots, p)$ indépendamment. Ce test s'appuie sur des résultats théoriques concernant les propriétés des estimateurs des moindres carrés. Avant de présenter la statistique du test, voici les principales propriétés des estimateurs $\hat{\beta}_k$ $(k = 0, 1, \ldots, p)$.

Tout d'abord, mentionnons que les estimateurs $\hat{\beta}_k$ issus de la méthode des moindres carrés sont non-biaisés et de variance minimum si les erreurs ε_i du modèle (2.2) sont indépendantes et identiquement distribuées avec une variance σ^2 (l'hypothèse de normalité n'est pas nécessaire ici). Ainsi, on a :

$$E\{\mathbf{b}\} = \begin{bmatrix} E\{\hat{\beta}_0\} \\ E\{\hat{\beta}_1\} \\ \vdots \\ E\{\hat{\beta}_p\} \end{bmatrix} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \boldsymbol{\beta} \tag{2.17}$$

De plus, on peut montrer (Neter *et al.*, 1985), sous ces hypothèses, que la matrice de variances-covariances des paramètres estimés est donnée par :

$$Var\{\mathbf{b}\} = \begin{bmatrix} Var\{\hat{\beta}_0\} & Cov\{\hat{\beta}_0,\hat{\beta}_1\} & \cdots & Cov\{\hat{\beta}_0,\hat{\beta}_p\} \\ Cov\{\hat{\beta}_1,\hat{\beta}_0\} & Var\{\hat{\beta}_1\} & \cdots & Cov\{\hat{\beta}_1,\hat{\beta}_p\} \\ \vdots & \vdots & & \vdots \\ Cov\{\hat{\beta}_p,\hat{\beta}_0\} & Cov\{\hat{\beta}_p,\hat{\beta}_1\} & \cdots & Var\{\hat{\beta}_p\} \end{bmatrix} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{2.18}$$

et que l'on peut l'estimer par :

$$s^2\{\mathbf{b}\} = \begin{bmatrix} s^2\{\hat{\beta}_0\} & s\{\hat{\beta}_0,\hat{\beta}_1\} & \cdots & s\{\hat{\beta}_0,\hat{\beta}_p\} \\ s\{\hat{\beta}_1,\hat{\beta}_0\} & s^2\{\hat{\beta}_1\} & \cdots & s\{\hat{\beta}_1,\hat{\beta}_p\} \\ \vdots & \vdots & & \vdots \\ s\{\hat{\beta}_p,\hat{\beta}_0\} & s\{\hat{\beta}_p,\hat{\beta}_1\} & \cdots & s^2\{\hat{\beta}_p\} \end{bmatrix} = \frac{SCE}{n-p-1}(\mathbf{X}'\mathbf{X})^{-1} \qquad (2.19)$$

Enfin, si l'on ajoute aux hypothèses cités plus haut que les erreurs ε_i sont distribuées selon une loi normale, on montre (Neter *et al.*, 1985) que $\left(\hat{\beta}_k - \beta_k\right)\Big/\sqrt{s^2\{\hat{\beta}_k\}}$ pour $k = 0, 1, 2, \ldots, p$ suit une loi de Student à $n - p - 1$ degrés de liberté.

Le test effectué par ***RéMuS*** sur chacun des paramètres $\beta_1, \beta_2, \ldots, \beta_p$ vérifie s'il existe une relation linéaire entre la variable dépendante et la variable explicative correspondante. Ces tests permettent donc d'examiner la pertinence de chacune des variables explicatives considérées. Plus précisément, pour une variable explicative X_k donnée, on test

$$H_0 : \beta_k = 0 \quad \text{contre} \quad H_1 : \beta_k \neq 0$$

à l'aide de la statistique t_k qui suit une loi de Student :

$$t_k = \frac{\hat{\beta}_k}{\sqrt{s^2\{\hat{\beta}_k\}}} \qquad (2.20)$$

La règle de décision pour effectuer ce test à un niveau de signification donné α est la suivante :

- Si $|t_k| \leq t_{n-p-1}(1-\alpha/2)$, on accepte H_0
- Sinon, on rejette H_0

où $t_{n-p-1}(1-\alpha/2)$ est le quantile de probabilité au non-dépassement $1-\alpha/2$ de la loi de Student à $n - p - 1$ degrés de liberté. Une table de ces quantiles est aussi disponible dans Neter *et al.* (1985). Elle est reproduite dans l'annexe B. ***RéMuS*** fournit pour chacune des variables explicatives la statistique calculée t_k ainsi que la probabilité au dépassement correspondante P_k (P-value). Un test sur le paramètre β_0 est aussi disponible. Ce test vérifie si la droite, le plan ou l'hyperplan passe par l'origine ($\beta_0 = 0$). La procédure est la même que pour les autres paramètres et ***RéMuS*** fournit la statistique calculée t_0 et la probabilité correspondante P_0.

Les résultats de la régression linéaire multiple sont présentés par ***RéMuS*** sous forme de tableaux. On y retrouve les estimations des paramètres, les statistiques des différents tests présentés dans cette section, les sommes des carrés nécessaires à l'interprétation des résultats ainsi que le coefficient de détermination. Cette représentation est illustrée à la Figure 2.2. Le guide de l'usager (Perron, 1993a) ou l'aide à l'écran du logiciel complète l'information de cette figure.

Résultats de la régression multiple

Paramètres	Estimation	Écart-type	Statistique	Prob. au dépass.
β_0	$\hat{\beta}_0$	$s^2\{\hat{\beta}_0\}$	t_0	P_0
β_1	$\hat{\beta}_1$	$s^2\{\hat{\beta}_1\}$	t_1	P_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_p	$\hat{\beta}_p$	$s^2\{\hat{\beta}_p\}$	t_p	P_p

Analyse de la variance

Source	Somme des carrés	Degrés de liberté	Somme des carrés moy.	Statistique	Prob. au dépass.
Modèle	SCR	p - 1	$SCR/(p-1)$	F	P
Erreur	SCE	n - p - 1	$SCE/(n-p-1)$		
Totale	SCT	n - 1			

$$R^2 = \frac{SCR}{SCT}$$

Figure 2.2. Les résultats de la régression tels que présentés dans ***RéMuS***.

2.5 Adéquation du modèle et analyse des résidus

Lorsqu'un modèle de régression est choisi pour une application donnée (par exemple la reconstitution de mesures de débits à un site), on ne peut généralement pas être certain à l'avance qu'il est approprié pour cette application. En effet, une ou plusieurs hypothèses du modèle comme la linéarité de la relation ou la normalité des termes d'erreur peuvent être inappropriées pour l'ensemble de données utilisé. Ainsi, il est important d'examiner la validité de ces postulats avant d'effectuer des analyses basées sur le modèle de régression multiple obtenu. Dans cette section, nous discutons de quelques méthodes graphiques et tests statistiques disponibles dans ***RéMuS*** pour valider le modèle.

Dans le modèle de régression multiple (2.2), on suppose que les ε_i sont indépendamment distribuées selon une loi normale de moyenne nulle et de variance σ^2. Si le modèle obtenu est approprié pour les observations considérées, les résidus e_i doivent vérifier les hypothèses faites sur les ε_i. Ceci est le principal fondement de l'analyse des résidus, une approche très utile permettant de juger de la validité du modèle de régression.

Comme on l'a vu plus haut, le résidu e_i est la différence entre la valeur observée y_i et la valeur correspondante $\hat{y}_i$ prédite par le modèle (éq. 2.7). On considère ici l'utilisation des résidus pour examiner quatre causes de non-respect des hypothèses du modèle de régression :

1. La fonction de régression n'est pas linéaire;
2. Les termes d'erreur ne possèdent pas une variance constante (hétéroscédasticité);
3. Les termes d'erreur ne sont pas indépendants;
4. Les termes d'erreur ne sont pas distribués selon une loi normale.

1. *Non-linéarité de la fonction de régression*

On peut généralement vérifier si une fonction de régression linéaire est appropriée pour les données en étudiant les graphiques mettant en relation les mesures de la variable dépendante en fonction des mesures des différentes variables explicatives considérées (Y vs X_k, k = 1, 2, ..., p). Ce type de graphique est disponible dans ***RéMuS***. Toutefois, le graphique des résidus en fonction des observations des variables explicatives (e vs X_k, k=1, 2, ..., p) possède en général quelques avantages. Premièrement, ce graphique peut être utilisé afin d'examiner d'autres facettes de

l'adéquation du modèle. Deuxièmement, pour certains ensembles de données, il arrive souvent que l'échelle des graphiques Y vs X_k permette difficilement d'examiner la forme de la relation. Il est alors beaucoup plus facile d'étudier le graphique des résidus.

La Figure 2.3(a) montre un exemple du type de relation entre les résidus et une variable X nous permettant de déduire qu'il existe une relation linéaire entre Y et X. En effet, si les résidus sont distribués selon une bande horizontale centrée en zéro, aucune tendance systématique ne ressort du graphique et la relation entre Y et X est linéaire. La Figure 2.3(b) donne un exemple type de l'effet d'une relation non-linéaire entre Y et X sur le même graphique des résidus. Dans ce cas, les résidus vont du négatif au positif de façon systématique. Ce problème pourrait être corrigé en transformant la variable explicative X (voir Section 2.6).

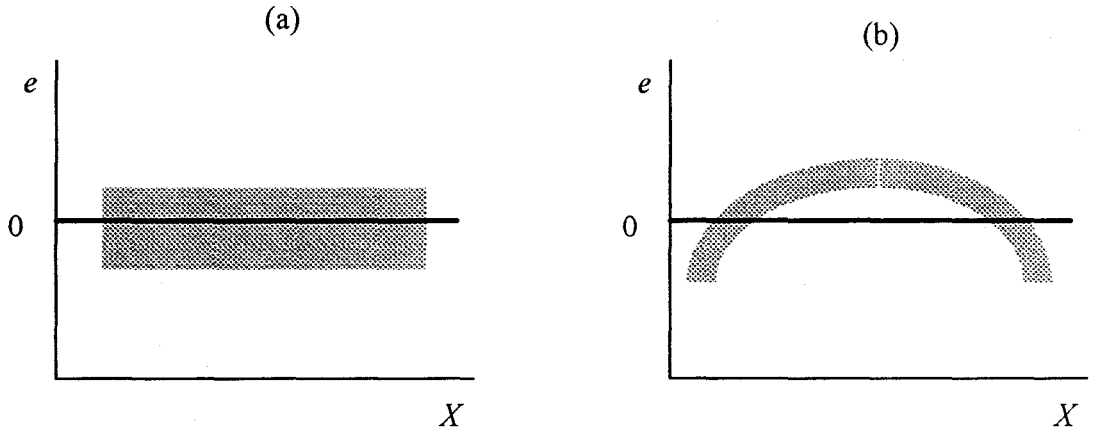


Figure 2.3. Exemples types de graphiques des résidus.

2. *Hétéroscédasticité*

Le graphique des résidus peut aussi être utilisé pour vérifier si la variance des termes d'erreur est constante. En régression multiple (plus d'une variable explicative), on suggère de tracer les résidus e_i en fonction des valeurs prédites $\hat{y}_i$ (ce type de graphique est disponible dans ***RéMuS***). La Figure 2.4 donne l'exemple d'un tel graphique lorsque la variance des résidus augmente avec les valeurs prédites. En effet, plus $\hat{y}_i$ augmente, plus la dispersion du nuage de points e_i est grande.

Dans la plupart des cas, les problèmes d'hétéroscédasticité sont de type conique comme le montre la Figure 2.4. Toutefois, on peut rencontrer des variances d'erreur qui croissent ou décroissent de

façon différente avec les valeurs prédites. Pour pallier ce problème, on peut transformer la variable dépendante Y (voir Section 2.6, transformation de Box-Cox).

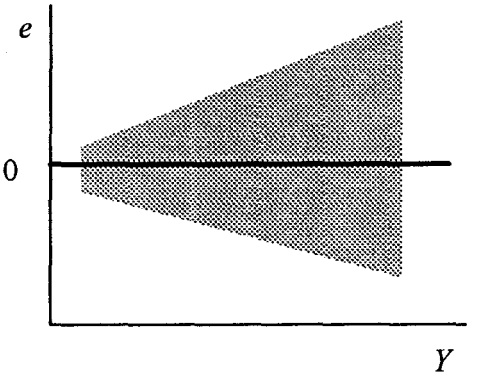


Figure 2.4. Exemple type d'un problème d'hétéroscédasticité.

3. *Dépendance des termes d'erreur.*

Dans bien des applications en hydrologie, les observations utilisées sont mesurées dans le temps. En particulier, pour le problème qui nous concerne, les mesures sont des débits mensuels moyens obtenus sur plusieurs années. Il est donc intéressant de tracer les résidus en fonction du temps afin d'examiner s'il existe une relation temporelle entre les termes d'erreur (autocorrélation). Si c'est le cas, les erreurs sont autocorrélées et l'hypothèse d'indépendance des erreurs dans le modèle de régression n'est pas respectée. La Figure 2.5 présente deux exemples typiques d'autocorrélation des termes d'erreur.

