Méthodologie statistique

M 2013/01

La régression quantile en pratique

Pauline Givord - Xavier D'Haultfœuille

Document de travail

Institut National de la Statistique et des Études Économiques

INSTITUT NATIONAL DE LA STATISTIQUE ET DES ÉTUDES ÉCONOMIQUES

Série des documents de travail « Méthodologie Statistique »

de la Direction de la Méthodologie et de la Coordination Statistique et Internationale

M 2013/01

La régression quantile en pratique

Pauline Givord* - Xavier D'Haultfœuille**

Les auteurs remercient les très nombreuses personnes qui ont contribué par leurs commentaires sur les premières versions successives de ce document à l'améliorer significativement, et tout particulièrement Didier Blanchet pour ses suggestions précieuses, ainsi que Cédric Afsa, Pascale Breuil, Elise Coudin, Jean-Michel Floch, Bertrand Garbinti, Marine Guillerm, Jérôme Le, Simon Quantin, Olivier Sautory et les participants des JMS 2012 et du séminaire de méthodologie statistique de l'Insee. Les auteurs restent seuls responsables des erreurs et approximations qui pourraient y demeurer.

* DMCSI
18, bd Adolphe Pinard - 75675 PARIS CEDEX 14

** CREST
15, bd Gabriel Péri - 92245 MALAKOFF CEDEX

Direction de la méthodologie et de la coordination statistique et internationale -Département des Méthodes Statistiques - Timbre L101 - 18, bd Adolphe Pinard - 75675 PARIS CEDEX - France -
Tél. : 33 (1) 41 17 66 01 - Fax : 33 (1) 41 17 61 97 - CEDEX - E-mail : L001-dg@insee.fr - Site Web Insee : http://www.insee.fr

La régression quantile en pratique

Résumé

Les régressions quantiles sont des outils statistiques dont l'objet est de décrire l'impact de variables explicatives sur une variable d'intérêt. Elles permettent une description plus riche que les régressions linéaires classiques, puisqu'elles s'intéressent à l'ensemble de la distribution conditionnelle de la variable d'intérêt et non seulement à la moyenne de celle-ci. En outre, elles peuvent être plus adaptées pour certains types de données (variables censurées ou tronquées, présence de valeurs extrêmes, modèles non linéaires...). Ce document propose une introduction pratique à ces outils, en insistant sur les détails de leur implémentation pratique par les logiciels statistiques standards (Sas, R, Stata). Il peut également être utilisé comme un guide d'interprétation d'études mobilisant ces méthodes, en s'appuyant sur les deux applications concrètes exposées en détail. Enfin, il présente, pour un public plus averti, des extensions récentes traitant en particulier du traitement de l'endogénéité (variables instrumentales, données de panel...).

Classification JEL : C1, C3.

Mots clés : Régression quantile, Quantile Treatment Effect, Variable instrumentale, Régression quantile avec des données de panel.

Abstract

Quantile regressions are statistical tools that describe the impact of explanatory variables on a variable of interest. They provide a more detailed picture than classic linear regression, as they focus on the entire conditional distribution of the dependent variable, not only on its mean. They are also more suited to some kind of data such as truncated and censored dependent variable, outcomes with fat-tailed distributions, nonlinear models... This document proposes a practical introduction to these tools, with a special interest on their implementation in standard statistical software (Sas, R, Stata). We also present in details two empirical applications, to help people interpreting studies that rely on these methods. Finally, we propose for more advanced readers recent extensions in particular on endogeneity issues (instrumental variables, panel data...).

KeyWords : Quantile Regression, Quantile Treatment Effect, Instrumental Variable Quantile Regression, Quantile Regression with panel data.

Introduction

La majorité des études économiques empiriques se concentrent sur la modélisation de la moyenne. Celle-ci apporte une information essentielle mais néanmoins limitée. Le revenu moyen n'informe pas, par exemple, sur la répartition plus ou moins inégale de ces revenus dans la population. Par exemple, on a observé sur les dernières années une très faible augmentation du revenu moyen, alors que le revenu du dernier décile et surtout du dernier percentile a fortement augmenté (voir les travaux de Landais, 2007, et pour une actualisation Solard, 2010). L'une des préconisations du rapport Stiglitz-Sen-Fitoussi appelle donc à sortir de la « dictature de la moyenne » en présentant plus souvent des analyses sur la répartition des revenus. L'intérêt d'analyser l'ensemble de la distribution de la variable d'intérêt et de ses déterminants ne se limite pas à la mesure des inégalités. Pour évaluer l'effet d'une politique publique, il est souvent pertinent d'aller au-delà des effets moyens de celle-ci. Il peut ainsi être socialement souhaitable de mettre en œuvre une politique éducative qui permet de réduire la proportion d'élèves en grande difficulté, même si elle n'a qu'un effet négligeable sur le niveau moyen de l'ensemble des élèves.

