Méthodes appliquées de détection et d'estimation de rupture dans des modèles de régression

Yacine Saidi

THESE

présentée à

l' Université Scientifique et Médicale de Grenoble

pour obtenir le grade de
DOCTEUR DE 3ème CYCLE
«Mathématiques appliquées»

par

Yacine SAIDI

METHODES APPLIQUEES DE DETECTION ET D'ESTIMATION DE RUPTURE DANS DES MODELES DE REGRESSION.

Thèse soutenue le 30 janvier 1986 devant la commission d'examen.

B. VAN CUTSEM	Président
F. BRODEAU	
D. DUBAN	Examinateurs
A. LE BRETON	
T. PHAM DINH	

UNIVERSITE SCIENTIFIQUE ET MEDICALE DE GRENOBLE

Année universitaire 1982-1983

Président de l'Université : M. TANCHE

MEMBRES DU CORPS ENSEIGNANT DE L'U.S.M.G.

(RANG A)

SAUF ENSEIGNANTS EN MEDECINE ET PHARMACIE

PROFESSEURS DE 1ère CLASSE

ARNAUD Paul	Chimie organique
ARVIEU Robert	Physique nucléaire I.S.N.
AUBERT Guy	Physique C.N.R.S.
AYANT Yves	Physique approfondie
BARBIER Marie-Jeanne	Electrochimie
BARBIER Jean-Claude	Physique expérimentale C.N.R.S. (labo de magnétisme)
BARJON Robert	Physique nucléaire I.S.N.
BARNOUD Fernand	Biosynthèse de la cellulose-Biologie
BARRA Jean-René	Statistiques - Mathématiques appliquées
BELORISKY Elie	Physique
BENZAKEN Claude (M.)	Mathématiques pures
BERNARD Alain	Mathématiques pures
BERTRANDIAS Françoise	Mathématiques pures
BERTRANDIAS Jean-Paul	Mathématiques pures
BILLET Jean	Géographie
BONNIER Jean-Marie	Chimie générale
BOUCHEZ Robert	Physique nucléaire I.S.N.
BRAVARD Yves	Géographie
CARLIER Georges	Biologie végétale
CAUQUIS Georges	Chimie organique
CHIBON Pierre	Biologie animale
COLIN DE VERDIERE Yves	Mathématiques pures
CRABBE Pierre (détaché)	C.E.R.M.O.
CYROT Michel	Physique du solide
DAUMAS Max	Géographie
DEBELMAS Jacques	Géologie générale
DEGRANGE Charles	Zoologie
DELOBEL Claude (M.)	M.I.A.G. Mathématiques appliquées
DEPORTES Charles	Chimie minérale
DESRE Pierre	Electrochimie
DOLIQUE Jean-Michel	Physique des plasmas
DUCROS Pierre	Cristallographie
FONTAINE Jean-Marc	Mathématiques pures
GAGNAIRE Didier	Chimie physique

.../...

GASTINEL Noël	Analyse numérique - Mathématiques appliquées
GERBER Robert	Mathématiques pures
GERMAIN Jean-Pierre	Mécanique
GIRAUD Pierre	Géologie
IDELMAN Simon	Physiologie animale
JANIN Bernard	Géographie
JOLY Jean-René	Mathématiques pures
JULLIEN Pierre	Mathématiques appliquées
KAHANE André (détaché DAFCO)	Physique
KAHANE Josette	Physique
KOSZUL Jean-Louis	Mathématiques pures
KRAKOWIAK Sacha	Mathématiques appliquées
KUPTA Yvon	Mathématiques pures
LACAZE Albert	Thermodynamique
LAJZEROWICZ Jeannine	Physique
LAJZEROWICZ Joseph	Physique
LAURENT Pierre	Mathématiques appliquées
DE LEIRIS Joël	Biologie
LLIBOUTRY Louis	Géophysique
LOISEAUX Jean-Marie	Sciences nucléaires I.S.N.
LOUP Jean	Géographie
MACHE Régis	Physiologie végétale
MAYNARD Roger	Physique du solide
MICHEL Robert	Minéralogie et pétrographie (géologie)
MOZIERES Philippe	Spectrométrie - Physique
OMONT Alain	Astrophysique
OZENDA Paul	Botanique (biologie végétale)
PAYAN Jean-Jacques (détaché)	Mathématiques pures
PEBAY PEYROULA Jean-Claude	Physique
PERRIAUX Jacques	Géologie
PERRIER Guy	Géophysique
PIERRARD Jean-Marie	Mécanique
RASSAT André	Chimie systématique
RENARD Michel	Thermodynamique
RICHARD Lucien	Biologie végétale
RINAUDO Marguerite	Chimie CERMAV
SENGEL Philippe	Biologie animale
SERGERAERT Francis	Mathématiques pures
SOUTIF Michel	Physique
VAILLANT François	Zoologie
VALENTIN Jacques	Physique nucléaire I.S.N.
VAN CUTSEN Bernard	Mathématiques appliquées
VAUQUOIS Bernard	Mathématiques appliquées
VIALON Pierre	Géologie

PROFESSEURS DE 2ème CLASSE

ADIBA Michel	Mathématiques pures
ARMAND Gilbert	Géographie

.../...

AURIAULT Jean-Louis	Mécanique
BEGUIN Claude (M.)	Chimie organique
BOEHLER Jean-Paul	Mécanique
BOITET Christian	Mathématiques appliquées
BORNAREL Jean	Physique
BRUN Gilbert	Biologie
CASTAING Bernard	Physique
CHARDON Michel	Géographie
COHENADDAD Jean-Pierre	Physique
DENEUVILLE Alain	Physique
DEPASSEL Roger	Mécanique des fluides
DOUCE Roland	Physiologie végétale
DUFRESNOY Alain	Mathématiques pures
GASPARD François	Physique
GAUTRON René	Chimie
GIDON Maurice	Géologie
GIGNOUX Claude (M.)	Sciences nucléaires I.S.N.
GUITTON Jacques	Chimie
HACQUES Gérard	Mathématiques appliquées
HERBIN Jacky	Géographie
HICTER Pierre	Chimie
JOSELEAU Jean-Paul	Biochimie
KERCKOVE Claude (M.)	Géologie
LE BRETON Alain	Mathématiques appliquées
LONGEQUEUE Nicole	Sciences nucléaires I.S.N.
LUCAS Robert	Physiques
LUNA Domingo	Mathématiques pures
MASCLE Georges	Géologie
NEMOZ Alain	Thermodynamique (CNRS - CRTBT)
OUDET Bruno	Mathématiques appliquées
PELMONT Jean	Biochimie
PERRIN Claude (M.)	Sciences nucléaires I.S.N.
PFISTER Jean-Claude (détaché)	Physique du solide
PIBOULE Michel	Géologie
PIERRE Jean-Louis	Chimie organique
RAYNAUD Hervé	Mathématiques appliquées
ROBERT Gilles	Mathématiques pures
ROBERT Jean-Bernard	Chimie physique
ROSSI André	Physiologie végétale
SAKAROVITCH Michel	Mathématiques appliquées
SARROT REYNAUD Jean	Géologie
SAXOD Raymond	Biologie animale
SOUTIF Jeanne	Physique
SCHOOL Pierre-Claude	Mathématiques appliquées
STUTZ Pierre	Mécanique
SUBRA Robert	Chimie
VIDAL Michel	Chimie organique
VIVIAN Robert	Géographie

Je remercie Monsieur **B. VAN CUTSEM**, Professeur à l' USMG pour l'honneur qu'il me fait en acceptant de présider ce jury.

J'adresse ma reconnaissance à Monsieur **F. BRODEAU**, Professeur à l' USSG d'avoir dirigé ce travail. Pour sa disponibilité et ses encouragements, je veux lui témoigner toute ma gratitude.

Je remercie vivement,

Monsieur **A. LE BRETON**, Professeur à l'USMG et Monsieur **T. PHAM DINH**, Chargé de recherches au Laboratoire TIM3 d'avoir accepté de critiquer la rédaction de ce travail et de participer à ce jury.

Monsieur **DUBAN**, Ingénieur à la division technique de EDF-Grenoble de l'intérêt qu'il porte à ce travail en acceptant de faire partie du jury.

J'adresse, un salut amical à tout ceux que mon séjour à la Tour de Mathématiques m'a fait connaître et apprécier.

Merci à **HELENE** et au Service de Reprographie de la Tour IRMA pour la frappe et le tirage de cette thèse.

TABLE DES MATIERES

INTRODUCTION

Lorsqu'on observe un phénomène sur une longue période, on peut se demander s'il est stable pour toute la période étudiée. Il peut arriver que certaines circonstances provoquent une altération du modèle à partir d'une certaine date. Cette altération se traduit par un changement de certains paramètres du modèle. Citons deux situations qui reflétent cette idée d'instabilité du modèle.

Le dérèglement d'une machine peut entraîner une augmentation de la proportion des objets défectueux dans une production. L'application d'un traitement sur un malade peut influer sur l'évolution de la maladie ou plus précisement sur certains de ses symptômes.

Dans beaucoup de domaines des situations comparables se produisent et viennent enrichir la classe des modèles que l'on dit avec rupture. Ainsi dans [22], un modèle de régression à deux droites pour rendre compte de la liaison entre la variable serum-créatinine, substance indiquant le niveau de fonctionnement du rein, et la variable temps permet d'apporter un éclaircissement sur le phénomène de rejet chez les personnes ayant subi une transplantation rénale.

De même en hydrologie, on soupçonne la présence d'une ou plusieurs ruptures, selon la structure des rivières, dans la liaison existant entre le débit et la hauteur d'une rivière (cf. chapitre 4).

L'analyse des modèles avec ruptures s'articule autour des deux axes : détection de la rupture et estimation des paramètres du modèle.

On veut être en mesure de dire, au vu de l'observation du phénomène étudié, s'il y a existence de rupture, autrement dit changement des paramètres du modèle, ou non. L'estimation se fait logiquement après avoir réglé le problème de détection.

Les modèles statistiques faisant l'objet de telles questions sont nombreux, citons : variables indépendantes équidistribuées, régression, séries chronologiques, processus continus....

L'article de Page [17] marque semble-t-il le point de départ de l'étude

des modèles avec rupture qui s'est par la suite énormément enrichie. Le problème était de détecter une diminution de la qualité dans une production d'objets. Ce que Page a traduit par un test de détection de changement de la moyenne dans une suite d'observations indépendantes équidistribuées. Beaucoup d'auteurs ont proposé par la suite et pour le même modèle différentes procédures de détection. On peut citer entre autres : Battacharya &Johnson [3]; Pettit [18](1)(2) pour les méthodes non paramètriques; Hawkins [13], Worsley [23] pour les méthodes de vraisemblance; Chernoff &Zacks [7], Sen &Srivastava [21] pour le point de vue bayésien. Plus récemment, Deshayes &Picard [8](1) ont, suite à une comparaison asymptotique de type Bahadur entre certains tests, conclu à la supériorité de la procédure basée sur le rapport des vraisemblances.

En parallèle, d'autres auteurs se sont intéressés au problème de rupture dans un modèle de régression. On peut citer :

Quandt [19](2), Brown &Al [5] et McCabe &Harrisson [6] pour les tests basés sur les résidus, Garbade[12], Deshayes &Picard [8](3) pour une approche du problème d'un point de vue asymptotique et Basseville [2] pour une revue.

Lorsque l'hypothèse d'une rupture pour un modèle de régression est retenue, on a envie de savoir avec exactitude où elle a lieu. Cela conduit alors à poser le problème de l'estimation dans un modèle de régression, "continu" ou non, à deux ou plusieurs phases. Dans le cas "continu", les points de "jonction" des phases sont supposés inconnus.

Cet aspect a donné lieu à des techniques d'estimation basées sur le critère des moindres carrés (Hudson [15]), ou la maximisation de la vraisemblance (Hinkley [14](1)(2)).

Par ailleurs, le comportement asymptotique des estimateurs est discuté par Hinkley dans les mêmes articles et par Feder dans [11].

Notre travail est consacré à l'étude des deux problèmes : détection et

estimation, dans un modèle de régression.

Dans une première partie, correspondant aux chapitres1 et 2, on traite de la détection de rupture dans un modèle de régression.

Au chapitre 1, on présente le test de la somme cumulée introduit par Brown &al. Après avoir construit la statistique du test à partir des résidus récursifs et étudié son comportement asymptotique sous l'hypothèse de stabilité du modèle, on discute de la qualité de l'approximation de la statistique de test par un mouvement brownien.

Dans le but d'appliquer la procédure de détection à un problème pratique, nous étudions, à l'aide de simulations, la détermination de la fonction frontière qui sert à définir la région critique du test.

On entreprend aussi une étude empirique de la puissance du test en faisant varier les deux paramètres qui caractérisent la rupture à savoir : son emplacement et son amplitude.

Le chapitre 2 est consacré à l'étude du test du rapport de vraisemblances maximales. Pour les deux cas, où la variance σ^2 du bruit du modèle est connue et inconnue, on détermine les statistiques du test et on étudie leur loi exacte sous l'hypothèse de stabilité du modèle sans pour autant arriver à des lois connues. On montre, pour un modèle de régression simple, comment on peut calculer numériquement la loi exacte de la statistique du test du rapport de vraisemblances.

Dans une deuxième partie du chapitre, une étude empirique est menée pour déterminer certains fractiles, dans le cas où σ^2 est inconnue, de la loi de la statistique du test du R.V.M. Les valeurs ainsi déterminées nous permettent d'appliquer le test dans le cadre de l'étude pratique mentionnée plus haut et dont nous donnons le détail au chapitre 4.

On termine le chapitre en étudiant la puissance du test. Ceci nous permet alors une comparaison avec le test de la somme cumulée.

Le chapitre 3 est consacré à l'estimation dans un modèle de régression à une rupture, respectant une contrainte de continuité. On présente dans ce chapitre, les techniques développées par Hinkley et basées sur la maximisation de la vraisemblance. On intégre dans l'algorithme d'estimation certaines propriétés inhérentes au procédé de détermination de l'estimateur du point d'intersection des deux phases de la régression. On montre aussi comment s'applique la méthode des multiplicateurs de Lagrange au problème de l'estimation dans un modèle de régression à plusieurs phases sous l'hypothèse que les points de "jonction" sont connus. Au dernier paragraphe de ce chapitre, on aborde le problème de la consistance des estimateurs des moindres carrés, qui se confondent avec ceux du maximum de vraisemblance lorsque les résidus sont gaussiens, dans un modèle de régression à une ou plusieurs ruptures.

Enfin nous développons au chapitre 4 une étude appliquée traitant d'un problème en hydrologie. Dans le but de tester l'existence d'un changement dans la liaison entre le Débit et la Hauteur d'eau pour une rivière donnée, nous appliquons les deux procédures de détection précédemment citées. Lorsque l'hypothèse de l'existence d'un changement (ou d'une rupture) est retenue, nous estimons, à l'aide de l'algorithme du chapitre 3, les phases du modèle expliquant la liaison entre les deux variables : Débit et Hauteur.

CHAPITRE II

TEST DE LA SOMME CUMULEE DES RESIDUS RECURSIFS. ETUDE THEORIQUE ET SIMULATION.

1. MODELE ET NOTATIONS :

On suppose observer n réalisations indépendantes $y_1, \ldots, y_n$ d'un phénomène dépendant du temps représenté par un modèle de régression du type :

$$(1) \qquad y_t = x_t^T \beta_t + e_t \quad , t = 1 \ldots n .$$

y_t désigne l'observation du phénomène à l'instant t.

x_t désigne l'observation à l'instant t de la variable explicative à valeurs dans $\mathbb{R}^k$. Pour toute la suite les observations x_t sont supposées déterministes comme dans un modèle linéaire.

$(e_t)_{1 \leq t \leq n}$ est une suite de variables aléatoires indépendantes, de même loi gaussienne, centrée, de variance σ^2, définies sur un même espace probabilisé $(\Omega, \mathcal{A}, P)$.

Pour tout $t \in \{1 \ldots n\}$, β_t est un paramètre, dépendant du temps, à valeurs dans $\mathbb{R}^k$.

Définition 1 :

On dit que le modèle (1) admet une rupture à l'instant r_0 si :

$$\beta_t = \beta^* \quad \text{pour } t \in \{1 \ldots r_0\}$$

$$\beta_t = \beta^{**} \text{ pour } t \in \{r_0+1, \ldots, n\} \quad \text{et} \quad \beta^* \neq \beta^{**}.$$

Pour toute la suite une procédure de détection de rupture dans le modèle (1) consiste en un test de stabilité dans le temps des paramètres de régression $(\beta_t)_{t=1 \ldots n}$. Les hypothèses à tester peuvent s'écrire :

(H_0) : hypothèse nulle. Il y a absence de rupture.

$$\beta_t = \beta, \text{ pour tout } t \in \{1, \ldots, n\}$$

(H_1) : il existe une rupture à l'instant r_0,

$$\beta_t = \beta^* \qquad 1 \leq t \leq r_0$$

$$\beta_t = \beta^{**} \qquad r_0+1 \leq t \leq n \qquad \beta^* \neq \beta^{**}.$$

L'instant de rupture r_0 est inconnu. On le suppose toutefois prendre ses valeurs dans (k, ... , n-k). Cette condition est nécessaire à l'identification des paramètres β^* et β^{**} sous l'hypothèse H_1.

On se propose comme nous l'avons annoncé dans l'introduction générale de présenter la procédure de la somme cumulée des résidus récursifs (cf. définition plus loin) due à Brown &al [5]. Nous avons choisi d'étudier cette procédure car elle est représentative, dans sa mise en oeuvre, d'autres procédures que l'on présente sous le label "cusum" (somme cumulée). Citons les tests dus à Mc Cabé & Harrisson [6] et P.K. Sen [cf. 2] basés tous les deux sur des sommes cumulées de résidus. De plus la détermination explicite des régions critiques de ces tests est basée sur une approximation des statistiques de somme cumulée par un mouvement brownien ou un pont brownien (pour la statistique de P.K. Sen). Il est donc intéressant de voir comment se fait cette approximation et quelles en sont les conséquences sur la validité du test.

On présente dans le paragraphe suivant la statistique de la somme cumulée des résidus récursifs. Après avoir introduit les résidus récursifs et présenté leurs propriétés, on construit la statistique du test et on étudie son comportement sous l'hypothèse nulle.

2. LE TEST DE LA SOMME CUMULEE DES RESIDUS RECURSIFS

2.1. Résidus récursifs

Pour tout indice j appartenant à (k,...,n), on définit sous l'hypothèse H_0, $\hat{\beta}_j$ comme étant l'estimateur aux moindres carrés du paramètre β basé sur

les j premières composantes $y_1, \ldots, y_j$. Le paramètre β étant à valeurs dans $\mathbb{R}^k$, il est nécessaire d'avoir au moins k observations pour l'estimer. En notant : $X_j^T = (x_1, \ldots, x_j)$ et $Y_j^T = (y_1, \ldots, y_j)$ on a :

$\hat{\beta}_j = (X_j^T X_j)^{-1} X_j^T Y_j$, $j = k, \ldots, n$, sous la condition que $(X_j^T X_j)^{-1}$ existe pour tout j, $j = k, \ldots, n$.

Définition 2 :

On appelle résidu récursif, ou innovation en filtrage, d'ordre j la variable aléatoire w_j définie par :

$$w_j = (y_j - x_j^T \hat{\beta}_{j-1}) / [1 + x_j^T (X^T_{j-1} X_{j-1})^{-1} x_j]^{1/2}, \quad j = k+1, \ldots, n$$

En se réferant à [5], on peut énoncer la proposition suivante.

Proposition 1 :

Sous l'hypothèse nulle H_0, les résidus récursifs $(w_j)_{j=k+1\ldots n}$ sont indépendants, gaussiens, de moyenne nulle et de variance σ^2.

L'utilisation des résidus $(w_j)_{j=k+1\ldots n}$ est d'un double intérêt. Ces résidus, comme le montre la proposition 1, ont un intérêt statistique évident puisqu'ils sont indépendants et de même loi sous H_0. Ce qui n'est pas le cas des résidus ordinaires. De plus, on peut les calculer "récursivement" d'où leur qualificatif. D'un point de vue pratique, ceci est un avantage appréciable. On peut les obtenir en se servant des relations.

$$(X^T_j X_j)^{-1} = (X^T_{j-1} X_{j-1})^{-1} - \frac{(X^T_{j-1} X_{j-1})^{-1} x_j x_j^T (X^T_{j-1} X_{j-1})^{-1}}{1 + x_j^T (X^T_{j-1} X_{j-1})^{-1} x_j}$$

$$\hat{\beta}_j = \hat{\beta}_{j-1} + (X^T_j X_j)^{-1} x_j (y_j - x_j^T \hat{\beta}_{j-1}) \qquad j = k+1, \ldots, n.$$

La première relation est due à Bartlett. Elle permet d'éviter le calcul direct des matrices inverses $(X^T_j X_j)^{-1}$ à chaque étape j.

Proposition 2 :

Si on note r_0 la vraie valeur de l'instant de rupture, les résidus $(w_j)_{j=k+1...n}$ restent, sous l'hypothèse H_1, indépendants, gaussiens et de même variance σ^2. Mais leur espérance est modifiée à partir de l'instant r_0+1.

En effet, sous l'hypothèse H_1 on a :

$$E(w_j) = \begin{cases} 0 & \text{pour } j = k+1, \ldots, r_0. \\ \dfrac{x_j^T (X^T_{j-1} X_{j-1})^{-1} (X^T_{r0} X_{r0})(\beta^{**} - \beta^{*})}{[1 + x_j^T (X^T_{j-1} X_{j-1})^{-1} x_j]^{1/2}} & j = r_0+1, \ldots, n. \end{cases}$$

L'existence d'une rupture à l'instant r_0 a donc un effet de biais sur les résidus $(w_j)_{j=k+1...n}$ à partir de l'instant r_0. Si l'on se base sur une statistique de somme cumulée de ces résidus récursifs la détection d'une rupture est attendue dès que cette statistique s'éloigne significativement de zéro.

2.2. La statistique de la somme cumulée:

Pour tout r appartenant à (k+1,...,n) on note $S_{r-k} = \sum_{j=k+1,r} w_j$ la somme cumulée des résidus récursifs jusqu'à l'instant r. On a le résultat suivant [6] :

Lemme 1 :

Sous l'hypothèse H_0, $(S_{r-k},\ k+1 \leq r \leq n)$ est un processus gaussien en temps discret, d'espérance nulle et de covariance donnée par :

$$E(S_{r-k}\, S_{m-k}) = (\min(r,m) - k)\,\sigma^2$$

$$k+1 \leq r \leq n \;;\; k+1 \leq m \leq n.$$

Le processus $(S_{r-k})_r$ ainsi caractérisé constitue l'élément de base pour la construction du test de la somme cumulée des résidus récursifs.

Pour tout n fixé, on associe au processus en temps discret $(S_{r-k})_r$ la fonction aléatoire Z_n que l'on précise ci-après. Celle-ci permet au passage de montrer un résultat de convergence qui servira à l'approximation du niveau du test de la somme cumulée.

Pour tout $\omega \in \Omega$, $Z_n(\omega)$ est un élément de $C_{[0,1]}$, l'espace des fonctions continues sur [0,1]. On note $Z_n(t)$ la valeur de la fonction $Z_n(\omega)$ au point t appartenant à [0,1].

(i) Pour les éléments (r-k)/(n-k) de [0,1], on pose :

$$Z_n\,((r-k)/(n-k)) = (1/\sigma\sqrt{n-k})\; S_{r-k}, \quad r = k+1 \ldots n$$

(ii) Pour les autres éléments de]0,1], on définit $Z_n(t)$ par interpolation linéaire :

si $t \in [(r-k-1/n-k), (r-k/n-k)]$, $r = k+1 \ldots n$, on note :

$Z_n(t) = ((r-k)/(n-k)-t)(n-k)Z_n((r-k-1)/n)+(t-((r-k-1)/n-k))(n-k)Z_n((r-k)/n)$

En utilisant (i), $Z_n(t)$ peut se mettre sous la forme suivante :

$$Z_n(t) = (1/\sigma\sqrt{n-k})\, S_{r-k-1}+((n-k)t/\sigma\sqrt{n-k}).w_r-((r-k-1)/\sigma\sqrt{n-k})w_r.$$

Sachant que $[(n-k)t] = r-k-1$, $Z_n(t)$ peut alors s'écrire :

$$Z_n(t) = (1/\sigma\sqrt{n-k})\, S_{[(n-k)t]}+((n-k)t - [(n-k)t]))/\sigma\sqrt{n-k}\; w_{[(n-k)t]+k+1}.$$

Pour t = 0, on pose : $Z_n(0,.) = 0.$

Les sommes cumulées $(S_{r-k})_r$ sont des sommes partielles de résidus récursifs dont on sait qu'ils sont, d'après la proposition 1, indépendants, identiquement distribués, de moyenne nulle et de variance σ^2.

En conséquence, en se basant sur le théorème de convergence de Donsker sur C[0,1] (cf. théorème 10.1 de Billingsley [4]) on établit le résultat :

Théorème 1 :

Sous l'hypothèse H_0, la suite des fonctions aléatoires Z_n, ainsi définies, converge en distribution vers le mouvement brownien W défini sur [0,1].