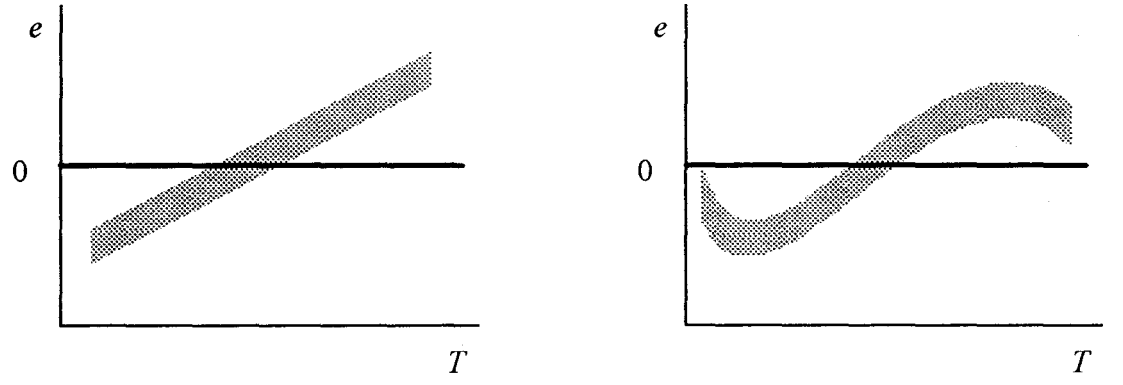


Figure 2.5. Exemples types d'un problème d'autocorrélation des erreurs.

On remarque en effet que les résidus négatifs sont associés aux premières mesures alors que les résidus positifs le sont aux données plus récentes. Dans ces deux exemples, une dépendance temporelle semble présente. Ainsi, l'hypothèse d'indépendance des termes d'erreurs n'est pas respectée et les modèles correspondant ne sont pas valides. Des modèles de série chronologique seraient alors plus adéquats (Fuller, 1976).

En plus de ce graphique, ***RéMuS*** effectue le test de Durbin et Watson (1950) sur les résidus. Ce test vérifie si l'autocorrélation d'ordre 1, ρ, est significative ($H_0: \rho = 0$ *contre* $H_1: \rho \neq 0$). La statistique calculée D fournie par le logiciel est définie de la manière suivante :

$$D = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{2.21}$$

où e_t est le résidu correspondant à la $i^{\text{ième}}$ mesures dans le temps. Durbin et Watson (1950) ont obtenu, pour un niveau de signification donné α, des bornes d_L et d_U telles qu'une valeur calculée de D située à l'extérieur de l'intervalle $[d_L; d_U]$ nous permet de conclure quant à la présence d'une autocorrélation positive significative ($\rho > 0$). La règle de décision pour tester l'autocorrélation positive est la suivante :

- Si $D > d_U$, on conclut $\rho = 0$
- Si $D < d_L$, on conclut $\rho > 0$
- Si $d_L \leq D \leq d_U$, on ne peut conclure

Si on désire plutôt vérifier la présence d'une autocorrélation négative ($\rho < 0$), on utilise la statistique $4 - D$. Pour le même niveau de signification α, la règle de décision est alors équivalente à celle qui permet de tester l'hypothèse d'une autocorrélation positive :

- Si $4 - D > d_U$, on conclut $\rho = 0$
- Si $4 - D < d_L$, on conclut $\rho < 0$
- Si $d_L \leq 4 - D \leq d_U$, on ne peut conclure

La table donnée à l'annexe C contient les bornes d_L et d_U pour différentes tailles d'échantillon ($15 \leq n \leq 100$), pour deux niveaux de signification ($\alpha = 5\%, 1\%$), et pour différents nombres de

variables explicatives ($1 \leq p \leq 5$). Le test de Durbin-Watson peut donc être effectué seulement si la taille d'échantillon est supérieure à 14, et le nombre de variables explicatives inférieur à 6. Prenons l'exemple d'un modèle de régression simple (p=1) dont les paramètres sont estimés à l'aide d'un échantillon de 20 observations, et tel que $D = 0.825$. Pour un niveau de signification de 1%, on trouve dans la table (annexe C) que $d_L = 0.95$ et $d_U = 1.15$. Puisque, dans ce cas, $D = 0.825 < d_L = 0.95$, on conclut que les termes d'erreurs possèdent un coefficient d'autocorrélation positif significatif au niveau de signification de 1%.

Pour plus de détail sur ce test, nous invitons le lecteur à consulter Durbin et Watson (1950).

4. *Non-normalité des termes d'erreur.*

La normalité des termes d'erreur d'un modèle de régression peut être examinée à l'aide de plusieurs types de graphiques et tests statistiques. Dans ***RéMuS,*** nous donnons la possibilité à l'usager de visualiser les résidus sur un papier de probabilité normal. Les résidus ordonnés sont alors tracés en fonction de leur probabilité empirique sur un graphique dont les axes sont gradués de sorte qu'une relation linéaire suggère que l'hypothèse de normalité est vérifiée alors qu'une relation non-linéaire indique plutôt la non-normalité des termes d'erreurs. Si les résidus ne semblent pas distribués selon une loi normale, on peut transformer la variable dépendante Y comme nous le verrons dans la section suivante.

2.6 Transformation de variables

Lorsque l'analyse des résidus indique qu'une ou plusieurs hypothèses sous-jacentes au modèle ne sont pas respectées pour les observations considérées, il est généralement possible de corriger ce problème en transformant les variables. Deux situations principales où une transformation des données s'avére utile peuvent être identifiées :

1. lorsque la relation entre la variable dépendante et les variables explicatives n'est pas linéaire;

2. lorsque les termes d'erreurs ne proviennent pas d'une loi normale et|ou leur variance n'est pas constante.

Dans le premier cas, on agira sur les variables explicatives alors que, dans le second cas, c'est la variable dépendante qui sera traitée. Ainsi, pour valider l'hypothèse de linéarité, on peut, suite à l'examen des graphiques de Y en fonction de X_k, transformer la ou les variables explicatives qui semblent liées à Y par une fonction non-linéaire qui est généralement identifiable visuellement. Les transformations de variables explicatives disponibles dans ***RéMuS*** sont $\ln X$, X^2, $1/X$, $1/X^2$ et $\sqrt{X}$. La régression multiple peut ensuite être appliquée aux nouvelles données ainsi transformées. Il est toutefois important d'examiner attentivement les résidus du modèle transformé car ce type de transformation ne garantit pas nécessairement que toutes les hypothèses sur les termes d'erreur soient vérifiées.

Dans le deuxième cas, c'est-à-dire lorsque les termes d'erreurs ne proviennent pas d'une loi normale et|ou leur variance n'est pas constante, la situation est plus complexe car aucune méthode graphique ne peut nous aider à choisir la bonne transformation. L'approche incorporée dans le logiciel ***RéMuS*** est celle de Box et Cox (1964). Dans cette méthode, on transforme la variable dépendante Y. On considère la famille des transformations de puissance proposée par Tukey (1957) et indicée par le paramètre λ. Ainsi, la variable dépendante transformée $Y^{(\lambda)}$ s'écrit:

$$Y^{(\lambda)} = \begin{cases} \dfrac{Y^{\lambda} - 1}{\lambda}, & \lambda \neq 0 \\ \ln Y & \lambda = 0 \end{cases} \tag{2.22}$$

Cette famille inclut implicitement les transformations usuelles telles, le logarithme, la racine carrée, le carré, etc., appliquées à la variable dépendante Y. Plus précisément, la méthode de Box et Cox (1964) consiste à estimer le paramètre λ à l'aide de la méthode du maximum de vraisemblance. Cette technique revient en fait à chercher la valeur de λ qui permet d'obtenir le modèle de régression ayant la plus petite somme des carrés de résidus *SCE*. Cette estimation donne la transformation de la variable dépendante qui devrait nous permettre de valider les hypothèses de normalité et d'homogénéité de la variance. Notons, toutefois, que cette méthode ne peut être appliquée que pour des valeurs positives de la variable dépendante. ***RéMuS*** donne la valeur estimée de λ ainsi que les bornes de l'intervalle de confiance à 95% pour ce paramètre. Cet intervalle peut être utilisé, soit pour vérifier si une transformation est nécessaire (si la valeur 1 appartient à l'intervalle, il n'est pas nécessaire de transformer la variable), soit pour permettre à l'usager d'arrondir la valeur obtenue afin de se ramener à une transformation usuelle ($\ln Y$, Y^2, $1/Y$, $1/Y^2$ et $\sqrt{Y}$) pour simplifier le modèle. Pour plus de détails concernant cette approche, nous invitons le lecteur à consulter l'article de Box et Cox (1964).

2.7 Reconstitution des données

L'objectif de ***RéMuS*** est non seulement de prédire une donnée de la variable dépendante, mais aussi de reconstituer cette variable à partir de la prédiction tout en conservant les principales caractéristiques statistiques de la série de données observées $y_1, y_2, \ldots, y_n$. C'est pourquoi nous employons ici le terme reconstitution plutôt que prédiction. Plus précisément, nous désirons estimer à l'aide du modèle de régression une série de valeurs y_l , $l = 1, 2, \ldots, n_l$, manquantes étant données les observations $x_{1l}, x_{2l}, \ldots, x_{pl}$ correspondantes des variables explicatives, et ce tout en conservant la moyenne et la variance des n observations $y_1, y_2, \ldots, y_n$. Pour ce faire, on calcule tout d'abord les valeurs prédites par le modèle :

$$\hat{y}_l = \hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}_k x_{lk} \quad , l = 1, 2, \ldots, n_l \tag{2.23}$$

Puisque les estimateurs des paramètres sont non-biaisés (Section 2.4), la moyenne des valeurs estimées $\hat{y}_l$ sera donc égale à celle des données observées. Toutefois, la variance ne peut être reproduite. Pour corriger cette lacune, Fiering (1963) propose d'ajouter à l'expression (2.23) un terme aléatoire défini comme suit :

$$\delta_l = u_l \sqrt{\frac{SCT}{n-p-1}\left(1-R^2\right)} = u_l \sqrt{\frac{SCE}{n-p-1}} \tag{2.24}$$

où u_l est un nombre aléatoire provenant d'une loi normale centrée-réduite. Notons que $SCE/(n-p-1)$ est un estimateur non-biaisé de σ^2, la variance de Y (Neter et Wasserman, 1985). Ainsi, les données reconstituées fournies par le logiciel ***RéMuS*** sont de la forme :

$$\tilde{Y}_l = \hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}_k x_{lk} + \delta_l \quad , l = 1, 2, \ldots, n_l \tag{2.25}$$

et possèdent, pour des valeurs fixées des variable explicatives, une variance égale à σ^2. Il est important de noter que la moyenne et la variance des données reconstituées coïncideront avec celles des données originales si et seulement si toutes les hypothèses du modèle de régression sont vérifiées. Il faut tout de même s'attendre, même si le modèle est valide, à un certain écart dû à la variabilité de l'échantillon. On doit aussi noter que si une transformation de la variable dépendante (méthode de Box-Cox) est effectuée, ce sont les caractéristiques statistiques des observations transformées qui seront conservées et non celles des données elles-mêmes. En effet, en appliquant la transformation inverse aux prédictions, on introduit un biais.

2.8 Validation de la qualité de la reconstitution

Pour apprécier la qualité de la reconstitution, ***RéMuS*** offre deux tests statistiques qui permettent la comparaison de la moyenne et de la variance des données reconstituées avec celles des données originales. Ces tests qui vérifient l'égalité des deux premiers moments sont le test de Wilcoxon (1945) et le test de Levene (1960). Ces deux tests sont non paramétriques, c'est-à-dire qu'ils ne reposent pas sur une hypothèse concernant la loi d'où proviennent les observations. Considérons les deux échantillons, l'un formé des observations mesurées $y_1, y_2, \ldots, y_n$, et l'autre des données reconstituées $\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_{n_l}$. Supposons de plus qu'ils proviennent respectivement de deux populations de moyenne et de variance (μ, σ^2) et $(\tilde{\mu}, \tilde{\sigma}^2)$. Le test de Wilcoxon vérifie les hypothèses :

$$H_0 : \mu = \tilde{\mu} \quad \text{contre} \quad H_1 : \mu \neq \tilde{\mu}$$

La statistique utilisée est donnée par :

$$Z = \frac{W - \frac{n(n+n_l+1)}{2} + \frac{1}{2}}{\sqrt{Var\{W\}}} \tag{2.26}$$

où :

- $W = \sum_{i=1}^{n+n_l} R_i \, s(R_i)$
- R_i est le rang correspondant à l'observation i de l'échantillon combiné de taille $n + n_l = N$
- $s(R_i) = \begin{cases} 0 & si\ R_i\ correspond\ à\ une\ donnée\ de\ l'échantillon\ des\ y_i \\ 1 & si\ R_i\ correspond\ à\ une\ donnée\ de\ l'échantillon\ des\ \tilde{y}_i \end{cases}$
- $Var\{W\} = \frac{nn_l(N+1)}{12} - \frac{nn_l \sum_{k=1}^{h} d_k^3 - d_k}{12N(N-1)}$, en supposant que nous avons dans l'échantillon combiné h groupes distincts contenant des observations égales (si

toutes les observations sont distinctes, on a $h = N$) et que le nombre d'observations égales dans chacun de ces groupes soient respectivement $d_1, d_2, \ldots, d_h$.

