Quantile de régression séquentielle : application à l'étude du mode d'action de composés chimiques

Ronan BUREAU [2] & Gilles DURRIEU [1] & Jonathan VILLAIN [1,2]

[1] *Laboratoire de Mathématiques de Bretagne Atlantique, Université de Bretagne Sud et UMR CNRS 6205, Campus de Tohannic, 56017 Vannes.*
gilles.durrieu@univ-ubs.fr et jonathan.villain@univ-ubs.fr
[2] *Centre d'Études et de Recherche sur le Médicament de Normandie, Université de Caen Basse Normandie, Caen.*
ronan.bureau@unicaen.fr

Résumé. Du fait de l'évolution constante du nombre de composés chimiques connus, nous proposons une méthode séquentielle robuste pour les modèles QSAR (Quantitative structure–activity relationship). Les méthodes de statistiques séquentielles sont adaptées à ce type de situation où l'objectif principal est la précision que l'on souhaite obtenir. La définition d'une règle d'arrêt nous indique si nous pouvons arrêter l'expérience après les premières mesures ou si nous devons continuer avec une nouvelle observation. Cette méthode est ensuite appliquée pour déterminer le mode d'action de composés chimiques dans le domaine de la chémoinformatique.

Mots-clés. Statistique séquentielle, régression quantile, mode d'action, QSAR.

Abstract. Due to the constant evolution of the number of known chemical compounds, we propose a robust sequential method for QSAR (Quantitative structure–activity relationship) models. The sequential statistical methods are adpated to this type of situation in which the main objective is the precision that we want to obtain. The definition of a stopping rule indicates if we can stop the experiment after the first measurements or if we must continue with a new observation. This method is then applied for determining the mode of action of chemicals compounds in chemoinformatics.

Keywords. Sequential statistic, quantile regression, mode of action, QSAR.

1 Introduction

L'apparition de la chimie nous a permis de comprendre de nombreux phénomènes ainsi que la découverte des éléments qui nous entourent. À l'heure actuelle, la recherche en chimie est toujours très active, ce qui a pour effet d'aboutir à la découverte de nouveaux composés chimiques. L'ensemble des composés chimiques connus est loin d'être exhaustif

et les avancées sur les médicaments et sur les composés chimiques évoluent de jour en jour. Afin de se faire une idée *a priori* sur les effets des composés, de nouveaux modèles QSAR (Quantitative structure–activity relationship) sont établis sur un nombre restreint de composés chimiques. Il est important de mettre en place une procédure statistique qui prend en compte l'arrivée séquentielle de nouveaux composés chimiques.

Nous construisons dans ce papier une méthode séquentielle pour les modèles QSAR. Dans la première partie, nous décrivons la procédure séquentielle pour un paramètre réel du vecteur de régression et par conséquent nous nous limitons à une seule variable explicative d'un modèle de régression (cadre unidimensionnel) et dans une seconde partie nous généralisons les résultats au cas multidimensionnel. Un aspect qui nous a motivé est le nombre et le mode d'acquisition des données. Tout d'abord, l'acquisition des données recueillies par les chimistes du centre d'étude et de recherche sur le médicament de Normandie de l'université de Caen est réalisée au fur et à mesure dans le temps. Les méthodes de statistique séquentielles sont tout à fait adaptées à ce type de situation où l'objectif principal est la précision que l'on souhaite obtenir. La définition d'une *règle d'arrêt* nous indique si nous pouvons arrêter l'expérience après les n premières mesures $X_1, \ldots, X_n$ ou si nous devons continuer avec une nouvelle observation X_{n+1}. Nous appliquons ensuite cette méthode en chémoinformatique.

2 Méthode séquentielle

Nous nous attachons à mettre en œuvre la méthodologie précédente sur le modèle linéaire suivant :

$$\boldsymbol{Y}_n = \boldsymbol{X}_n \boldsymbol{\beta} + \boldsymbol{\varepsilon}_n \tag{1}$$

où pour tout $n \geq 1$, $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)'$ est le vecteur des observations, $\boldsymbol{X}_n$ est une matrice connue de dimension $n \times p$ ayant pour lignes $\boldsymbol{x}_i' \in \mathbb{R}^p$, $i = 1, \ldots, n$, $\boldsymbol{\varepsilon}_n = (\varepsilon_1, \ldots, \varepsilon_n)'$ est un vecteur d'erreurs indépendantes et identiquement réparties (i.i.d.), de fonction de répartition F inconnue et de médiane nulle ($F^{-1}(1/2) = 0$) et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ désigne le vecteur inconnu des paramètres de régression à estimer.