Le test de la somme cumulée des résidus récursifs consiste alors à vérifier, sous l'hypothèse H_0, le caractère approximativement brownien du processus $(S_r/\sigma\sqrt{n-k})_{r=k+1\ldots n}$.

2.3. Comment utilise-t-on la statistique de la somme cumulée :

On rejette l'hypothèse H_0, de stabilité des paramètres de régression du modèle, si la fonction aléatoire $Z_n(t)$ en valeur absolue dépasse une certaine fonction frontière f(t) fixée, continue et positive sur l'intervalle [0,1].

Afin de définir parfaitement le test il faut donc faire le choix a priori de la fonction frontière f(t). Un choix approprié peut être fait si on a une information sur l'emplacement du point de rupture. Il peut arriver en effet, dans certaines études pratiques, d'être renseigné sur la localisation probable du point de rupture (situé en début, milieu ou en fin d'observation par exemple).

Si aucune information n'est disponible un choix possible pour la frontière est la famille des paraboles $f_\lambda(t) = \lambda\sqrt{t}$ si on s'appuie sur le fait

qu'asymptotiquement pour tout $t \in [0,1]$, $Z_n(t)$ est centré et de variance égale à t.

Pour ce type de frontière, la région critique du test de la somme cumulée est [il existe $t \in [0,1]$ tel que $|Z_n(t)| > f_\lambda(t)$]. Pour n assez grand la probabilité pour que $|Z_n(t)|$ ait dépassé la frontière à un instant donné, appartenant à [0,1] est la même pour n'importe quel instant t de l'intervalle [0,1]. Dans ce sens certains auteurs parlent de l'optimalité de la famille des frontières paraboles.

Mais le choix de la famille de paraboles, $f_\lambda(t) = \lambda\sqrt{t}$, pose un problème technique. On ne dispose pas dans la littérature du calcul de la probabilité de la région critique dans ce cas.

Cette contrainte a conduit Brown &al. à se restreindre à des fonctions frontières affines. Ils ont choisi la famille des droites tangentes à la parabole $a\sqrt{2t}$ au point $t = 1/2$. Les droites ainsi choisies ont pour équation : $f_a(t) = at + a/2$.

Le choix des droites permet en effet le calcul de la probabilité du franchissement de la frontière par le mouvement brownien. Lorsque ce type de frontière est choisi, la région critique du test s'écrit :

$R_n(a)$ = [il existe t de [0,1] tel que : $|Z_n(t)| > at + a/2$]

On note $\alpha_n(a)$ la probabilité sous l'hypothèse H_0 de la région $R_n(a)$.

$$\alpha_n(a) = P_{H0}(R_n(a))$$

En se basant sur le théorème 1 de convergence, on en déduit :

$$\lim_{n \to +\infty} \alpha_n(a) = \alpha(a),$$

où $\alpha(a)$ = P(il existe t, $t \in [0,1]$ tel que $|W(t)| > at + a/2$)

On sait, en se reportant au travail de Durbin [10], calculer en fonction du paramètre a la probabilité de franchissement d'une droite, sur l'intervalle [0,1], par un mouvement brownien.

Autrement dit la quantité :

$$P \text{ (il existe } t \in [0,1], \text{ t.q } W(t) > at + a/2).$$

En effet, en appliquant un résultat de Durbin (cf. lemme 3 de [10]) on a :

$$P \text{ (il existe } t \in [0,1], \text{ tel que } W(t) > at + a/2) = (1-\Phi(3a/2))+\exp(-a^2).\Phi(a/2).$$

avec $$\Phi(u) = (1/\sqrt{2\Pi}) . \int_{]-\infty,u]} \exp(-(1/2)x^2)\, dx.$$

On en déduit sachant la définition de $\alpha(a)$:

$$\alpha(a) = 2[(1-\Phi(3a/2))+\exp(-a^2).\Phi(a/2)]$$

La détermination de la valeur critique a, de la région $R_n(a)$, associé à un seuil de signification donné $\alpha_n(a)$ se fait, asymtotiquement, en approximant $\alpha_n(a)$ par $\alpha(a)$. On se fixe le seuil $\alpha(a)$ et on en déduit la valeur critique a en se servant de l'identité ci-dessus.

Pour les valeurs des seuils de signification habituellement utilisés on a le tableau1, ci-dessous, des valeurs critiques a.

α (a)	0.01	0.05	0.10
a	2.286	1.896	1.700

Tableau 1 :

Valeurs des seuils théoriques et des paramètres a, de la frontière droite at+a/2, associés.

Dans la pratique, le test de la somme cumulée des résidus récursifs, avec pour frontière la droite f(t) = at + a/2, se fera comme suit :

On rejette l'hypothèse H_0 dès que l'on rencontre un instant r, $k+1 \leq r \leq n$, tel que :

$$|S_{r-k}| / \sigma\sqrt{n-k} > a\,(r-k/n-k) + a/2.$$

En résumé, sous l'hypothèse H_0, le processus discret $(S_{r-k}/\sigma\sqrt{n-k},\ r=k+1,n)$ est approximé par le processus continu Z_n qui converge en distribution vers le mouvement brownien standard. La détermination de la valeur critique de la région de rejet, du test de la somme cumulée basé sur le processus discret $(S_{r-k}/\sigma\sqrt{n-k},\ (r=k+1,\ \dots\ ,\ n))$ se fait asymptotiquement en utilisant la convergence de Z_n vers W.

2.4 Comportement de la somme cumulée sous l'hypothèse H_1 :

On veut étudier le comportement asymptotique de la statistique de la somme cumulée sous H_1.

Le modèle de référence reste le même :

$$(1) \qquad y_i = x_i^T \beta_i + e_i,\ i = 1 \dots n.$$

Le point de vue asymptotique dans un modèle avec rupture consiste à considérer une suite d'hypothèses $H_1(n)$ dépendant de n.

$$H_1(n) : \begin{cases} \beta_i = \beta^* & 1 \le i \le r_0(n) \\ \beta_i = \beta^{**} & r_0(n)+1 \le i \le n \end{cases}$$

où $r_0(n)$ est le point de rupture, dépendant de n, et tel que

$$\lim_n (r_0(n)/n) = t_0 \ ;\ t_0 \in]0,1[.$$

L'amplitude $|\beta^{**}-\beta^*|$ peut dépendre ou non de n. Dans notre cas on la considère fixe et indépendante de n.

Remarquons que ce point de vue est équivalent au suivant : on suppose que l'on se fixe [0,1] comme l'intervalle d'observation.

Pour tout n fixé, on suppose observer un processus continu $(y(t),\ 0 \le t \le 1)$, basé sur le modèle (1), aux différents instants : 1/n, 2/n, ... $r_0(n)/n$, ...

tel que $\lim_n (r_0(n)/n) = t_0 \ ;\ t_0 \in]0,1[$.

L'hypothèse de rupture se traduit par :

$$y_t = (x_t^T \beta^* + e_t)\, \mathbb{1}_{[0,to]}(t) + (x_t^T \beta^{**} + e_t)\, \mathbb{1}_{]to,1]}(t).$$

Ainsi lorsque n est fixé, on confond de ce point de vue l'observation $y_{i/n}$ issue du processus continu et l'observation y_i issue du processus en temps discret et ce pour tout $i \in (1,\ldots,n)$

Cette remarque correspond au cas où le phénomène étudié est réellement continu et que nous l'observons en des instants discrets. Faire croître n revient à subdiviser [0,1] de plus en plus finement et donc à observer le phénomène continu de plus en plus finement aussi.

On retient pour la suite la première approche.

On sait que la présence d'une rupture à l'instant $r_0(n)$ a un effet de biais sur les résidus récursifs $(W_j,\ k+1 \leq j \leq n)$ à partir de $j = r_0(n) + 1$.

Les propriétés des résidus récursifs sous H_1 énoncées dans la proposition 2 entraînent que le processus $(S_{r-k} / \sigma\sqrt{n-k},\ k+1 \leq r \leq n)$ sous H_1 est gaussien et de même structure de covariance que sous H_0. Seule son espérance est modifiée puisque un biais est introduit à partir de l'instant de rupture.

Le biais ainsi provoqué dépend à la fois de l'amplitude $|\beta^{**} - \beta^*|$ de la rupture et des observations de la variable de régression.

La dépendance du biais vis à vis de la variable de régression est gênante. On ne peut espérer évaluer le biais, même asymptotiquement, sans faire d'hypothèses a priori sur les observations $(x(i),\ i = 1 \ldots n)$.

Plusieurs hypothèses peuvent être envisagées de sorte que le calcul du biais devienne possible et que celui-ci ne dépende plus, asymptotiquement, des $(x(i),\ i = 1 \ldots n)$.

On présente un cas particulier où l'hypothèse faite sur la variable de régression conduit à l'évaluation du biais asymptotique du processus $Z_n(t)$

associé à $\{S_{r-k}/\sigma\sqrt{n-k}\ ,\ k+1 \leq r \leq n\}$. La connaissance du biais de $Z_n(t)$ permet ensuite de conclure à la consistance du test.

On se place dans l'hypothèse où la variable de régression est à valeurs dans $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Pour tout $i = 1...n$, on a $x_i = i$.

Ce choix correspond au cas où les observations de la variable de régression sont régulierement espacées, par exemple lorsque la régression se fait par rapport à la variable temps.

Proposition 3 :

On a : $E(Z_n(t)) = 0$ pour $t \in [0, t_0]$

et $\lim_n E(Z_n(t))/n^{3/2} = t_0^2\,((t-t_0)/t)(b^{**}-b^{*})/\sigma$ pour $t \in]t_0, 1]$.

Démonstration

On a :

$$E(W_j) = \begin{cases} 0 & \text{si } j \leq [nt_0] \\ \dfrac{(x_j^T (X_{j-1}^T X_{j-1})^{-1} (X^T_{[nto]} X_{[nto]})(\beta^{**}-\beta^{*}))}{[1+x_j^T (X_{j-1}^T X_{j-1})^{-1} x_j]^{1/2}} & j > [nt_0] \end{cases}$$

$\beta^{*}, \beta^{**} \in \mathbb{R}$; $r_0(n) = [nt_0]$, $t_0 \in [0,1]$.

Pour n tendant vers $+\infty$ et pour tout $j > [nt_0]$, on a les équivalences suivantes :

$$(X_{j-1}^T X_{j-1})^{-1} \simeq 3/j3 \ ; \ (X^T_{[nto]} X_{[nto]}) \simeq 1/3\,[n\,t_0]^3$$

Ce qui implique :

$$E(W_j) \simeq (j\,[n\,t_0]^3 / j^3)(\beta^{**}-\beta^{*})/[1+3/j]^{1/2},$$

$$\text{i.e.} \quad E(W_j) \simeq ([n\,t_0]^3/j^2)(\beta^{**}-\beta^{*}).$$

En se reportant à la définition de $Z_n(t)$, on a : pour tout $t > t_0$,

$$\lim_n E(Z_n(t))/n^{3/2} = \lim_n (1/n^{3/2}\sqrt{n-1}\sigma) \sum_{j=[nt0]+1,[nt]} E(W_j).$$

Il vient, suite à l'équivalence établie pour $E(W_j)$:

$$\lim_n E(Z_n(t))/n^{3/2}$$

$$= \lim_n (1/n^{3/2}\sqrt{n-1}) \sum_{j=[nt0]+1,[nt]} ([nt_0]^3/j^2)(\beta^{**}-\beta^*)/\sigma$$

$$= \lim_n n\, t_0^3 \left[\sum_{j=[nt0]+1,[nt]} 1/j^2 \right] (\beta^{**}-\beta^*)/\sigma$$

On a :

$$\sum_{j=[nt0]+1,[nt]} 1/j^2 = (1/[nt_0] - 1/[nt]) + o(1/n^2)$$

D'où :

$$\lim_n E(Z_n(t))/n^{3/2} = \lim_n (n\, t_0^3 n(t-t_0))/(n^2 t\, t_0)(\beta^{**}-\beta^*)/\sigma$$

$$= t_0^2 (1-t_0/t)(\beta^{**}-\beta^*)/\sigma, \text{ pour } t > t_0.$$

Par ailleurs, puisque $E(W_j) = 0$ pour $j \leq [n\, t_0]$, il est facile d'en déduire $E(Z_n(t)) = 0$ pour $t \leq t_0$.□

Il découle alors de la proposition 3 les remarques suivantes concernant le lien entre le biais de $Z_n(t)$ et l'emplacement de point de rupture t_0 :

Si $t_0 \in]0,1[$, le biais de $Z_n(t)$, en valeur absolue, tend vers $+\infty$ lorsque $n \to +\infty$ et ce pour tout t tel que $t > t_0$.

Si t_0 est aux bords de l'intervalle [0,1], la rupture se trouve en début ou en fin, le biais de $Z_n(t)$ sous H_1 est le même que sous H_0. Il est donc nul pour tout t. On s'attend alors que pour de tels instants de rupture, la statistique du test de la somme cumulée ne soit pas un détecteur efficace.

La proposition 3 permet aussi de déduire la consistance du test dans le cas où l'instant de rupture t_0 appartient à $]0,1[$.

Proposition 4

Sous la même hypothèse concernant le régresseur à valeurs dans $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, le test de la somme cumulée est consistant dès que l'instant t_0, avec $[nt_0] = r_0(n)$, appartient à $]0,1[$.

Preuve :

On sait déjà que si $t_0 \in]0,1[$, $|EZ_n(t)|$ tend vers $+\infty$ quand $n \to +\infty$, pour tout $t > t_0$.

D'autre part la variance de $Z_n(t)$ est asymptotiquement bornée puisqu'elle converge vers t, quel que soit t dans $[0,1]$.

Ainsi il existe $K > 0$ tel que $\text{var } Z_n(t) \leq K$.

On choisit alors $\varepsilon > 0$ et $A > 0$ tel que $4K/A^2 \leq \varepsilon$. On remarque qu'on a toujours l'inclusion suivante :

$$(|EZ_n(t)| > 3A/2) \cap (|Z_n(t) - EZ_n(t)| < A/2) \subset (|Z_n(t)| > A)$$

Ils'ensuit l'inégalité :

$$P(|Z_n(t)| \leq A) \leq P(|EZ_n(t)| \leq 3A/2) + P(|Z_n(t) - EZ_n(t)| \geq A/2)$$

Le premier terme du membre de droite tend vers 0 quand $n \to +\infty$ et pour tout $t > t_0$.

En appliquant l'inégalité de Tchebychev au second terme, il vient :

$$\text{Il existe } n_0 \in \mathbb{N} \text{ tel que } \forall n \geq n_0$$

$$P(|Z_n(t) - EZ_n(t)| \geq A/2) \leq 4 \text{ var } (Z_n(t))/A^2 \leq 4K/A^2 \leq \varepsilon$$

D'où alors :

$\forall\ \varepsilon > 0$, $\forall A > 0$ tel que $4K/A^2 \leq \varepsilon$, il existe $n_0 \in \mathbb{N}$, tel que $\forall n \geq n_0$, on ait :

$$P(|Z_n(t)| \leq A) \leq \varepsilon \text{ et donc } P(|Z_n(t)| > A) > 1 - \varepsilon.$$

Pour conclure à la consistance du test, remarquons que la puissance du test est évaluée par :

$$P_{H1} \text{ (il existe t, } t \in [0,1] \text{ tel que } |Z_n(t)| > f(t))$$

La fonction frontière f(t) étant choisie bornée sur [0,1], il vient alors :

$$\lim_{n \to +\infty} P_{H1} \text{ (il existe t, } t \in [0,1] \text{ tel que } |Z_n(t)| > f(t)) = 1.$$

3. Remarques sur l'approximation de la statistique de la somme cumulée par un processus brownien. Etude par simulation du seuil de signification du test.

L'étude qui précéde suppose que σ^2, variance du bruit gaussien, est connue. Or dans la pratique, celle-ci sera toujours inconnue. Il nous faut donc l'estimer. On retiendra dans notre étude l'estimation suivante :
pour tout r = k+1 ... n, on note

$$\hat{\sigma}^2(r) = (1/n-k-2)\left[\sum_{j=k+1,r}(w(j) - w_{k,r})^2 + \sum_{j=r+1,n}(w(j)-w_{r,n})^2\right]$$

où $\quad w_{k,r} = (1/r-k)\sum_{j=k+1,r} w(j)$; $w_{r,n} = (1/n-r)\sum_{j=r+1,n} w(j)$.

La statistique de test devient alors : $\{(S_{r-k}/\hat{\sigma}(r)\sqrt{n-k}),\ k+1 \leq r \leq n\}$ et on rejettera H_0 dès que l'on aura un instant r, $k+1 \leq r \leq n$, tel que :

$$|S_{r-k}|/\hat{\sigma}(r).\sqrt{n-k} > a((r-k)/n-k) + a/2.$$

On peut remarquer que lorsque σ^2 est supposée inconnue, le test perd sa propriété récursive car l'estimation de σ^2 nécessite le calcul de tous les résidus récursifs {w(i), i = k+1 ... n}.
On aurait pu, pour conserver la propriété récursive du test, penser à estimer σ^2 par $(1/r-k-1)\sum_{i=k+1,r}(w(i) - w_{k,r})^2$ pour $k+1 \leq r \leq n$, mais cette estimation, ne prenant pas en compte assez d'information, risque d'être très trompeuse.

3.1. Etude empirique du seuil de signification du test de la somme cumulée :

On se propose d'estimer, à partir de la simulation d'un modèle de régression, le seuil de signification du test de la somme cumulée basé sur le processus discret $\{S_{r-k} / \hat{\sigma}(r).\sqrt{n-k}\}$.

Les estimations ainsi obtenues nous permettront d'avoir une idée sur la validité et la qualité de l'approximation brownienne.

3.1.1. Simulation du modèle et résultats :

On rappelle que le seuil théorique du test de la somme cumulée des résidus récursifs est donné - suite à l'approximation par le mouvement brownien - par :

$$\alpha_T = P \text{ (il existe } t \in [0,1] \text{ tel que } |W(t)| > f(t))$$

Lorsque la fonction frontière choisie est $f(t) = at + a/2$, on a (cf.tableau 1):

$\alpha_T = 0.01$ pour $a = 2.286$; $\alpha_T = 0.05$ pour $a = 1.896$;

$\alpha_T = 0.10$ pour $a = 1.700$.

Le seuil du test est la probabilité de rejeter l'hypothèse H_0 alors qu'elle est vraie. Lorsque σ^2 est inconnue, on a :

$$\alpha = P_{H0}(\text{il existe } r,\ k+1 \leq r \leq n, \text{ tel que } |S_{r-k}|/\hat{\sigma}(r)\sqrt{n-k} > [a(r-k)/(n-k)+a/2])$$

La simulation consiste à choisir un modèle de régression sans rupture (on se met sous l'hypothèse H_0) et à faire n_e expériences de ce modèle dans le but d'estimer α. Pour chaque valeur de a, on pose :

$E_a = \{$ il existe $r,\ k+1 \leq r \leq n$, tel que $|S_{r-k}|/\hat{\sigma}(r).\sqrt{n-k} > [a(r-k)/(n-k)+a/2]\}$.

Le seuil estimé par simulation est alors :

$$\hat{\alpha} = \{\text{nombre de fois où } E_a \text{ se réalise}\} / n_e .$$

La simulation porte sur le modèle de régression :

(2) $$y_t = c + b\,x_t + e_t, \quad t = 1 \ldots n$$

$(e_t)_{1 \le t \le n}$ sont des observations i.i.d de loi N(0,1) (bruit du modèle)

$(x_t)_{1 \le t \le n}$, n-observations indépendantes de même loi N(0,25) (variable de régression du modèle).

c, b sont les paramètres du modèle que l'on fixe pour toute la simulation.

Une expérience consiste à engendrer n observations indépendantes de loi N(0,1) et n observations indépendantes de loi N(0,25). Pour ce faire nous avons utilisé la méthode de Box et Muller. Elle permet d'engendrer deux variables indépendantes, t_1 et t_2, de loi $N(m,\sigma^2)$ à partir de deux autres variables, R et v, de lois respectives, χ_2^2 et uniforme sur [0,1], par les formules :

$$t_1 = m + \sigma \sqrt{R}.\,\mathrm{Cos}\,(2\Pi v)$$

$$t_2 = m + \sigma \sqrt{R}.\,\mathrm{Sin}\,(2\Pi v)$$

Par ailleurs si u est une variable de loi uniforme sur [0,1], la variable -2 Log u suit une loi de χ_2^2. Les deux formules peuvent s'écrire alors :

$$t_1 = m + \sigma \sqrt{-2\,\mathrm{Log}\,u}\,\mathrm{Cos}\,(2\Pi v)$$

$$t_2 = m + \sigma \sqrt{-2\,\mathrm{Log}\,u}\,\mathrm{Sin}\,(2\Pi v)$$

Elles permettent alors d'engendrer t_1 et t_2 à partir de u et v, toutes deux des lois uniformes sur [0,1].

Nous avons calculé $\hat{\alpha}$ pour certaines valeurs de n, choisies pour représenter des types d'échantillon. Le nombre d'expériences effectuées diffère selon la taille de l'échantillon.

Nous avons réalisé :

- 4000 expériences pour n = 30 (représentatif de petits échantillons)

- 4000 expériences pour n = 60, 2000 expériences pour n = 100 (taille représentatives "d'échantillon moyen")
- 1000 expériences pour n = 200 et n = 500 ("grands échantillons")

Les seuils sont donnés en ‰ dans le tableau 2.

n \ α_T	100	50	10
30	52.5	24	3
60	68	30	4
100	75	32.5	6.5
200	70	39	7
500	34	35	9

Tableau 2 : Estimations du seuil de signification du test de la somme cumulée avec la fonction frontière f(t) = at+a/2.

3.1.2. Commentaire. Résultats complémentaires.

Il apparait au vu du tableau 2 que les seuils théoriques 1%, 5% et 10% sont, dans tous les cas, sous estimés. Ceci est une conséquence logique de l'approximation, asymptotique, de la statistique du test de la somme cumulée par le mouvement brownien sur [0,1]. On peut d'ailleurs remarquer que l'erreur d'estimation (écart entre α_T et $\hat{\alpha}$) semble se réduire lorsque la taille de l'échantillon augmente.

Dans le cas des petits échantillons (n = 30), par exemple, les estimations $\hat{\alpha}$ sont pratiquement la moitié des seuils supposés. L'approche par le mouvement brownien a donc le désavantage d'exagérer le seuil du test de la somme cumulée. Les valeurs du paramètre a (cf. tableau1) sont trop grandes pour les seuils que l'on se fixe. Elles privilégient l'hypothèse

d'absence de rupture H_0 et conduisent nécessairement à l'affaiblissement de la puissance du test qui est l'indicateur du pouvoir de détection du test. Cette remarque, confirmée par les essais de simulation, avait été avancée par Anderson dans la discussion de [5].

Ces conclusions nous ont conduit naturellement à rechercher, par de nombreux essais de simulation, les valeurs du paramètre a, de la fonction frontière at + a/2, qui conduisent aux estimations les plus proches des seuils fixés a priori à 1%, 5% et 10%.

Pour une valeur fixée de a, on calcule l'estimateur $\hat{\alpha}$ correspondant. On retient la valeur de a qui donne un $\hat{\alpha}$ aussi proche que possible de la valeur supposée du seuil.

La nature de la région critique du test :

[il existe r, $k+1 \leq r \leq n$, tel que $|S_{r-k}| / \sigma(r).\sqrt{n-k} > [(a(r-k)/n-k)+a/2]$] nous oblige, pour trouver les "bonnes " valeurs de a, à procéder par des essais multiples, en choisissant à chaque essai une nouvelle valeur de a ajustée selon l'estimation $\hat{\alpha}$ obtenue à l'essai précédent.

On donne les valeurs de a ainsi que les estimations α correspondantes dans les cas : n = 30, n = 60 et n = 100. Les estimations sont obtenues à partir de 1000 expériences dans chaque cas.

n = 30		n = 60		n = 100	
a	α	a	α	a	α
1.500	0.107	1.570	0.106	1.620	0.098
1.700	0.052	1.760	0.05	1.800	0.053
2.120	0.009	2.120	0.011	2.120	0.01

α \ n	30	60	100
0.01	2.120	2.120	2.120
0.05	1.700	1.760	1.800
0.10	1.500	1.570	1.620

Tableau 3 : Valeurs du paramètre a, de la fonction frontière affine f(t) = at +a/2, ajustées par simulation.

On voit que les valeurs de a conduisant à des estimations proches des seuils supposés sont inférieures aux valeurs du tableau 1. L'utilisation des valeurs de a, obtenues par le biais du processus brownien, pourra se faire dans la pratique pour des échantillons de taille supérieure ou égale à 500 car dans ce cas l'erreur commise sur la probabilité de rejeter H_0 à tort n'est pas trop grande. Par contre elle n'est pas du tout conseillée pour de petits et moyens échantillons. En conséquent les valeurs de a, obtenues par simulation, peuvent être d'un grand intérêt pratique. Il sera jugé préférable d'utiliser ces valeurs (pour des échantillons de petite et moyenne taille) plutôt que les valeurs de a du tableau 1.