La règle de décision pour effectuer ce test à un niveau de signification donné α est la suivante :

- Si $|Z| \leq z(1-\alpha/2)$, on accepte H_0
- Sinon, on rejette H_0

où $z(1-\alpha/2)$ est le quantile de probabilité au non-dépassement $1-\alpha/2$ de la loi normale centrée-réduite. Lehmann (1975, Chap. 1) donne les détails concernant le test de Wilcoxon.

Le test de Levene (1960) examine l'hypothèse d'égalité des variances des deux échantillons, c'est-à-dire :

$$H_0 : \sigma^2 = \tilde{\sigma}^2 \quad \text{contre} \quad H_1 : \sigma^2 \neq \tilde{\sigma}^2$$

Conover et al. (1980) ont montré à l'aide de simulations que ce test est le plus efficace parmi 50 procédures considérées pour comparer les variances. Pour alléger la présentation du test, nous notons y_{1j} la jième donnée observée et y_{2j} la jième donnée reconstituée. Pour effectuer ce test, on calcule d'abord pour chaque échantillon les écarts en valeurs absolues des observations par rapport à la médiane :

$$EC_{ij} = \left| y_{ij} - Med_i \right|, \qquad i = 1, 2 \tag{2.27}$$

La statistique du test de Levene (1960) est alors donnée par :

$$L = \frac{(N-2)\left[n\left(\overline{EC_1} - \overline{EC}\right)^2 + n_l\left(\overline{EC_2} - \overline{EC}\right)^2 \right]}{\sum_{j=1}^{n}\left(EC_{1j} - \overline{EC_1}\right)^2 + \sum_{j=1}^{n_l}\left(EC_{2j} - \overline{EC_2}\right)^2} \tag{2.28}$$

où $\overline{EC_1}$, $\overline{EC_2}$ et $\overline{EC}$ sont respectivement les moyennes des écarts pour l'échantillon des données observées, pour l'échantillon des données reconstituées et pour l'échantillon combiné.

La règle de décision pour effectuer ce test à un niveau de signification donné α est la suivante :

- Si $L \leq F_{1,n-n_l-2}(1-\alpha)$, on accepte H_0
- Sinon, on rejette H_0

où $F_{1,n-n_l-2}(1-\alpha)$ est le quantile de probabilité au non-dépassement 1 - α de la loi de Fisher à 1 et $(n-n_l-2)$ degrés de liberté.

RéMuS donne pour ces deux tests la statistique calculée ainsi que la probabilité au dépassement. Si la moyenne et/ou la variance des données observées sont significativement différentes de celles des données reconstituées pour un niveau de signification α choisi à priori, il est alors souhaitable de revoir le modèle utilisé. Le choix du niveau de signification revient à l'usager. Rappelons que α est la probabilité de rejeter H_0 alors que cette hypothèse est vraie.

CHAPITRE III

REGRESSION RIDGE

3.1 La multicollinéarité

Lorsqu'on analyse un ensemble de données à l'aide d'une régression multiple, on désire généralement étudier la nature et la signification des relations qui existent entre la variable dépendante et les variables explicatives. On peut entre autres se demander :

1. Quelle est l'importance relative de chacune des variables explicatives?
2. Quel est l'effet d'une variable explicative donnée sur la variable dépendante?
3. Est-ce qu'une variable explicative peut être retranchée de l'analyse étant donné son effet négligeable sur la variable dépendante?
4. Est-ce qu'une variable non inclue dans le modèle peut être ajoutée à celui-ci?

On peut répondre à ces questions assez facilement si les variables explicatives sont indépendantes (non-corrélées entre elles). En effet, dans ce cas les tests sur les paramètres ainsi que l'analyse des résidus du modèle présentés au chapitre précédent peuvent être utilisés avec efficacité. Toutefois, il n'en est pas de même si les variables explicatives sont corrélées, c'est-à-dire s'il y a *multicollinéarité*.

En présence de *multicollinéarité*, quelques problèmes typiques peuvent survenir :

- Enlever ou ajouter une variable explicative au modèle peut changer radicalement la valeur des paramètres estimés des autres variables;

- Les estimations des paramètres peuvent avoir une variance très élevée et donc être imprécises, et parfois même erronées;

- Les paramètres peuvent ne pas être statistiquement significatifs alors qu'une relation entre la variable dépendante et les variables explicatives existe.

Nous n'expliciterons pas ici les fondements théoriques du problème de *multicollinéarité* qui sont décrits en détail dans Neter *et al.* (1985) ou Draper et Smith (1966). Mentionnons seulement que l'estimation du vecteur des paramètres $\boldsymbol{\beta}$ fait intervenir l'inversion de la matrice $\mathbf{X'X}$, et que cette matrice devient de plus en plus difficile à inverser à mesure que la corrélation entre les variables explicatives augmente. Ainsi, si ces intercorrélations sont grandes, les éléments de la matrice inversée peuvent être erronés et engendrer une instabilité dans l'estimation des paramètres. À la limite, s'il existe une corrélation parfaite entre deux variables explicatives, la matrice $\mathbf{X'X}$ sera singulière, donc non-inversible, et les paramètres ne pourront être déterminés.

Peu de méthodes formelles ont été développées pour détecter la présence d'un problème sérieux de multicollinéarité. Toutefois, on peut soupconner la présence de multicollinéarité lorsqu'on observe:

1. De grands changements dans l'estimation des paramètres lorsqu'on ajoute ou on retire une variable du modèle;

2. Des résultats non-significatifs lors d'un test concernant le paramètre d'une variable explicative pertinente pour le modèle;

3. Des paramètres estimés qui sont négatifs (positifs) alors que théoriquement, ou physiquement, on s'attend à obtenir des paramètres positifs (négatifs).

4. De grands coefficients de corrélation entre les variables explicatives.

Lorsque le modèle est utilisé uniquement à des fins de prédictions, les effets de la multicollinéarité sur l'estimation des paramètres peuvent être négligés (Neter et al., 1985, Chap. 11). Par contre, si l'étude de l'effet des paramètres sur la variable dépendante (analyse fonctionnelle) présente un intérêt pour l'utilisateur, des mesures correctives doivent être envisagées. Différentes approches correctives ont été proposées afin de tenir compte du problème de *multicollinéarité*. La première, l'approche heuristique, est basée sur l'expérience de l'utilisateur, sur sa connaissance des causes de la multicollinéarité et du problème à l'étude. Cette approche, essentiellement subjective, repose sur un choix judicieux des variables explicatives qui seront utilisées dans le modèle. On propose,

entre autres, de limiter le plus possible le nombre de variables explicatives, d'éviter l'utilisation de variables explicatives fortement corrélées entre elles, d'utiliser les variables les plus pertinentes pour le problème à traiter. D'autres approches interviennent directement sur l'estimation des paramètres en modifiant la méthode des moindres carrés. C'est le cas, en particulier, de la régression ridge qui est disponible dans le logiciel ***RéMuS*** et qui est présentée dans ce qui suit.

3.2 Le modèle de régression ridge et l'estimation des paramètres

La régression ridge (Hoerl et Kennard, 1970a) est une des méthodes correctives qui ont été proposées pour tenir compte du problème de *multicollinéarité* en modifiant l'estimation des paramètres par les moindres carrés. Les estimateurs issus de la régression ridge sont biaisés mais plus stables que ceux obtenus par la régression multiple lorsqu'il y a *multicollinéarité*. Dans ces circonstances, le biais n'est donc pas nécessairement une lacune. En effet, un estimateur qui possède un faible biais mais qui d'autre part est plus précis qu'un autre estimateur non-biaisé, peut être avantageusement utilisé puisque sa probabilité d'être proche de la valeur cible inconnue est supérieure.

Pour présenter la régression ridge, il est préférable de reparamétriser le modèle de régression multiple (2.2). Cette transformation, qui consiste à standardiser les observations, n'est qu'un changement d'écriture du modèle et n'affecte en rien les résultats. Considérons tout d'abord les écart-types des $(p + 1)$ variables :

$$s_Y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \qquad \text{et} \qquad s_k = \sqrt{\frac{\sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}{n-1}}, \quad \text{pour } k = 1, 2, \ldots, p$$

Les nouvelles observations standardisées sont données par :

$$y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{y_i - \bar{y}}{s_Y}\right) \qquad \text{et} \qquad x_{ik}^* = \frac{1}{\sqrt{n-1}}\left(\frac{x_{ik} - \bar{x}_k}{s_k}\right), \quad \text{pour } k = 1, 2, \ldots, p$$

Le modèle de régression reparamétrisé s'écrit alors :

$$y_i^* = \beta_1^* x_{i1}^* + \beta_2^* x_{i2}^* + \ldots + \beta_p^* x_{ip}^* + \varepsilon_i \qquad (3.1)$$

et on montre aisément (Neter, et al, 1985) que les paramètres β_k du modèle original (2.2) peuvent être obtenus en fonction des paramètres standardisés β_k^* ; on a :

$$\beta_k = \left(\frac{s_Y}{s_k}\right)\beta_k^* \text{ , pour } k = 1, 2, ..., p \quad \text{et} \quad \beta_0 = \bar{y} - \sum_{k=1}^{p} \beta_k \bar{x}_k \tag{3.2}$$

On peut aussi montrer que les nouvelles matrices $\mathbf{X'X}$ et $\mathbf{X'Y}$ qui interviennent dans le calcul des estimateurs sont respectivement la matrice des corrélations entre les variables explicatives, et la matrice des corrélations entre la variable dépendante et les variables explicatives. Pour faire une distinction avec le modèle (2.2), on note ces matrices respectivement $\mathbf{r}_{XX}$ et $\mathbf{r}_{YX}$. Les estimateurs obtenus par la méthode des moindres carrés, s'écrivent alors sous forme matricielle de la façon suivante :

$$\mathbf{b}^* = \mathbf{r}_{XX}^{-1}\mathbf{r}_{YX} \tag{3.3}$$

où $\mathbf{r}_{XX}^{-1}$ est l'inverse de la matrice des corrélations entre les variables explicatives.

Les problèmes d'inversion de la matrice $\mathbf{X'X}$ dans le modèle (2.2) causés par la présence de multicollinéarité peuvent maintenant être transposés à la matrice $\mathbf{r}_{XX}$. En effet, plus les intercorrélations sont grandes (plus $\mathbf{r}_{XX}$ s'éloigne de la matrice identité), plus les éléments de la matrice inverse peuvent être erronés ce qui engendrent une instabilité dans l'estimation des paramètres.

Les paramètres standardisés estimés par la régression ridge sont obtenus en introduisant une constante $k \geq 0$ dans les équations normales de la méthode des moindres carrés de la façon suivante :

$$(\mathbf{r}_{XX} - k\mathbf{I}_p)\mathbf{b}^R = \mathbf{r}_{YX} \tag{3.4}$$

où $\mathbf{b}^R = (b_1^R, b_2^R, ..., b_p^R)$ est le vecteur des paramètres standardisés estimés par la régression ridge et $\mathbf{I}_p$ est la matrice identité de dimension $p \times p$. On additionne donc une constante k à chacun des éléments situés sur la diagonale de la matrice $\mathbf{r}_{XX}$. Ceci a pour effet de faciliter l'inversion de cette matrice. La solution de ce système d'équation donne les nouveaux estimateurs qui dépendent maintenant de la constante k :

$$\mathbf{b}^R = (\mathbf{r}_{XX} - k\mathbf{I}_p)^{-1}\mathbf{r}_{YX} \tag{3.5}$$

La constante k traduit l'importance du biais des estimateurs. Si k = 0, l'équation (3.5) est équivalente à (3.3) et les estimateurs obtenus par la régression ridge correspondent aux estimateurs non-biaisés des moindres carrés ordinaires. Lorsque $k > 0$, les estimateurs sont biaisés mais plus stables que ceux des moindres carrés ordinaires. Pour retrouver les estimateurs non-standardisés, il suffit d'utiliser l'équation (3.2). On peut ensuite effectuer une analyse des résidus, tranformer les données si nécessaire et enfin reconstituer des données manquantes à l'aide de la régression ridge en procédant de la même manière qu'en régression multiple (sections 2.5, 2.6 et 2.7).