En 1978, Koenker et Basset ont proposé les quantiles de régression. On appelle θ-quantile de régression toute solution du problème de minimisation

$$\hat{\boldsymbol{\beta}}(\theta) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\theta \left(Y_i - \boldsymbol{x}_i' \boldsymbol{\beta} \right) \tag{2}$$

où $\theta \in [0,1]$ et $\rho_\theta(x) = x\left(\theta - \mathbb{I}(x < 0)\right)$ et $\mathbb{I}(\mathcal{P})$ prend la valeur 1 ou 0 selon que la condition $\mathcal{P}$ est vérifiée ou non. Un cas particulier de cette classe d'estimateurs (obtenu pour $\theta = 1/2$) est l'estimateur L^1 ou la régression médiane qui s'obtient par résolution du problème de minimisation (2). Dans Briollais et Durrieu (2014), une revue des quantiles de régression est donnée dans le domaine de la génomique. Nous donnons dans le théorème 1, le comportement asymptotique de l'estimateur $\hat{\boldsymbol{\beta}}(\theta)$ pour $\theta \in [0,1]$.

Théorème 1. *Sous les conditions de régularité adéquates, nous avons quand* $n \to \infty$ *:*

$$\sqrt{n}\left(\widehat{\boldsymbol{\beta}}(\theta) - \boldsymbol{\beta}\right) \xrightarrow{D} N_p\left(0, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right), \tag{3}$$

où

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \frac{\theta\left(1-\theta\right)}{f^2(Q(\theta))}\left(\boldsymbol{X'X}\right)^{-1}, \tag{4}$$

$q(\theta) = 1/f(Q(\theta))$ *est la densité du quantile et* $Q(\theta) = F^{-1}(\theta)$*.*

La variance asymptotique (4) dépendant de la densité de probabilité des erreurs (inconnue), nous avons besoin de "bons" estimateurs de la variance asymptotique. Il est possible pour estimer $q(\theta)$ de procéder par une estimation de type histogramme ou en utilisant un estimateur non paramétrique à noyau de la densité du quantile Dodge et Jurečková (1995) et Durrieu et Briollais (2009) qui est défini sous des conditions de régularité adéquates pour $0 < \theta < 1$ par

$$\widehat{Z}_n(\theta) = \frac{1}{\nu_n^2}\int_0^1 \widehat{\beta}_1^n(w)\, k\left(\frac{\theta - w}{\nu_n}\right) dw, \tag{5}$$

où ν_n désigne la taille de la fenêtre et $k(.)$ une fonction noyau. Il a été prouvé en particulier que quand $n \to \infty$:

$$\sqrt{n\nu_n}\left(\widehat{Z}_n(\theta) - q(\theta)\right) = O_p(1) \text{ et } \sqrt{n\nu_n}\left(\widehat{Z}_n(\theta) - q(\theta)\right) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, q^2(\theta)\,\overline{K}\right), \tag{6}$$

où $\overline{K} = \int K^2(x)\,dx$ avec $K(x) = \int_{-\infty}^{x} k(y)\,dy$.

Il est aussi possible de préciser le comportement asymptotique de cet estimateur dans le cas non identiquement distribué Koenker (2005). D'autres estimateurs ont aussi été proposés pour ce problème : l'estimateur de Huber-Eicker-White sandwich, l'estimateur de rang Koenker (2005) et Gutenbrunner et al. (1993), des méthodes boostraps Kocherginsky et al. (2005).

2.1 Cas univarié

Dans ce cas, on fixe une précision $d > 0$ et un seuil α dans l'intervalle $]0, 1[$. Nous présentons la méthode décrite par Durrieu et Briollais (2009). Notre objectif est de construire un intervalle de confiance I_n pour β_1, basé sur un estimateur robuste, telle que sa longueur L_n satisfait à

$$L_n \leq 2d, \tag{7}$$

et qui vérifie

$$P_F\left(\beta_1 \in I_n\right) \geq 1 - \alpha. \tag{8}$$

Bien entendu, avec de telles conditions, n ne peut être fixé *a priori* et nous sommes naturellement placés dans le cadre d'une procédure séquentielle. Stein (1945) a prouvé que, dans le cas où $\boldsymbol{X} = (X_1, \ldots, X_n)'$ est un vecteur gaussien, ce type d'intervalle peut

être construit par une procédure à deux pas. Plus tard, Chow et Robbins (1965) ont proposé une procédure séquentielle dans le cas d'une population avec variance finie. Pour le modèle linéaire, des méthodes similaires ont été construites par Ghosh et Sen (1972) en utilisant des statistiques de rang pour estimer le paramètre de régression et par Jurečková (1991) et Jurečková et Sen (1996). Nous décrivons ici la procédure de construction d'un intervalle de confiance d'un paramètre de régression dans un modèle linéaire lorsque la fonction de répartition des erreurs est supposée inconnue et un estimateur L^1 du coefficient de régression est choisi. Plus précisément, nous nous mettons dans le cas de la régression médiane.