4. ETUDE PAR SIMULATION DE LA PUISSANCE DU TEST DE LA SOMME CUMULEE :

Nous avons voulu voir quelle serait la conséquence du choix du paramètre a de la fonction frontière f(t) = at + a/2 sur la puissance du test de la somme cumulée. La puissance du test est donnée par la probabilité de rejeter l'hypothèse H_0 alors qu'elle est fausse. Si on la note β, on a :

$$\beta = P_{H1}[\text{il existe } r,\ k+1 \leq r \leq n,\ \text{tel que } |S_{r-k}| / \sigma(r).\sqrt{n-k} > [(a(r-k)/n-k)+a/2]]$$

Nous avons donc simulé un modèle de régression avec rupture :

$$(3) \qquad y_t = \begin{cases} c + b_0 x_t + e_t \ , \ t = 1, \ldots, r_0 \\ c + b_1 x_t + e_t \ , \ t = r_0 + 1, \ldots, n \end{cases}$$

$(e_t)_{1 \leq t \leq n}$ et $(x_t)_{1 \leq t \leq n}$ sont définis de la même façon que dans le modèle (2). Les paramètres de la simulation sont : n, taille de l'échantillon, r_0, instant de la rupture et $\Delta = |b_1 - b_0|$, amplitude de la rupture.

Remarquons que pour le modèle (3), les observations $(x_t)_{1 \leq t \leq n}$ de la variable de régression sont rangées dans l'ordre croissant :

$$x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_n.$$

On considère à nouveau l'événement E_a :

E_a = [il existe r, k+1 ≤ r ≤ n, tel que $|S_{r-k}| / \sigma(r).\sqrt{n-k} > [(a(r-k)/n-k)+a/2]$.

Pour prendre en compte l'influence des deux paramètres r_0 et Δ, instant et amplitude de la rupture, nous avons opéré comme suit : pour r_0 et Δ fixés, on réalise n_e simulations du modèle (3). La puissance du test est estimée par : $\hat{\beta}$ = (nombre de fois où E_a se réalise sous H_1)/n_e.

Nous avons estimé la puissance pour les tailles d'échantillon suivantes : n = 30, n = 60 et n = 100. Dans chaque cas, on fait varier l'emplacement r_0 de la rupture ainsi que son amplitude Δ.

Pour chaque couple (r_0, Δ), on dispose, dans les tableaux 4, 5 et 6 plus loin, deux puissances estimées :

- en première ligne, l'estimateur $\hat{\beta}_1$ avec les valeurs a du tableau 1;
- en deuxième ligne, l'estimateur $\hat{\beta}_2$ avec les valeurs a du tableau 3 (valeurs estimées par simulation).

L'amplitude Δ est prise en fonction d'une unité d'amplitude δ =

$(25.n)^{-1/2}$. Le choix de δ est inspiré de la méthode de simulation retenue par Garbade [12] et reprise par Mc Cabe &Harrisson [6] à des fins de comparaison de certains résultats de simulation.

Pour le calcul effectif de $\hat{\beta}_1$ et $\hat{\beta}_2$, la valeur de a dont on se sert, est donnée dès que l'on se fixe le seuil de signification α du test.

Les puissances sont estimées à partir de 100 expériences et elles sont données en % dans les différents tableaux.

	$r_0 = 5$		$r_0 = 15$		$r_0 = 25$	
α en % →	1	10	5	10	5	10
$\Delta = \delta$	1 5	5 16	2 6	6 12	1 4	4 10
$\Delta = 10\delta$	16 28	28 48	5 16	16 30	0 3	3 9
$\Delta = 50\delta$	58 94	94 100	76 92	92 99	0 3	3 32

n = 30

$\delta \simeq 3.651.\ 10^{-2}$

---

n = 60 $\quad \delta \simeq 2.581.\ 10^{-2}$

	$r_0 = 15$			$r_0 = 30$			$r_0 = 45$			$r_0 = 55$		
α en % →	1	5	10	1	5	10	1	5	10	1	5	10
$\Delta = \delta$	1 1	4 7	8 12	1 1	4 6	7 10	0 0	5 6	6 10	0 0	0 5	5 10
$\Delta = 10\delta$	10 16	31 44	46 55	1 4	9 11	14 23	1 3	6 11	12 21	0 1	2 7	8 11
$\Delta = 50\delta$	100 100	100 100	100 100	95 100	100 100	100 100	73 88	100 100	100 100	0 0	0 2	5 17

	$r_0 = 25$			$r_0 = 50$			$r_0 = 75$			$r_0 = 95$		
α en % →	1	5	10	1	5	10	1	5	10	1	5	10
$\Delta = \delta$	2 2	3 7	12 13	0 1	2 3	5 9	0 0	1 2	2 6	0 1	4 6	7 9
$\Delta = 10\delta$	14 24	34 42	51 57	3 6	17 19	24 28	3 5	12 15	21 26	0 2	4 4	6 12
$\Delta = 50\delta$	100 100	100 100	100 100	100 100	100 100	100 100	99 99	100 100	100 100	1 3	6 8	14 19

$n = 100 \quad \delta \simeq 2\,10^{-2}$

TABLEAU 4,5,6 : Puissances estimées du test de la somme cumulée.

Commentaire :

Une comparaison des puissances estimées $\hat{\beta}_1$ et $\hat{\beta}_2$ permet de conclure : les valeurs de a ajustées par simulation (cf tableau 3) procurent au test de la somme cumulée un meilleur pouvoir de détection de la rupture.

Les écarts assez significatifs entre les puissances estimées $\hat{\beta}_1$ et $\hat{\beta}_2$, principalement pour les tailles 30 et 60, confortent l'hypothèse selon laquelle " l'approximation brownienne n'est pas souhaitable pour de telles tailles d'échantillon".

On peut remarquer que lorsque l'amplitude Δ est fixée, la puissance dans les deux cas, diffère selon l'emplacement r_0 de la rupture. On s'aperçoit, par exemple, que le test de la somme cumulée détecte beaucoup mieux les ruptures situées en début d'observation que celle proches de la fin. A notre avis, ceci se produit pour au moins deux raisons :

- Lorsque une rupture est proche de la fin, il n'y a pas assez d'observations après la rupture, ou pour être plus exact assez de résidus récursifs biaisés, pour permettre à la statistique de la somme cumulée d'être "suffisamment grande" pour dépasser la frontière fixée a priori. Par contre lorsque la rupture a lieu au début, c'est l'effet inverse qui se produit.

- Le type de fonction frontière a aussi son influence. Il aurait été plus naturel de prendre pour frontière la parabole a. $\sqrt{2t}$. Un problème technique conduit à remplacer cette dernière par sa tangente au point 1/2. Cette droite est en tout point supérieure à la parabole. Ce qui a pour conséquence de rendre le franchissement encore plus difficile pour la statistique de la somme cumulée.

L'instant de franchissement de la droite par la statistique de la somme cumulée, lorsqu'il existe, peut être considéré comme un estimateur de l'instant r_0 simulé. Nous nous sommes donc intéressés à ces instants de franchissement lorsque les trois paramètres n, Δ, r_0 varient. Il ressort que dans tous les cas, la détection en pratique se fait avec un grand retard par rapport à la date supposée r_0. L'estimation est donc fortement biaisée. Cela explique la faible performance du test lorsque l'instant de rupture est simulé en fin d'observation, même pour une amplitude de la rupture assez grande

On peut espérer rectifier cette lacune du test en choisissant d'autres fonctions frontières surtout si une information a priori existe sur l'emplacement de la rupture. C'est la particularité du problème qui pourrait décider du choix a priori de la frontière.

CHAPITRE III

LA PROCEDURE DU RAPPORT DE VRAISEMBLANCES MAXIMALES.

1. INTRODUCTION :

Certains auteurs comme [13], [23], [8](2) ont étudié, à l'aide du test de rapport de vraisemblances, le problème de rupture de moyenne dans une suite de variables aléatoires indépendantes et de même loi gaussienne.

Le problème étudié consiste à tester une hypothèse H_0 de stabilité de la moyenne contre une hypothèse H_1 traduisant un changement de la moyenne. Les hypothèses H_0 et H_1 s'écrivent :

H_0 : les observations $Z_1 \dots Z_n$ sont indépendantes, de même loi $N(\theta,\sigma^2)$.

H_1 : il existe un indice r_0 inconnu, $1 \leq r_0 \leq n-1$ tel que

$Z_1 \dots Z_{r0}$, indépendantes, de loi $N(\theta_1, \sigma^2)$;

$Z_{r0+1} \dots Z_n$, indépendantes, de loi $N(\theta_2, \sigma^2)$; $\quad \theta_1 \neq \theta_2$.

Les paramètres r_0, θ, θ_1 et θ_2 sont supposés inconnus. La variance σ^2 est supposée connue ou inconnue.

Pour ce modèle, Hawkins [13] et Worsley [23] ont pu déterminer la loi exacte, sous l'hypothèse H_0, de la statistique du R.V.M. pour les deux cas : σ^2 connue et inconnue. La loi est d'obtention difficile et nécessite un calcul numérique fastidieux à chaque fois que la taille n de l'échantillon varie.

Deshayes et Picard, dans [8](2) et pour le même modèle, ont abordé le problème du point de vue de la théorie asymptotique locale. Les hypothèses H_0 et H_1 ne sont pas considérées fixes. On les fait se rapprocher à mesure que n augmente. L'amplitude de la rupture, à savoir $|\theta_2 - \theta_1|$, est prise égale à $d/\sqrt{n}$ où d est une grandeur fixée. Ils montrent, sous des conditions de régularité, la convergence, dans l'espace $C(\mathbb{R}^2 \times [0,1])$ des fonctions continues définies sur $\mathbb{R}^2 \times [0,1]$, du processus de vraisemblance à double indice (cf. [8](2)) vers un processus continu qui s'écrit comme fonction du

mouvement brownien standard et de la matrice d'information de Fisher associée au paramètre (θ,σ^2). Les deux indices du processus de vraisemblance sont : le paramètre inconnu (θ,σ^2) et le paramètre (r/n), r étant l'instant possible de la rupture et n la taille de l'échantillon.

Cette approche met bien en relief les problèmes posés par une rupture située aux bords. Le résultat de convergence ainsi obtenu permet alors à ses auteurs de déduire le comportement limite des estimateurs du maximum de vraisemblance des paramètres du modèle.

Bien que le modèle retenu pour notre étude (cf. Chap.I,§1) soit différent de celui présenté plus haut, il nous a paru important de donner certains repères pour situer les contributions de différents auteurs. De plus le modèle de régression étant une généralisation du modèle défini par un n-échantillon d'une loi de moyenne θ et de variance σ^2, il semble naturel d'adopter une telle présentation.

Dans le cadre du modèle de régression, on peut citer Quandt [19](2) et Deshayes et Picard [8](3) pour la détection de rupture à l'aide des méthodes de vraisemblance. Ces derniers étendent au modèle de régression les résultats asymptotiques établis dans [8](2).

Dans [19](2), Quandt étudie la conjecture selon laquelle "Asymptotiquement la statistique du test du R.V.M suit une loi du χ^{2}" et conclut à son rejet suite à une étude empirique. En fait l'irrégularité de la vraisemblance, due au fait que l'instant de rupture soit un paramètre discret inconnu, ne permet pas d'utiliser les résultats classiques sur la loi asymptotique de la statistique du rapport des vraisemblances maximales.

L'objet de ce chapitre est de voir comment s'appliquent les méthodes de vraisemblance au modèle de régression.

Pour les deux cas : σ^2 connue et inconnue, on s'intéresse au comportement de la statistique du R.V.M. sous l'hypothèse H_0 de stabilité du modèle.

Ainsi, on met bien en relief les difficultés liées au calcul de la loi exacte de la statistique du test. On montre comment les résultats de Hawkins [13], sur la loi exacte de la statistique du R.V.M., peuvent être adaptés dans le cas du modèle de régression.

Suite à une étude empirique, des résultats sur la capacité du test du R.V.M. à détecter une rupture, sont présentés.

Enfin, une comparaison est faite avec le test de la somme cumulée (cf chapitre 1) à l'aide des mêmes résultats empiriques.

2. LA STATISTIQUE DU RAPPORT DES VRAISEMBLANCES MAXIMALES

Notons que le modèle et les notations sont ceux déjà utilisés au chapitre1 et décrits au paragraphe 1 du même chapitre. De même, les hypothèses H_0 et H_1 sont celles présentées au chapitre 1.

2.1. Cas de la variance σ^2 connue :

Sous l'hypothèse H_0, d'absence de rupture, on note $L(\beta)$ la fonction de vraisemblance :

$$L(\beta) = (1/\sqrt{2\Pi}\,\sigma)^n \exp\left[(-1/2\sigma^2)\left[\sum_{t=1,n} (y_t - x_t^T\beta)^2\right]\right]$$

L'estimateur du maximum de vraisemblance de β est :

$$\hat{\beta} = \left(\sum_{t=1,n} x_t x_t^T\right)^{-1} \sum_{t=1,n} x_t y_t$$

Sous l'hypothèse $H_1(r)$, d'existence d'une rupture en un instant r tel que $k \leq r \leq n-k$, on note $L(r,\beta_1,\beta_2)$ la fonction de vraisemblance.

$L(r,\beta_1, \beta_2) =$

$$(1/\sqrt{2\Pi}\sigma)^n \exp\left[-(1/2\sigma^2)\left[\sum_{t=1,r} (y_t - x_t^T\beta_1)^2 + \sum_{t=r+1,n} (y_t - x_t^T\beta_2)^2\right]\right].$$

On note, sous l'hypothèse $H_1(r)$, $\hat{\beta}_1(r)$, $\hat{\beta}_2(r)$ les estimateurs du maximum

de vraisemblance des paramètres b_1 et b_2 :

$$\hat{\beta}_1(r) = (\Sigma_{t=1,r} x_t x_t^T)^{-1} \Sigma_{t=1,r} x_t y_t ;$$

$$\hat{\beta}_2(r) = (\Sigma_{t=r+1,n} x_t x_t^T)^{-1} \Sigma_{t=r+1,n} x_t y_t.$$

Ainsi le rapport des vraisemblances maximisées relatif aux hypothèses H_0 et $H_1(r)$ est une fonction du seul paramètre r.

La statistique du test du R.V.M de H_0 contre H_1 est alors le maximum sur r, $k \leq r \leq n-k$, de la fonction $L^*(r)$ définie par :

$$L^*(r) = (Max_{H1(r)} L(r,\beta_1,\beta_2) / Max_{H0} L(\beta)) = (L(r, \hat{\beta}_1(r), \hat{\beta}_2(r) / L(\hat{\beta})),$$

On rejette l'hypothèse H_0 de stabilité du modèle si :

$$| Max_{k \leq r \leq n-k} L^*(r)| > c_\alpha,$$

où la valeur critique c_α est définie par :

$$P_{H0} (| Max_{k \leq r \leq n-k} L^*(r)| > c_\alpha) = \alpha.$$

α, désigne le seuil de signification du test qui est supposé fixé a priori.

2.1.1 Calcul de la statistique du test du R.V.M.

Le rapport des vraisemblances maximisées relatif aux hypothèses H_0 et $H_1(r)$ s'écrit :

$$L^*(r) = \exp\{(1/2\sigma^2)[\Sigma_{t=1,n}(y_t - x_t^T\hat{\beta})^2 - \Sigma_{t=1,r}(y_t - x_t^T\hat{\beta}_1(r))^2 - \Sigma_{t=r+1,n}(y_t - x_t^T\hat{\beta}_2(r)^2]\}.$$

Si on pose : $L_1(r) = 2\sigma^2 \log L^*(r)$, le test peut être basé sur la statistique :

$Max_{k \leq r \leq n-k} [L_1(r)]$.

Pour tout $r \in \{k,..., n-k\}$, on pose :

$$V_{0,r} = \sum_{t=1,r} x_t x_t^T ; \quad V_{r,n} = \sum_{t=r+1,n} x_t x_t^T.$$

On a alors :

Lemme 1

La statistique $L_1(r)$ peut s'écrire :

pour tout $r \in \{k, ..., n-k\}$

$$L_1(r) = (\hat{\beta}_2(r) - \hat{\beta}_1(r))^T V_{0,r} V_{0,n}^{-1} V_{r,n} (\hat{\beta}_2(r) - \hat{\beta}_1(r)).$$

Démonstration

Rappelons que

$$L_1(r) = \sum_{t=1,n}(y_t - x_t^T\hat{\beta})^2 - \sum_{t=1,r}(y_t - x_t^T\hat{\beta}_1(r))^2 - \sum_{t=r+1,n}(y_t - x_t^T\hat{\beta}_2(r))^2$$

Les estimateurs $\hat{\beta}$, $\hat{\beta}_1(r)$ et $\hat{\beta}_2(r)$ sont liés par la relation :

$$V_{0,n}\hat{\beta} = V_{0,r}\hat{\beta}_1(r) + V_{r,n}\hat{\beta}_2(r).$$

En reportant la valeur de $\hat{\beta}$, en fonction de $\hat{\beta}_1(r)$ et $\hat{\beta}_2(r)$, dans le premier terme de $L_1(r)$ et en le développant, on montre que :

$$\sum_{t=1,n}(y_t - x_t^T\hat{\beta})^2 = \sum_{t=1,r}(y_t - x_t^T\hat{\beta}_1(r))^2 + \sum_{t=r+1,n}(y_t - x_t^T\hat{\beta}_2(r))^2$$
$$+ (\hat{\beta}_2(r) - \hat{\beta}_1(r))^T V_{0,r} V_{0,n}^{-1} V_{r,n} (\hat{\beta}_2(r) - \hat{\beta}_1(r)).$$

Le résultat annoncé découle alors de cette identité. □

On énonce un autre résultat qui nous sert pour la suite :

Lemme 2 : La matrice $V_{0,r}V_{0,n}^{-1}V_{r,n}$ est symétrique.

Démonstration :

les matrices $V_{0,r}$, $V_{0,n}^{-1}$, $V_{r,n}$ étant toutes symétriques, il suffit de montrer que l'on a : $V_{0,r}V_{0,n}^{-1}V_{r,n} = V_{r,n}V_{0,n}^{-1}V_{0,r}$.

Considérons le membre de gauche :

$$V_{0,r}V_{0,n}^{-1}V_{r,n} = (V_{0,n} - V_{r,n})V_{r,n}^{-1}(V_{0,n} - V_{0,r})$$

$$= V_{0,n}V_{0,n}^{-1}V_{0,n} - V_{r,n}V_{0,n}^{-1}V_{0,n} - V_{0,n}V_{0,n}^{-1}V_{0,r} + V_{r,n}V_{0,n}^{-1}V_{0,r}$$

$$= V_{0,n} - V_{r,n} - V_{0,r} + V_{r,n}V_{0,n}^{-1}V_{0,r}.$$

Ce qui se simplifie pour donner l'égalité cherchée.

2.1.2. Comportement de $\max_{k \le r \le n-k} L_1(r)$ sous l'hypothèse H_0

Sous l'hypothèse H_0, pour tout $r \in (k,\ldots,n-k)$ les estimateurs $\hat{\beta}_1(r)$ et $\hat{\beta}_2(r)$ sont indépendants, gaussiens, d'espérance β et de matrices de covariance respectives $\sigma^2 V_{0,r}^{-1}$ et $\sigma^2 V_{r,n}^{-1}$. Le vecteur $(\hat{\beta}_2(r)-\hat{\beta}_1(r))$ est alors gaussien, centré et de matrice de covariance $\sigma^2 (V_{0,r}^{-1} + V_{r,n}^{-1})$.

Pour tout $r \in (k,\ldots,n-k)$, la loi de la statistique $L_1(r)$ est donnée par la proposition suivante :

Proposition 1 :

Sous l'hypothèse H_0, pour tout r, $k \le r \le n-k$, la statistique $L_1(r)/\sigma^2$ suit une loi du χ^2 à k degrés de liberté.

Preuve :

Pour tout $r \in (k,\ldots,n-k)$, la variable aléatoire $L_1(r)$ est une forme quadratique de vecteur gaussien.

Pour déduire sa loi, on se base sur le théorème 1 de Barra [1] (p.88).

Le lemme 2 nous assure que $V_{0,r}V_{0,n}^{-1}V_{r,n}$ est une matrice symétrique d'ordre k. On vérifie, pour être dans les conditions d'application du dit théorème1, que la matrice produit : $(V_{0,r}V_{0,n}^{-1}V_{r,n})(\sigma^2(V_{0,r}^{-1}+V_{r,n}^{-1}))$ est égale à σ^2 Id où Id désigne la matrice identité. Pour le voir, il suffit de développer la dite matrice produit en utilisant la symétrie de

$V_{0,r}V_{0,n}^{-1}V_{r,n}$. Une application directe du théorème 1 [1](p.88) nous conduit au résultat :

Pour tout r, $k \leq r \leq n-k$, $L_1(r)$ suit la loi $\Gamma(k/2, 1/2\sigma^2)$.

Si l'on suppose que l'instant de rupture est un paramètre connu et prend pour valeur r_0, le test du R.V.M. pour tester la stabilité du modèle de régression serait basé sur la statistique $L_1(r_0)/\sigma^2$ laquelle suit, sous l'hypothèse de stabilité, une loi du χ^2_k.

Si l'instant de rupture r_0 est inconnu et c'est le cas qui nous intéresse, le test du R.V.M est alors basé sur la statistique : $\max_{k\leq r\leq n-k} (L_1(r)/\sigma^2)$.

La statistique du test consiste donc en le maximum de (n-2k+1) variables, de même loi χ^2_k, non indépendantes. La dépendance entre les différentes variables $(L_1(r)/\sigma^2, r = k... n-k)$ a pour conséquence de rendre le calcul de la loi exacte de $\max_{k\leq r\leq n-k} (L_1(r)/\sigma^2)$ très complexe.

On montre plus loin comment on peut obtenir la loi de cette statistique dans le cas d'un modèle où la variable de régression est à valeurs dans $\mathbb{R}$.

On aborde à présent le cas où la variance σ^2 est inconnue.

2.2. Cas de la variance σ^2 inconnue :

Les vraisemblances sous H_0 et sous H_1 deviennent ici des fonctions du paramètre σ.

Le rapport des vraisemblances maximisées relatif aux hypothèses H_0 et $H_1(r)$ s'écrit :

$$L(r) = \frac{\max_{H1(r)} L(r, \beta_1, \beta_2, \sigma)}{\max_{H0} L(\beta, \sigma)} = \frac{L(r, \hat{\beta}_1(r), \hat{\beta}_2(r), \hat{\sigma}_1(r))}{L(\hat{\beta}, \hat{\sigma}_0)}$$

où $\hat{\sigma}^2_0$ et $\hat{\sigma}^2_1(r)$ désignent les estimateurs du maximum de vraisemblance de σ^2 respectivement sous H_0 et sous $H_1(r)$.

$$\hat{\sigma}^2_0 = (1/n) \sum_{t=1,n} (y_t - x_t^T \hat{\beta})^2 \ ;$$

$$\hat{\sigma}^2_1(r) = (1/n) \left[\sum_{t=1,r} (y_t - x_t^T \hat{\beta}_1(r))^2 + \sum_{t=r+1,n} (y_t - x_t^T \hat{\beta}_2(r))^2 \right]$$

Le rapport L(r) devient après avoir remplacé les estimateurs par leur expression :

$$L(r) = \frac{(\sqrt{2\Pi}\,\hat{\sigma}_0)^n \ \exp(-n/2)}{(\sqrt{2\Pi}\,\hat{\sigma}_1(r))^n \ \exp(-n/2)} = (\hat{\sigma}_0 / \hat{\sigma}_1(r))^n.$$

Maximiser $[L(r)_{k \le r \le n-k}]$ est équivalent à maximiser $[(\hat{\sigma}^2_0/\hat{\sigma}^2_1(r))_{k \le r \le n-k}]$

Il s'ensuit qu'en utilisant le développement de $n\hat{\sigma}^2_0$,

$$n\hat{\sigma}^2_0 = \sum_{t=1,n} (y_t - x_t^T \hat{\beta})^2 = \sum_{t=1,r} (y_t - x_t^T \hat{\beta}_1(r))^2$$

$$+ \sum_{t=r+1,n} (y_t - x_t^T \hat{\beta}_2(r))^2 + (\hat{\beta}_2(r) - \hat{\beta}_1(r)) V_{0,r} V_{0,n}^{-1} V_{r,n} (\hat{\beta}_2(r) - \hat{\beta}_1(r)),$$

le rapport $(\hat{\sigma}^2_0/\hat{\sigma}^2_1(r))$ peut se mettre sous la forme :

$$(\hat{\sigma}^2_0/\hat{\sigma}^2_1(r)) = 1 + \frac{(\hat{\beta}_2(r) - \hat{\beta}_1(r))^T V_{0,r} V_{0,n}^{-1} V_{r,n} (\hat{\beta}_2(r) - \hat{\beta}_1(r))}{n\hat{\sigma}^2_1(r)}$$

Si on note

$$L_2(r) = \frac{(\hat{\beta}_2(r) - \hat{\beta}_1(r))^T V_{0,r} V_{0,n}^{-1} V_{r,n} (\hat{\beta}_2(r) - \hat{\beta}_1(r))}{n\hat{\sigma}^2_1(r)} = \frac{L_1(r)}{n\hat{\sigma}^2_1(r)},$$

maximiser$_{k \le r \le n-k}$ $[(\hat{\sigma}^2_0/\hat{\sigma}^2_1(r))]$ est équivalent à maximiser$_{k \le r \le n-k}$ $[L_2(r)]$.