Par ailleurs, dans certains cas la moyenne conditionnelle s'avère difficile à modéliser. Cela peut être le cas en présence de valeurs extrêmes ou aberrantes (dues par exemple à des erreurs de mesures), auxquelles la moyenne est bien plus sensible que les quantiles. Lorsque la distribution de la variable d'intérêt est très étalée (par exemple la distribution de revenus, dans laquelle des revenus très élevés peuvent parfois être observés), la moyenne pourra être très variable en fonction de l'échantillon utilisé. L'estimation de la moyenne est également compromise en présence de données censurées, c'est-à-dire lorsqu'on n'observe la variable d'intérêt qu'au-delà ou en deçà d'un seuil fixe. Par exemple, pour des raisons de confidentialité, les données individuelles de revenus sont parfois diffusées en écrêtant ceux qui sont supérieurs à un certain niveau. Il n'est pas possible d'estimer la moyenne ou la moyenne conditionnelle d'une variable censurée de la sorte, sauf à faire des hypothèses paramétriques sur la distribution de cette variable au-dessus du seuil. En revanche, en-dessous de ce niveau, les quantiles de la variable censurée coïncident avec ceux de la variable d'intérêt (cf. Buchinsky, 1994, pour une application à l'évolution des revenus aux Etats-Unis). La régression quantile est un outil dont dispose l'économètre pour répondre à ces limites inhérentes à la moyenne. Elle permet d'avoir une description plus précise de la distribution d'une variable d'intérêt conditionnelle à ses déterminants qu'une simple régression linéaire, qui se focalise sur la moyenne conditionnelle. Si son principe est ancien, elle a connu récemment un regain d'intérêt. On en trouvera un exemple récent d'utilisation sur données françaises dans Charnoz et al. (2011) qui étudient les déterminants des inégalités de salaires ou Cornec (2010) pour l'utilisation de ces outils pour la prévision de la conjoncture économique (voir également la note de conjoncture de l'Insee de décembre 2011). Un ensemble de procédures préprogrammées en font aujourd'hui un outil simple d'utilisation. L'objet de ce document est de préciser les principes de cette méthode et l'utilisation qui peut en être faite. Le lecteur intéressé pourra trouver une présentation plus détaillée en anglais dans Koenker (2005), ou d'autres introductions, également en anglais, dans Koenker & Hallock (2001) et Cade & Noon (2003).

Ce document est destiné à plusieurs usages. Les lecteurs souhaitant simplement disposer d'un guide d'interprétation pour les articles mettant en œuvre ce type de méthode tireront surtout profit, après la section 1 qui présente les principaux intérêts de la régression quantile, de la dernière partie qui détaille deux applications à partir d'exemples

concrets. Un lecteur néophyte souhaitant mettre en œuvre des régressions quantiles sous sa forme classique complétera cette lecture par la partie 2 qui rappelle le principe de l'estimation par régressions quantiles, les propriétés statistiques des estimateurs obtenus et les procédures disponibles dans les logiciels statistiques standards. Enfin, un lecteur plus averti trouvera dans la partie 3 plusieurs extensions utiles, dont en particulier la prise en compte de l'endogénéité ainsi que des régressions quantiles non linéaires. Les parties les plus avancées en termes économétriques et pouvant être sautées en première lecture sont suivies d'un astérisque.

1 L'intérêt des régressions quantiles

Pour poser les notations, nous nous intéressons à une variable aléatoire Y, de fonction de répartition F_Y ($F_Y(y) = P(Y \leq y)$). Rappelons que le quantile d'ordre τ est par définition $q_\tau(Y) = \inf\{y : F_Y(y) \geq \tau\}$. Si F_Y est continue, on a la définition usuellement utilisée $P(Y < q_\tau(Y)) = \tau$ [1]. Les quantiles les plus couramment utilisés sont la médiane ($\tau = 0,5$), les premier et dernier déciles ($\tau = 0,1$ et $\tau = 0,9$), et les premier et dernier quartiles ($\tau = 0,25$ et $\tau = 0,75$). Dans le langage courant, il y a parfois confusion entre la valeur du quantile et les personnes pour lesquelles la valeur de Y se situe en-dessous de ce quantile. Par exemple, on parle du « premier décile » pour désigner les 10% de la population les moins riches. Cette désignation est incorrecte. Le premier décile désigne stricto sensu le seuil de revenus en-dessous duquel 10% exactement de la population se situe.