3.3 Détermination de k

On peut montrer (Hoerl et Kennard, 1970a), qu'à mesure que la constante k augmente, le biais de $\mathbf{b}^R$ augmente mais sa variance diminue. On peut montrer aussi qu'il existe toujours une valeur de k telle que l'estimateur $\mathbf{b}^R$ possède un écart quadratique moyen *ECM* inférieur à celui de l'estimateur des moindres carrés ordinaire **b**. Toutefois, il est difficile de choisir la constante k puisque la valeur optimale varie d'un ensemble de données à l'autre. Une méthode utilisée couramment pour déterminer la constante k, repose sur l'examen du graphique des traces. Ce graphique est une représentation simultanée des estimations des p paramètres standardisés $b_1^R, b_2^R, \ldots, b_p^R$ obtenus pour différentes valeurs de la constante k, généralement comprises entre 0 et 1. La Figure 3.1 donne un exemple type d'un graphique des traces pour un modèle à trois variables explicatives.

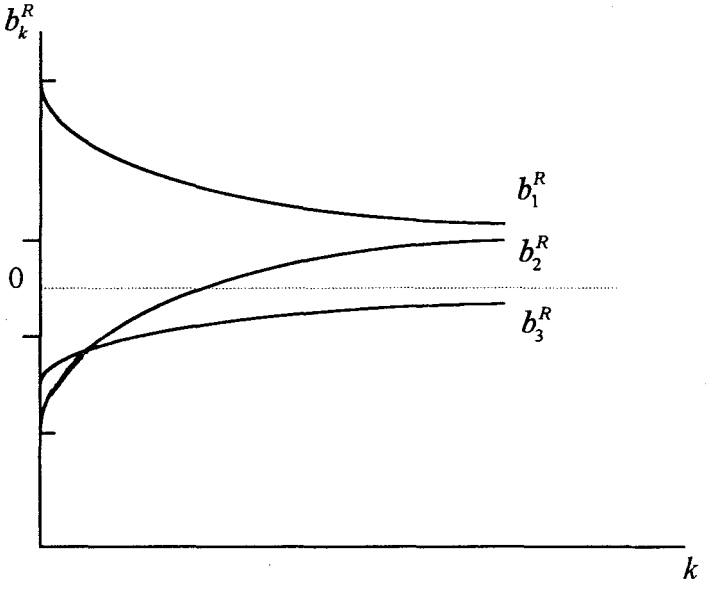


Figure 3.1. Exemple type d'un graphique des traces pour un modèle à 3 variables.

En général, les valeurs des paramètres estimés peuvent fluctuer considérablement lorsque k est proche de zéro (voir Figure 3.1) et peuvent même changer de signe. Toutefois, ces fluctuations s'atténuent graduellement et les valeurs des paramètres varient de moins en moins à mesure que k augmente. Ainsi, en pratique, on examine le graphique des traces et on choisit graphiquement la plus petite valeur de k qui correspond à une zone de stabilité des courbes associées à chacun des paramètres (Hoerl et Kennard, 1970b). Toutefois, Vinod (1976) a montré que cette procédure peut amener à surestimer la valeur de k. Cet auteur a donc proposé, comme méthode complémentaire, une procédure automatique permettant de déterminer la valeur du paramètre k. Cette procédure utilise l'indice ISRM défini par :

$$ISRM = \sum_{i=1}^{p} \left[\frac{(\lambda_i - k)^2}{\sum_{j=1}^{p} \frac{\lambda_j}{\lambda_j + k}} - 1 \right]^2 \qquad (3.6)$$

où $\lambda_1, \lambda_2, \ldots, \lambda_p$ sont les valeurs propres de la matrice $\mathbf{r}_{XX}$. Cet indice est nul si les variables explicatives sont non-corrélées. Vinod (1976) suggère d'utiliser le k qui correspond à la valeur minimale de l'indice *ISRM*.

Dans ***RéMuS***, lors d'une régression ridge, l'usager doit tout d'abord choisir la valeur du paramètre k. Pour ce faire, il peut soit visualiser le graphique des traces et inscrire la valeur choisie dans le champ réservé à cette fin, soit demander une sélection automatique selon l'indice de Vinod (1976). Une fois son choix arrêté, il peut effectuer la régression. Les résultats sont alors présentés sous forme de tableaux comme pour la régression multiple ordinaire (Section 2.5). Le logiciel présente la valeur de k utilisée, les tests sur les paramètres, l'analyse de la variance ainsi que le coefficient de détermination R^2. Notons que R^2 diminue à mesure que k augmente. En effet, la somme des carrés des erreurs *SCE* est minimum lorsque $k = 0$ (solution des moindres carrés), mais elle augmente à mesure que croît la constante k alors que la somme des carrés totale *SCT* reste fixe. Ceci met en évidence que la stabilité des coefficients est obtenue par la régression ridge au prix d'une perte de variance expliquée qui augmente avec le paramètre k.

3.4 Remarque

Bien que conçue initialement pour corriger les effets néfastes de la *multicollinéarité*, la régression ridge peut aussi être utilisée comme méthode de sélection de variables dans un modèle de régression multiple classique. En effet, la représentation graphique des traces peut s'avérer un outil simple et efficace pour éliminer des variables explicatives de moindre importance pour le modèle. Hoerl et Kennard (1970b) ont suggéré d'éliminer les variables dont la trace est instable et dont le coefficient tend rapidement vers zéro (Figure 3.2a). De plus, il est conseillé d'éliminer les variables possédant une trace stable mais dont le paramètre est proche de zéro (Figure 3.2b). Enfin, l'intérêt de variables explicatives dont la trace est instable mais qui ne tendent pas nécessairement vers zéro peut être mise en doute (Figure 3.2c). Utilisée dans ce contexte, la régression ridge devient un outil complémentaire à l'approche corrective heuristique brièvement présentée à la Section 3.1.

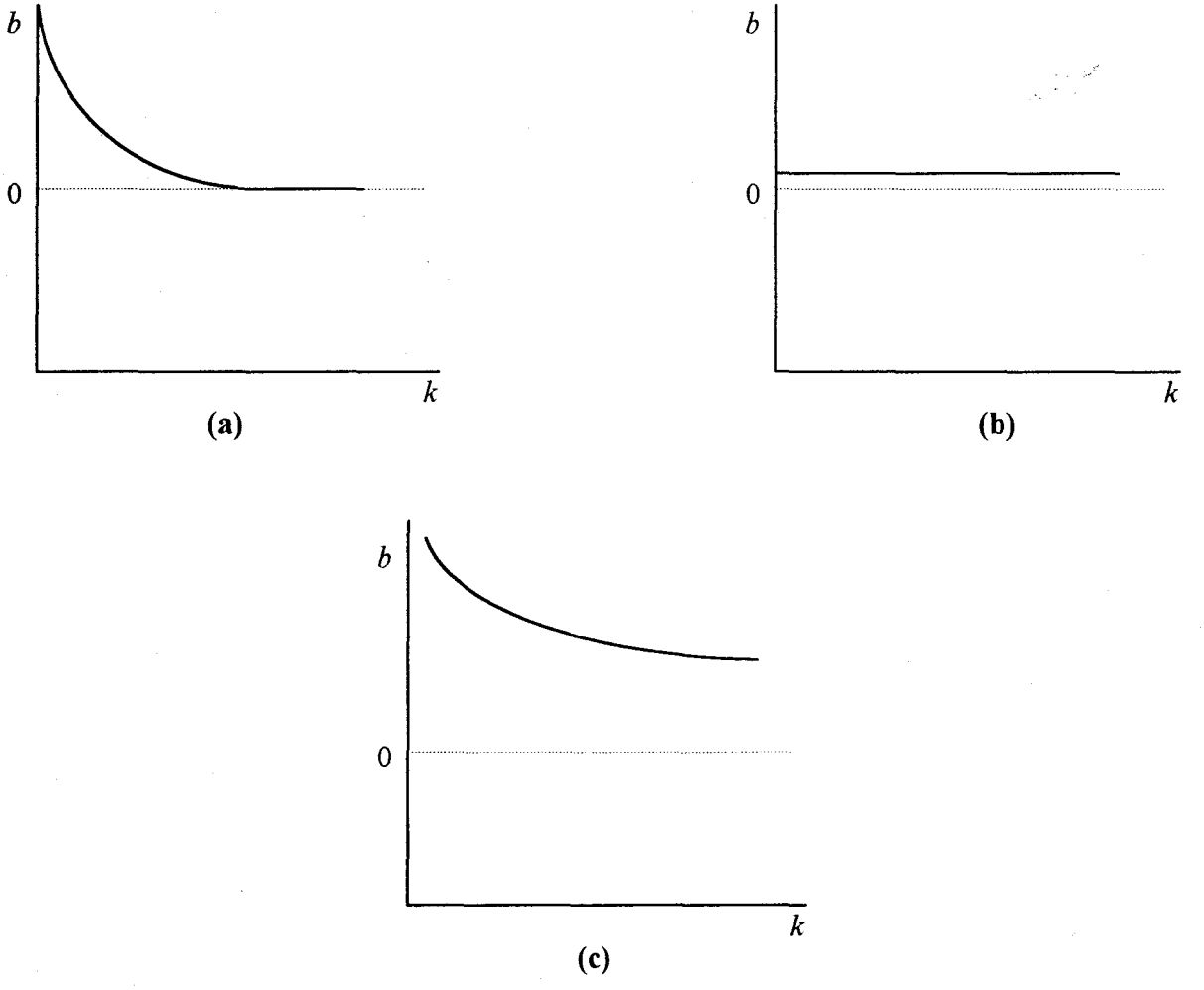


Figure 3.2. Exemples types du comportement des traces de variables à rejeter .

CHAPITRE IV

REGRESSION PAS A PAS

4.1 Choix des variables

L'objectif de la régression multiple est d'obtenir, à partir d'un ensemble de variables explicatives, une fonction de régression faisant intervenir des variables explicatives X_k significatives pour expliquer la variation de la variable dépendante Y. Ceci permet d'obtenir une équation de régression possédant un coefficient de détermination R^2 élevé, ce qui correspond à une petite somme des carrés des résidus.

L'approche la plus sûre pour atteindre cet objectif est d'effectuer la régression en entrant progressivement les variables explicatives une à la fois dans le modèle. On commence par la variable qui possède la plus grande corrélation avec la variable dépendante. On effectue ensuite une étude complète des résidus (Section 2.5). On peut alors tracer les résidus du modèle obtenu en fonction de chacune des variables explicatives qui n'ont pas encore été retenues dans le modèle. Si aucun de ces graphiques ne révèle une relation entre les résidus et l'une des variables explicatives inutilisées, on peut conclure qu'une seule variable explicative est utile. Si au contraire on observe une relation entre les résidus et l'une des variables explicatives, cela signifie que cette dernière est probablement pertinente. On peut alors l'ajouter au modèle et vérifier plus précisément l'apport de cette variable à l'explication de la variabilité totale de Y en examinant d'une part la valeur du coefficient de détermination, et d'autre part, la statistique t_k (Section 2.4). Il est aussi important de s'assurer que la valeur du paramètre de la première variable reste stable suite à l'inclusion de la nouvelle variable explicative. En effet, si par exemple la première variable devient non-significative, il y a probablement un problème de multicollinéarité entre les variables explicatives. Si l'analyse des résidus et les résultats des tests montrent que le modèle à deux variables n'est pas encore satisfaisant, on peut continuer la procédure en ajoutant une nouvelle

variable. Il est important de noter que l'ajout de variables non-significatives n'améliore pas la précision du modèle. De plus, un grand nombre de variables augmente les chances d'avoir des problèmes de multicollinéarité. Il convient donc de choisir un modèle contenant un nombre raisonnable de variables explicatives.

La procédure que nous venons de décrire est sûre mais lourde à appliquer. C'est pourquoi, plusieurs méthodes automatiques de sélection progressives de variables ont été proposées. Celles-ci peuvent s'avérer fort utiles lorsqu'on désire effectuer rapidement plusieurs régressions, ou lorsque l'ensemble de variables explicatives disponibles est grand. Toutefois, ce type de méthode automatique possède quelques lacunes. En effet, les critères utilisés pour entrer une variable dans le modèle ne tiennent pas compte des hypothèses de base telles que la normalité, l'indépendance et l'homogénéité de la variance des résidus. De plus, si les résultats de chacune des étapes de la sélection ne sont pas examinés attentivement, les problèmes de multicollinéarité peuvent souvent passer inaperçus.

Dans ***RéMuS*** , nous avons incorporé une méthode de sélection automatique des variables explicatives fréquemment utilisée en régression multiple. Il s'agit de la régression pas à pas (Stepwise regression, Neter *et al.*, 1985) que nous présentons dans ce qui suit.