Le test du R.V.M. dans le cas où σ^2 est inconnue est alors basé sur la statistique : $\max_{k \le r \le n-k} L_2(r)$.

2.2.1. Comportement de $\max_{k \le r \le n-k} L_2(r)$ sous l'hypothèse H_0 :

Nous procédons comme dans le cas de la variance connue. On veut connaître pour tout r, $k \le r \le n-k$, la loi de la variable aléatoire $L_2(r)$.

On sait (cf. proposition 1) que pour tout r, $k \le r \le n-k$, $(L_1(r)/\sigma^2)$ suit une loi du χ^2_k. Par ailleurs, on montre que la variable $(n\hat{\sigma}^2_1(r)/\sigma^2)$ suit, sous l'hypothèse H_0, une loi du χ^2_{n-2k}.

En effet,

$$n\sigma^2_1(r) = \sum_{t=1,r} (y_t - x_t^T \hat{\beta}_1(r))^2 + \sum_{t=r+1,n} (y_t - x_t^T \hat{\beta}_2(r))^2$$

Si on pose :

$$X_1^T(r) = (x_1 \dots x_r) \; ; \; X_2^T(r) = (x_{r+1} \dots x_n) \; ;$$

$$Y_1^T(r) = (y_1 \dots y_r) \; ; \; Y_2^T(r) = (y_{r+1} \dots y_n) \; ;$$

Il est alors évident que :

$$n\hat{\sigma}^2_1(r) = \|Y_1(r) - X_1(r)\hat{\beta}_1(r)\|^2 + \|Y_2(r) - X_2(r)\hat{\beta}_2(r)\|^2$$

En appliquant des résultats classiques sur le modèle linéaire à chaque phase du modèle, on en déduit que sous l'hypothèse H_0 :

$\|Y_1(r) - X_1(r)\hat{\beta}_1(r)\|^2$ et $\|Y_2(r) - X_2(r)\hat{\beta}_2(r)\|^2$ suivent respectivement des lois du χ^2 à (r-k) d.d.l et (n-r-k) d.d.l. De plus, ces deux variables sont indépendantes. Il s'ensuit que $n\hat{\sigma}^2_1(r)/\sigma^2$ suit bien un χ^2_{n-2k}.

Pour obtenir la loi du rapport $L_1(r)/n\hat{\sigma}^2_1(r)$, il nous reste à montrer l'indépendance des deux variables $L_1(r)$ et $n\hat{\sigma}^2_1(r)$.

Lemme 3 :

Sous l'hypothèse H_0 et pour tout $r \in \{k, \dots, n-k\}$, les variables $L_1(r)$ et $n\hat{\sigma}^2_1(r)$ sont indépendantes.

Démonstration :

On montre par un calcul simple que $L_1(r)$ peut aussi s'écrire :

$$L_1(r) = \| X_1(r)(\hat{\beta}_1(r) - \hat{\beta}) \|^2 + \| X_2(r)(\hat{\beta}_2(r) - \hat{\beta}) \|^2 .$$

Ainsi on pose :

$$Z_1(r)^T = (Y_1(r) - X_1(r)\hat{\beta}_1(r) \quad Y_2(r) - X_2(r)\hat{\beta}_2(r))$$

$$Z_2(r)^T = (X_1(r)(\hat{\beta}_1(r) - \hat{\beta}) \quad X_2(r)(\hat{\beta}_2(r) - \hat{\beta}))$$

On a bien sûr :

$$n\hat{\sigma}^2{}_1(r) = \| Z_1(r) \|^2 \qquad \text{et} \qquad L_1(r) = \| Z_2(r) \|^2 .$$

Montrer l'indépendance de $n\hat{\sigma}^2{}_1(r)$ et $L_1(r)$ revient à montrer celle des deux vecteurs $Z_1(r)$ et $Z_2(r)$. $Z_1(r)$ et $Z_2(r)$ étant des vecteurs gaussiens, centrés, à valeurs dans $\mathbb{R}^n$, il suffit alors de montrer que : $E(Z_1(r)Z_2(r)^T)$ est nulle .

La matrice $E(Z_1(r)\, Z_2(r)^T)$ d'ordre n, s'écrit sous la forme :

$$E(Z_1(r)\, Z_2(r)^T) = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

où A, B, C et D sont les matrices de dimensions respectives (r,r), (r,n-r), (n-r,r) et (n-r,n-r) suivantes :

$$A = E[(Y_1(r) - X_1(r)\hat{\beta}_1(r))\ (X_1(r)(\hat{\beta}_1(r) - \hat{\beta}))^T] ;$$

$$B = E[(Y_1(r) - X_1(r)\hat{\beta}_1(r))\ (X_2(r)(\hat{\beta}_2(r) - \hat{\beta}))^T] ;$$

$$C = E[(Y_2(r) - X_2(r)\hat{\beta}_2(r))\ (X_1(r)(\hat{\beta}_1(r) - \hat{\beta}))^T] ;$$

$$D = E[(Y_2(r) - X_2(r)\hat{\beta}_2(r))\ (X_2(r)(\hat{\beta}_2(r) - \hat{\beta}))^T] ;$$

Il suffit de montrer que la matrice A est nulle. En effet, des arguments analogues permettent de montrer aussi la nullité des matrices B, C et D. On omet pour la suite de noter l'indice r.

$$A = E(Y_1\hat{\beta}_1{}^T X_1{}^T) - E(Y_1\hat{\beta}^T X_1{}^T) - X_1 E(\hat{\beta}_1\hat{\beta}_1{}^T)X_1{}^T + X_1 E(\hat{\beta}_1\hat{\beta}^T)X_1{}^T$$

Chaque terme est évalué séparément. Nous aurons à utiliser les identités suivantes :

$$\hat{\beta}_1 = V_{0,r}{}^{-1} X_1{}^T Y_1 \; ; \hat{\beta}_2 = V_{r,n}{}^{-1} X_2{}^T Y_2 \; ; \quad V_{0,n}\hat{\beta} = V_{0,r}\hat{\beta}_1 + V_{r,n}\hat{\beta}_2$$

Ainsi :

$$E(Y_1\hat{\beta}_1{}^T X_1{}^T) = E(Y_1 Y_1{}^T) X_1 V_{0,r}{}^{-1} X_1{}^T.$$

On montre que sous H_0 : $E(Y_1 Y_1{}^T) = X_1\beta\beta^T X_1{}^T + \sigma^2 \mathbf{1}_r$, où $\mathbf{1}_r$ désigne la matrice identité d'ordre r. Il vient alors,

$$E(Y_1\hat{\beta}_1{}^T X_1{}^T) = X_1\beta\beta^T X_1{}^T + \sigma^2 X_1 V_{0,r}{}^{-1} X_1{}^T.$$

Par ailleurs :

$$E(Y_1\hat{\beta}^T X_1{}^T) = E(Y_1(\hat{\beta}_1{}^T V_{0,r} + \hat{\beta}_2{}^T V_{r,n}) V_{0,n}{}^{-1} X_1{}^T)$$

$$= E(Y_1\hat{\beta}_1{}^T)V_{0,r}V_{0,n}{}^{-1}X_1{}^T + E(Y_1\hat{\beta}_2{}^T)V_{r,n}V_{0,n}{}^{-1}X_1{}^T.$$

En remplaçant $E(Y_1\hat{\beta}_1{}^T)$ par son expression et en se servant de l'indépendance entre Y_1 et $\hat{\beta}_2$, on arrive aprés calcul à :

$$E(Y_1\hat{\beta}^T X_1{}^T) = X_1\beta\beta^T X_1{}^T + \sigma^2 X_1 V_{0,n}{}^{-1} X_1{}^T.$$

D'autre part, puisque $var(\hat{\beta}_1) = \sigma^2 V_{0,r}$ et $E(\hat{\beta}_1\hat{\beta}_1{}^T) = var(\hat{\beta}_1) + E(\hat{\beta}_1)E(\hat{\beta}_1{}^t)$, on en conclut : $X_1 E(\hat{\beta}_1\hat{\beta}_1{}^T) X_1{}^T = \sigma^2 X_1 V_{0,r}{}^{-1} X_1{}^T + X_1\beta\beta^T X_1{}^T$.

Il nous reste à calculer le dernier terme $X_1 E(\hat{\beta}_1\hat{\beta}^t)X_1{}^T$.

$$E(\hat{\beta}_1\hat{\beta}^t) = E(\hat{\beta}_1\,(\hat{\beta}_1{}^T V_{0,r} + \hat{\beta}_2{}^T V_{r,n})V_{0,n}{}^{-1})$$

$$= E(\hat{\beta}_1\hat{\beta}_1{}^T)V_{0,r}V_{0,n}{}^{-1} + E(\hat{\beta}_1\hat{\beta}_2{}^t)V_{r,n}V_{0,n}{}^{-1}$$

$\hat{\beta}_1$ et $\hat{\beta}_2$ étant indépendants, un calcul simple permet d'arriver à :

$$X_1 E(\beta_1\beta^T)X_1{}^T = \sigma^2 X_1 V_{0,n}{}^{-1} X_1{}^T + X_1\beta\,\beta^T X_1{}^T.$$

Il est alors évident de conclure à la nullité de la matrice A.

On peut à présent déduire la loi sous H_0 de $L_2(r)$.

Proposition 2 :

Pour tout r, $k \leq r \leq n-k$, $(n-2k)/k$. $L_2(r)$ suit sous l'hypothèse H_0 une loi de Fisher (k, n-2k).

Preuve :

Le résultat découle de l'indépendance de $L_1(r)$ et $n\sigma^2_1$ et du fait que :

pour tout r, $k \leq r \leq n-k$: $L_1(r)/\sigma^2 \sim \chi^2_k$

$n\sigma^2_1(r)/\sigma^2 \sim \chi^2_{n-2k}$.

On a donc montré quelle est la loi de la statistique du test du R.V.M. lorsque la variance σ^2 est inconnue et lorsque l'instant de la rupture est supposé connu.

Lorsque l'instant de la rupture, r_0, est inconnu le test peut être basé sur la statistique : $\max_{k\leq r\leq n-k}[(n-2k/k)(L_1(r)/n\hat{\sigma}^2_1(r))]$. La loi de cette statistique sous H_0 est alors celle du maximum de (n-2k+1) variables aléatoires non indépendantes et de même loi Fisher-Snedecor (k, n-2k).

Les difficultés liées au calcul de la loi exacte des deux statistiques du R.V.M nous ont incité à déterminer, par une étude empirique, certains fractiles de la statistique du R.V.M dans le cas de la variance σ^2 inconnue. Les fractiles que nous obtenons nous servent pour mener une étude empirique de la puissance du test. De même nous nous servons de ces fractiles pour utiliser le test du R.V.M dans le cadre d'une application pratique (cf. chapitre 4).

Nous montrons cependant comment on peut calculer numériquement la loi de la statistique du R.V.M lorsque la variable de régression du modèle est scalaire.

3. ETUDE DANS UN MODELE DE REGRESSION SIMPLE :

3.1. Modèle et résultats théoriques :

3.1.1. Notations du modèle et statistique du R.V.M :

L'étude présentée dans ce paragraphe s'applique au modèle suivant :

$$(1)\begin{cases} y_t = x_t \beta_t + e_t,\ t = 1 \dots n \\ (e_t)_{1\le t\le n} \text{ est une suite de variables aléatoires} \\ \qquad \text{indépendantes de même loi } N(0,\sigma^2), \\ \text{le paramètre } \beta_t \text{ est à valeurs dans } \mathbb{R}, \text{ pour tout} \\ t \in \{1, \dots, n\}. \end{cases}$$

Les hypothèses H_0 et H_1 traduisant respectivement la stabilité des coefficients $(\beta_t)_{1\le t\le n}$ et l'existence d'une rupture à l'instant inconnu r_0 s'écrivent :

$$(H_0) \qquad y_t = x_t \beta + e_t,\ t = 1, \dots, n$$

$$(H1) \qquad y_t = \begin{cases} x_t \beta_1 + e_t,\ t = 1, \dots, r_0 \\ x_t \beta_2 + e_t,\ t = r_0+1, \dots, n \end{cases}$$

r_0 : instant inconnu de la rupture. $|\beta_2-\beta_1|$: amplitude de la rupture.

On présente certains résultats théoriques sur la statistique du R.V.M associée aux hypothèses H_0 et H_1, dans le cas où la variance σ^2 est connue.

En reprenant les mêmes notations qu'au paragraphe 2, on a :

$$V_{0,r} = \sum_{t=1,r} x_t^2 ;\quad V_{r,n} = \sum_{t=r+1,n} x_t^2, \qquad 1 \le r \le n-1.$$

Les estimateurs du m.v. des paramètres β, β_1 et β_2 sont alors :

$$\hat{\beta} = (\sum_{t=1,n} x_t y_t)/V_{0,n} ;\quad \hat{\beta}_1(r) = (\sum_{t=1,r} x_t y_t)/V_{0,r} ;$$

$$\hat{\beta}_2(r) = (\sum_{t=r+1,n} x_t y_t)/V_{r,n}.$$

On obtient le résultat suivant comme cas particulier du lemme 1 :

pour tout r, $1 \leq r \leq n-1$

$$L(r) = [\,((V_{0,r}\, V_{r,n})/V_{0,n})^{1/2}(\hat{\beta}_2(r) - \hat{\beta}_1(r))]^2$$

Par conséquent si l'on note $|Z(r)| = ((V_{0,r}V_{r,n})/V_{0,n})^{1/2}|\hat{\beta}_2(r) - \hat{\beta}_1(r)|/\sigma$ pour tout r, $1 \leq r \leq n-1$, le test du R.V.M est alors basé sur la statistique $\max_{1\leq r\leq n-1} |Z(r)|$.

Ainsi, on rejette l'hypothèse H_0, au seuil fixé α, si :

$$\max_{1\leq r\leq n-1} |Z(r)| > c_\alpha \quad \text{avec} \quad P_{H0}\,(\max_{1\leq r\leq n-1} |Z(r)| > c_\alpha) = \alpha.$$

3.1.2. Etude du processus $[Z(r),\ 1 \leq r \leq n-1]$ sous H_0 :

On peut voir aisèment que pour tout r, $1 \leq r \leq n-1$, Z(r) est, sous l'hypothèse H_0 une variable gaussienne, centrée et réduite. Le comportement du processus $(Z(r))_{1\leq r\leq n-1}$ sous H_0 est donné par la proposition suivante :

Proposition 3

Sous H_0, $[Z(r),\ 1 \leq r \leq n-1]$ est un processus gaussien, centré, de fonction de covariance :

$$E(Z(r)\, Z(s)) = (\,(V_{r,n}\, V_{0,s})/(V_{s,n}\, V_{0,r})\,)^{1/2};\ s < r.$$

De plus $[Z(r),\ 1 \leq r \leq n-1]$ est un processus de Markov.

Démonstration

Pour $s < r$

(1) $E(Z(s)Z(r)) = ((V_{0,r}V_{r,n}V_{0,s}V_{r,n})^{1/2}/\sigma^2 V_{0,n})$

$$\cdot\, E[(\hat{\beta}_2(s) - \hat{\beta}_1(s))(\hat{\beta}_2(r) - \hat{\beta}_1(r))]$$

On montre que pour $s < r$:

$$E\,(\hat{\beta}_1(s)\hat{\beta}_1(r)) = \beta^2 + \sigma^2/V_{0,r}\ ;\ \ E\,(\hat{\beta}_1(s)\hat{\beta}_2(r)) = \beta^2$$

$$E(\hat{\beta}_2(s)\hat{\beta}_1(r)) = \beta^2 + \sigma^2(V_{s,r}/(V_{s,n}V_{0,r}))\ ;$$

$$E(\hat{\beta}_2(s)\hat{\beta}_2(r)) = \beta^2 + \sigma^2/V_{s,n}$$

Il suffit alors de développer le membre de droite de **(1)** et de remplacer les 4 espérances par leur valeur pour arriver à :

$$E(Z(s)\,Z(r)) = \left((V_{r,n}V_{0,s})/(V_{s,n}V_{0,r}) \right)^{1/2} \qquad s < r.$$

Pour montrer que le processus $(Z(r),\ 1 \leq r \leq n-1)$ est de Markov, on utilise une caractérisation d'un processus gaussien markovien : (cf Doob [9] p.233)

$$\text{Pour } s < m < r \quad \rho(Z(s), Z(r)) = \rho(Z(s), Z(m)).\ \rho(Z(m), Z(r))$$

où $\rho(Z(s), Z(r))$ est la corrélation linéaire entre $Z(s)$ et $Z(r)$.

Il est alors aisé de vérifier cette propriété pour le processus $(Z(r))_{1 \leq r \leq n-1}$ étudié.□

On se propose, en s'appuyant sur la proposition 3 et en s'inspirant de l'étude de Hawkins [13], de déterminer, sous H_0, la loi de la statistique $\max_{1 \leq r \leq n-1} |Z(r)|$.

On pose :

$$\varphi(u,a,b) = (1/\sqrt{2\Pi b})\ \exp\left[-1/2\ ((u-a)^2/b)\right]$$

$$h_1(u,v) = 1 \text{ si } u,v \geq 0$$

$$h_r(u,v) = P(|Z(i)| < v\ ,\ i = 1 \ldots r-1\ /\ Z(r) = u),\ u,\ v \geq 0,\ 2 \leq r \leq n-1$$

$$g(u) = \sum_{r=1,n-1} h_r(u,u)\ h_{n-r}(u,u).$$

Théorème 1 (Hawkins [13])

La densité de la variable $\max_{1 \leq r \leq n-1} |Z(r)|$ est :

$$f(u) = 2\ \varphi(u,0,1)\ g(u)\ ,\ \text{pour } u \geq 0.$$

Pour connaître complètement la loi de $\max_{1 \leq r \leq n-1} |Z(r)|$, il est nécessaire de pouvoir calculer la fonction $g(u)$. On adapte à notre étude un résultat de

Hawkins relatif au calcul des fonctions $h_r(u,v)$.

On note $\rho(r)$ la corrélation linéaire entre les variables Z(r) et Z(r-1).

$$\rho(r) = ((V_{r,n}V_{0,r-1})/(V_{r-1,n}\,V_{0,r}))^{1/2}$$

On a alors :

Théorème 2 :

Les fonctions $h_r(u,v)$, $1 \leq r \leq n-1$, pour $u,v \geq 0$ vérifient :

$h_1(u,v) = 1$

$$h_r(u,v) = \int_{[0,v]} h_{r-1}(z,v) \left[\varphi(z,\rho(r)u, 1-\rho^2(r)) + \varphi(z, -\rho(r)\,u, 1-\rho^2(r)) \right] dz,$$

$r \geq 2$.

Il est alors possible de calculer numériquement les fractiles de la loi de la statistique du test du R.V.M. On se sert pour cela de l'algorithme de calcul des fonctions h_r et de la densité de $\max_{1\leq r\leq n-1} |Z(r)|$. On peut remarquer cependant que la loi dépend fortement des observations de la variable de régression. Cette dépendance se manifeste à travers les différentes corrélations $\rho(r)$. Elle est un inconvénient majeur puisque elle oblige à reprendre la procédure de calcul de la loi de $\max_{1\leq r\leq n-1} |Z(r)|$ à chaque fois que l'on change la variable de régression.

Nous avons pu mettre au point un programme permettant de calculer, en tout point, la fonction de répartition de la statistique $\max_r |Z(r)|$. Les paramètres d'entrée sont alors : n, la taille de l'échantillon ; $x_1 \ldots x_n$, les observations de la variable de régression et le ou les points où l'on veut calculer la fonction de répartition. La procédure fait intervenir des calculs d'intégrales par approximation (méthode des trapèzes) à deux niveaux. D'abord pour le calcul de toutes les fonctions $(h_r,\ 1 \leq r \leq n-1)$, ensuite pour le calcul de la fonction de répartition.

Cela a pour conséquence - si l'on veut obtenir de bonnes approximations - de rendre la procédure très longue lorsque n est grand.

Nous avons testé le programme, pour n = 4 et lorsque les observations de la variable de régression sont x(i) = i/n , i = 1 ... n.

On donne les valeurs de $F(u) = P(\max_{1 \le r \le n-1} |Z(r)| \le u)$ calculées pour différentes valeurs de u.

u	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
F(u)	0.0636	0.3316	0.6445	0.8350	0.9215	0.9476	0.9538	0.955

Pour n = 2, on retrouve la loi de la valeur absolue, d'une variable N(0,1). Une comparaison entre les valeurs, de la fonction de répartition, calculées par le programme et celles données par les tables statistiques nous a assuré de la bonne marche de la procédure.

Au paragraphe suivant, on présente les résultats d'une étude empirique concernant certains fractiles de la statistique du test du R.V.M dans le cas où σ^2 est inconnue. Les valeurs critiques estimées par simulation permettent alors, entre autres, de mener une étude détaillée de la puissance du test.

3.2 Calcul par simulation des valeurs critiques. Etude empirique de la puissance :

La simulation a porté sur le modèle de régression simple (1) décrit au paragraphe 3.1.1. Le modèle a été simulé dans les deux cas correspondant aux hypothèses H_0 et H_1.

La statistique sur laquelle a porté l'étude est $\max_{1 \le r \le n-1} |Z(r)|$, avec

$$Z(r) = ((V_{0,r}V_{r,n})/V_{0,n})^{1/2}(\hat{\beta}_2(r) - \hat{\beta}_1(r))/\hat{\sigma}(r).$$

où $\hat{\sigma}^2(r)$ est un estimateur sans biais de σ^2.

$$\hat{\sigma}^2(r) = 1/n-2\ [\sum_{t=1,r}(y_t - x_t\hat{\beta}_1(r))^2 + \sum_{t=r+1,n}(y_t - x_t\hat{\beta}_2(r))^2].$$

3.2.1. Valeurs critiques empiriques de $\max_{1\le r\le n-1} |Z(r)|$:

Pour certaines valeurs de n, nous avons réalisé 1000 expériences à partir du modèle (1), du § 3.1.1., sous l'hypothèse H_0, suivant la même méthode de simulation qu'au chapitre1. Ainsi, nous avons obtenu 1000 réalisations de la statistique $\max_{1\le r\le n-1} |Z(r)|$ sous l'hypothèse H_0. Grâce à un algorithme de tri nous les avons rangées dans un ordre croissant.

Pour les trois seuils de signification usuels α = 1% ; 5% ; 10% nous avons pris respectivement pour valeurs critiques $\hat{c}_\alpha$ associées les $990^{\text{ième}}$, $950^{\text{ième}}$ et $900^{\text{ième}}$ des 1000 réalisations ordonnées.

α \ n	20	25	30	35	40	45	50	60	200
0.01	3.696	3.649	3.513	3.493	3.515	3.441	3.447	3.463	3.502
0.05	2.853	2.879	2.877	2.873	2.847	2.802	2.868	2.926	2.961
0.10	2.516	2.560	2.539	2.589	2.528	2.515	2.548	2.625	2.687

Tableau 1 : Valeurs critiques estimées - sur 1000 expériences - de la Statistique du P.V.M., $(\max |Z_1(r)|,\ 1 \le r \le n-1)$

Les $\hat{c}_\alpha$ ainsi calculées sont des estimateurs des valeurs critiques théoriques c_α qui vérifient : $P_{H0}\,(\max_{1\leq r\leq n-1} |Z(r)| > c_\alpha) = \alpha$.

On remarque que les valeurs critiques $\hat{c}_\alpha$, associées à un α donné a priori, ne varient pas beaucoup suivant la taille n de l'échantillon. La proximité des valeurs $\hat{c}_\alpha$, correspondant à n = 60 et n = 200, nous suggère d'utiliser dans la pratique ces valeurs critiques lorsque la taille de l'échantillon étudié dépasse 60.

3.2.2. Quelques résultats empiriques sur la puissance du test du R.V.M :

On rappelle que la puissance théorique du test du R.V.M est donnée par :

$$\nu = P_{H1}\,(\max_{1\leq r\leq n-1} |Z(r)| > c_\alpha)$$

où c_α, désigne la valeur critique théorique définie par :

$P_{H0}\,(\max_{1\leq r\leq n-1} |Z(r)| > c_\alpha) = \alpha$ et α seuil de signification du test.

En se servant des valeurs critiques $\hat{c}_\alpha$ - estimateurs des valeurs critiques théoriques c_α - nous estimons empiriquement la puissance théorique ν. L'hypothèse H_1 définie au paragraphe 3.1.1. dépend des deux paramètres r_0 et $\Delta = |\beta_2 - \beta_1|$, respectivement emplacement et amplitude de la rupture.