1.1 Modéliser la dépendance de la distribution de la variable d'intérêt aux explicatives

Les régressions quantiles tentent d'évaluer comment les quantiles conditionnels $q_\tau(Y|X)$, définis par $q_\tau(Y|X) = \inf\left\{y : F_{Y|X}(y) \geq \tau\right\}$, se modifient lorsque les déterminants $X \in \mathbb{R}^p$ de la variable d'intérêt varient. Il n'y a pas de raison en effet de supposer que l'impact d'une de ces caractéristiques X_k soit le même aux différents quantiles de la distribution conditionnelle de Y. On peut en trouver une illustration dans les classiques courbes de croissance utilisées dans les carnets de santé. Elles montrent comment la distribution du poids, ou de la taille, varie en fonction de l'âge. Plus précisément, elles représentent certains percentiles (traditionnellement les $3^{\text{ème}}$, $25^{\text{ème}}$, $75^{\text{ème}}$ et $97^{\text{ème}}$) de ces distributions conditionnelles à l'âge (voir graphique 1). Elles permettent ainsi de vérifier que la croissance d'un enfant est « normale » en le situant dans la distribution correspondant à son âge. A partir de quinze mois, on constate que les différents percentiles des poids augmentent de manière à peu près linéaire avec l'âge. Mais les taux de croissance correspondants diffèrent suivant le percentile : les droites ne sont pas parallèles.

1. Cette égalité n'est cependant pas vraie dans le cas général, voir l'annexe ainsi que pour quelques propriétés des quantiles utilisées par la suite.

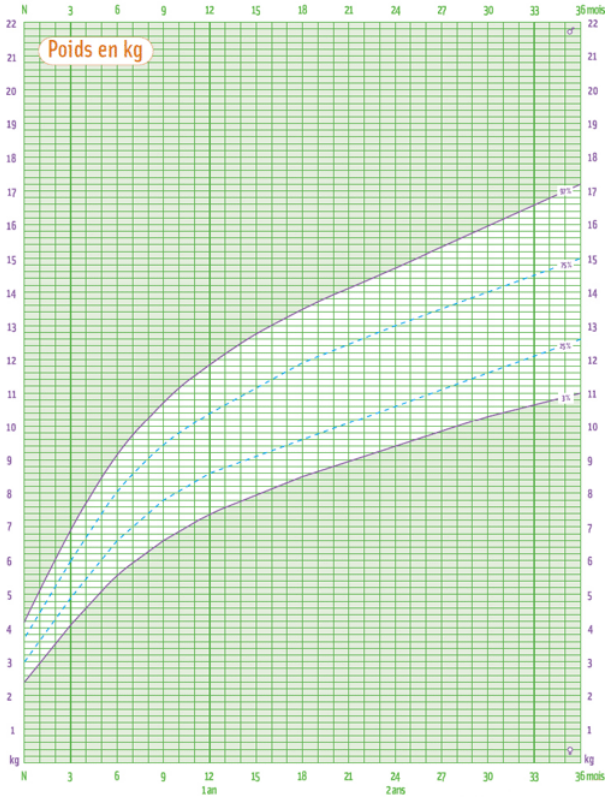


FIGURE 1 – Courbe de croissance du poids selon l'âge.

Ce type de modélisation graphique est possible et utile lorsque l'on s'intéresse à un seul déterminant, mais atteint vite ses limites pour étudier simultanément l'effet de plusieurs caractéristiques sur la variable d'intérêt. Les régressions quantiles permettent justement d'étudier ce cadre multivarié : plus précisément, elles tentent de déterminer comment les quantiles de la distribution conditionnelle $F_{Y|X}$ varient en fonction des variables explicatives X.

Dans la régression quantile standard, on suppose que ces quantiles de la distribution conditionnelle ont une forme linéaire :

$$q_\tau(Y|X) = X'\beta_\tau \tag{1.1}$$

où à chaque τ correspond donc un vecteur de coefficients $\beta_\tau = (\beta_{1\tau}, ..., \beta_{p\tau})'$ correspondant aux p variables explicatives (dont la constante) $X = (1, X_2, ...X_p)'$ [2]. Pour la suite, il peut être utile de remarquer que cette expression peut s'écrire de manière équivalente :

$$Y = X'\beta_\tau + \epsilon_\tau, \text{ avec } q_\tau(\epsilon_\tau|X) = 0 \tag{1.2}$$

La condition 1.1 est à rapprocher de celle effectuée dans la régression linéaire standard, dans laquelle on modélise la moyenne conditionnelle de la variable d'intérêt Y comme une expression linéaire des variables explicatives X : $E(Y|X) = X'\beta$. Une différence

2. Cette dépendance linéaire n'exclut pas une dépendance plus compliquée des quantiles par rapport à certaines variables explicatives : par exemple, dans l'exemple des courbes de croissance, on voit que les quantiles conditionnels à l'âge ne sont pas linéaires. En revanche, cette dépendance serait probablement bien approchée en utilisant non pas l'âge en niveau mais en logarithme, ou encore en utilisant une forme polynomiale de cette variable.

importante est qu'ici, on autorise les coefficients à différer d'un quantile à l'autre. Ceci apporte une information supplémentaire qui ne ressort pas d'une simple régression linéaire. Pour bien comprendre les implications de ce dernier point, considérons quelques exemples.