4.2 Régression pas à pas

La méthode de régression pas à pas permet de calculer une série d'équations de régression où, à chaque étape, une variable explicative est ajoutée ou retranchée du modèle selon un critère de sélection. Le critère d'entrée ou de sortie d'une variable explicative est un rapport de sommes des carrés que l'on note $F^*(.)$, et que l'on compare à une valeur théorique critique établie a priori. Ce critère permet d'évaluer l'effet de l'ajout d'une nouvelle variable explicative sur la contribution de la ou des variables explicatives déjà contenues dans le modèle. Si cet effet est significatif, la nouvelle variable est gardée dans le modèle. De plus, ce critère nous permet d'évaluer alternativement l'apport de chaque variable explicative comme si elle était la toute dernière variable à être introduite dans l'équation. Si cet apport n'est pas significatif, la variable correspondante est alors retranchée de l'équation de régression. La sélection se termine lorsqu'aucune variable explicative ne peut être ajoutée ou retranchée de l'équation de régression.

Voici donc les différentes étapes effectuées par la régression pas à pas :

1. La procédure débute en ajustant une régression simple (une seule variable explicative) à l'aide des p variables considérées pertinentes a priori pour le modèle. Pour chacun des modèles de régression simple, la statistique $F^*(k)$ suivante est calculée :

$$F^*(k) = (n-2)\frac{SCR(X_k)}{SCE(X_k)}, \qquad k = 1, \ldots, p \tag{4.1}$$

où $SCR(X_k)$ et $SCE(X_k)$ sont respectivement les sommes des carrés de la régression et des résidus correspondant au modèle incluant X_k (Section 2.4). La variable explicative qui possède la plus grande valeur $F^*(k)$ est alors la première variable candidate pour le modèle. Si cette statistique est supérieure à une valeur théorique F_a spécifiée par l'usager (voir Section 4.3), la variable X_k est entrée dans le modèle. Toutefois, si la valeur de la statistique est inférieure, la régression pas à pas est terminée et aucun lien n'existe entre Y et les autres variables explicatives.

2. Supposons maintenant que la variable X_j a été entrée dans le modèle à l'étape 1. La régression pas à pas ajuste maintenant tous les modèles à deux variables explicatives dont l'une d'elles est X_j. Pour chacun de ces modèles de régression, la statistique suivante est calculée :

$$F^*(j,k) = (n-3)\frac{SCR(X_j, X_k) - SCR(X_j)}{SCE(X_j, X_k)}, \qquad k = 1, \ldots, j-1, j+1, \ldots, p \tag{4.2}$$

où $SCR(X_j, X_k)$ et $SCE(X_j, X_k)$ sont respectivement la somme des carrées de la régression et des résidus du modèle à deux variables, et $SCR(X_j)$ la somme des carrés de la régression du modèle à une variable obtenue à l'étape 1. Cette statistique mesure donc l'apport de la nouvelle variable X_k considérée dans le modèle. On peut montrer (Neter *et al.*, 1985) que cette statistique est équivalente à celle qui vérifie si le paramètre associé à la variable X_k est nul dans le modèle à deux variables (statistique t_k, Section 2.4). Comme dans la première étape, la variable explicative correspondant à la plus grande valeur de $F^*(j,k)$ est retenue. Cette valeur est ensuite comparée à F_a. Si elle est supérieure, la variable correspondante est ajoutée au modèle de l'étape 1, sinon, la sélection est terminée et une régression simple avec X_j est jugée satisfaisante.

3. Supposons que la variable X_i a été ajoutée au modèle à l'étape 2. La régression pas à pas examine maintenant si l'une des deux variables explicatives du modèle doit être retranchée. Pour effectuer cette étape, on calcule la contribution de chacune des variables explicatives actuellement dans l'équation de régression comme si chacune d'elle était la dernière variable introduite dans le

modèle. Nous devons alors calculer une statistique $F^*(i,j)$ pour chaque variable. À ce stade-ci de la procédure, toutefois, il n'y a qu'un seul calcul à effectuer puisqu'une seule variable a été introduite dans le modèle avant la présente étape. Il s'agit de la variable X_j entrée à l'étape 1. La statistique utilisée est la suivante :

$$F^*(i,j) = (n-3)\frac{SCR(X_i, X_j) - SCR(X_i)}{SCE(X_i, X_j)} \tag{4.3}$$

S'il existait plusieurs variables explicatives dans l'équation avant l'introduction de la plus récente variable, il faudrait alors calculer cette statistique pour chacune des variables entrée dans le modèle auparavant. Parmi toutes ces valeurs obtenues, nous retenons la variable explicative correspondant à la plus petite valeur $F^*(i,j)$ et nous la comparons à une valeur théorique critique fixée a priori F_r (voir Section 4.3). Si elle est inférieure, la variable correspondante est retranchée, sinon elle demeure dans l'équation.

4. Supposons enfin que la variable X_j n'est pas rejetée de telle sorte que le modèle retenu pour l'instant possède deux variables explicatives X_j et X_i. Alors la régression pas à pas examine premièrement la possibilité d'ajouter une nouvelle variable au modèle, et ensuite si l'une des variables entrées aux étapes précédentes peut être retranchée. Cette procédure est effectuée jusqu'à ce qu'aucune variable explicative puisse être introduite ou retranchée.

Une fois l'ensemble de variables sélectionné par la régression pas à pas, ***RéMuS*** présente les résultats sous forme de tableaux de la même façon que pour la régression multiple (Figure 2.1). L'analyse des résidus peut être ensuite effectuée comme à la Section 2.5. Finalement, ***RéMuS*** permet de reconstituer les données à l'aide de l'approche présentée à la Section 2.7.

4.3 Remarque

Les valeurs critiques d'entrée et de sortie des variables explicatives, F_a et F_r , sont habituellement établies en fonction du risque α (niveau de signification des tests), du nombre d'observations et du nombre potentiel de variables explicatives p. Puisque le nombre de degrés de libertés associés à la somme des carrés des résidus varie avec le nombre de variables dans le modèle (donc d'une étape à l'autre lors de la procédure), ces valeurs critiques ne peuvent pas être interprétées précisément en terme de probabilité. Toutefois, on suggère souvent d'utiliser $F_a = F_r = F_{1,n-p-1}(1-\alpha)$, où

$F_{1,n-p-1}(1-\alpha)$ est le quantile de probabilité au non-dépassement 1 - α de la loi de Fisher à *1* et (*n-p-1*) degrés de liberté. Une table de ces quantiles est disponible dans plusieurs livres de statistique, en particulier dans Neter *et al.* (1985). Nous l'avons reproduite dans l'annexe A. Par exemple, $F_a = F_r = 4$ peut correspondre au quantile $F_{1,60}(0.95)$, donc à un niveau de signification de 5% pour chacun des tests effectués lors la procédure pour n = 70 données et p = 9 variables explicatives potentielles.

CHAPITRE V

REGRESSION MULTIDIMENSIONNELLE

5.1 Problématique

Nous avons présenté aux chapitres précédents trois types de régressions multiples qui permettent de reconstituer des données manquantes d'une variable dépendante. Lorsqu'on désire, par exemple, reconstituer des débits mensuels moyens d'un ensemble de q sites voisins (plusieurs variables dépendantes), on peut effectuer q analyses de régressions indépendamment. On calcule alors les prévisions de chaque modèle, et on leur ajoute un terme d'erreur (Section 2.7) qui permet de conserver la moyenne et la variance des q séries observées. Cette approche, dite *en parallèle* (q régressions indépendantes), ne permet de conserver, lors de la reconstitution, que ces caractéristiques statistiques (moyenne et variance). Les résultats ne reproduisent en aucune façon les liens (corrélations) qui pourraient exister entre les q sites à reconstituer. Cette lacune peut engendrer des erreurs non-négligeables et une perte d'information lorsque les résultats sont utilisés pour prendre une décision dans un contexte régional (Bernier, 1971).

Pour éviter ce type de problème et pour tirer le maximum d'information des données observées, les débits reconstitués des q sites doivent traduire les inter-relations existant entre elles. Une solution consiste à utiliser un modèle multidimensionnel qui considère l'ensemble des q variables dépendantes comme un tout. Cette solution, que nous nommons régression multidimensionnelle, est programmée dans le logiciel ***RéMuS*** et permet de conserver la structure de corrélation lors de la reconstitution simultanée de plusieurs variables dépendantes (par exemple, les débits de sites situés dans une même région). En effet, d'une part les paramètres estimés de ce modèle traduisent les liens qui peuvent exister entre les variables dépendantes, et d'autre part, les prévisions issues de ce modèle multidimensionnel, auxquelles on ajoute des termes d'erreur conjointement distribués selon une loi

normale multidimensionnelle, reflètent la structure de corrélation entre les variables dépendantes. Cette méthode est toutefois sujette à deux contraintes qui peuvent restreindre son utilisation en pratique :

les q variables dépendantes doivent être mises en relation avec les p mêmes variables explivatives;

la méthode n'utilise que les données concomitantes aux q variables dépendantes.

Ce chapitre présente le modèle de régression multidimensionnelle qui s'avère être une généralisation de la régression multiple. Nous y donnons les principaux résultats théoriques utilisés dans ***RéMuS***.

5.2 Le modèle de régression multidimensionnelle

Considérons q variables dépendantes $Y_1, Y_2, \ldots, Y_q$ et p variables explicatives indépendantes $X_1, X_2, \ldots, X_p$. Soient les n mesures correspondantes $y_{1i}, y_{2i}, \ldots, y_{qi}$ et $x_{1i}, x_{2i}, \ldots, x_{pi}$, $i = 1, 2, \ldots, n$ (par exemple, les débits mesurés sur n années concomitantes à $p + q$ sites). On suppose toujours que les réalisations $x_{1i}, x_{2i}, \ldots, x_{pi}$ des variables explicatives $X_1, X_2, \ldots, X_p$ sont connues exactement. Alors, le modèle de régression multidimensionnelle s'écrit, sous forme matricielle, de la façon suivante :

$$\underset{q\times n}{\mathbf{Y}} = \underset{q\times(p+1)}{\boldsymbol{\beta}}\ \underset{(p+1)\times n}{\mathbf{X}} + \underset{q\times n}{\boldsymbol{\varepsilon}} \tag{5.1}$$

où :

$$\underset{q\times n}{\mathbf{Y}} = \begin{bmatrix} y_{11} & y_{12} & y_{13} & \cdots & y_{1n} \\ y_{21} & y_{22} & y_{23} & \cdots & y_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ y_{q1} & y_{q2} & y_{q3} & \cdots & y_{qn} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \cdots & \mathbf{y}_n \end{bmatrix}$$

$$\underset{(p+1)\times n}{\mathbf{X}} = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{12} & x_{13} & \cdots & x_{1n} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{p1} & x_{p2} & x_{p3} & \cdots & x_{pn} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_n \end{bmatrix}$$

$$\underset{q\times(p+1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_{10} & \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \beta_{20} & \beta_{21} & \beta_{22} & \cdots & \beta_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ \beta_{q0} & \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_0 & \boldsymbol{\beta}_1 & \cdots & \boldsymbol{\beta}_p \end{bmatrix}$$

$$\underset{q\times n}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \varepsilon_{13} & \cdots & \varepsilon_{1n} \\ \varepsilon_{21} & \varepsilon_{22} & \varepsilon_{23} & \cdots & \varepsilon_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ \varepsilon_{q1} & \varepsilon_{q2} & \varepsilon_{q3} & \cdots & \varepsilon_{qn} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 & \boldsymbol{\varepsilon}_2 & \cdots & \boldsymbol{\varepsilon}_n \end{bmatrix}$$

Ainsi, la matrice $\mathbf{Y}$ contient n vecteurs colonne $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ dont les q éléments correspondent aux mesures des q variables dépendantes pour une période donnée (par exemple, les débits de chacun des sites à reconstituer mesurés sur une année). La matrice $\mathbf{X}$ contient n vecteurs colonne $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ de $p + 1$ éléments. Le premier élément de chacun de ces vecteurs est égal à 1 alors que les autres correspondent aux mesures des p variables explicatives pour une période donnée. La matrice $\boldsymbol{\beta}$ renferme les $p + 1$ vecteurs colonne $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p$. Le premier vecteur contient les q ordonnées à l'origine associées à chacune des variables dépendantes. Chacun des p autres vecteurs contient les q paramètres associés à chaque variable dépendante pour une variable explicative donnée ($\beta_{ij}, j \neq 0$, est le paramètre de la variable explicative X_j pour la variable dépendante Y_i).