L'étude empirique de la puissance consiste alors à voir comment se comporte le test du R.V.M - à travers sa puissance - lorsque les paramètres r_0 et Δ varient.

Pour chaque couple (r_0, Δ), nous réalisons 100 expériences du modèle (1) sous H_1. $\hat{\nu}$, estimateur de ν, est alors donné par la fréquence de dépassement de la valeur $\hat{c}_\alpha$ (fixée dès que l'on se donne le seuil α) par les 100 réalisations, sous l'hypothèse H_1, de la statistique $\max_{1\leq r\leq n-1} |Z(r)|$.

Le calcul empirique de la puissance - pour différentes valeurs de r_0 et Δ a été fait pour les tailles d'échantillon suivantes : 20, 30, 50 et 60. Une comparaison entre le test du R.V.M et celui de la somme cumulée (cf. chapitre 1) peut se faire en se basant sur les puissances calculées dans les deux cas où n prend les valeurs 30 et 60.

On présente les différents résultats obtenus. Les puissances empiriques consignées dans les tableaux sont en %.

n = 20

r_0	5			10			15		
Δ \ α %	1	5	10	1	5	10	1	5	10
δ	2	10	16	2	11	16	2	8	13
10 δ	85	96	98	80	95	98	83	98	99

n = 30

r_0	5			10			15			20			25			28		
Δ \ α %	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10
δ	2	6	17	1	7	15	2	10	14	1	8	12	2	8	17	0	6	12
10 δ	88	96	98	92	99	99	93	98	100	90	96	99	88	93	99	70	81	95

n = 50

r_0	5			15			25			35			45			48		
α % / Δ	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10
δ	0	4	7	1	6	14	2	9	18	2	10	17	2	8	12	1	8	9
$10\,\delta$	90	95	98	95	98	99	97	100	100	96	98	100	88	97	99	55	77	85

n = 60

r_0	5			15			30			45			55		
α % / Δ	1	5	10	1	5	10	1	5	10	1	5	10	1	5	10
δ	2	5	11	2	7	13	2	7	12	1	5	12	1	4	13
$10\,\delta$	76	91	95	92	98	100	97	99	100	96	99	99	81	91	94

Une analyse rapide des différents tableaux fait ressortir deux faits importants :

- le test du R.V.M détecte assez faiblement les très petites ruptures (ie. lorsque $\Delta = \delta = (25n)^{-1/2}$). Ceci est assez attendu dans la mesure où pour une telle amplitude ($\Delta = \delta$) les hypothèses H_0 et H_1 sont très proches l'une de l'autre.

- Dès que l'on augmente l'amplitude de la rupture, en passant à $\Delta = 10\delta$ (pour n = 60, l'amplitude est alors $\Delta = |\beta_2 - \beta_1| \simeq 0.258$), le test du R.V.M détecte la rupture d'une façon remarquable.

Si l'on s'intéresse aux ruptures provoquées en fin d'échantillon (voir par exemple les puissances pour les couples (n,r_0) suivants : (30,28), (50,48) et (60,55)) on s'aperçoit que la détection se fait moins bien que pour les autres ruptures. Mais on peut dire - même si la puissance est moins bonne dans ces cas - que le test du R.V.M reste très bon comparativement au test de la somme cumulée qui conduit dans de pareils cas (rupture en fin) à des puissances très faibles (cf.tableaux 4, 5, 6 ch.1). Une explication à ce meilleur comportement du test du R.V.M pour les ruptures situées en fin d'échantillon repose sur le fait que l'estimation de la rupture est meilleure et que par conséquent le retard à la détection est beaucoup moins important pour le test du R.V.M que pour celui de la somme cumulée.

On peut remarquer que Deshayes et Picard dans [9](1) avaient conclu dans une étude comparative, à la supériorité du test du R.V.M pour détecter un changement de moyenne dans une suite d'observations gaussiennes indépendantes. En se basant sur les puissances obtenues (pour n = 30 et n = 60) pour les deux tests on peut dire que pour de très petites ruptures les deux procédures semblent se valoir et dans ce cas il est utile de les utiliser toutes les deux. Mais pour des ruptures plus importantes on privilégiera sans conteste le test du R.V.M.

CHAPITRE III

ESTIMATION DANS UN MODELE DE REGRESSION A DEUX PHASES

1. MODELE ET NOTATIONS.

On considère le modèle de régression :

$$y_i = \begin{cases} \alpha_0 + \beta_0 x_i + e_i & \text{pour } x_i \leq \gamma \\ \alpha_1 + \beta_1 x_i + e_i & \text{pour } x_i > \gamma \end{cases}$$

$(e_i)_{i=1....n}$ est une suite de variables aléatoires indépendantes définies sur un même espace probabilisé $(\Omega,\mathcal{A},\mathcal{P})$,à valeurs dans $(\mathbb{R},\mathcal{B}_{\mathbb{R}})$ de même loi gaussienne centrée et de variance σ^2 inconnue.

$(x_i)_{i=1....n}$ sont les observations de la variable de régression x. On les suppose ordonnées et distinctes :

$$x_1 < x_2 <.....< x_n.$$

On suppose la fonction de régression continue ce qui se traduit par la contrainte non linéaire suivante :

(c) $$(\alpha_0-\alpha_1) + (\beta_0-\beta_1)\gamma = 0$$

On suppose tous les paramètres du modèle inconnus et on pose $\theta^t = (\eta^t,\gamma)$ où $\eta^t = (\alpha_0,\beta_0,\alpha_1,\beta_1)$.

On peut remarquer qu'il est possible d'écrire le modèle en y intégrant la contrainte (c). Les deux phases deviennent alors des fonctions non linéaires par rapport aux paramètres :

$$y_i = \begin{cases} \alpha + \beta_0(x_i-\gamma) + e_i & \text{pour } x_i \leq \gamma \\ \alpha + \beta_1(x_i-\gamma) + e_i & \text{pour } x_i > \gamma \end{cases}$$

Nous retiendrons cependant la première forme du modèle car elle est naturellement adaptée aux développements qui suivent dans le texte.

L'objet de ce chapitre est de traiter l'estimation, par la méthode du maximum de vraisemblance, des différents paramètres du modèle. Ce travail ne constitue pas en fait une nouveauté. Il se veut être une synthèse de résultats obtenus par des auteurs comme Hudson [15] et Hinkley [14](1)(3). Remarquons tout de même que Quandt [19](1) a étudié un tel modèle de régression sans toutefois s'imposer l'hypothèse de continuité de la régression.

On introduit quelques notations :

pour tout $r \in \{1,\ldots,n\}$,

$$x_{0,r} = r^{-1} \sum_{i=1,r} x_i \quad ; S_{xy}(o,r) = \sum_{i=1,r} (x_i - x_{0,r})(y_i - y_{0,r}) ;$$

pour tout $r \in \{2,\ldots,n-2\}$,

$$x_{r,n} = (n-r)^{-1} \sum_{i=r+1,n} x_i ;$$

$$S_{xy}(r,n) = \sum_{i=r+1,n} (x_i - x_{0,n-r})(y_i - y_{0,n-r}).$$

2. ESTIMATION PAR LE MAXIMUM DE VRAISEMBLANCE DE γ

2.1 Le problème d'estimation de γ

On note $L(y,\theta,\sigma)$ la fonction de vraisemblance associée au modèle étudié. Elle s'écrit :

$L(y,\theta,\sigma) =$

$$(2\pi\sigma^2)^{-n/2} \exp\left[-(2\sigma^2)^{-1}\left[\sum_{x_i \leq \gamma}(y_i - \alpha_0 - \beta_0 x_i)^2 + \sum_{x_i > \gamma}(y_i - \alpha_1 - \beta_1 x_i)^2\right]\right]$$

Estimer les paramètres θ et σ^2 par le maximum de vraisemblance consiste bien sûr à maximiser $L(y,\theta,\sigma)$ par rapport à θ et σ, sous la contrainte non linéaire sur θ.

On peut vérifier sans trop de difficulté que la contrainte (c) assure la continuité de la fonction de vraisemblance par rapport à θ. Par contre la fonction $L(y,\theta,\sigma)$ n'est pas dérivable, par rappport à γ, aux différents points x_i.

Il s'ensuit que la méthode classique de dérivation de la vraisemblance, pour obtenir les estimateurs du m.v., n'est pas applicable dans ce cadre. Il est donc nécessaire, afin d'estimer γ, de contourner cette difficulté.

L'idée servant de base à l'estimation de γ est la suivante :

Sans être trop restrictif, on peut faire l'hypothèse (H) que γ appartient à l'intervalle $[x_2, x_{n-2}[$. En pratique, cela veut dire que nous disposons d'au moins deux observations dans chaque phase. Si une procédure de détection a montré l'existence d'une rupture ne se trouvant pas aux bords, alors l'hypothèse H est des plus réalistes.

Sous cette hypothèse, compte tenu la nature ordonnée des x_i, γ se trouve entre deux observations consécutives de la variable de régression.

Ce point de vue introduit alors naturellement la famille de sous-hypothèses $(H(r))_{r = 2,\ldots,n-2}$ où H(r) est l'hypothèse selon laquelle le paramètre γ appartient à $[x_r, x_{r+1}[$.

Sous H(r) la vraisemblance s'écrit :

$$L_r(y,\theta,\sigma^2)=(1/2\Pi\sigma^2)^{n/2}.\exp\{-(1/2\sigma^2)[\sum_{i=1,r}(y_i-\alpha_0-\beta_0x_i)^2 + \sum_{i=r+1,n}(y_i-\alpha_1-\beta_1x_i)^2]\}$$

Pour tout $r \in \{2, \ldots, n-2\}$, sous l'hypothèse H(r) et pour γ fixé, on note $\hat{\alpha}_0(r)$, $\hat{\beta}_0(r)$, $\hat{\alpha}_1(r)$, $\hat{\beta}_1(r)$ et $\hat{\sigma}^2(r)$, en omettant la dépendance par rapport à γ, les estimateurs du maximum de vraisemblance des paramètres α_0, β_0, α_1, β_1 et σ^2.

On sait les obtenir en appliquant la méthode des multiplicateurs de Lagrange (la méthode est développée au §3 dans un cas semblable) puisque γ étant fixé, la contrainte n'est plus non linéaire mais est linéaire.

L'estimateur $\hat{\sigma}^2(r)$ de σ^2 s'écrit :

$$\hat{\sigma}^2(r) = (1/n)\left[\sum_{i=1,r}(y_i-\hat{\alpha}_0(r)-\hat{\beta}_0(r)\,x_i)^2 + \sum_{i=r+1,n}(y_i-\hat{\alpha}_1(r)-\hat{\beta}_1(r)x_i)^2\right]$$

Ainsi la vraisemblance $L_r(y,\theta,\sigma^2)$ maximisée par rapport à η et σ^2 s'écrit :

$$L_r(\gamma) = (1/2\Pi\hat{\sigma}^2(r))^{n/2}\exp(-n/2) \quad \text{pour tout } r \in \{2, \ldots, n-2\}.$$

Il s'ensuit que pour tout $r \in \{2, \ldots, n\}$, maximiser $L_r(\gamma)$ par rapport à γ revient à minimiser $n\hat{\sigma}^2(r)$ que l'on note $S_r^2(\gamma)$.

$n^{-1}S_r^2(\gamma)$ est la variance résiduelle du modèle de régression à deux phases contraintes de se rencontrer au point d'abscisse γ. On peut écrire, suite à un calcul de substitution, en se servant des notations introduites au début :

$$S_r^2(\gamma) = \sum_{i=1,n}(y_i - y_{0,n})^2 - [r(n-r)/n]\,(y_{0,r} - y_{r,n})^2$$

$$+ \hat{\beta}_0(r)\left\{[r(n-r)/n]\,(y_{r,n} - y_{0,r})(x_{0,r}-\gamma) - S_{xy}(0,r)\right\}$$

$$+\hat{\beta}_1(r)\left\{[r(n-r)/n]\,(y_{0,r} - y_{r,n})(x_{r,n}-\gamma) - S_{xy}(r,n)\right\}.$$

Il est facile de voir que l'estimateur du maximum de vraisemblance de γ, sous l'hypothèse (H), est la valeur de γ qui réalise :

$$\min_r\,(\min_{H(r)} S_r^2(\gamma)).$$

Ceci nous conduit alors à étudier les fonctions $\{S_r^2(\gamma),\ 2 \leq r \leq n-2\}$.

$S_r^2(\gamma)$ a une expression assez compliquée puisqu'elle fait intervenir les estimateurs sous contrainte. L'idée est donc de se ramener, par le biais d'un calcul, à étudier d'une façon équivalente d'autres fonctions notées $T_r(\gamma)$ dont la manipulation est beaucoup plus aisée.

Pour ce faire, on introduit S_0^2, qui est n fois la variance résiduelle du

modèle de régression à une seule droite :

$$S_0^2 = \Sigma_{i=1,n} (y_i - y_{0,n})^2 - (S_{xy}(o,n) / S_{xx}(o,n)).$$

Pour tout $r \in \{2, \ldots, n-2\}$, on note $\tilde{\alpha}_0(r)$, $\tilde{\beta}_0(r)$, $\tilde{\alpha}_1(r)$ et $\tilde{\beta}_1(r)$ les estimateurs du maximum de vraisemblance, sans contrainte de continuité, des paramètres α_0, β_0, α_1 et β_1. Ces estimateurs ne dépendent que du paramètre r et ils s'obtiennent aisément.

$$\tilde{\beta}_0(r) = S_{xy}(o,r)/S_{xx}(o,r) \; ; \; \tilde{\alpha}_0(r) = y_{0,r} - \tilde{\beta}_0(r)\, x_{0,r}$$

$$\tilde{\beta}_1(r) = S_{xy}(r,n)/S_{xx}(r,n) \; ; \; \tilde{\alpha}_1(r) = y_{r,n} - \tilde{\beta}_1(r)\, x_{r,n}$$

On en déduit un estimateur naturel de γ :

$$\tilde{\gamma}(r) = (\tilde{\alpha}_0(r) - \tilde{\alpha}_1(r))/(\tilde{\beta}_1(r) - \tilde{\beta}_0(r))$$

La procédure d'estimation de γ est basée principalement sur ces estimateurs sans contrainte comme on le voit plus loin.

Pour tout $r \in \{2, \ldots, n-2\}$ on pose : $T_r(\gamma) = S_0^2 - S_r^2(\gamma)$.

On a bien sûr l'équivalence : pour $\gamma \in H(r)$, $\min_\gamma S_r^2(\gamma) \Leftrightarrow \max_\gamma T_r(\gamma)$.

On pose :

$$C(r) = S_{xx}(o,r)\, S_{xx}(r,n) + [r(n-r)/n]\, (x_{0,r}^2\, S_{xx}(r,n) + x^2_{r,n}\, S_{xx}(o,r));$$

$$D(r) = [r(n-r)/n]\, (x_{0,r}\, S_{xx}(r,n) + x_{r,n}\, S_{xx}(o,r));$$

$$E(r) = [r(n-r)/n]\, (S_{xx}(o,r) + S_{xx}(r,n)).$$

La fonction $T_r(\gamma)$ peut s'écrire (cf. Hinkley [14](1)) :

$$T_r(\gamma) = \frac{(C(r) - D(r)\,(\tilde{\gamma}(r) + \gamma) + E(r)\,\tilde{\gamma}(r)\,\gamma)^2\; (\tilde{\beta}_0(r) - \tilde{\beta}_1(r))^2}{(C(r) - 2D(r)\gamma + E(r)\,\gamma^2).S_{xx}(o,n)}$$

2.2. Etude de la fonction $T_r(\gamma)$ et détermination de γ.

Pour tout $r \in \{2, \ldots, n-2\}$, $T_r(\gamma)$ est une fonction de γ définie sur tout $\mathbb{R}$.

Le dénominateur $(C(r) - 2D(r)\gamma + E(r)\gamma^2)$ est strictement positif pour toute valeur de γ sur $\mathbb{R}$. En effet il a même signe que E(r) puisque le discriminant $(D(r)^2-E(r)\,C(r))$ est strictement négatif.

Comme rapport de deux polynômes en γ, $T_r(\gamma)$ est une fonction continue, indéfiniment dérivable sur tout $\mathbb{R}$. L'étude de la fonction $T_r(\gamma)$ a pour but de rendre le calcul de l'estimateur de maximum de vraisemblance de γ facilement accessible.

La dérivée première de $T_r(\gamma)$ est :

$$(\partial T_r(\gamma) / \partial\gamma)= \frac{2(C(r) - D(r)(\tilde{\gamma}(r) + \gamma) + E(r)\tilde{\gamma}(r)\,\gamma)(D(r)^2- E(r)\,C(r)(\gamma-\tilde{\gamma}(r))(\tilde{\beta}_0(r)-\tilde{\beta}_1(r))^2}{(C(r) - 2D(r)\,\gamma + E(r)\gamma^2)^2 \qquad S_{xx}(o,n)}$$

Elle s'annule deux fois aux points γ_0 et γ_1 :

$\gamma_0 = \tilde{\gamma}(r)$ et $\gamma_1 = (C(r) - D(r)\,\tilde{\gamma}(r))/(D(r)- E(r)\,\tilde{\gamma}(r))$

Les valeurs de la dérivée seconde en ces deux points sont :

$$\frac{\partial^2 T_r}{\partial\gamma^2}(\gamma_0) = 2\,\frac{(\tilde{\beta}_0(r)-\tilde{\beta}_1(r))^2}{S_{xx}(o,n)}\,\frac{(D(r)^2- E(r).C(r))}{(E(r)\,\tilde{\gamma}(r)^2 - 2D(r)\tilde{\gamma}(r) + C(r))}$$

$$\frac{\partial^2 T_r}{\partial\gamma^2}(\gamma_1) = 2\,\frac{(\tilde{\beta}_0(r)-\tilde{\beta}_1(r))^2}{S_{xx}(o,n)}\,\frac{(D(r)^2- E(r)C(r))\,(E(r)\tilde{\gamma}(r) - D(r)(\gamma_1-\tilde{\gamma}(r))}{(E(r)\gamma_1{}^2 - 2D(r)\gamma_1 + C(r))^2}$$

On sait que : $D(r)^2 - E(r)\,C(r) < 0$

et $(E(r)\,\gamma^2 - 2D(r)\,\gamma + C(r)) > 0$, pour tout valeur de γ.

Il s'ensuit alors : $\partial^2 T_r/\partial\gamma^2\,(\tilde{\gamma}(r)) < 0$.

$\tilde{\gamma}(r)$ est donc un maximum pour la fonction $T_r(\gamma)$ sur $\mathbb{R}$. En fait on montre qu'il est unique. Montrons que $[\,\partial^2 T_r/\,\partial\gamma^2\,](\gamma_1) > 0$. Ce qui assure que γ_1 est le minimum de $T_r(\gamma)$ sur tout $\mathbb{R}$. On peut voir que :

$$\text{signe}\,[\,\partial^2 T_r/\,\partial\gamma^2\,](\gamma_1) = \text{signe}\,((D(r) - E(r)\tilde{\gamma}(r))\,(\gamma_1 - \tilde{\gamma}(r))).$$

Or on montre : $((D(r) - E(r)\tilde{\gamma}(r))(\gamma_1 - \tilde{\gamma}(r)) = E(r)\tilde{\gamma}(r)^2 - D(r)\tilde{\gamma}(r) + C(r)$.

Le signe de $[\partial^2 T_r/\partial\gamma^2](\gamma_1)$ est donc bien positif.

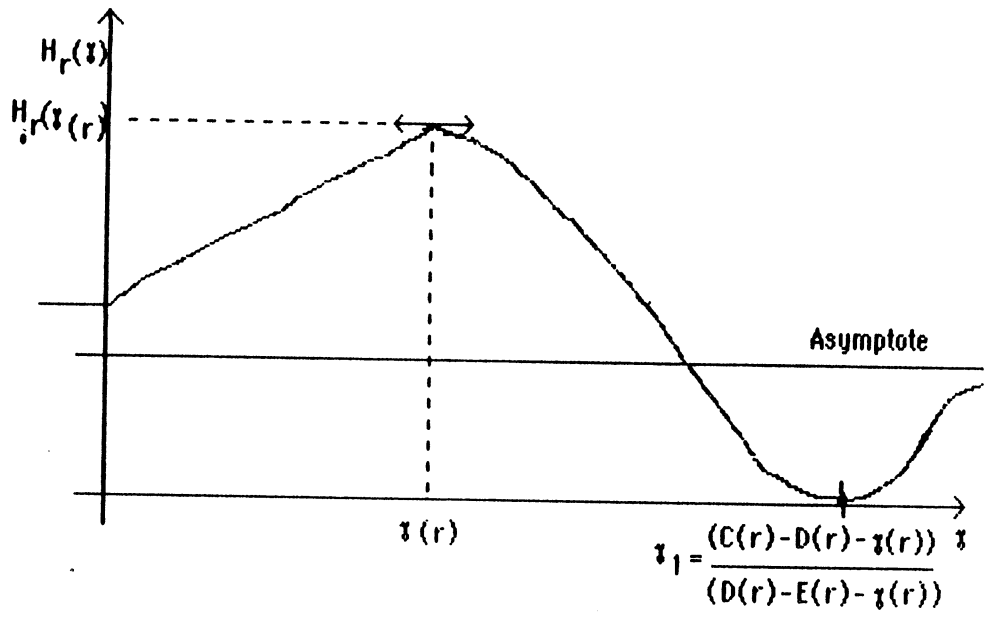
En résumé la fonction $T_r(\gamma)$ est définie sur tout $\mathbb{R}$, à valeurs dans $\mathbb{R}^+$. Elle a pour extrémum $\tilde{\gamma}(r)$ et γ_1. Le graphe de $T_r(\gamma)$ admet une asymptote horizontale d'équation :

$$y = \frac{(E(r)\tilde{\gamma}(r) - (D(r))^2}{E(r)} \cdot \frac{(\tilde{\beta}_0(r) - \tilde{\beta}_1(r))^2}{S_{xx}(o,n)}$$

on présente les variations de $T_r(\gamma)$ dans le tableau suivant :

γ	$-\infty$		$\gamma(r)$		γ_1		$+\infty$
$H'_r(\gamma)$	+		0	-	0		+
$H_r(\gamma)$	asymp	↗	$H_r(\gamma(r))$	↘	0	↗	asymp

Le graphe de $T_r(\gamma)$ est le suivant (avec des valeurs fictives) :



Avec : $T_r(\tilde{\gamma}(r)) = (C(r) - 2D(r)\tilde{\gamma}(r) + E(r)\tilde{\gamma}(r)^2)[(\tilde{\beta}_0(r) - \tilde{\beta}_1(r)^2)/S_{xx}(o,n)]$

$T_r(\gamma_1) = 0.$

Il ressort de cette étude que pour tout $r \in \{2, \dots, n-2\}$, $T_r(\gamma)$ a pour maximum, sur $\mathbb{R}$ tout entier, la valeur $\tilde{\gamma}(r)$. En se basant sur ce résultat et sur la définition de l'estimateur du maximum de vraisemblance de γ, il s'ensuit les deux cas de figure notées (1) et (2) :

(1): $\hat{\gamma}$ peut être égal à un certain $\tilde{\gamma}(r_0)$ qui réalise nécessairement les conditions (i) et (ii) suivantes :

(i): $\tilde{\gamma}(r_0)$ appartient à l'intervalle $[x_{r0}, x_{r0+1}[$.

(ii): $T_{r0}(\tilde{\gamma}(r_0))$ est le maximum parmi les quantités $T_r(\tilde{\gamma}(r))$ pour $\tilde{\gamma}(r) \in [x_r, x_{r+1}[$ et $2 \leq r \leq n-2$.

(2): $\hat{\gamma}$ peut coïncider avec une certaine observation x_{r1}. L'indice r_1 étant tel que : $\tilde{\gamma}(r_1) \notin [x_{r1}, x_{r1+1}[$ et $\tilde{\gamma}(r_1-1) \notin [x_{r1-1}, x_{r1}[$.

A la suite de l'étude qui précède, on présente l'algorithme de calcul de l'estimateur du m.v de γ.

2.3. Algorithme de calcul de $\hat{\gamma}$:

On donne deux remarques découlant de l'étude des fonctions $T_r(\gamma)$ et utiles pour simplifier la recherche de l'estimateur du maximum de vraisemblance de γ.

Remarque 1 : Si pour un indice r donné, on a $x_r < \tilde{\gamma}(r) < x_{r+1}$ alors les deux éventualités $\hat{\gamma} = x_r$ et $\hat{\gamma} = x_{r+1}$ sont à écarter.

Remarque 2 : (Elle est implicitement contenue dans (2) plus haut).

Pour que x_r soit une solution envisageable, il faut nécessairement que : $\tilde{\gamma}(r-1) \notin [x_{r-1}, x_r[$ et $\tilde{\gamma}(r) \notin [x_r, x_{r+1}[$.