1.1.1 Le modèle de translation simple

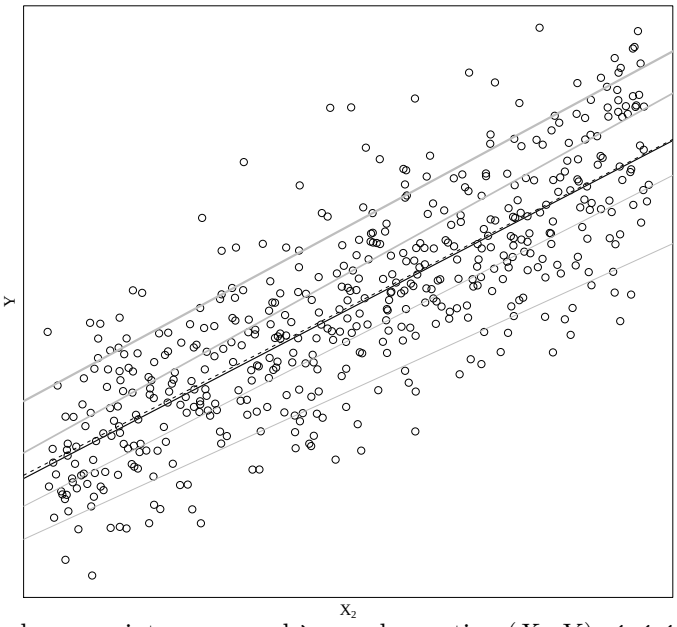
Le premier exemple suppose que les variables explicatives n'ont d'impact que sur la moyenne de la variable d'intérêt (et pas sur sa variance par exemple). Il s'agit du modèle de translation linéaire :

$$Y = X'\gamma + \varepsilon \tag{1.3}$$

où ε est indépendant de X et de moyenne nulle. Sous cette hypothèse, les résidus sont en particulier homoscédastiques (i.e., $V(\varepsilon|X) = \sigma^2$). Dans ce modèle, les distributions conditionnelles $F_{Y|X=x}$ sont parfaitement parallèles lorsque x varie : les différents quantiles conditionnels dépendent donc linéairement de X, $q_\tau(Y|X) = X'\gamma + q_\tau(\varepsilon)$. On est donc bien dans le cadre de l'hypothèse 1.1, mais ici seul le coefficient correspondant à la constante varie en fonction de τ : $\beta_{1,\tau} = \gamma_1 + q_\tau(\varepsilon)$, tandis que $\beta_{k,\tau} = \gamma_k$ pour $k > 1$.

On en trouvera une illustration dans le graphique 2, dans le cas simple d'une régression univariée où il on ne s'intéresse qu'à une seule variable explicative X_2 en sus de la constante. Dans ce cas, les droites correspondant aux régressions quantiles sont des lignes parallèles : elles ont toutes comme pente β_2 correspondant à la variable X_2, on parle donc d'*homogénéité des pentes*.

Cela signifie également que les coefficients $(\beta_{k,\tau})_{k=2,\dots,p}$ sont les mêmes que ceux correspondant à la modélisation de la moyenne conditionnelle $E(Y|X) = X'\gamma$ pour tout τ. Les résultats obtenus par régression quantile ou par une régression linéaire estiment donc les mêmes paramètres, par des méthodes différentes. Il peut parfois être intéressant d'utiliser les estimateurs des régressions quantiles plutôt que ceux des moindres carrés ordinaires. Ils sont plus robustes par exemple à la présence de valeurs aberrantes, et peuvent être plus précis pour certaines distributions de ε (cf. section 2.3 ci-dessous). Enfin, comme on ignore en général que le vrai modèle est un modèle de translation, ils peuvent être utilisés pour tester cette restriction, qui implique que les estimateurs de régressions quantiles effectuées pour différents τ doivent être très proches.



Lecture : chaque point correspond à une observation (X_2, Y) généré par le processus $Y = \beta_1 + \beta_2 X_2 + \epsilon$, avec $\epsilon \hookrightarrow \mathcal{N}(0, \sigma)$. Les droites correspondent aux droites de régressions quantiles pour les déciles d'ordre 1, 3, 7 et 9 (en gris), la médiane (en noir) et à la droite de régression linéaire classique (en noir pointillé).