Enfin, la matrice $\boldsymbol{\varepsilon}$ contient les n vecteurs $\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_n$ des q termes d'erreur. On suppose que chacun de ces vecteurs de termes d'erreur est distribué selon une loi normale multidimensionnelle de moyenne nulle et de matrice variances-covariances $\boldsymbol{\Sigma}$:

$$\boldsymbol{\varepsilon}_i \approx N\left(\underset{q\times 1}{\mathbf{0}}, \underset{q\times q}{\boldsymbol{\Sigma}}\right), \quad \forall i$$

Il existe alors, pour ce modèle, une corrélation non-nulle entre les termes d'erreur correspondant à différentes variables dépendantes.

5.3 Estimation des paramètres et inférence

Pour estimer les paramètres de la régression multidimensionnelle, c'est-à-dire déterminer les éléments $\hat{\beta}_{ij}$ de la matrice estimée $\mathbf{B}$:

$$\underset{q\times(p+1)}{\mathbf{B}} = \begin{bmatrix} \hat{\beta}_{10} & \hat{\beta}_{11} & \hat{\beta}_{12} & \cdots & \hat{\beta}_{1p} \\ \hat{\beta}_{20} & \hat{\beta}_{21} & \hat{\beta}_{22} & \cdots & \hat{\beta}_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ \hat{\beta}_{q0} & \hat{\beta}_{q1} & \hat{\beta}_{q2} & \cdots & \hat{\beta}_{qp} \end{bmatrix} = \begin{bmatrix} \mathbf{b}_0 & \mathbf{b}_1 & \cdots & \mathbf{b}_p \end{bmatrix}$$

on utilise encore une fois la méthode des moindres carrés. On peut montrer (Srivastava et Carter, 1983) que l'estimateur de la matrice $\boldsymbol{\beta}$ est donné par :

$$\begin{aligned} \mathbf{B} &= \begin{bmatrix} \mathbf{b}_0 & \mathbf{b}_1 & \cdots & \mathbf{b}_p \end{bmatrix} \\ &= \mathbf{Y}\mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1} \end{aligned} \tag{5.2}$$

Les estimateurs $\hat{\beta}_{ij}$ de β_{ij} ont la propriété d'être conjointement distribués selon une loi normale multidimensionnelle avec une moyenne et une matrice de variances-covariances telles que :

$$E\{\hat{\beta}_{ij}\} = \beta_{ij}, \qquad Var\{\hat{\beta}_{ij}\} = \sigma_{ii}a_{ij} \qquad \text{et} \qquad Cov\{\hat{\beta}_{ij}, \hat{\beta}_{kl}\} = \sigma_{ik}a_{jl} \tag{5.3}$$

où a_{ij} et a_{jl} sont des éléments de la matrice $(\mathbf{X}\mathbf{X}')^{-1}$ et, σ_{ii} et σ_{ij}, sont des éléments de la matrice $\boldsymbol{\Sigma}$. On remarque que les estimateurs des paramètres sont non-biaisés et, qu'ils sont corrélés d'une variable dépendante à l'autre, ce qui traduit la structure de corrélation qui existe entre les variables dépendantes. Si on avait effectué plusieurs régressions multiples de façon indépendante, cette corrélation aurait été nulle, c'est-à-dire que pour deux variables dépendantes Y_i et Y_k, on aurait :

$$Cov\{\hat{\beta}_{ij}, \hat{\beta}_{kj}\} = 0, \quad j = 0, 1, \ldots, p \tag{5.4}$$

Une fois que la valeur des paramètres pour chacune des variables explicatives est obtenue, il est intéressant de les tester. Dans le logiciel ***RéMuS***, on teste, pour chacune des variables explicatives, si le vecteur des paramètres correspondant est égal au vecteur nul **0**. Pour une variable explicative donnée X_k, ce test examine les hypothèses suivantes :

$$H_0 : \beta_{1k} = \beta_{2k} = \cdots = \beta_{qk} = 0 \quad \text{contre} \quad H_1 : \textit{au moins un } \beta_{ik} \neq 0$$

Pour effectuer ce test, on utilise la statistique F_k :

$$F_k = \frac{(n-p-q-1)}{q} \frac{(1-U_k)}{U_k} \tag{5.5}$$

où :

- $U_k = \dfrac{|\mathbf{V}_k|}{|\mathbf{V}_k + \mathbf{W}_k|}$, $|\mathbf{X}|$ désignant le déterminant d'une matrice $\mathbf{X}$;
- $\mathbf{V}_k = \mathbf{I}_q \mathbf{Y}\left[\mathbf{I}_n - \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\right]\mathbf{Y}'\mathbf{I}_q$, $\mathbf{I}_n$ étant la matrice identité de dimension n;
- $\mathbf{W}_k = \mathbf{I}_q \mathbf{B}\mathbf{C}_k\left[\mathbf{C}_k'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{C}_k\mathbf{X}\right]^{-1}\mathbf{C}_k'\mathbf{B}'\mathbf{I}_q$, $\mathbf{C}_k$ étant un vecteur colonne de dimension $p+1$ dont le $k$ième élément est égal à 1 et tous les autres éléments sont nuls.

La règle de décision pour effectuer ce test à un niveau de signification donné α est la suivante :

- Si $F_k \leq F_{q,n-p-q+1}(1-\alpha)$, on accepte H_0
- Si $F_k > F_{q,n-p-q+1}(1-\alpha)$, on rejette H_0

où $F_{q,n-p-q+1}(1-\alpha)$ est le quantile de probabilité au non-dépassement 1 - α de la loi de Fisher à p et n-p-q+1 degrés de liberté. Un exposé théorique complet sur ce test est donné dans Srivastava et Carter (1983). En plus de fournir la valeur des statistiques F_k (k = 0, 1, ..., p) calculées, ***RéMuS*** donne la probabilité au dépassement P_k associée à cette valeur (P-value).

Ce test permet donc de vérifier la pertinence d'une variable explicative X_k. F_k est en quelque sorte une généralisation de la statistique t_k utilisée en régression multiple (éq. 2.20). L'acceptation de l'hypothèse nulle signifie qu'il n'y a pas de relation entre la variable explicative X_k et les q

variables dépendantes à un niveau de signification α. Il est alors inutile de conserver cette variable dans le modèle. Toutefois, si on rejette cette hypothèse, c'est qu'il existe une relation entre X_k et au moins une des variables dépendantes, et il est préférable de la garder dans l'équation.

Les résultats de la régression multidimensionnelle sont présentés par ***RéMuS*** sous forme de tableaux. On y retrouve les estimations des paramètres ainsi que les statistiques F_k des tests sur les paramètres accompagnées de leur probabilité au dépassement P_k. Cette représentation est illustrée à la figure 5.1. Le guide de l'usager (Perron, 1993a) ou l'aide à l'écran du logiciel complètent l'information de cette figure.

Régression Multidimensionnelle

Paramètres estimés

		Var. Ind			
Var. Dép.	*Constante*	X_1	X_2	• • •	X_p
Y_1	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	• • •	$\hat{\beta}_{1p}$
Y_2	$\hat{\beta}_{20}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	• • •	$\hat{\beta}_{2p}$
⋮	⋮	⋮	⋮	⋮	⋮
Y_q	$\hat{\beta}_{q0}$	$\hat{\beta}_{q1}$	$\hat{\beta}_{q2}$	• • •	$\hat{\beta}_{qp}$

Tests sur les paramètres

Statist. :	F_0	F_1	F_2	• • •	F_p
Prob. :	P_0	P_1	P_2	• • •	P_p

Figure 5.1. Les résultats de la régression multidimensionnelle présentés dans ***RéMuS***.

5.4 Reconstitution des données

Nous désirons, pour une période donnée l (par exemple une année donnée) où la valeur des variables dépendantes est manquante, reconstituer le vecteur $\mathbf{y}_l = (y_{1l}, y_{2l}, \ldots, y_{ql})$ à l'aide du modèle de régression multidimensionnelle étant données les observations $\mathbf{x}_l = (x_{1l}, x_{2l}, \ldots, x_{pl})$ correspondantes des variables explicatives. On veut, de plus, que cette reconstitution préserve la moyenne, la variance et la structure de corrélation observées dans les séries originales des variables dépendantes (matrice **Y**). Le principe utilisé ici est le même qu'en régression multiple (Section 2.7). En effet, on calcule tout d'abord la prévision $\hat{\mathbf{y}}_l$ pour ensuite ajouter à chacun des éléments de ce vecteur un terme aléatoire tiré d'une loi normale multidimensionnelle. On obtient alors le vecteur reconstitué $\tilde{\mathbf{y}}_l$. Plus précisément, on a l'expression suivante :

$$\underset{q\times 1}{\tilde{\mathbf{y}}_l} = \underset{q\times(p+1)}{\mathbf{B}}\,\underset{(p+1)\times 1}{\mathbf{x}_l} + \underset{q\times 1}{\boldsymbol{\delta}_l} = \begin{bmatrix} \hat{\beta}_{10} & \hat{\beta}_{11} & \hat{\beta}_{12} & \cdots & \hat{\beta}_{1p} \\ \hat{\beta}_{20} & \hat{\beta}_{21} & \hat{\beta}_{22} & \cdots & \hat{\beta}_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ \hat{\beta}_{q0} & \hat{\beta}_{q1} & \hat{\beta}_{q2} & \cdots & \hat{\beta}_{qp} \end{bmatrix} \cdot \begin{bmatrix} x_{1l} \\ x_{2l} \\ \vdots \\ x_{pl} \end{bmatrix} + \begin{bmatrix} \delta_{1l} \\ \delta_{2l} \\ \vdots \\ \delta_{ql} \end{bmatrix} = \begin{bmatrix} \tilde{y}_{1l} \\ \tilde{y}_{2l} \\ \vdots \\ \tilde{y}_{ql} \end{bmatrix} \tag{5.6}$$

où le vecteur $\boldsymbol{\delta}_l$ contient q nombres aléatoires conjointement distribués selon une loi normale de moyenne nulle et de matrice variances-covariances $\boldsymbol{\Sigma}_\delta$ définie comme suit :

$$\underset{q\times q}{\boldsymbol{\Sigma}_\delta} = \frac{1}{n-p}\mathbf{Y}\left[\mathbf{I}_n - \mathbf{X}'(\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\right]\mathbf{Y}' \tag{5.7}$$

Dans ***RéMuS***, pour obtenir le vecteur aléatoire δ_l , on procède par génération de nombres aléatoires. L'approche utilisée est décrite en détail dans Devroye (1986, p. 563-566) et Law et Kelton (1982, p. 269).

L'ajout de ce vecteur aléatoire (utilisé entre autres dans Bernier, 1971) permet de conserver la moyenne, la variance et les corrélations existant entre les séries de variables dépendantes, en autant que les hypothèses du modèle soient vérifiées. En effet, les différentes remarques faites à ce sujet pour la régression multiple (Section 2.7) s'appliquent aussi à la régression multidimensionnelle.

REFERENCES BIBLIOGRAPHIQUES

Beard, L.R. (1971). *HEC-4 Monthly Streamflow Simulation.* The Hydrologic Engineering Center, Corps of Engineers, US Army, Davis, California 95616.

Bernier, J. (1971). Modèles probabilistes à variables hydrologiques multiples et hydrologie synthétique. *International Symposium on Mathematical Models in Hydrology*, Warsaw.

Box, G.E.P., Cox, D.R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Ser. B*, 211-252.

Conover, W.J., Johnson, M.E. et Johnson, M.M. (1981). A comparative study of tests for homogeneity of variances with applications to the outer continental shelf bidding data. *Technometrics*, **23**, 351-361.

Devroye, L. (1986). *Non-uniform random variate generation.* Springer-Verlag, New York.

Draper, N.R. et Smith, H. (1966). *Applied Regression Analysis.* Wiley, New York.

Durbin, J. et Watson, G.S. (1950). Testing for serial correlation in least square regression I. *Biometrika*, **37**, 409-428.

Feiring, M.B. (1963). *Use of Correlation to Improve Estimates of the Mean and Variance.* United States Geological Survey, Professionnal Paper 434C.

Fuller, W.A. (1976). *Introduction to Statistical Time Series.* Wiley, New York.

Hoerl, A.E. et Kennard, R.W. (1970a). Ridge regression: Biased estimate for non orthogonal problems. *Technometrics*, **12**, 55-67.

Hoerl, A.E. et Kennard, R.W. (1970b). Ridge regression: Applications to non orthogonal problems. *Technometrics*, **12**, 69-82.

Johnson, L.W. et Riess, R.D. (1982). *Numerical Analysis*. Addison-Wesley.

Law, A.M. et Kelton, W. (1982). *Simulation Modeling and Analysis*. McGraw-Hill, Inc.

Lehmann, E.L. (1975). *Nonparametrics: Statistical Methods Based on Ranks*. Holden Day, California.

Levene, H. (1960). Robust tests for the equality of variances, in *Contributions to Probability and Statistics*, ed. I. Olkin. Palo Alto, Stanford University Press, 278-292.

Neter, J., Wasserman, W. et Kutner, M.H. (1985). *Applied Linear Statistical Models*. Irwin, Homewood, Illinois.