Description de l'algorithme :

Pour tout r = 2, ... , n-2, on détermine les estimateurs sans contrainte $\tilde{\beta}_0(r)$, $\tilde{\alpha}_0(r)$, $\tilde{\beta}_1(r)$ et $\tilde{\alpha}_1(r)$. On en déduit alors $\tilde{\gamma}(r)$ par :

$$\tilde{\gamma}(r) = (\tilde{\alpha}_0(r) - \tilde{\alpha}_1(r)) / (\tilde{\beta}_1(r) - \tilde{\beta}_0(r)).$$

Pour tout r = 2, ... , n-2, tel que $\tilde{\gamma}(r) \in [x_r, x_{r+1}[$, on calcule $T_r(\tilde{\gamma}(r))$. On déduit le maximum parmi toutes ces valeurs que l'on note par exemple Max1.

Pour tout r = 2, ... , n-2, tel que $\tilde{\gamma}(r-1) \notin [x_{r-1}, x_r[$ et $\tilde{\gamma}(r) \notin [x_r, x_{r+1}[$, on calcule les différents $T_r(x_r)$ et on en déduit leur maximum : Max2.

Il reste alors à comparer Max1 et Max2 pour en déduire l'estimateur de γ. On présente en annexe le programme permettant le calcul des estimateurs du maximum de vraisemblance de γ et des autres paramètres du modèle.

3. Estimateurs du M.V. des autres paramètres du modèle :

Les estimateurs du M.V. des paramètres α_0, β_0, α_1, β_1 et σ^2 dépendent bien sûr de la valeur prise par l'estimateur de γ. On sait que deux possibilités se présentent pour cet estimateur. Il en est de même pour ceux de α_0, β_0, α_1, β_1 et σ^2.

(a) $\hat{\gamma} = \tilde{\gamma}(r_0)$, $x_{r0} \leq \tilde{\gamma}(r_0) < x_{r0+1}$. Dans ce cas les estimateurs du M.V. s'obtiennent directement : $(\hat{\alpha}_0, \hat{\beta}_0, \hat{\alpha}_1, \hat{\beta}_1) = (\tilde{\alpha}_0(r_0), \tilde{\beta}_0(r_0), \tilde{\alpha}_1(r_0), \tilde{\beta}_1(r_0))$

$$\hat{\sigma}^2 = (1/n)\left[\sum_{i=1,\, r0}(y_i - \tilde{\alpha}_0(r_0) - \tilde{\beta}_0(r_0)x_i)^2 + \sum_{i=r0+1,\, n}(y_i - \tilde{\alpha}_1(r_0) - \tilde{\beta}_1(r_0)x_i)^2\right]$$

(b). $\hat{\gamma} = x_{r1}$. Ce cas nécessite quelques développements. L'estimation des paramètres du modèle se fait comme si on connaissait l'abscisse du point d'intersection.

Le paramètre r_1 est connu. La contrainte liant les paramètres n'est plus non linéaire mais est linéaire.

La méthode du multiplicateur de Lagrange permet de résoudre le problème de l'estimation. En effet, il s'agit de :

minimiser, par rapport à η, $Q(\eta)$;

$$Q(\eta) = \sum_{i=1,r1} (y_i-\alpha_0-\beta_0 x_i)^2 + \sum_{i=r1+1,n}(y_i-\alpha_1-\beta_1 x_i)^2],$$

sous la contrainte linéaire : $\eta^t s = 0$

où $s^t = (1, x_{r1}, -1, -x_{r1})$.

On pose :

$$X_1{}^t = \begin{pmatrix} 1 & \ldots\ldots & 1 \\ x_1 & \ldots & x_{r1} \end{pmatrix} \qquad X_2{}^t = \begin{pmatrix} 1 & \ldots\ldots & 1 \\ x_{r1+1} & \ldots & x_n \end{pmatrix}$$

$$Y_1{}^t = (y_1 \ldots\ldots y_{r1}) \quad ; \quad Y_2{}^t = (y_{r1+1} \ldots\ldots y_n)$$

$$X = \begin{pmatrix} X_1 & 0 \\ 0 & X_2 \end{pmatrix} \quad ; \quad Y^t = (Y_1{}^t, Y_2{}^t).$$

L'estimateur sans contrainte du M.V. de η s'écrit :

$$\tilde{\eta} = (X^t X)^{-1} X^t Y.$$

Cet estimateur, en général, ne vérifie pas la contrainte.

On pose : $\tilde{\eta}^t s = u$ et $\eta_1{}^t = (\alpha_0, \beta_0)$ et $\eta_2{}^t = (\alpha_1, \beta_1)$,

On a : $Q(\eta) = (Y_1 - X_1\eta_1)^t (Y_1 - X_1\eta_1) + (Y_2 - X_2\eta_2)^t (Y_2 - X_2\eta_2)$.

Minimiser $Q(\eta)$ sous la contrainte $\eta^t s = 0$ revient par la méthode de Lagrange à résoudre le système formé par les trois équations :

$\partial Q_1(\eta)/\partial \eta_1 = 0$; $\partial Q_1(\eta)/\partial \eta_2 = 0$; $\partial Q_1(\eta)/\partial \lambda = 0$

où : $Q_1(\eta) = Q(\eta) + 2\lambda\eta^t s$. En notant : $s^t = (s_1, s_2)$, les trois équations s'écrivent :

$$\partial Q_1(\eta)/\partial \eta_1 = 2(X_1{}^t X_1)\eta_1 - 2 Y_1{}^t X_1 + 2 \lambda . s_1 = 0.$$

$$\partial Q_1(\eta)/\partial \eta_2 = 2(X_2{}^t X_2)\eta_1 - 2 Y_2{}^t X_2 + 2 \lambda . s_2 = 0$$

$$\partial Q1(\eta) / \partial \lambda = 2(\eta_1{}^t s_1 + \eta_2{}^t s_2) = 0.$$

Les deux premières équations nous donnent :

$$\hat{\eta}_1 = \tilde{\eta}_1 - (X_1{}^tX_1)^{-1}\lambda s_1 \quad \text{et} \quad \hat{\eta}_2 = \tilde{\eta}_2 - (X_2{}^tX_2)^{-1}\lambda s_2.$$

En substituant ces deux expressions dans la troisième équation on déduit la valeur de λ : $\lambda = (\tilde{\eta}^t s)/(s^t(X^tX)^{-1}s)$.

En posant $v = s^t(X^tX)^{-1}s$, il vient, $\lambda = (u/v)$ et

$$\hat{\eta}_1 = \tilde{\eta}_1 - (u/v)(X_1{}^tX_1)^{-1}s_1$$

$$\hat{\eta}_2 = \tilde{\eta}_2 - (u/v)(X_2{}^tX_2)^{-1}s_2$$

On montre au paragraphe suivant comment la méthode des multiplicateurs de Lagrange peut s'appliquer au problème de l'estimation des paramètres d'un modèle de régression à plusieurs phases soumises à des contraintes linéaires.

4. Estimation dans un modèle à p phases, p> 2, lorsque les points de rupture sont connus.

On considère le modèle de régression à p phases linéaires :

$$y_i = \begin{cases} f_1(\alpha_1, x_i) + e_i, & 1 \leq i \leq j_1 \\ f_2(\alpha_2, x_i) + e_i, & j_1 \leq i \leq j_2 \\ \cdots\cdots\cdots\cdots & \cdots\cdots\cdots \\ f_p(\alpha_p, x_i) + e_i, & j_{p-1} \leq i \leq n. \end{cases}$$

$(e_i)_{1\leq i\leq n}$ est une suite de variables aléatoires, définies sur le même espace probabilisé $(\Omega, \mathcal{A}, P)$, de même loi $N(0, \sigma^2)$.

Pour tout j, j = 1, ... , p, $f_j(.,.)$ est une fonction continue par rapport à x et α_j et linéaire par rapport au paramètre α_j. On suppose que la suite $(j_1, j_2, \cdots, j_{p-1})$ est connue et que le modèle de régression est continue.

On a ainsi les contraintes, au nombre de (p-1) :

$$(1) \quad \begin{cases} f_j(\alpha_j, x_i) = f_{j+1}(\alpha_{j+1}, x_i), & j = 1, \cdots, p-1 \\ \text{et } i = j_1, j_2, \cdots, j_{p-1}. \end{cases}$$

L'estimation aux moindres carrés de $\alpha_1, \ldots, \alpha_p$, qui se confond avec le maximum de vraisemblance puisque les erreurs $(e_i)_{1\le i\le n}$ sont normales, résulte de la minimisation, sous les (p-1) contraintes linéaires (1), de :

$$Q(\alpha) = \sum_{i=1,j_1}(Y_i - f_1(\alpha_1, x_i))^2 + \ldots + \sum_{i=j_{p-1},n}(Y_i - f_p(\alpha_p, x_i))^2,$$

.Ce qui revient par la méthode des multiplicateurs de Lagrange à minimiser :

$$L(\alpha,\lambda_1, \ldots, \lambda_p) = Q(\alpha) + 2\lambda_1(f_1(\alpha_1, x_{j_1}) - f_2(\alpha_2, x_{j_1})) + \ldots$$

$$+ 2\lambda_{p-1}(f_{p-1}(\alpha_{p-1}, x_{j_{p-1}}) - f_p(\alpha_p, x_{j_{p-1}})).$$

Remarquons que les (p-1) contraintes linéaires (1) peuvent s'écrire, puisque les différentes fonctions $f_j(\alpha_j, x)$ sont linéaires, comme suit :

$$(1) \Leftrightarrow ({\alpha_j}^t\, {\alpha_{j+1}}^t)\, s_j = 0,\ j = 1, \ldots, p-1\ ;\ {s_j}^t = ({s_{j1}}^t\, {s_{j2}}^t).$$

Ainsi si pour j, j= 1, ... , p, α_j est à valeurs dans $\mathbb{R}^k$, s_j est à valeurs dans $\mathbb{R}^{2k}$. Avec des notations évidentes pour Y_j et X_j, j = 1, ... , p, on a:

$$L(\alpha,\lambda) = \|Y_1 - X_1\alpha_1\|^2 + \ldots + \|Y_p - X_p\alpha_p\|^2 + 2\lambda_1({\alpha_1}^t\, {\alpha_2}^t).s_1 + \ldots$$

$$+ 2\lambda_{p-1}({\alpha^t}_{p-1}\, {\alpha^t}_p).s_{p-1}.$$

Il s'agit alors de résoudre :

$$(\partial L(\alpha,\lambda))/\,\partial\alpha_j = 0, \text{ pour } j = 1, \ldots, p.$$

sous les contraintes $({\alpha_j}^t\, {\alpha^t}_{j+1}).s_j = 0,\ j = 1, \ldots, p-1.$

Ce qui donne le système :

$$(\partial L(\alpha_1,\lambda))/\,\partial\alpha_1 = 2({X_1}^t X_1)\alpha_1 - 2{X_1}^t Y_1 + 2\lambda_1 s_{11} = 0$$

$$(\partial L(\alpha_j,\lambda))/\,\partial\alpha_j = 2({X_j}^t X_j)\alpha_j - 2{X_j}^t Y_j + 2\lambda_{j-1}\, s_{j-1,2} + 2\lambda_j\, s_{j1} = 0$$

pour j =2, ... , p-1

$$(\partial L(\alpha_p,\lambda))/\,\partial\alpha_p = 2({X_p}^t X_p)\alpha_p - 2{X_p}^t Y_p + 2\lambda_{p-1} s_{p-1,2} = 0$$

qui s'écrit aussi

$$(S)\begin{cases} \hat{\alpha}_1 = \tilde{\alpha}_1 - \lambda_1 . (X_1^t X_1)^{-1} s_{11} = 0 \\ \hat{\alpha}_j = \tilde{\alpha}_j - \lambda_{j-1} (X_j^t X_j)^{-1} s_{j-1,2} - \lambda_j (X_j^t X_j)^{-1} s_{j1} \ , \ j = 2, \dots, p-2 \\ \hat{\alpha}_p = \tilde{\alpha}_p - \lambda_{p-1} (X_p^t X_p)^{-1} s_{p-1,2} \end{cases}$$

où les $(\tilde{\alpha}_j)_{1 \le j \le p}$ sont les estimateurs sans contraintes des $(\alpha_j)_{1 \le j \le p}$

$$\tilde{\alpha}_j = (X_j^t X_j)^{-1} X_j^t Y_j$$

Il faut donc calculer les $(\lambda_j)_{j=1, \dots, p-1}$ à partir du système (S) et en respectant les contraintes linéaires :

$$(\hat{\alpha}^t_j \, \hat{\alpha}^t_{j+1}) \, s_j = 0 \ , \ j = 1, \dots, p-1.$$

Les estimateurs sans contraintes $(\tilde{\alpha}_j)$ ne vérifient pas, en général, les contraintes. On pose ainsi :

$$(\tilde{\alpha}^t_j \, \tilde{\alpha}^t_{j+1}) \, s_j = u_j \ , \ j = 1, \dots, p-1, \ u_j \in \mathbb{R}.$$

Le système (S) et les contraintes linéaires appliquées aux estimateurs $(\hat{\alpha}_j)_{j=1 \dots p}$ donnent alors le système, noté (S1) , à (p-1) équations et (p-1) inconnues $\lambda_1, \dots, \lambda_{p-1}$:

$$(S1)\begin{cases} (\lambda_1 (s_1^t C_1^{-1} s_1) + \lambda_2 s_{12}^t (X_2^t X_2)^{-1} s_{21} = u_1 \\ (\lambda_{j-1} (s_{j1}^t (X_j^t X_j)^{-1} s_{j-1,2}) + \lambda_j s_j^t C_j^{-1} s_j \\ \qquad + \lambda_{j+1} s_{j2}^t (X^t_{j+1} X_{j+1})^{-1} s_{j+1,1} = u_j \\ \qquad \text{pour } j = 2, \dots, p-2 \\ (\lambda_{p-2} (s^t_{p-2,2} (X^t_{p-1} X_{p-1})^{-1} s_{p-1,1} + \lambda_{p-1} s^t_{p-1} C^{-1}_{p-1} s_{p-1} = u_{p-1} \end{cases}$$

où les matrices C_j^{-1}, $j = 1 \dots p-1$ sont : $C_j^{-1} = \begin{pmatrix} (X_j^t X_j)^{-1} & 0 \\ 0 & (X^t_{j+1} X_{j+1})^{-1} \end{pmatrix}$

Ainsi les multiplicateurs de Lagrange $(\lambda_j)_{j=1 \dots p}$ sont obtenus par

résolution du système linéaire (S1). Une fois les $(\lambda_j)_j$ calculés, on en déduit les estimateurs $(\hat{\alpha}_j)_{j=1...p}$ grâce au système (S).

Le système linéaire (S1) peut s'écrire matriciellement :

$$A\lambda = u \; ; \; \lambda^t = (\lambda_1 \dots \lambda_{p-1}) \; ; \; u^t = (u_1 \dots u_{p-1})$$

La matrice A est une matrice carrée (p-1, p-1), dont les coefficients sont connus et donnés par :

$$A = (a_{ij}) \; i = 1, \dots, p-1 \; , \; j = 1, \dots, p-1$$

Les termes diagonaux de A, (a_{ii}) i = 1, ..., p-1 sont :

$$a_{ii} = s_i^t C_i^{-1} s_i$$

Par ailleurs :

$$a_{12} = s^t_{12} (X_2^t X_2)^{-1} s_{21} \; ; \; a_{1j} = 0 \quad \text{pour } j = 3, \dots, p-1$$

$$a_{p-1,p-2} = s^t_{p-2,2} (X^t_{p-1} X_{p-1})^{-1} s_{p-1,1} \; ; \; a_{p-1,j} = 0 \text{ pour } j = 1, \dots, p-3$$

pour i = 2, ..., p-2, on a :

$$a_{i,i-1} = s^t_{i1} (X_i^t X_i)^{-1} s_{i-1,2} \; ; \; a_{i,i+1} = s^t_{i2} (X_{i+1}^t X_{i+1})^{-1} s_{i+1,1}$$

$$a_{ij} = 0 \text{ , pour } j = 1, \dots, p-1, \; j \neq i, \; j \neq i-1, \; j \neq i+1.$$

La matrice A étant parfaitement définie, on pourra alors dans la pratique utiliser un programme pour résoudre en λ le système linéaire : $A\lambda = u$.

Signalons qu'au chapitre 4 , nous présentons une application pratique, des résultats présentés dans ce chapitre, sur des données concrètes fournies par la division technique de EDF-Grenoble.

5. Existence et consistance de la suite des estimateurs des moindres carrés dans un modèle de régression à une rupture.

5.1. Notations et hypothèses.

. On introduit un formalisme différent de celui introduit au début de ce chapitre.

On considère le modèle de régression défini par :

$$y_{ti} = f_{ti}(\theta^0) + e_{ti}, \quad t_i \in [0,1], i \in \mathbb{N}^*$$

où $(e_{ti})_{i\in\mathbb{N}^*}$ est une suite de variables aléatoires définies sur un même espace probabilisé $(\Omega,\mathcal{A},P)$ supposées indépendantes et de même loi, de moyenne nulle et de variance σ^2 inconnue.

En posant : $\theta^t = (\tau, \alpha^t)$; $\alpha^t = (a, b_0, b_1)$, on définit pour tout $t \in [0,1]$ et pour tout θ dans Θ, Θ sous ensemble de $\mathbb{R}^4$ précisé plus loin, $f_t(\theta)$ comme suit :

$$f_t(\theta) = \begin{cases} f_0(t-\tau,\alpha) = a + b_0(t-\tau) & \text{si } 0 \le t \le \tau \\ \\ f_1(t-\tau,\alpha) = a + b_1(t-\tau) & \text{si } \tau \le t \le 1. \qquad b_0 \neq b_1 \end{cases}$$

α prend ses valeurs dans $\mathcal{F}$, sous ensemble compact de $\mathbb{R}^3$, que l'on définit comme suit :

$\mathcal{F} = \{(a,b_0,b_1) \in \mathbb{R}^3 \ / \ \max(|a|, |b_0|, |b_1|) \le \mu, |b_1 - b_0| \ge \delta\}$, μ et δ étant deux constantes positives.

Si une étude préalable a montré l'existence probable d'une rupture (cf. chapitres1 et 2), on peut supposer l'existence de deux réels positifs β_1, β_2 tels que : $0 < \beta_1 \le \tau \le \beta_2 < 1$.

L'ensemble Θ est alors défini comme le produit cartésien $[\beta_1, \beta_2] \times \mathcal{F}$.

$$\Theta = \{\theta = (\tau, \alpha^t) \in \mathbb{R}^4, \ \tau \in [\beta_1,\beta_2] \text{ et } \alpha^t \in \mathcal{F}\}.$$

Ainsi Θ est un sous ensemble compact de $\mathbb{R}^4$.

Pour tout $t \in [0,1]$, $f_t(.)$ est une application continue du sous ensemble Θ de $\mathbb{R}^4$ dans $\mathbb{R}$. Pour tout $\theta \in \Theta$, $f.(\theta)$ est une application continue de $[0,1]$ dans $\mathbb{R}$, indéfiniment dérivable sauf au point $t = \tau$ où la dérivée première est discontinue.

θ^0 est la vraie valeur du paramètre θ, supposée appartenir à l'intérieur du sous ensemble Θ.

Ce paragraphe traite de l'existence et des propriétés de consistance de l'estimateur des moindres carrés du paramètre θ.

Hinkley dans [14](1), a traité pour le même modèle, mais sous une hypothèse gaussienne pour les $(e_{ti})_i$, le problème du comportement asymptotique de l'estimateur du maximum de vraisemblance du paramètre θ. Cependant les méthodes utilisées, pour établir la normalité asymptotique notamment, sont assez empiriques et peu rigoureuses. On peut citer aussi Feder [11] pour une généralisation de l'étude à un modèle à plusieurs points de rupture et à des fonctions non nécessairement affines. Mais ce travail comporte des hypothèses théoriques difficilement vérifiables même pour des cas simples, comme celui que nous traitons.
Nous avons choisi de nous inspirer d'une étude de Jenrrich [16] pour montrer d'une façon simple l'existence et la consistance presque-sûre de la suite des estimateurs des moindres carrés du paramètre θ.

La propriété suivante, notée (P), facile à établir pour notre modèle,est utile pour conclure à la consistance presque-sûre de la suite des estimateurs des moindres carrés (θ_n).

(P) : si pour θ_1, θ_2 dans Θ on a $f_t(\theta_1) = f_t(\theta_2)$ pour tout $t \in [0,1]$ alors il s'ensuit : $\theta_1 = \theta_2$.

On fait l'hypothèse suivante sur la répartition des observations $(t_i)_{i \in \mathbb{N}^*}$.

On pose : $F_n(t) = n^{-1}$ (nombre des éléments de $(t_i)_{i=1,\dots,n} \leq t$).

(H) La suite $(F_n)_n$ converge ponctuellement vers une fonction F continue, strictement croissante sur [0,1] et telle que F(0) = 0.

Remarque : la propriété (P) devient une hypothèse à faire dans le cas d'un modèle de régression où les deux phases ne sont plus des droites (ex: parabole, sinus, cosinus, etc...). Les résultats obtenus, dans notre étude, pour (θ_n) restent valables un modèle à deux phases quelconques sous réserve que (P) soit vérifiée.

Remarquons aussi que (H) est vérifiée pour la suite $(t_i = i/n, i=1,..,n)$ ou pour toute suite réalisant une dichotomie de l'intervalle [0,1].

5.2. Existence et consistance.

L'estimation de θ^0 par la méthode des moindres carrés est obtenue par la minimisation, pour n fixé, sur Θ de :

$$S_n(\theta,y) = n^{-1} \sum_{i=1,n} (y_{ti} - f_{ti}(\theta))^2$$

avec $y^t = (y_{t1}, \dots , y_{tn})$, élément de $\mathbb{R}^n$.

Pour θ fixé dans Θ, $S_n(\theta,.)$ est une application mesurable de $(\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ dans $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$. Pour y fixé dans $\mathbb{R}^n$, $S_n(.,y)$ est une fonction continue sur Θ puisque chaque fonction f_{ti} l'est. Ainsi on peut appliquer le lemme 2 de [16] ie : il existe une application mesurable $\theta(y)$ de $\mathbb{R}^n$ dans Θ telle que pour tout $y \in \mathbb{R}^n$: $S_n(\theta(y),y) = \inf S_n(\theta,y)$, $\theta \in \Theta$.

Pour la suite on notera $S_n(\theta)$ au lieu de $S_n(\theta,y)$. L'existence d'un estimateur des moindres carrés, que l'on note θ_n, relatif à S_n est ainsi prouvée.

On obtient la consistance p.s. des estimateurs θ_n et $S_n(\theta_n)$ de θ^0 et σ^2 par le

Théorème :

Si l'hypothèse (H) portant sur la distribution des $(t_i)_i$ est vérifiée, on a :

p.s $\quad \lim_n \theta_n = \theta^0 \quad$ et $\quad \lim_n S_n(\theta_n) = \sigma^2$.

Démonstration :

C'est une application du théorème 6 de [16] lequel assure la convergence presque-sûre de θ_n et de $S_n(\theta_n)$ vers θ^0 et σ^2.

On se contente de montrer, en les rappelant au passage, comment les hypothèses de ce théorème, notées (a) et (b) dans [16], et qui conduisent au résultat, sont satisfaites pour notre modèle.

(a) On se donne le modèle

$$y_t = f_t(\theta^0) + e_t,\ t = 1,\ldots,n.$$

où les fonctions $f_t(.)$ sont continues sur un sous ensemble compact Θ d'un espace euclidien et les $(e_t)_t$ sont indépendants, de même loi centrée et de variance σ^2 inconnue.

(b) Pour tout θ dans Θ, $\text{Lim}\ n^{-1}\Sigma_{i=1,n}\, f_{ti}(\theta)^2$ existe.

Et $Q(\theta) = \text{Lim}\ n^{-1}\Sigma_{i=1,n}\,(f_{ti}(\theta) - f_{ti}(\theta^0))^2$, admet pour minimum unique θ^0.

(a) est bien sûr vérifiée et il reste à montrer que (b) l'est aussi.

Compte tenu de (H) et des propriétés de $f_t(\theta)$, on a par le théorème de Helly- Bray :

uniformément en θ dans Θ

$$\lim_n n^{-1} \sum_{i=1,n} (f_{ti}(\theta) - f_{ti}(\theta^0))^2 = \int_{[0,1]} (f_t(\theta) - f_t(\theta^0))^2\, dF(t) = Q(\theta).$$

On veut montrer que $Q(\theta)$ admet θ^0 pour minimum unique.

On a, $Q(\theta) \geq 0$. $Q(\theta)$ n'est nulle que si $f_t(\theta) = f_t(\theta^0)$ pour presque tout $t \in [0,1]$, et donc pour tout $t \in [0,1]$ puisque $f.(\theta)$ est continue.