FIGURE 2 – Exemple de données distribuées selon un modèle de translation

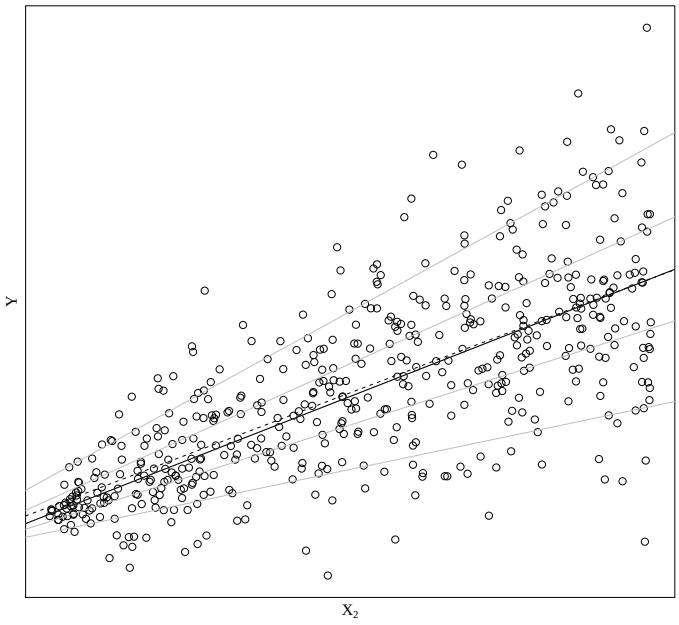
1.1.2 Le modèle de translation-échelle

Le deuxième exemple, un peu plus général que le modèle de translation, suppose que ces déterminants ont non seulement un impact sur la moyenne mais aussi sur la variance de la variable d'intérêt. Ces modèles, appelés « translation-échelle » , correspondent à une certaine forme d'hétéroscédasticité :

$$Y = X'\gamma + (X'\theta)\varepsilon \tag{1.4}$$

avec encore une fois ε indépendant de X, de moyenne nulle et $X'\theta > 0$. Dans un tel modèle, la dispersion de la variable dépendante conditionnelle à X est plus importante pour certaines valeurs de X. Un exemple classique est celui des salaires, qui sont plus dispersés pour les diplômés du supérieur que pour les personnes sans diplôme (cf. l'exemple de la partie 4.1). Le modèle de translation-échelle correspondant à l'équation (1.4) implique que $q_\tau(Y|X) = X'(\gamma + q_\tau(\varepsilon)\theta)$. Ainsi, l'hypothèse (1.1) est bien vérifiée, avec $\beta_\tau = \gamma + q_\tau(\varepsilon)\theta$. L'impact des variables explicatives ne sera pas le même pour les différents quantiles, et il n'y a plus homogénéité des pentes. Dans l'exemple des salaires, l'effet du diplôme est faible pour les premiers quantiles ($\beta_{k,\tau}$ petit pour τ proche de 0) et plus fort pour les derniers quantiles ($\beta_{k,\tau}$ grand pour τ proche de 1).

Le modèle de translation-échelle est illustré par le graphique 3, dans le cas univarié. Ici les pentes $\beta_{2,\tau}$ correspondant aux différentes régressions quantiles sont croissantes avec τ (soit $\alpha_2 > 0$), ce qui traduit une dispersion d'autant plus grande que X_2 est élevé (comme dans l'exemple du diplôme). Cette information ne ressort pas d'une régression linéaire standard, qui se contente d'estimer le coefficient γ : on a en effet toujours $E(Y|X) = X'\gamma$.



Lecture : chaque point correspond à une observation (X_2, Y) généré par le processus $Y = \beta_1 + \beta_2 X_2 + \theta X_2 \epsilon$, avec $\epsilon \hookrightarrow \mathcal{N}(0, \sigma)$. Les droites correspondent aux droites de régressions quantiles pour les déciles d'ordre 1, 3, 7 et 9 (en gris), la médiane (en noir) et à la droite de régression linéaire classique (en noir pointillé).

FIGURE 3 – Exemple de données distribuées selon un modèle de translation échelle

*1.1.3 Un cadre général : le modèle à coefficients aléatoires**

Le modèle à coefficients aléatoires généralise ces modèles. Il s'écrit :

$$Y = X'\beta_U, \quad U \text{ indépendant de } X \text{ et de loi uniforme sur } [0,1], \tag{1.5}$$

où la fonction $u \mapsto x'\beta_u$ est strictement croissante pour tout x [3]. Dans ce modèle, U peut s'interpréter comme une composante individuelle inobservée qui positionne l'individu dans la distribution de Y. Par exemple si on veut modéliser le niveau de salaire, U correspondrait à une productivité « intrinsèque » du salarié, qui implique en particulier que le niveau d'études ou d'autres caractéristiques ont des effets qui varient d'une personne à l'autre. Ce modèle à coefficients aléatoires repose sur des hypothèses très flexibles sur la dépendance en U, qui peut être non linéaire. Il généralise les deux exemples précédents. Le modèle de translation linéaire correspond à un cas où le coefficient correspondant à la variable explicative $k > 1$, $\beta_{k,U}$, est indépendant de U. Dans le modèle de translation-échelle, on a $\beta_U = \gamma + q_U(\varepsilon)\theta$.