Perron, H. et al. (1993a). Guide de l'usager de ***RéMuS***.

Perron, H. et al. (1993b). Guide du programmeur de ***RéMuS*** .

Srivastava, M.S. et Carter, E.M. (1983). *An Introduction to Applied Multivariate Statistics*. North-Holland, New York.

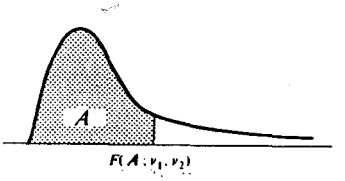
Tukey, J.W. (1957). On the comparative anatomy of transformations. *Annals of Mathematical Statistics*, **28**, 602-632.

Vinod, H. D. (1976). Application of new ridge regression methods to a study of Bell system scale economies. *Journal of the American Statistical Association*, **71**, 835-841.

Wilcoxon, F. (1945). Individual comparison by ranking methods. *Biometrics* , **1**, 80-83.

ANNEXE A

TABLE DE LA LOI DE FISHER



$$F(A;\nu_1,\nu_2)=\frac{1}{F(1-A;\nu_2,\nu_1)}$$

Den. df	A	Numerator df 1	2	3	4	5	6	7	8	9
1	.50	1.00	1.50	1.71	1.82	1.89	1.94	1.98	2.00	2.03
	.90	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9
	.95	161	200	216	225	230	234	237	239	241
	.975	648	800	864	900	922	937	948	957	963
	.99	4,052	5,000	5,403	5,625	5,764	5,859	5,928	5,981	6,022
	.995	16,211	20,000	21,615	22,500	23,056	23,437	23,715	23,925	24,091
	.999	405,280	500,000	540,380	562,500	576,400	585,940	592,870	598,140	602,280
2	.50	0.667	1.00	1.13	1.21	1.25	1.28	1.30	1.32	1.33
	.90	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38
	.95	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4
	.975	38.5	39.0	39.2	39.2	39.3	39.3	39.4	39.4	39.4
	.99	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4
	.995	199	199	199	199	199	199	199	199	199
	.999	998.5	999.0	999.2	999.2	999.3	999.3	999.4	999.4	999.4
3	.50	0.585	0.881	1.00	1.06	1.10	1.13	1.15	1.16	1.17
	.90	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24
	.95	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
	.975	17.4	16.0	15.4	15.1	14.9	14.7	14.6	14.5	14.5
	.99	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3
	.995	55.6	49.8	47.5	46.2	45.4	44.8	44.4	44.1	43.9
	.999	167.0	148.5	141.1	137.1	134.6	132.8	131.6	130.6	129.9
4	.50	0.549	0.828	0.941	1.00	1.04	1.06	1.08	1.09	1.10
	.90	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94
	.95	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
	.975	12.2	10.6	9.98	9.60	9.36	9.20	9.07	8.98	8.90
	.99	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7
	.995	31.3	26.3	24.3	23.2	22.5	22.0	21.6	21.4	21.1
	.999	74.1	61.2	56.2	53.4	51.7	50.5	49.7	49.0	48.5
5	.50	0.528	0.799	0.907	0.965	1.00	1.02	1.04	1.05	1.06
	.90	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32
	.95	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
	.975	10.0	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68
	.99	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2
	.995	22.8	18.3	16.5	15.6	14.9	14.5	14.2	14.0	13.8
	.999	47.2	37.1	33.2	31.1	29.8	28.8	28.2	27.6	27.2
6	.50	0.515	0.780	0.886	0.942	0.977	1.00	1.02	1.03	1.04
	.90	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96
	.95	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
	.975	8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52
	.99	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98
	.995	18.6	14.5	12.9	12.0	11.5	11.1	10.8	10.6	10.4
	.999	35.5	27.0	23.7	21.9	20.8	20.0	19.5	19.0	18.7
7	.50	0.506	0.767	0.871	0.926	0.960	0.983	1.00	1.01	1.02
	.90	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72
	.95	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
	.975	8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82
	.99	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
	.995	16.2	12.4	10.9	10.1	9.52	9.16	8.89	8.68	8.51
	.999	29.2	21.7	18.8	17.2	16.2	15.5	15.0	14.6	14.3

Den. df	A	Numerator df 10	12	15	20	24	30	60	120	∞
1	.50	2.04	2.07	2.09	2.12	2.13	2.15	2.17	2.18	2.20
	.90	60.2	60.7	61.2	61.7	62.0	62.3	62.8	63.1	63.3
	.95	242	244	246	248	249	250	252	253	254
	.975	969	977	985	993	997	1,001	1,010	1,014	1,018
	.99	6,056	6,106	6,157	6,209	6,235	6,261	6,313	6,339	6,366
	.995	24,224	24,426	24,630	24,836	24,940	25,044	25,253	25,359	25,464
	.999	605,620	610,670	615,760	620,910	623,500	626,100	631,340	633,970	636,620
2	.50	1.34	1.36	1.38	1.39	1.40	1.41	1.43	1.43	1.44
	.90	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.48	9.49
	.95	19.4	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
	.975	39.4	39.4	39.4	39.4	39.5	39.5	39.5	39.5	39.5
	.99	99.4	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5
	.995	199	199	199	199	199	199	199	199	200
	.999	999.4	999.4	999.4	999.4	999.5	999.5	999.5	999.5	999.5
3	.50	1.18	1.20	1.21	1.23	1.23	1.24	1.25	1.26	1.27
	.90	5.23	5.22	5.20	5.18	5.18	5.17	5.15	5.14	5.13
	.95	8.79	8.74	8.70	8.66	8.64	8.62	8.57	8.55	8.53
	.975	14.4	14.3	14.3	14.2	14.1	14.1	14.0	13.9	13.9
	.99	27.2	27.1	26.9	26.7	26.6	26.5	26.3	26.2	26.1
	.995	43.7	43.4	43.1	42.8	42.6	42.5	42.1	42.0	41.8
	.999	129.2	128.3	127.4	126.4	125.9	125.4	124.5	124.0	123.5
4	.50	1.11	1.13	1.14	1.15	1.16	1.16	1.18	1.18	1.19
	.90	3.92	3.90	3.87	3.84	3.83	3.82	3.79	3.78	3.76
	.95	5.96	5.91	5.86	5.80	5.77	5.75	5.69	5.66	5.63
	.975	8.84	8.75	8.66	8.56	8.51	8.46	8.36	8.31	8.26
	.99	14.5	14.4	14.2	14.0	13.9	13.8	13.7	13.6	13.5
	.995	21.0	20.7	20.4	20.2	20.0	19.9	19.6	19.5	19.3
	.999	48.1	47.4	46.8	46.1	45.8	45.4	44.7	44.4	44.1
5	.50	1.07	1.09	1.10	1.11	1.12	1.12	1.14	1.14	1.15
	.90	3.30	3.27	3.24	3.21	3.19	3.17	3.14	3.12	3.11
	.95	4.74	4.68	4.62	4.56	4.53	4.50	4.43	4.40	4.37
	.975	6.62	6.52	6.43	6.33	6.28	6.23	6.12	6.07	6.02
	.99	10.1	9.89	9.72	9.55	9.47	9.38	9.20	9.11	9.02
	.995	13.6	13.4	13.1	12.9	12.8	12.7	12.4	12.3	12.1
	.999	26.9	26.4	25.9	25.4	25.1	24.9	24.3	24.1	23.8
6	.50	1.05	1.06	1.07	1.08	1.09	1.10	1.11	1.12	1.12
	.90	2.94	2.90	2.87	2.84	2.82	2.80	2.76	2.74	2.72
	.95	4.06	4.00	3.94	3.87	3.84	3.81	3.74	3.70	3.67
	.975	5.46	5.37	5.27	5.17	5.12	5.07	4.96	4.90	4.85
	.99	7.87	7.72	7.56	7.40	7.31	7.23	7.06	6.97	6.88
	.995	10.2	10.0	9.81	9.59	9.47	9.36	9.12	9.00	8.88
	.999	18.4	18.0	17.6	17.1	16.9	16.7	16.2	16.0	15.7
7	.50	1.03	1.04	1.05	1.07	1.07	1.08	1.09	1.10	1.10
	.90	2.70	2.67	2.63	2.59	2.58	2.56	2.51	2.49	2.47
	.95	3.64	3.57	3.51	3.44	3.41	3.38	3.30	3.27	3.23
	.975	4.76	4.67	4.57	4.47	4.42	4.36	4.25	4.20	4.14
	.99	6.62	6.47	6.31	6.16	6.07	5.99	5.82	5.74	5.65
	.995	8.38	8.18	7.97	7.75	7.65	7.53	7.31	7.19	7.08
	.999	14.1	13.7	13.3	12.9	12.7	12.5	12.1	11.9	11.7

Den. df	A	Numerator df 1	2	3	4	5	6	7	8	9
8	.50	0.499	0.757	0.860	0.915	0.948	0.971	0.988	1.00	1.01
	.90	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56
	.95	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
	.975	7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36
	.99	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
	.995	14.7	11.0	9.60	8.81	8.30	7.95	7.69	7.50	7.34
	.999	25.4	18.5	15.8	14.4	13.5	12.9	12.4	12.0	11.8
9	.50	0.494	0.749	0.852	0.906	0.939	0.962	0.978	0.990	1.00
	.90	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44
	.95	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
	.975	7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03
	.99	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
	.995	13.6	10.1	8.72	7.96	7.47	7.13	6.88	6.69	6.54
	.999	22.9	16.4	13.9	12.6	11.7	11.1	10.7	10.4	10.1
10	.50	0.490	0.743	0.845	0.899	0.932	0.954	0.971	0.983	0.992
	.90	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35
	.95	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
	.975	6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78
	.99	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
	.995	12.8	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97
	.999	21.0	14.9	12.6	11.3	10.5	9.93	9.52	9.20	8.96
12	.50	0.484	0.735	0.835	0.888	0.921	0.943	0.959	0.972	0.981
	.90	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21
	.95	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
	.975	6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44
	.99	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
	.995	11.8	8.51	7.23	6.52	6.07	5.76	5.52	5.35	5.20
	.999	18.6	13.0	10.8	9.63	8.89	8.38	8.00	7.71	7.48
15	.50	0.478	0.726	0.826	0.878	0.911	0.933	0.949	0.960	0.970
	.90	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09
	.95	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
	.975	6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12
	.99	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
	.995	10.8	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54
	.999	16.6	11.3	9.34	8.25	7.57	7.09	6.74	6.47	6.26
20	.50	0.472	0.718	0.816	0.868	0.900	0.922	0.938	0.950	0.959
	.90	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96
	.95	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
	.975	5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84
	.99	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
	.995	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96
	.999	14.8	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24
24	.50	0.469	0.714	0.812	0.863	0.895	0.917	0.932	0.944	0.953
	.90	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91
	.95	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
	.975	5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70
	.99	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
	.995	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69
	.999	14.0	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80