Il vient alors, d'après (P) : $\theta = \theta^0$ et $Q(\theta^0) = 0$.□

CHAPITRE IV

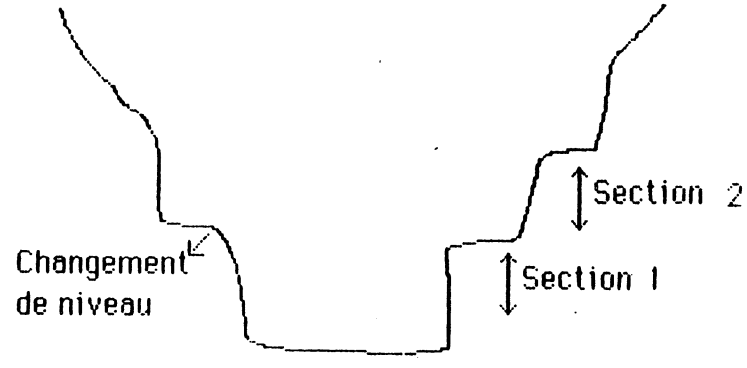
DETECTION DE RUPTURE ET ESTIMATION SUR DES DONNEES D'HYDROLOGIE

On dispose pour une station donnée d'un cours d'eau, d'un nuage de points dont les coordonnées sont la hauteur de la rivière et son débit donné en m^3/s. Les données (hauteur, débit) sont appelées des jaugeages. La hauteur d'une rivière peut être déterminée de deux façons : soit à l'aide d'un capteur de pression, soit par un système de flotteur.

L'évaluation du débit semble être plus complexe. Partant de la relation : débit = vitesse × surface, on recueille, à l'aide d'un système de mesure, la vitesse en plusieurs points d'une surface donnée. La vitesse globale est ensuite déduite. Le débit est alors le produit de cette vitesse par la surface de la section retenue.

L'objectif est alors, au vu des observations recueillies et aimablement mises à notre disposition par Mr Duban responsable à la division technique de EDF-Grenoble, d'estimer la relation existant entre le débit et la hauteur d'une rivière. Cela permet alors, sachant la relation estimée, de déduire le débit associé à une hauteur fixée a priori. Un tableau de ces données est appelé un barème de tarage. C'est donc l'établissement de ce barème qui intéresse particulièrement les utilisateurs. Il semblerait que la liaison statistique entre le débit et la hauteur ne soit pas stable et qu'elle varie selon les valeurs prises par la variable Hauteur. Cette instabilité peut être due, entre autres comme nous l'a expliqué Mr Duban, à la structure du lit de la rivière qui peut correspondre au schéma suivant.

<u>Exemple de section du lit d'une rivière</u> :



Une rupture dans la liaison entre le débit et la hauteur correspond alors au passage d'une section à une autre. Le nombre de ruptures, et donc de sections, variant avec la structure de la rivière.

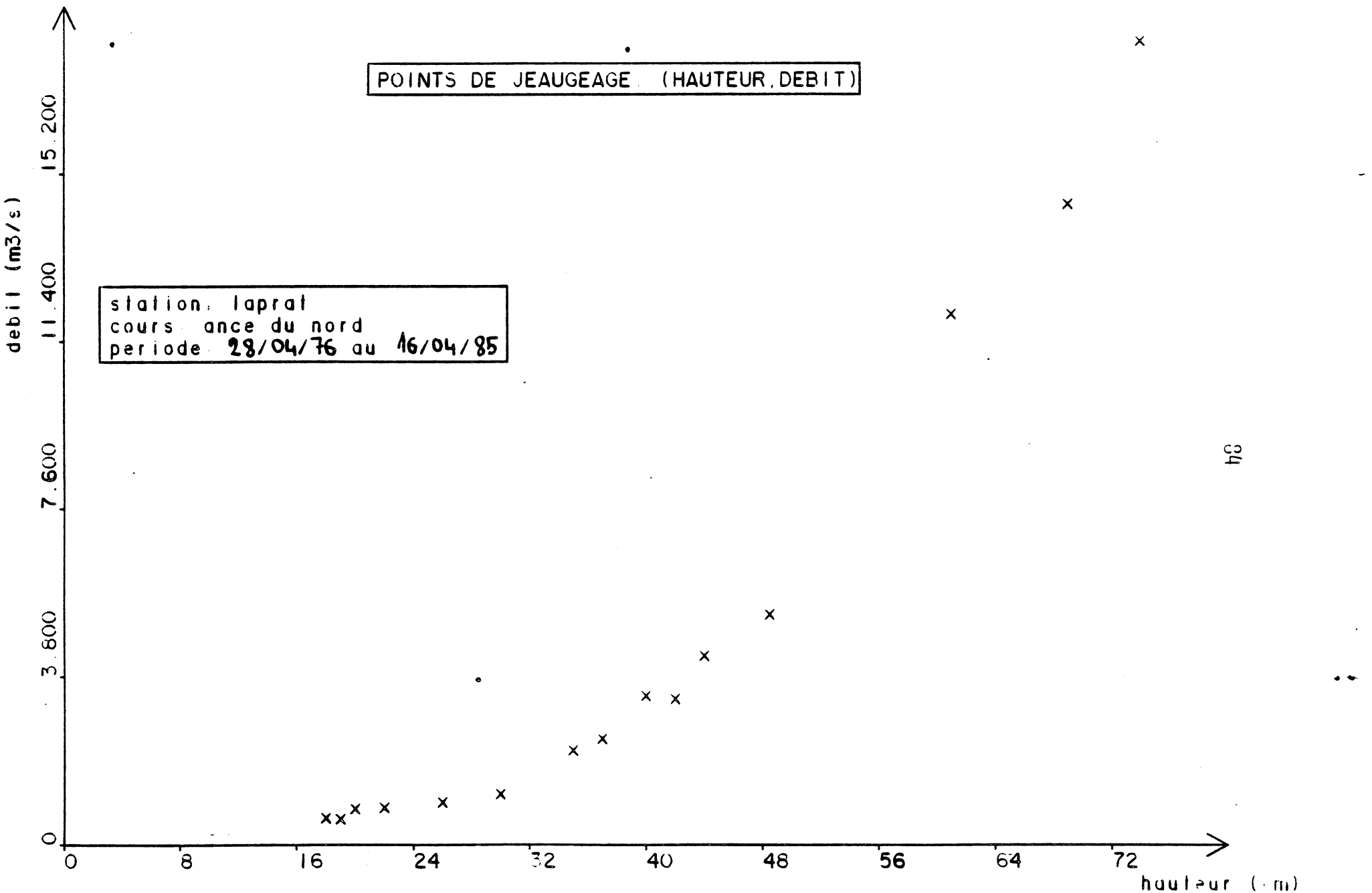
Il est donc légitime, au lieu d'ajuster directement une seule courbe au nuage de points, de vouloir tester, sur la base des observations disponibles, cette hypothèse de rupture.

Sur des données correspondant à 4 stations, nous avons testé à l'aide des deux tests : somme cumulée et rapport de vraisemblances, l'hypothèse de l'existence d'un changement dans la liaison entre la hauteur et le débit.

L'estimation des deux phases de la liaison, sous l'hypothèse que la rupture ait été retenue, est traitée pour chaque station.

Les valeurs critiques des tests sont celles obtenues par simulation dans les chapitres 1 et 2.

Pour chaque station nous représentons le nuage de points : (hauteur, débit).


POINTS DE JEAUGEAGE (HAUTEUR, DEBIT)
station: laprat
cours: ance du nord
periode: 28/04/76 au 16/04/85
debit (m3/s)
0
3.800
7.600
11.400
15.200
0
8
16
24
32
40
48
56
64
72
hauteur (cm)

PRESENTATION DES RESULTATS :

1) Station : LAPRAT ; cours d'eau : ANCE DU NORD ;
période de jaugeage : 28/04/76 au 16/04/85.
Nombre d'observations : 15.

Le test de la somme cumulée détecte une rupture, pour les trois seuils α = 1%, 5% et 10%, à la 11^{ième} observation. La statistique du test vaut : 5.3016 et les valeurs de la fonction frontière sont respectivement pour les seuils 1%, 5% et 10% : 2.6907, 2.1576 et 1.9038. Etant donné ces résultats, on peut dire qu'une rupture est bien détectée.

Le test du rapport de vraisemblances détecte lui aussi une rupture à la 12^{ième} observation et ce pour les trois seuils considérés : 1%, 5% et 10%.

Sous cette hypothèse de rupture, l'estimation des deux droites liées par une contrainte de continuité a conduit aux résultats suivants.

L'estimation de l'abscisse du point d'intersection est : 0.6131 ; la rupture a lieu entre la 13^{ième} et la 14^{ième} observation ;

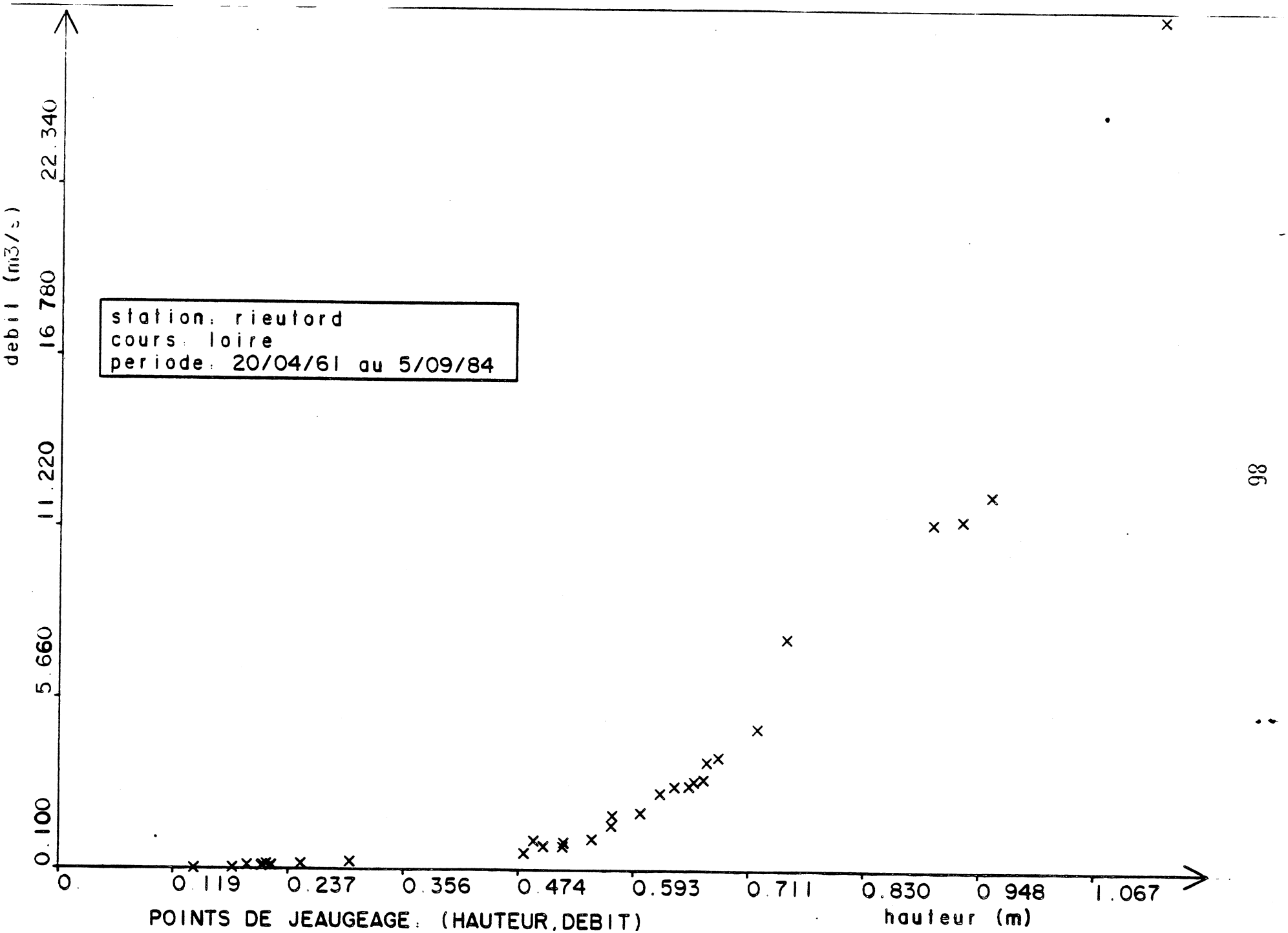
Les deux droites (D1) et (D2) sont :

$$(D1) \quad y = -4.4954 + 21.701\,x$$

$$(D2) \quad y = -36.559 + 73.999\,x$$

l'ordonnée du point d'intersection est donc $\simeq$ 8.809.

On remarque que l'estimation du point de rupture se situe en fin d'observation ou presque. On peut ne pas être satisfait dans ce cas par la qualité de l'estimation, surtout pour les paramètres de la deuxième droite. En effet seules deux observations, les 14ième et 15ième, sont prises en compte dans l'estimation de la deuxième phase. Il est donc utile de disposer d'observations supplémentaires pour cette station afin d'être plus sûr des estimations obtenues.


station: rieutord
cours: loire
periode: 20/04/61 au 5/09/84
debit (m3/s)
0.100
5.660
11.220
16.780
22.340
0
0.119
0.237
0.356
0.474
0.593
0.711
0.830
0.948
1.067
hauteur (m)
POINTS DE JEAUGEAGE: (HAUTEUR, DEBIT)

2) Station : RIEUTORD, cours d'eau : LOIRE ;
période de jaugeage : 20/04/61 au 05/09/84.
Nombre d'observations : 31.

Le test de la somme cumulée s'avère impuissant pour détecter une rupture et accepte donc l'hypothèse de stabilité aux trois seuils α = 1%, 5% et 10%. Cela est assez surprenant compte tenu de la forme du nuage qui fait apparaître tout de même une instabilité dans la liaison entre le débit et la hauteur.

Le test du Rapport de vraisemblance arrive à détecter une rupture à la $29^{ième}$ observation pour les seuils α = 1%, 5% et 10%. La détection est donc assez tardive et cela peut expliquer pourquoi le test de la somme cumulée n'a pas été un bon indicateur dans ce cas.

L'hypothèse de l'existence d'une rupture est donc retenue et l'estimation des deux phases donne lieu aux résultats suivants :

l'abscisse du point d'intersection est : 0.82926 ;

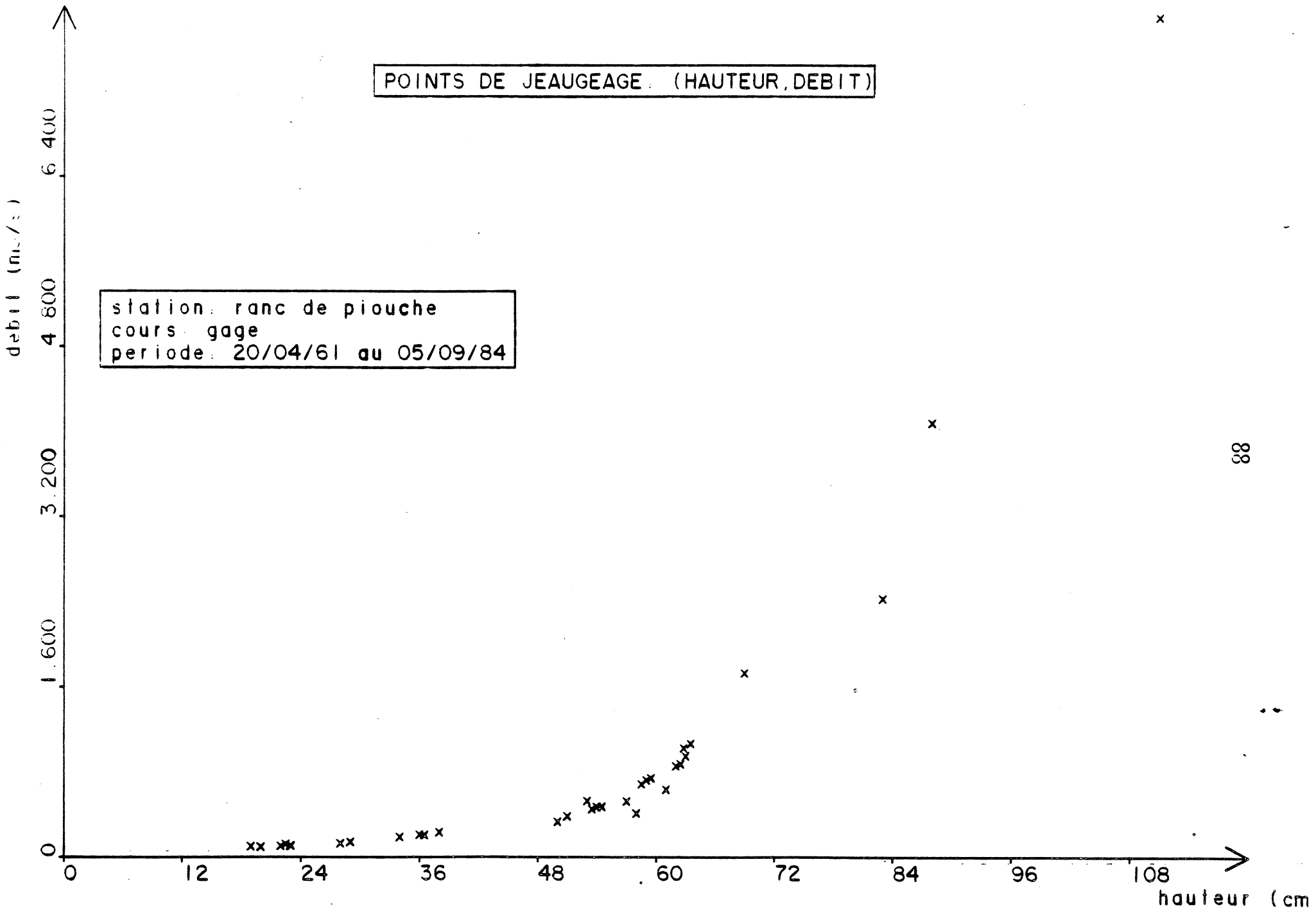
La rupture a lieu entre la $27^{ième}$ et la $28^{ième}$ observation et les deux droites estimées (D1) et (D2) sont :

$$(D1)\quad y = -1.6835 + 7.2198\ x$$

$$(D2)\quad y = -57.467 + 74.488\ x$$

Le point d'intersection a pour ordonnée : $\simeq$ 4.302.

On peut remarquer, d'une façon générale, que les estimateurs de l'instant de rupture obtenus par le biais des tests et celui obtenu par l'estimation des deux phases sont différents. Cela est simplement dû au fait que les modèles de régression avec rupture sous-jacents ne sont pas exactement les mêmes. En effet, lorsque l'instabilité du modèle est retenue, nous appliquons la procédure d'estimation à un modèle de régression, à deux phases linéaires, qui tient compte d'une contrainte de continuité liant les paramètres du modèle.


POINTS DE JEAUGEAGE (HAUTEUR, DEBIT)
station: ranc de piouche
cours: gage
periode: 20/04/61 au 05/09/84
débit (m3/s)
0
1.600
3.200
4 800
6 400
0
12
24
36
48
60
72
84
96
108
hauteur (cm
88

3) Station : RANC DE PIOUCHE ; cours d'eau : GAGE ;
période de jaugeage : 20/04/61 au 05/09/84.
Nombre d'observations : 34.

Le test de la somme cumulée détecte une rupture pour les trois seuils utilisés.

Pour α = 5% et α = 10%, la rupture a lieu à la 30^{ième} observation. La statistique vaut : 2.6433 et les valeurs de la fonction frontière sont respectivement : 2.390 et 2.109. Pour α = 1%, il y a détection de rupture à la 31^{ième} observation. La statistique du test est : 5.698 et la frontière prend pour valeur : 3.047.

De même le test du rapport de vraisemblances détecte une rupture à la 32^{ième} observation aux seuils α = 1%, 5% et 10%.

L'estimation a conduit aux résultats suivants :

Le point d'intersection a pour abscisse : 0.7377 et la rupture a lieu entre les 31^{ième} et 32^{ième} observations ;

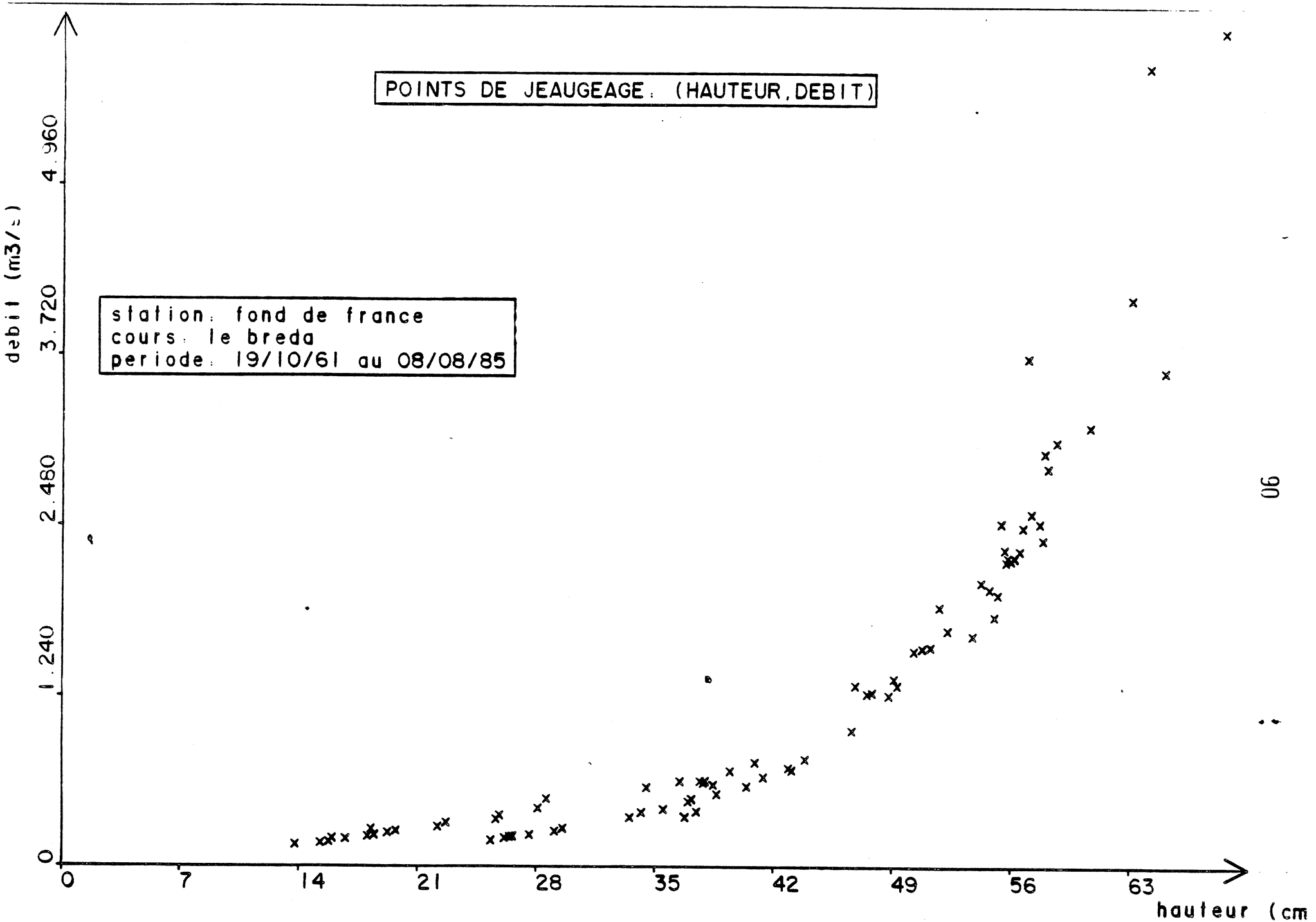
Les deux droites sont :

$$(D1) : y = -0.4292 + 1.985\ x$$

$$(D2) : y = -12.7089 + 18.630\ x$$

l'ordonnée du point d'intersection est alors : $\simeq$ 1.035.

Comme pour la première station, la rupture se produisant en fin d'observation, l'estimation de la deuxième phase ne se fait qu'à l'aide de trois observations uniquement.


POINTS DE JEAUGEAGE: (HAUTEUR,DEBIT)
station: fond de france
cours: le breda
periode: 19/10/61 au 08/08/85
debit (m3/s)
0
1.240
2.480
3.720
4.960
0
7
14
21
28
35
42
49
56
63
hauteur (cm

4)Station : FOND DE FRANCE ; cours d'eau : LE BREDA ;
période de jaugeage : 19/10/61 au 08/08/85.
Nombre d'observations : 80.

Cette station présente l'avantage de disposer d'un assez grand nombre d'observations comparativement aux autres stations. On s'attend ainsi, à ce que l'estimation des deux phases, si l'hypothèse de l'existence d'une rupture est retenue, soit de meilleure qualité que les précédentes. Il semble, au vu du nuage, que le changement de la liaison linéaire entre le débit et la hauteur n'ait pas lieu en fin d'observation et on peut espérer, vu la taille de l'échantillon, que les tests aussi le détectent avec peu de retard.

Ne disposant pas pour le test de la somme cumulée des valeurs critiques pour n = 80, nous avons testé la procédure avec successivement les valeurs critiques estimées pour n = 60 et n = 100. L'idée est que, pour un seuil α donné, les valeurs critiques pour n = 60 et n = 100 servent d'encadrement à celle associée à n = 80.