Ce modèle à coefficients aléatoires fournit une interprétation intéressante du coefficient β_τ dans (1.1), à condition de faire une hypothèse supplémentaire. Il implique en effet que si

3. Ce modèle vérifie bien l'hypothèse (1.1) puisque par indépendance de U et X et croissance de $u \mapsto x'\beta_u$,

$$P(Y \leq x'\beta_\tau | X = x) = P(x'\beta_U \leq x'\beta_\tau) = P(U \leq \tau) = \tau.$$

l'on modifie marginalement la variable observable X indépendamment de l'effet individuel U, l'effet sur la variable d'intérêt Y est égal à β_U. Ainsi, β_τ correspond à l'effet marginal de X pour les individus au $\tau^{\text{ième}}$quantile de la distribution des caractéristiques inobservées U. Si l'on s'intéresse par exemple à la modélisation des salaires en fonction du niveau d'études, la composante de $\beta_{0.5}$ correspondant au nombre d'années d'études mesurera l'effet d'une petite augmentation de celui-ci pour les salariés dont la productivité intrinsèque U se situe à un niveau médian. Cette interprétation n'est cependant correcte que lorsque cette productivité intrinsèque est la même quel que soit le niveau d'études. Autrement dit, il faut supposer que les salariés « médians » en termes de productivité individuelle pour un niveau d'études particulier le sont encore pour un autre niveau d'études. Cette condition, appelée l'hypothèse d'*invariance de rang* dans la littérature, est forte.

1.1.4 Interprétation des régressions quantiles

La manière dont les distributions conditionnelles se modifient en fonction des variables explicatives renvoie à plusieurs questions. La première est simplement de décrire comment les quantiles conditionnels se modifient en fonction de ces déterminants, sans tenter de savoir si ce sont des personnes « comparables » aux différents quantiles conditionnels. Les régressions quantiles sont des outils développés pour répondre à cette question. A condition de faire l'hypothèse simple mais restrictive d'invariance des rangs, elles peuvent permettre de répondre à une deuxième question, un peu plus précise, qui est de déterminer quelle est la variation de la variable d'intérêt correspondant à une variation marginale d'un de ces déterminants, pour les personnes qui se trouvent à un certain niveau de la distribution conditionnelle de la variable d'intérêt Y. En général, les régressions quantiles ne donnent pas d'éléments pour répondre à une question encore différente, qui est d'estimer la distribution des effets de ce déterminant X sur la variable d'intérêt Y.

Pour bien comprendre que les réponses à ces trois questions peuvent être différentes, il peut être utile de recourir à un exemple très simple. Supposons qu'on s'intéresse à l'effet d'une variable binaire X sur une variable Y. La « population » à laquelle on s'intéresse est composée de cinq types en proportion identique, qu'on identifiera par les lettres de A à E. Suivant la valeur de X, ces personnes peuvent avoir des valeurs différentes de Y. Lorsque $X = 0$, les personnes ont, suivant leur type, $Y^A = 1$, $Y^B = 2$, $Y^C = 4$, $Y^D = 5$, ou $Y^E = 9$. Lorsque $X = 1$, la variable d'intérêt prend les valeurs $Y^A = 4$, $Y^B = 6$, $Y^C = 5$, $Y^D = 11$ ou $Y^E = 10$, suivant les types. Les effets individuels correspondant à un passage de $X = 0$ à $X = 1$ sont donc $\Delta Y^A = 3$, $\Delta Y^B = 4$, $\Delta Y^C = 1$, $\Delta Y^D = 6$ et $\Delta Y^E = 1$. Une régression quantile d'ordre 0.5 (*regression médiane*) de Y sur X mesure l'écart entre la médiane de la distribution de Y conditionnelle à $X = 0$ et la médiane de la distribution conditionnelle à $X = 1$. Il vaut donc ici 2 (6- 4). Cette valeur est différente de l'effet pour les individus médians quand $X = 0$ (i.e., tels que $Y = q_{0.5}(Y|X = 0)$) de passer de $X = 0$ à $X = 1$: ces individus sont ceux de types C, pour lesquels on a $\Delta Y^C = 1$.

Cette différence vient du fait qu'ici, les individus ne sont pas ordonnés (en termes de Y) de la même manière lorsque $X = 0$ et $X = 1$. Les individus de type C sont « devant » ceux de type B lorsque $X = 0$ mais derrière eux lorsque $X = 1$. L'hypothèse d'invariance des rangs n'est donc pas vérifiée, et on ne peut pas interpréter le coefficient de la régression quantile d'ordre 0.5 comme l'effet d'un passage de $X = 0$ à $X = 1$ pour les individus médians quand $X = 0$. Enfin, ces deux valeurs sont encore différentes de la médiane de la distribution des effets individuels ΔY, qui vaut 3. Ceci est lié à la non linéarité

des quantiles. La médiane de la distribution de ΔY ne correspond pas à la différence des médianes des distributions de la variable d'intérêt, ou en termes mathématiques, $q_{0.5}(\Delta Y) \neq q_{0.5}(Y|X=1) - q_{0.5}(Y|X=1)$.