Le schéma de construction repose essentiellement sur deux étapes :

— **Première étape** : nous déterminons deux estimateurs $\hat{\beta}_1$ et $\widehat{\Xi}_n$ tel que $I_n = \left[\hat{\beta}_1 - \widehat{\Xi}_n, \hat{\beta}_1 + \widehat{\Xi}_n\right]$ soit un intervalle de confiance de β_1 de coefficient de confiance $1-\alpha$. Cette construction se fait de manière classique en utilisant les propriétés des estimateurs (normalité asymptotique de $\hat{\beta}_1$ et consistance de $\widehat{\Xi}_n$).

— **Deuxième étape** : en nous plaçant maintenant dans le contexte de l'analyse séquentielle, nous étudions la variable d'arrêt N_d qui correspond au plus petit entier $n \geq n_0$ telle que la longueur de I_{N_d} est inférieure ou égale à $2d$. Nous discuterons l'introduction du paramètre n_0 et le choix de sa taille ultérieurement.

2.2 Cas multivarié

Nous considérons maintenant le vecteur des paramètres de régression $\boldsymbol{\beta} \in \mathbb{R}^p$, avec $p > 1$. D'après (1), nous obtenons :

$$\sqrt{n}\left(\hat{\beta}(\theta) - \beta\right) \xrightarrow[n\to+\infty]{D} \mathcal{N}_p\left(0, \boldsymbol{\Sigma}_\theta\right).$$

On en déduit que la forme quadratique

$$n\left(\hat{\beta}(\theta) - \beta\right)' \boldsymbol{\Sigma}_\theta^{-1}\left(\hat{\beta}(\theta) - \beta\right)$$

suit une loi du χ^2 à p degrés de liberté où $\hat{\beta}(\theta)$ désigne l'estimateur du quantile de régression de $\beta \in \mathbb{R}^p$ pour un échantillon de taille n.

La région de confiance de niveau de confiance $(1-\alpha)\%$ pour $\alpha \in [0,1]$ est donc :

$$RC_n(\alpha) = \left\{\beta \in \mathbb{R}^p : n\left(\hat{\beta}(\theta) - \beta\right)' \boldsymbol{\Sigma}_\theta^{-1}\left(\hat{\beta}(\theta) - \beta\right) \leq \chi^2_{p,1-\alpha}\right\}$$

où $\chi^2_{p,1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une loi du χ^2 à p degrés de liberté.

Il faut maintenant, comme dans le cas unidimensionnel, définir une variable d'arrêt. Pour cela, nous considérons l'ellipsoïde de confiance de niveau $1-\alpha$, $\alpha \in [0,1]$, défini par

$$n\left(\hat{\beta}(\theta)-\beta\right)' \mathbf{\Sigma}_\theta^{-1}\left(\hat{\beta}(\theta)-\beta\right) = \chi^2_{p,1-\alpha}. \tag{9}$$

Comme la matrice $\mathbf{\Sigma}_\theta$ est inconnue, nous l'estimons à partir des estimateurs de la densité du quantile $q(\theta)$ de type noyau et de ce fait la forme quadratique (9) suit une distribution T^2 de Hotelling de paramètres p et $n-1$, notée $T^2_{p,n-1}$ qui a la propriété d'être associée à la distribution de Fisher par

$$T^2_{p,n-1} = \frac{p(n-1)}{(n-p)} F_{p,n-p}.$$

Nous notons $\widehat{\mathbf{\Sigma}}_\theta$ l' estimateur de $\mathbf{\Sigma}_\theta$. Nous en déduisons :

$$\frac{n-p}{p(n-1)} n\left(\hat{\beta}(\theta)-\beta\right)' \widehat{\mathbf{\Sigma}}_\theta^{-1}\left(\hat{\beta}(\theta)-\beta\right)$$