α = 10%, la détection se fait différemment selon la valeur critique utilisée.
Détection de rupture à la 47^{ième} observation (pour n = 60) ;
Statistique calculée : 1.8522 ; valeur de la frontière : 1.6706.
Détection de rupture à la 48^{ième} observation (pour n = 100) ;
Statistique calculée : 1.9347 ; valeur de la frontière : 1.7861.

α = 5%, détection de rupture à la 49^{ième} observation ;
Statistique calculée : 2.1809 ;
valeur de la frontière : 1.9630 (pour n = 60) ;
valeur de la frontière : 2.0076 (pour n = 100) ;

$\alpha = 1\%$, détection de rupture à la $50^{ième}$ observation ;

Statistique calculée : 2.4014 ;

valeur de la frontière : 2.3917 (pour n = 60 et pour n = 100) ;

On peut ainsi, au vu de ces résultats, accepter l'existence d'une rupture.

Pour le test du rapport de vraisemblance, nous avons pris pour valeurs critiques celles obtenues pour n = 60. Le test confirme bien la présence d'une rupture et la détecte, beaucoup plus tardivement que le test de la somme cumulée, à la $76^{ième}$ observation aux seuils α = 1%, 5% et 10%.

L'estimation sous l'hypothèse d'existence d'une rupture a donné les résultats suivants :

l'estimation du point d'intersection des deux droites est : 0.50667 ;

La rupture a lieu entre la $51^{ième}$ et la $52^{ième}$ observation et les deux droites sont :

$$(D1) \quad y = -0.4077 + 2.8279\,x$$

$$(D2) \quad y = -10.871 + 23.478\,x$$

l'ordonnée du point d'intersection est : $\simeq$ 1.0251.

Comment expliquer cette différence, assez nette, entre les instants de détection issus des deux tests ? Il faut remarquer que la philosophie des deux procédures n'est pas la même. En ce sens que la statistique de la somme cumulée réagit dés qu'elle rencontre une rupture assez importante et de ce fait détecte vraisemblablement la première rupture lorsque il en existe plusieurs. La statistique du rapport de vraisemblances, par contre, fait apparaître la rupture la plus importante et pas nécessairement la première si le phénoméne étudié comporte plus d'une rupture. Il s'ensuit que pour cette station, l'hypothèse que l'on pourrait retenir est qu'il y a deux changements, dans la liaison linéaire entre le débit et la hauteur, et non pas un seul.

Au vu de l'estimation des deux phases, on a retenu l'existence d'une première rupture entre la 51^{ième} et la 52^{ième} observation et conclu donc à l'hypothèse de stabilité pour les 51 premières observations. Pour les 29 restantes, nous avons testé la conjecture de la présence d'une deuxième rupture à l'aide du test de la somme cumulée.

Ainsi une rupture est décelée à la 26^{ième} observation, autrement dit à la 77ième lorsque toutes les observations sont comptées, aux deux seuils supposés α = 1% et α = 5%.

La statistique calculée est : 3.5688 et les valeurs de la frontière sont respectivement, pour α = 1% et α = 5%, 3.0229 et 2.4240. Il est donc tout à fait conseillé, pour cette station, de retenir un modèle de régression linéaire à trois phases distinctes.

La procédure d'estimation étant mise au point pour le cas de deux phases uniquement, nous avons été conduit, plutôt que d'estimer les trois phases globalement comme il aurait été plus juste, à estimer les deux dernières phases à partir des 29 dernières observations. Cela a pour conséquence évidente que les deux premières phases ne se rencontrent plus entre la 51ième et la 52ième.

On obtient ainsi les équations des trois phases linéaires entre le débit y et la hauteur x :

$$(D1)\quad y_i = -0.4077 + 2.8279\, x_i \qquad 1 \leq i \leq 51\ ;$$

$$(D2)\quad y_i = -3.2376 + 9.4213\, x_i \qquad 52 \leq i \leq 59\ ;$$

$$(D3)\quad y_i = -13.153 + 27.227\, x_i \qquad 60 \leq i \leq 80\ ;$$

Le deuxième point de rupture, estimé entre les 59ième et 60ième observation, a pour coordonnées : (0.5568, 2.0081).

A l'issue de cette application, on peut considérer que les deux tests utilisés ont, dans l'ensemble, bien réagi pour rendre compte de la présence

d'une rupture dans la liaison linéaire entre le débit et la hauteur. De plus leur utilisation en parallèle a été bénéfique comme nous l'avons vue dans le cas de la station Fond de France.

De même les régressions estimées entre le débit et la hauteur peuvent servir de support à l'établissement d'un barème de tarage avec cependant quelques réserves pour les stations Laprat et Ranc de Piouche, en raison de la remarque concernant la qualité de l'estimation de la deuxième phase.

Nous avons voulu voir, suite à une suggestion de Mr Duban, quel serait le comportement des deux procédures de détection pour tester un changement de liaison non nécessairement linéaire entre le débit et la hauteur.

Pour la station Fond de France, au vu de l'allure du nuage et compte tenu de certains résultats empiriques concernant cette station, nous avons pensé tester la stabilité d'une liaison du type :

(1) $Q = a H^p$,

où Q et H désignent respectivement les variables Débit et Hauteur.

Si on prend le logarithme des observations $(Q_i, H_i)_{i=1...n}$, on se ramène à tester la stabilité d'une régression linéaire. On applique alors les deux procédures aux observations $(\log Q_i, \log H_i)_{i=1...n}$.

Le test de la somme cumulée retient l'hypothèse d'un changement dans la liaison linéaire entre log Q et log H.

L'utilisation des valeurs critiques pour n = 60 et n = 100 conduit au même résultat.

Aux trois seuils α = 1%, 5% et 10%, le test détecte une rupture à la $16^{\text{ième}}$ observation.

La statistique prend pour valeur : 1.5426 et les valeurs respectives de la frontière sont : 1.4676, 1.2184 et 1.0869 pour n = 60 ; 1.4676, 1.2461 et 1.1215 pour n = 100.

Le test du rapport de vraisemblance confirme l'existence d'un changement puisqu'il détecte pour les trois seuils une rupture à la 12ième observation.

Nous acceptons alors, suite à ces résultats, l'hypothèse de l'instabilité de la liaison entre log Q et log H.

On se retrouve donc sous l'hypothèse suivante :

$$(2) \qquad Q_i = f(H_i) = \begin{cases} a_1 H_i^{p_1} & i = 1, \ldots, r \\ a_2 H_i^{p_2} & i = r+1, \ldots, n. \end{cases}$$

$$\text{et} \qquad H_r \leq \gamma < H_{r+1}.$$

γ est le point d'intersection des deux phases du modèle. Le paramètre r est supposé inconnu.

Pour estimer les différents paramètres, on transforme (2). Ainsi :

$$(2) \Leftrightarrow (2') \quad \log Q_i = \begin{cases} \log a_1 + p_1 \log H_i & i = 1, \ldots, r \\ \log a_2 + p_2 \log H_i & i = r+1, \ldots, n. \end{cases}$$

$$\text{et} \qquad \log H_r \leq \log \gamma < \log H_{r+1}.$$

En appliquant l'algorithme d'estimation, présenté au chapitre III, au modèle (2'), on arrive aux résultats suivants.

Les deux courbes du modèle (2) se rencontrent entre la 26ième et la 27ième observation.

Le point d'intersection γ a pour coordonnées : (0.358, 0.38907).

Les paramètres des deux phases de (2) sont :

$$(a_1, p_1) = (0.96519, 0.8845)$$

$$(a_2, p_2) = (21.4727, 3.9046).$$

BIBLIOGRAPHIE

Nous avons classé les articles traitant des problèmes de rupture selon les deux aspects : estimation et test de détection. Nous faisons suivre ceux traitant de l'aspect estimation (resp. test) par (E) (resp. (T)), et ceux traitant des deux par (E.T).

[1] BARRA, J.R. : (1971)
Notions fondamentales de statistiques mathématiques. DUNOD.

[2] BASSEVILLE, M. : (1982)
Contributions à la détection sequentielle de ruptures de modèles statistiques.
Thèse d'état. RENNES. (E.T)

[3] BATTACHARYYA, G.K. and JOHNSON, R.A. : (1968)
Non-parametric tests for shift at an unknown time point.
Ann. Math. Statist., 39, 1731-43. (T)

[4] BILLINGLSEY, P. : (1968)
Convergence of probability measures.
J. Wiley.

[5] BROWN, R.L., DURBIN, J. and EVANS, J.M. : (1975)
Techniques for testing the constancy of regression relationships over time.
J Royal Statistical Soc. B, Vol 37, 149-192. (T)

[6] Mc CABE, B.P.M. and HARRISSON M.J. : (1980)
Testing the constancy of regression relationships over time using least squares residuals.
Applied Statistics 29, n°2, 142-148. (T)

[7] CHERNOFF, H. and ZACKS, S. : (1964)
Estimating the current mean of a normal distribution which is subjected to changes in time.
Ann. Math. Statist. , 35, 999-1018. (E.T)

[8] DESHAYES, J. et PICARD, D. :
(1) (1979)
Tests de rupture de régression : comparaison asymptotique.
Preprint de l'Université d'ORSAY. 79 T 23. (T)
(2) (1980)
Testing for a change-point in statistical models. Preprint Université PARIS SUD; Département Mathématiques E.R.A. 532, "Statistique Appliquée". 1-50. (E.T)

(3) (1982)
Ruptures dans les modèles de régression. Lois asymptotiques des tests et estimateurs du maximum de vraisemblance.
Preprint d'Orsay. T 20. (E.T)

[9] DOOB, J.L. : (1953).
Stochastic Processes.
Wiley.

[10] DURBIN, J. : (1971)
Boundary crossing probabilities for the brownian notion and Poisson processes and techniques for computing the power of the Kolmogorov-Smirnov test.
Journal of applied Probability 8, 431-453.

[11] FEDER, P.I. : (1975)
On asymptotic distribution theory in segmented regression problems. Identified case.
Annals of Statistics. Vol 3, N° 1, 49-83. (E)

[12] GARBADE, K. : (1977)
Two methods for examining the stability of regression coefficients.
J.A.S.A. ; Vol 72, n° 357, 54-63. (T)

[13] HAWKINS, D.M. : (1977)
Testing a sequence of observations for a shift in location.
J.A.S.A. ; Vol 72, n° 357, 180-186. (T)

[14] HINKLEY, D.V. :
(1) (1969)
Inference about the intersection in two-phase regression.
Biometrika 56, 495-504. (E)

(2) (1970)
Inference about the change point in a sequence of random variables.
Biometrika 57, 1-17. (E)

(3) (1971)
Inference in two-phase regression.
J.A.S.A. ; Vol 6, n° 336 ; 736-743 (E)

[15] HUDSON, D.J. : (1966)
Fitting segmented curves whose join-points have to be estimated.
J.A.S.A. 61 ; 1097-1129. (E)

[16] JENNRICH, R.I. : (1969)
Asymptotic properties of non-linear least squares estimators.
The annals of Math. Statist. Vol 40, n° 2, 633-643.

[17] PAGE, E.S. : (1955)
A test for a change in a parameter occuring at an unknown point.
Biometrika, 42, 523-6. (E.T)

[18] PETTIT, A.N. :
(1) (1979)
A non-parametric approach to the change-point problem.
Applied Statistics, 28, n° 2, 126-135. (T)

(2)(1980)
A simple cumulative sum type statistic for the change point problem with zero-one observations.
Biometrika, 67, n° 1, 79-84 (T)

[19] QUANDT, R.E.
(1) (1958)
The estimation of the parameters of a linear regression system obeying two separate regimes.
J.A.S.A. 53, 873-880. (E)

(2) (1960)
Tests of the hypothesis that a linear regression system obeys two separate regimes.
J.A.S.A . 55, 324-330. (T)

[20] SAPORTA, G. : (1978)
Théories et méthodes de la statistique.
Technip.

[21] SEN, A. and SRIVASTAVA, M.S. : (1975)
On tests for detecting change in mean.
Annals of statistics, 3, 98-108. (T)

[22] SMITH, A.F.M. and COOK, D.G. (1980)
Straight lines with a change point : a bayesian analysis of some renal transplant data.
Applied Statistics, 29, n° 2, 180-189 (E.T)

[23] WORSLEY, K. J. : (1979)
On the likelihood ratio test for a shift in location of normal populations.
J.A.S.A. , Vol 74, n° 366, 365-367. (T)

ANNEXE.

```
PROGRAM ESTIMATION ;

          {Ce programme calcule les estimateurs du maximum de vraisemblance
         des parametres d'un modele de regression avec rupture.}

label     1;
type      vec = array[1,.100] of real;
var
          x,y,mx,my,mxs,mys,xx,xy,xxs,xys,a0,b0,a1,b1,g,c,d,e:vec;
          i,j,n,r,m1,m2:integer;u,v,gmax,xmax,a,b,f,h,u0,v0,u1,v1:real;
          f1:text;chen:string;

procedure interm;
var       i,j:integer;
begin
          mx[1]:=0;my[1]:=0;xx[1]:=0;xy[1]:=0;
          mxs[1]:=0;mys[1]:=0;xxs[1]:=0;xys[1]:=0;
          for i:=1 to n do
                    begin
                              for j:=1 to i do
                                        begin
                                                  mx[i]:=mx[i]+x[j];
                                                  my[i]:=my[i]+y[j]
                                        end;
                              mx[i]:=mx[i]/i;
                              my[i]:=my[i]/i;
                              for j:=1 to i do
                                        begin
                                                  xx[i]:=xx[i]+sqr(x[j]-mx[i]);
                                                  xy[i]:=xy[i]+(x[j]-mx[i])*(y[j
]-my[i])
                                        end;
                    end;
          for i:=1 to n-1 do
                    begin
                              for j:=i+1 to n  do
                                        begin
                                                  mxs[i]:=mxs[i]+x[j];
                                                  mys[i]:=mys[i]+y[j]
                                        end;
                              mxs[i]:=mxs[i]/(n-i);
                              mys[i]:=mys[i]/(n-i);
                              for j:=i+1 to n do
                                        begin
                                                  xxs[i]:=xxs[i]+sqr(x[j]-mxs[i]);
                                                  xys[i]:=xys[i]+(x[j]-mxs[i])*(y[j]-
mys[i])
                                        end;
                    end;
end;

          procedure simplif;
          var       i:integer;
begin
          interm;
          {calcul des estimateurs sans contraintes}

          for i:=2 to n-2 do
                    begin
                              b0[i]:=xy[i]/xx[i];a0[i]:=my[i]-(mx[i]*b0[i]);
```

```
                                b1[i]:=xys[i]/xxs[i];a1[i]:=mys[i]-(mxs[i]*b1[i]);
                                g[i]:=(a1[i]-a0[i])/(b0[i]-b1[i]);
c[i]:=((n*xx[i]*xxs[i])+(i*(n-i)*((sqr(mx[i])*xxs[i])+(sqr(mxs[i])*xx[i]))))/n;
d[i]:=(((mx[i]*xxs[i])+(mxs[i]*xx[i]))*(i*(n-i)))/n;
e[i]:=((xx[i]+xxs[i])*(i*(n-i)))/n;
                        end;
end;
        procedure extreme;
        begin
                gmax:=0;m1:=0;
             u:=sqr((c[2]-(d[2]*(g[2]+x[2]))+(e[2]*g[2]*x[2]))*(b0[2]-b1[2]))
;
                v:=(c[2]-(2*d[2]*x[2])+(e[2]*sqr(x[2])))*xx[n];
                xmax:=u/v;m2:=2;
                if ( g[n-2] < x[n-1] ) and ( g[n-2] >= x[n-2] )  then
                        begin
                                u:=c[n-2]-(2*d[n-2]*g[n-2])+(e[2]*sqr(g[
n-2]));
                                v:=(sqr(b0[n-2]-b1[n-2]))/xx[n];
                                gmax:=u*v;
                                m1:=n-2
                        end;
        end;

{******* PROGRAMME PRINCIPAL *******}
begin
        {recherche de l'estimateur de gamma}
        write('entrer le nom du fichier');readln(chen);
        assign(f1,chen);reset(f1);
        writeln('entrer n');readln(n);
        for i:=1 to n do
        begin
                readln(f1,x[i],y[i]);
                x[i]:=ln(x[i]);y[i]:=ln(y[i])
        end;
        simplif;
        extreme;
        i:=2;
        while i < n-2  do
                begin
                        if ( g[i] < x[i+1] ) and ( g[i] >= x[i] )  then
                                begin
                                        u:=c[i]-(2*d[i]*g[i])+(e[i]*sq
1:
r(g[i]));
                                        v:=(sqr(b0[i]-b1[i]))/xx[n];
                                        u:=u*v;
                                        if u > gmax  then
                                                begin
                                                        gmax:=u;
                                                        m1:=i
                                                end;
                                        i:=i+1;
                                end
                        else
                begin
                   if ( g[i+1] >= x[i+2] ) or ( g[i+1] < x[i+1] )  then
                begin
                u:=sqr(c[i+1]-(d[i+1]*(g[i+1]+x[i+1]))+(e[i+1]*g[i+1]*x[i+1]
));
                u:=u*sqr(b0[i+1]-b1[i+1]);
```

```
                v:=(c[i+1]-(2*d[i+1]*x[i+1])+(e[i+1]*sqr(x[i+1])))*xx[n];
                u:=u/v;
                if u > xmax  then
                        begin
                                xmax:=u;
                                m2:=i+1
                        end;
                        i:=i+2
                end else
                begin
                        i:=i+1;
                        goto 1
                end
                end
        end;
        writeln('gmax = ',gmax,' m1 = ',m1);writeln;
        writeln('xmax = ',xmax,' m2 = ',m2);writeln;
        if gmax > xmax   then
                begin
                writeln('1 estimation de gamma est :',g[m1]);writeln;
writeln('la rupture est entre la ',m1,'ieme et la ',m1+1,'ieme observation');
                writeln;
                writeln('a0^ = ',a0[m1],'      b0^ = ',b0[m1]);writeln;
                writeln('a1^ = ',a1[m1],'      b1^ = ',b1[m1])
                end  else
        begin
                u:=(a0[m2]-a1[m2])+(b0[m2]-b1[m2])*x[m2];
                v:=(n/(m2*(n-m2)))+((sqr(mx[m2]-x[m2]))/xx[m2])+((sqr(mxs[m2
]-x[m2]))/xxs[m2]);u:=u/v;
                v:=(1/m2)+((sqr(mx[m2]))/xx[m2])-((mx[m2]*x[m2])/xx[m2]);
                u0:=a0[m2]-(u*v);v0:=b0[m2]-(u*((x[2]-mx[m2])/xx[m2]));
                v:=((mxs[m2]*x[m2])/xxs[m2])-((sqr(mxs[m2]))/xxs[m2])-(1/(n-
m2));
                u1:=a1[m2]-(u*v);v1:=b1[m2]-(u*((mxs[m2]-x[m2])/xxs[m2]));
                writeln('1 estimation de gamma est :',x[m2]);writeln;
        writeln('la rupture a lieu a la ',m2,'ieme observation');
                writeln;writeln;
                writeln('a0^ = ',u0,'       b0^ = ',v0);writeln;
                writeln('a1^ = ',u1,'       b1^ = ',v1)
        end
end.
```

```
PROGRAM FRACTIL;

{Ce programme calcule les fractiles de la loi de max|Z(r)|,1<=r<=n-1 (Page 50).
 Il fait appel aux fonctions h(r) qui sont calculees par le module ITERE.}

CONST     PI=3.1415926535897;
TYPE      LIST= ARRAY[1..20] OF REAL;
VAR       X:LIST;I,M,N,K,K1,K2,P,PP:INTEGER;V1,V2,P1,A2:REAL;
          F,G:TEXT;

EXTERNAL FUNCTION H(R:INTEGER;U,V:REAL):REAL;

FUNCTION REPART(V1,V2:REAL):REAL;
VAR       Z,SZ:REAL; L:INTEGER;

FUNCTION G(U:REAL):REAL;
VAR       J:INTEGER;S,S1:REAL;

          BEGIN
                    S:=0;
                    S1:=2*(1/SQRT(2*PI))*EXP(-(SQR(U))/2);
                    K:=(N DIV 2);K2:=N-(2*K);
                    IF (K2=0)  THEN
                              BEGIN
                                        FOR J:=1 TO K-1  DO
                                        S:=S+2*H(J,U,U)*H(N-J,U,U);
                                        S:=S+SQR(H(K,U,U));
                                        S:=S*S1;G:=S;
                                END ELSE
                                  BEGIN
                                        FOR J:=1 TO K  DO
                                        S:=S+2*H(J,U,U)*H(N-J,U,U);
                                        S:=S*S1;G:=S;
                                  END;
          END;
BEGIN
          A2:=((V2-V1)/PP);Z:=0;
          FOR L:=1 TO (PP-1) DO
          BEGIN
                    SZ:=G(V1+(L*A2));
                    Z:=Z+(2*SZ);
          END;
          Z:=Z+G(V1)+G(V2);Z:=(A2/2)*Z;
          WRITELN('REPART:= ',Z);
          REPART:=Z;
END;

BEGIN (* P.P *)
          ASSIGN(F,'B:POINTS');RESET(F);ASSIGN(G,'B:PROBA');REWRITE(G);
          WRITELN('TAILLE DES DONNEES :');READLN(N);
          WRITELN('ENTRER LES',N,'VALEURS DE X');
          FOR I:=1 TO N DO  READ(X[I]);
          WRITELN('ENTRER P ET PP ASSEZ GRANDS');READLN(P,PP);
          READLN(F,M);
          WRITELN(G,'CALCUL DE FRACTILES DE LA LOI DU MAX  POUR N:= ',N);
          WRITELN(G);
          WRITELN(G,'LES PAS CHOISIS :',P,'ET',PP);
          WRITELN(G,'   V                              P(V)');
          FOR I:=1 TO M DO
                    BEGIN
```

```
                                READLN(F,V1,V2);
                                P1:=REPART(V1,V2);
                                WRITELN(G,V1,' ',V2,'    ',P1)
                        END;
CLOSE(G,K1);
END.


MODULE ITERE;

TYPE    LIST=ARRAY [1..20] OF REAL;
VAR       P:EXTERNAL INTEGER;X:EXTERNAL LIST;N:EXTERNAL INTEGER;A1:REAL;
FUNCTION H(R:INTEGER;U,V:REAL):REAL;
VAR
          I,J:INTEGER;T:REAL;

FUNCTION F(R:INTEGER;C,D:REAL):REAL;
CONST
          PI=3.1415926535897;
VAR
          Z1,Z2,Z3,Z4,Z5,Z,B,B1,B2:REAL;
          L:INTEGER;
BEGIN

          Z1:=0;Z2:=0;
          FOR L:=1 TO R DO Z1:=Z1+SQR(X[L]);
          FOR L:=R+1 TO N DO Z2:=Z2+SQR(X[L]);
          Z3:=Z1-SQR(X[R]);Z4:=Z2+SQR(X[R]);
          Z5:=((Z2*Z3)/(Z1*Z4));Z:=(1-Z5);
          B:=SQRT(2*PI*Z);B:=(1/B);
          B1:=EXP(-(SQR(C-((SQRT(Z5))*D))/(2*Z)));
          B2:=EXP(-(SQR(C+((SQRT(Z5))*D))/(2*Z)));
          F:=(B1+B2)*B;
END;

BEGIN
          A1:=(V/P);
          IF   R=1   THEN  H:=1
        ELSE
          BEGIN
                    T:=0;
                    FOR I:=1 TO P-1 DO T:=T+2*H(R-1,A1*I,V)*F(R,A1*I,U);
                    T:=T+ H(R-1,V,V)*F(R,V,U);
                    T:=(A1/2)*T;H:=T;
          END;
END;
MODEND.
```

AUTORISATION DE SOUTENANCE

DOCTORAT 3ème CYCLE, DOCTORAT-INGENIEUR, DOCTORAT USMG

Vu les dispositions de l'arrêté du 16 avril 1974,

Vu les dispositions de l'arrêté du 5 juillet 1984,

Vu les rapports de M. F. BRADEAU

M. .....

M. Yacine SAIDI est autorisé à présenter une thèse en vue de l'obtention du titre de Docteur 3e cycle en Mathématiques Appliquées

Grenoble, le 20 DEC. 1985

Le Président de l'Université Scientifique et Médicale

M. TANCHE

Résumé de thèse:

Nous étudions deux procédures - somme cumulée des résidus récursifs et rapport des vraisemblances maximales - de détection de rupture dans un modèle de régression, en vue de leur application à des problèmes concrèts. Nous menons une étude expérimentale par simulation, afin de cerner le comportement de ces deux méthodes de détection de rupture.
Le problème de l'estimation, par le maximum de vraisemblance, dans un modèle de régression à une rupture est traité.
Une application des méthodes étudiées sur des données d'hydrologie est présentée.

Mots clefs :

Résidus récursifs. Somme cumulée. Rapport de vraisemblances maximales. Modèle de régression. Rupture. Simulation. Test de détection.