Ces remarques permettent de bien cadrer les usages qui peuvent être faits, ou non, des résultats d'une régression quantile. Il est utile de faire le lien avec ceux obtenus par une régression linéaire classique. L'objet principal de celle-ci est de modéliser la manière dont la moyenne conditionnelle varie en fonction de déterminants : sur notre exemple, cela correspondrait à 3 ($= 7,2 - 4,2$). Du fait de la linéarité de la moyenne, la différence des moyennes conditionnelles correspond également à l'effet moyen de l'augmentation de X, c'est-à-dire à la moyenne de ΔY . En revanche, cette différence ne correspond pas à ce que gagnerait à passer de $X = 0$ à 1 une personne dont la valeur de Y est proche de la moyenne conditionnelle lorsque $X = 0$ (dans notre exemple, il s'agit encore de C). Contrairement aux régressions quantiles, la condition d'invariance des rangs ne suffit pas ici pour obtenir une telle interprétation.

1.2 Des estimateurs plus adaptés à certaines situations

Même dans le cadre du modèle de translation simple (1.3), dans lequel les coefficients correspondant à une régression linéaire et à une régression quantile sont les mêmes, il peut être préférable d'utiliser cette dernière. La première raison est qu'elle est robuste aux valeurs aberrantes ou à des erreurs très dispersées. Intuitivement, cette propriété est due au fait que les quantiles sont moins sensibles que la moyenne à la présence de valeurs très grandes. Considérons tout d'abord le cas des valeurs aberrantes. Supposons que la variable Y^* vérifie le modèle de translation simple (1.3), mais que dans de très rares cas, les données observées ne correspondent pas à la variable d'intérêt Y^* mais à une variable erronée, éventuellement corrélée avec les variables explicatives X. Formellement, on observe $Y = AX'\delta + (1 - A)Y^*$, où A est une variable inobservée valant 1 lorsque Y est aberrant, 0 sinon, avec $P(A = 1|X, \varepsilon) = p$ petit. Une régression linéaire de la variable observée (avec erreur) Y sur X donnera une estimation biaisée de notre paramètre d'intérêt γ, puisqu'elle sera égale (pour un échantillon de taille tendant vers l'infini) à $\gamma_{MCO} = \gamma + p(\delta - \gamma)$ [4]. Si δ est très différent de γ, le terme de biais peut être important même lorsque la probabilité d'observer des observations erronées p est faible. En revanche, on peut montrer que si $X'\delta$ est très grand, l'estimateur de l'effet de X_k ($k > 1$) obtenu par une régression quantile vaut bien γ_k [5]. En d'autres termes, la présence de valeurs aberrantes n'affecte pas les résultats de la régression quantile, sauf les coefficients de la constante.

Dans un même ordre d'idée, toujours dans le cas du modèle de translation simple, les résultats obtenus par régression quantile seront plus précis en général lorsque les résidus sont très dispersés. Un exemple extrême est celui où ε n'a pas d'espérance, ce qui se produit lorsque ε peut prendre des valeurs très grandes avec une probabilité importante. Cette situation n'est pas si rare en pratique. Les lois de Cauchy et certaines lois de Pareto (utilisées pour modéliser les hauts salaires ou les patrimoines), par exemple, n'ont pas

4. On néglige ici les erreurs d'estimation liées au fait que l'échantillon est de taille finie. Autrement dit, nos résultats doivent se comprendre comme étant les valeurs limites des estimateurs lorsque la taille de l'échantillon tend vers l'infini.

5. Ceci est lié au fait que le vrai modèle est supposé être un modèle de translation simple. Dans le cas plus général où Y^* vérifierait (1.1) pour tout τ , on obtiendrait $\beta_{\frac{\tau}{1-p}}$.

d'espérance. Dans ce cas, l'estimateur des moindres carrés ordinaires n'est pas convergent : même pour des échantillons énormes, il pourra prendre des valeurs très différentes du vrai paramètre β. A l'inverse, l'estimateur obtenu par régression quantile sera convergent.

Par ailleurs, une propriété importante des quantiles est qu'ils sont invariants par une transformation monotone : si g est une fonction croissante continue à gauche, on a $q_\tau(g(Y)) = g(q_\tau(Y))$ (voir l'annexe pour une preuve). Cette propriété n'est bien sûr pas vérifiée par l'espérance. Ceci rend les régressions quantiles plus naturelles et simples à utiliser dans des modèles non-linéaires [6] comme les modèles à censure fixe (cf. la partie 3.4 pour davantage de détails), ou les modèles de durées (voir par exemple Bilias & Koenker, Fitzenberger & Wilke, 2001, 2005). On peut trouver aussi des applications aux modèles de comptage (Machado & Silva, 2005).

2 Principes statistiques et mise en œuvre pratique

2.1 Définition de l'estimateur et propriétés statistiques*

Pour bien comprendre le principe des régressions quantiles, il est utile de détailler comment on peut estimer les quantiles d'une variable d'intérêt Y à partir d'un échantillon $(Y_i)_{i=1...n}$ de variables supposées i.i.d. La manière la plus intuitive de calculer l'estimateur standard $\widehat{q}_\tau(Y)$ consiste à ordonner ces n variables, le quantile d'ordre τ étant fourni par la $[n\tau^{ieme}]$ observation où $[n\tau]$ est le plus petit entier supérieur ou égal à $n\tau$. Mais il est plus utile, pour le passage aux régressions quantiles, de remarquer qu'on a également [7] (voir l'annexe pour une preuve) :

$$\widehat{q}_\tau(Y) = \arg\min_b \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - b). \tag{2.1}$$

où $\rho_\tau(.)$ est une « fonction test » définie par $\rho_\tau(u) = (\tau - \mathbb{1}\{u < 0\})\, u$. Par exemple, pour $\tau = 1/2$, c'est-à-dire si l'on s'intéresse à la médiane, la fonction test correspond simplement à la (demi-) valeur absolue. La solution du programme de minimisation ci-dessus correspond alors bien à la médiane.