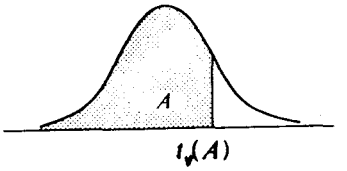
Den. df	Λ	Numerator df								
		10	12	15	20	24	30	60	120	∞
8	.50	1.02	1.03	1.04	1.05	1.06	1.07	1.08	1.08	1.09
	.90	2.54	2.50	2.46	2.42	2.40	2.38	2.34	2.32	2.29
	.95	3.35	3.28	3.22	3.15	3.12	3.08	3.01	2.97	2.93
	.975	4.30	4.20	4.10	4.00	3.95	3.89	3.78	3.73	3.67
	.99	5.81	5.67	5.52	5.36	5.28	5.20	5.03	4.95	4.86
	.995	7.21	7.01	6.81	6.61	6.50	6.40	6.18	6.06	5.95
	.999	11.5	11.2	10.8	10.5	10.3	10.1	9.73	9.53	9.33
9	.50	1.01	1.02	1.03	1.04	1.05	1.05	1.07	1.07	1.08
	.90	2.42	2.38	2.34	2.30	2.28	2.25	2.21	2.18	2.16
	.95	3.14	3.07	3.01	2.94	2.90	2.86	2.79	2.75	2.71
	.975	3.96	3.87	3.77	3.67	3.61	3.56	3.45	3.39	3.33
	.99	5.26	5.11	4.96	4.81	4.73	4.65	4.48	4.40	4.31
	.995	6.42	6.23	6.03	5.83	5.73	5.62	5.41	5.30	5.19
	.999	9.89	9.57	9.24	8.90	8.72	8.55	8.19	8.00	7.81
10	.50	1.00	1.01	1.02	1.03	1.04	1.05	1.06	1.06	1.07
	.90	2.32	2.28	2.24	2.20	2.18	2.16	2.11	2.08	2.06
	.95	2.98	2.91	2.84	2.77	2.74	2.70	2.62	2.58	2.54
	.975	3.72	3.62	3.52	3.42	3.37	3.31	3.20	3.14	3.08
	.99	4.85	4.71	4.56	4.41	4.33	4.25	4.08	4.00	3.91
	.995	5.85	5.66	5.47	5.27	5.17	5.07	4.86	4.75	4.64
	.999	8.75	8.45	8.13	7.80	7.64	7.47	7.12	6.94	6.76
12	.50	0.989	1.00	1.01	1.02	1.03	1.03	1.05	1.05	1.06
	.90	2.19	2.15	2.10	2.06	2.04	2.01	1.96	1.93	1.90
	.95	2.75	2.69	2.62	2.54	2.51	2.47	2.38	2.34	2.30
	.975	3.37	3.28	3.18	3.07	3.02	2.96	2.85	2.79	2.72
	.99	4.30	4.16	4.01	3.86	3.78	3.70	3.54	3.45	3.36
	.995	5.09	4.91	4.72	4.53	4.43	4.33	4.12	4.01	3.90
	.999	7.29	7.00	6.71	6.40	6.25	6.09	5.76	5.59	5.42
15	.50	0.977	0.989	1.00	1.01	1.02	1.02	1.03	1.04	1.05
	.90	2.06	2.02	1.97	1.92	1.90	1.87	1.82	1.79	1.76
	.95	2.54	2.48	2.40	2.33	2.29	2.25	2.16	2.11	2.07
	.975	3.06	2.96	2.86	2.76	2.70	2.64	2.52	2.46	2.40
	.99	3.80	3.67	3.52	3.37	3.29	3.21	3.05	2.96	2.87
	.995	4.42	4.25	4.07	3.88	3.79	3.69	3.48	3.37	3.26
	.999	6.08	5.81	5.54	5.25	5.10	4.95	4.64	4.48	4.31
20	.50	0.966	0.977	0.989	1.00	1.01	1.01	1.02	1.03	1.03
	.90	1.94	1.89	1.84	1.79	1.77	1.74	1.68	1.64	1.61
	.95	2.35	2.28	2.20	2.12	2.08	2.04	1.95	1.90	1.84
	.975	2.77	2.68	2.57	2.46	2.41	2.35	2.22	2.16	2.09
	.99	3.37	3.23	3.09	2.94	2.86	2.78	2.61	2.52	2.42
	.995	3.85	3.68	3.50	3.32	3.22	3.12	2.92	2.81	2.69
	.999	5.08	4.82	4.56	4.29	4.15	4.00	3.70	3.54	3.38
24	.50	0.961	0.972	0.983	0.994	1.00	1.01	1.02	1.02	1.03
	.90	1.88	1.83	1.78	1.73	1.70	1.67	1.61	1.57	1.53
	.95	2.25	2.18	2.11	2.03	1.98	1.94	1.84	1.79	1.73
	.975	2.64	2.54	2.44	2.33	2.27	2.21	2.08	2.01	1.94
	.99	3.17	3.03	2.89	2.74	2.66	2.58	2.40	2.31	2.21
	.995	3.59	3.42	3.25	3.06	2.97	2.87	2.66	2.55	2.43
	.999	4.64	4.39	4.14	3.87	3.74	3.59	3.29	3.14	2.97

Den. df	A	Numerator df 1	2	3	4	5	6	7	8	9
30	.50	0.466	0.709	0.807	0.858	0.890	0.912	0.927	0.939	0.948
	.90	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85
	.95	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
	.975	5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57
	.99	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
	.995	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45
	.999	13.3	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39
60	.50	0.461	0.701	0.798	0.849	0.880	0.901	0.917	0.928	0.937
	.90	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74
	.95	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
	.975	5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33
	.99	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
	.995	8.49	5.80	4.73	4.14	3.76	3.49	3.29	3.13	3.01
	.999	12.0	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69
120	.50	0.458	0.697	0.793	0.844	0.875	0.896	0.912	0.923	0.932
	.90	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68
	.95	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96
	.975	5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22
	.99	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56
	.995	8.18	5.54	4.50	3.92	3.55	3.28	3.09	2.93	2.81
	.999	11.4	7.32	5.78	4.95	4.42	4.04	3.77	3.55	3.38
∞	.50	0.455	0.693	0.789	0.839	0.870	0.891	0.907	0.918	0.927
	.90	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63
	.95	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88
	.975	5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11
	.99	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41
	.995	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62
	.999	10.8	6.91	5.42	4.62	4.10	3.74	3.47	3.27	3.10

Den. df	A	Numerator df 10	12	15	20	24	30	60	120	∞
30	.50	0.955	0.966	0.978	0.989	0.994	1.00	1.01	1.02	1.02
	.90	1.82	1.77	1.72	1.67	1.64	1.61	1.54	1.50	1.46
	.95	2.16	2.09	2.01	1.93	1.89	1.84	1.74	1.68	1.62
	.975	2.51	2.41	2.31	2.20	2.14	2.07	1.94	1.87	1.79
	.99	2.98	2.84	2.70	2.55	2.47	2.39	2.21	2.11	2.01
	.995	3.34	3.18	3.01	2.82	2.73	2.63	2.42	2.30	2.18
	.999	4.24	4.00	3.75	3.49	3.36	3.22	2.92	2.76	2.59
60	.50	0.945	0.956	0.967	0.978	0.983	0.989	1.00	1.01	1.01
	.90	1.71	1.66	1.60	1.54	1.51	1.48	1.40	1.35	1.29
	.95	1.99	1.92	1.84	1.75	1.70	1.65	1.53	1.47	1.39
	.975	2.27	2.17	2.06	1.94	1.88	1.82	1.67	1.58	1.48
	.99	2.63	2.50	2.35	2.20	2.12	2.03	1.84	1.73	1.60
	.995	2.90	2.74	2.57	2.39	2.29	2.19	1.96	1.83	1.69
	.999	3.54	3.32	3.08	2.83	2.69	2.55	2.25	2.08	1.89
120	.50	0.939	0.950	0.961	0.972	0.978	0.983	0.994	1.00	1.01
	.90	1.65	1.60	1.55	1.48	1.45	1.41	1.32	1.26	1.19
	.95	1.91	1.83	1.75	1.66	1.61	1.55	1.43	1.35	1.25
	.975	2.16	2.05	1.95	1.82	1.76	1.69	1.53	1.43	1.31
	.99	2.47	2.34	2.19	2.03	1.95	1.86	1.66	1.53	1.38
	.995	2.71	2.54	2.37	2.19	2.09	1.98	1.75	1.61	1.43
	.999	3.24	3.02	2.78	2.53	2.40	2.26	1.95	1.77	1.54
∞	.50	0.934	0.945	0.956	0.967	0.972	0.978	0.989	0.994	1.00
	.90	1.60	1.55	1.49	1.42	1.38	1.34	1.24	1.17	1.00
	.95	1.83	1.75	1.67	1.57	1.52	1.46	1.32	1.22	1.00
	.975	2.05	1.94	1.83	1.71	1.64	1.57	1.39	1.27	1.00
	.99	2.32	2.18	2.04	1.88	1.79	1.70	1.47	1.32	1.00
	.995	2.52	2.36	2.19	2.00	1.90	1.79	1.53	1.36	1.00
	.999	2.96	2.74	2.51	2.27	2.13	1.99	1.66	1.45	1.00

ANNEXE B

TABLE DE LA LOI DE STUDENT



	A						
ν	.60	.70	.80	.85	.90	.95	.975
1	0.325	0.727	1.376	1.963	3.078	6.314	12.706
2	0.289	0.617	1.061	1.386	1.886	2.920	4.303
3	0.277	0.584	0.978	1.250	1.638	2.353	3.182
4	0.271	0.569	0.941	1.190	1.533	2.132	2.776
5	0.267	0.559	0.920	1.156	1.476	2.015	2.571
6	0.265	0.553	0.906	1.134	1.440	1.943	2.447
7	0.263	0.549	0.896	1.119	1.415	1.895	2.365
8	0.262	0.546	0.889	1.108	1.397	1.860	2.306
9	0.261	0.543	0.883	1.100	1.383	1.833	2.262
10	0.260	0.542	0.879	1.093	1.372	1.812	2.228
11	0.260	0.540	0.876	1.088	1.363	1.796	2.201
12	0.259	0.539	0.873	1.083	1.356	1.782	2.179
13	0.259	0.537	0.870	1.079	1.350	1.771	2.160
14	0.258	0.537	0.868	1.076	1.345	1.761	2.145
15	0.258	0.536	0.866	1.074	1.341	1.753	2.131
16	0.258	0.535	0.865	1.071	1.337	1.746	2.120
17	0.257	0.534	0.863	1.069	1.333	1.740	2.110
18	0.257	0.534	0.862	1.067	1.330	1.734	2.101
19	0.257	0.533	0.861	1.066	1.328	1.729	2.093
20	0.257	0.533	0.860	1.064	1.325	1.725	2.086
21	0.257	0.532	0.859	1.063	1.323	1.721	2.080
22	0.256	0.532	0.858	1.061	1.321	1.717	2.074
23	0.256	0.532	0.858	1.060	1.319	1.714	2.069
24	0.256	0.531	0.857	1.059	1.318	1.711	2.064
25	0.256	0.531	0.856	1.058	1.316	1.708	2.060
26	0.256	0.531	0.856	1.058	1.315	1.706	2.056
27	0.256	0.531	0.855	1.057	1.314	1.703	2.052
28	0.256	0.530	0.855	1.056	1.313	1.701	2.048
29	0.256	0.530	0.854	1.055	1.311	1.699	2.045
30	0.256	0.530	0.854	1.055	1.310	1.697	2.042
40	0.255	0.529	0.851	1.050	1.303	1.684	2.021
60	0.254	0.527	0.848	1.045	1.296	1.671	2.000
120	0.254	0.526	0.845	1.041	1.289	1.658	1.980
∞	0.253	0.524	0.842	1.036	1.282	1.645	1.960

	A						
ν	.98	.985	.99	.9925	.995	.9975	.9995
1	15.895	21.205	31.821	42.434	63.657	127.322	636.590
2	4.849	5.643	6.965	8.073	9.925	14.089	31.598
3	3.482	3.896	4.541	5.047	5.841	7.453	12.924
4	2.999	3.298	3.747	4.088	4.604	5.598	8.610
5	2.757	3.003	3.365	3.634	4.032	4.773	6.869
6	2.612	2.829	3.143	3.372	3.707	4.317	5.959
7	2.517	2.715	2.998	3.203	3.499	4.029	5.408
8	2.449	2.634	2.896	3.085	3.355	3.833	5.041
9	2.398	2.574	2.821	2.998	3.250	3.690	4.781
10	2.359	2.527	2.764	2.932	3.169	3.581	4.587
11	2.328	2.491	2.718	2.879	3.106	3.497	4.437
12	2.303	2.461	2.681	2.836	3.055	3.428	4.318
13	2.282	2.436	2.650	2.801	3.012	3.372	4.221
14	2.264	2.415	2.624	2.771	2.977	3.326	4.140
15	2.249	2.397	2.602	2.746	2.947	3.286	4.073
16	2.235	2.382	2.583	2.724	2.921	3.252	4.015
17	2.224	2.368	2.567	2.706	2.898	3.222	3.965
18	2.214	2.356	2.552	2.689	2.878	3.197	3.922
19	2.205	2.346	2.539	2.674	2.861	3.174	3.883
20	2.197	2.336	2.528	2.661	2.845	3.153	3.849
21	2.189	2.328	2.518	2.649	2.831	3.135	3.819
22	2.183	2.320	2.508	2.639	2.819	3.119	3.792
23	2.177	2.313	2.500	2.629	2.807	3.104	3.768
24	2.172	2.307	2.492	2.620	2.797	3.091	3.745
25	2.167	2.301	2.485	2.612	2.787	3.078	3.725
26	2.162	2.296	2.479	2.605	2.779	3.067	3.707
27	2.158	2.291	2.473	2.598	2.771	3.057	3.690
28	2.154	2.286	2.467	2.592	2.763	3.047	3.674
29	2.150	2.282	2.462	2.586	2.756	3.038	3.659
30	2.147	2.278	2.457	2.581	2.750	3.030	3.646
40	2.123	2.250	2.423	2.542	2.704	2.971	3.551
60	2.099	2.223	2.390	2.504	2.660	2.915	3.460
120	2.076	2.196	2.358	2.468	2.617	2.860	3.373
∞	2.054	2.170	2.326	2.432	2.576	2.807	3.291

ANNEXE C

TABLE DE DURBIN ET WATSON

Level of Significance $\alpha = .05$

	$p-1=1$		$p-1=2$		$p-1=3$		$p-1=4$		$p-1=5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

Level of Significance $\alpha = .01$

	$p-1=1$		$p-1=2$		$p-1=3$		$p-1=4$		$p-1=5$	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.44	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65