suit une loi de Fisher de paramètres p et $(n-p)$ et par conséquent l'ellipsoïde de confiance de niveau $(1-\alpha)$, $\alpha \in [0,1]$ s'écrit :

$$RC_n(\alpha) = \left\{\beta \in \mathbb{R}^p : n\left(\hat{\beta}(\theta)-\beta\right)' \widehat{\mathbf{\Sigma}}_\theta^{-1}\left(\hat{\beta}(\theta)-\beta\right) \leq \frac{p(n-1)}{n(n-p)} F_{1-\alpha,p,n-p}\right\}$$

où $F_{1-\alpha,p,n-p}$ est le quantile d'ordre $(1-\alpha)$ d'une loi de Fisher à p et $(n-p)$ degrés de liberté.

La longueur du plus grand axe vaut :

$$\frac{2}{\sqrt{\dfrac{n(n-p)}{p(n-1)} \dfrac{\Phi_{\min}\left(\widehat{\mathbf{\Sigma}}_\theta^{-1}\right)}{F_{1-\alpha,p,n-p}}}},$$

où $\Phi_{\min}(.)$ désigne la fonction donnant la valeur propre minimale. Afin d'imposer que chaque composante du vecteur $\boldsymbol{\beta}$ se trouve dans un intervalle de confiance de longueur au plus $2d$, il faut que :

$$\frac{n(n-p)}{p(n-1)} \geq \frac{F_{1-\alpha,p,n-p}}{d^2\, \Phi_{\min}\left(\widehat{\mathbf{\Sigma}}_\theta^{-1}\right)}.$$

Ainsi, la variable d'arrêt N_d s'écrit :

$$N_d = \min\left\{n \geq n_0 : \frac{n(n-p)}{p(n-1)} \geq \frac{F_{1-\alpha,p,n-p}}{d^2\, \Phi^\star_{\min}\left(\widehat{\mathbf{\Sigma}}_\theta^{-1}\right)}\right\}, \tag{10}$$

où

$$\Phi^{\star}_{\min}\left(\widehat{\mathbf{\Sigma}}_{\theta}^{-1}\right) = \min\left\{\Phi^{\star}_{\min}\left(\widehat{\mathbf{\Sigma}}_{\theta}^{-1}\right), \varepsilon'_n\right\},$$

avec $(\varepsilon'_n)_{n\geq 1}$, une suite de nombres réels positifs tendant vers l'infini. Ainsi,

$$N_d = \min\left\{n \geq n_0 : \Phi^{\star}_{\min}\left(\widehat{\mathbf{\Sigma}}_{\theta}^{-1}\right) \geq \frac{F_{1-\alpha,p,n-p}}{d^2}\,\frac{p(n-1)}{n(n-p)}\right\}.$$

est bien une variable d'arrêt. Cette procédure séquentielle nous permettra en utilisant différents quantiles de déterminer avec un nombre minimum de composés si la narcose est non polaire ou polaire ou si une réactivité non spécifique est présente.

Bibliographie

Briollais L. and Durrieu G. (2014) Application of quantile regression to recent genetic and -omic studies, Human Genetics, 133, 951-966.

Chow Y. S. and Robbins H. (1965) On the asymptotic theory of fixed-width sequential confidence intervals for the mean, The Annals of Mathematical Statistics, 36(2), 457–462.

Dodge Y. and Jurečková J. (1995) Estimation of quantile density function based on regression quantiles, Statistics and Probability Letters, 23, 73-78.

Dodge Y. and Jurečková J. (1991) Flexible l-estimation in the linear model, Computational statistics and data analysis, 12(2), 211–220.

Durrieu G. and Briollais L. (2009) Sequential determination of sample size for robust linear regression : application to microarray experimental designs, Journal of American Statistical Association, 104, 650-660.

Gutenbrunner C.J., Jurečková J., Koenker R., Portnoy S. (1993) Tests of linear hypotheses based on regression rank scores, Journal of non Parametric Statistics, 2, 307–333.

Jurečková J. and Sen P.K. (1996) Robust statistical procedures : asymptotics and interrelations, John Wiley & Sons, New York.

Kocherginsky M., He X. and Mu Y. (2005) Practical confidence intervals for regression quantiles, Journal of Computational and Graphical Statistics, 14, 41–55.

Koenker R. (2005) Quantile Regression, *Econometric Society Monographs*, Cambridge University Press, New York.

Stein C. (1945) A two-sample test for a linear hypothesis whose power is independent of the variance, The Annals of Mathematical Statistics, 16(3), 243–258.