L'intérêt de cette définition est de s'étendre simplement au cadre conditionnel qui nous intéresse, où l'on modélise le quantile conditionnel de la variable d'intérêt Y comme une fonction des variables explicatives X. Il suffit en effet de remplacer $\widehat{q}_\tau(Y)$ et b dans (2.1) par respectivement $q_\tau(Y|X)$ et une fonction $b(X)$. Dans le cas des régressions quantiles classiques, on peut se limiter aux fonctions linéaires puisqu'on suppose que $q_\tau(Y|X) = X'\beta_\tau$. On a alors :

$$\beta_\tau = \arg\min_\beta E[\rho_\tau(Y - X'\beta)]. \tag{2.2}$$

On peut noter l'analogie avec le modèle de régression linéaire classique, qui modélise l'espérance conditionnelle de Y par une forme linéaire en X : $E(Y|X) = X'\beta_0$. L'espérance

6. Cette propriété signifie aussi qu'on pourra facilement déduire l'effet marginal d'un déterminant du salaire (par exemple) sur un de ses quantiles conditionnels à partir d'une régression quantile modélisant le log du salaire. On aura en effet $q_\tau(W|X) = \exp(X\beta_\tau)$ où β_τ correspond au coefficient estimé par la régression quantile de $\log(W)$ sur X.

7. En toute rigueur, il n'y a pas toujours unicité au programme de minimisation $\min_a E[\rho_\tau(Y - a)]$, cf. l'annexe pour une discussion. On néglige ici ces complications.

d'une variable aléatoire pouvant être obtenue par $E(Y) = \arg\min_a E[(Y-a)^2]$, le coefficient β_0 est défini par $\beta_0 = \arg\min_\beta E[(Y - X'\beta)^2]$. La fonction de perte quadratique qui est utilisée dans une régression linéaire par les moindres carrés ordinaires est donc remplacée, dans la régression quantile, par la fonction test $\rho_\tau(.)$. Celle-ci augmentant de manière linéaire et non quadratique avec le résidu, les très grands écarts sont beaucoup moins pénalisés, ce qui explique la robustesse de la régression quantile aux valeurs extrêmes ou aberrantes.

Cette estimation peut se faire pour tout quantile d'ordre τ, où $\tau \in [0,1]$. Il existe donc en principe une infinité de régressions quantiles possibles. En pratique, le nombre de quantiles qu'on estime dépendra de la taille de l'échantillon. Il est bien entendu illusoire de tenter d'approcher très finement une distribution avec un nombre fini d'observations : le nombre de quantiles empiriques distincts sera probablement restreint [8] . Le choix de modéliser l'ensemble des percentiles ou simplement les quartiles et la médiane dépendra non seulement du degré de précision souhaitée pour décrire la distribution mais aussi des données disponibles.

2.2 Algorithmes utilisés

Il n'existe pas de solution explicite à (2.1), si bien qu'il faut résoudre ce programme numériquement. Un problème est que la fonction objectif n'est ni différentiable (la fonction ρ_τ n'est pas dérivable en 0) ni strictement convexe. Les algorithmes standards tels que celui de Newton Raphson ne peuvent donc pas être utilisés directement. Cependant, on peut reformuler (2.1) comme un programme linéaire :

$$\min_{(\beta,u,v)\in\mathbb{R}^p\times\mathbb{R}^{2n}_+} \tau\mathbf{1}'u + (1-\tau)\mathbf{1}'v \quad s.t.\ \mathbf{X}\beta + u - v - \mathbf{Y} = 0,$$

où $\mathbf{X} = (X_1, ..., X_n)'$, $\mathbf{Y} = (Y_1, ..., Y_n)'$ et $\mathbf{1}$ est un vecteur de 1 de taille n. La méthode du simplexe a été jusqu'à récemment la méthode la plus classique pour résoudre ce type de problèmes linéaires. Cependant, elle devient coûteuse en temps de calcul lorsque le nombre d'observations augmente et elle n'est donc indiquée que pour de petits échantillons. Pour des échantillons plus conséquents, les méthodes de points intérieurs sont plus performantes pour résoudre ces programmes linéaires (Portnoy & Koenker, 1997).

2.3 Propriétés asymptotiques de l'estimateur et estimation de la précision*

Les propriétés asymptotiques de $\widehat{\beta}_\tau$ sont délicates à établir car, contrairement à l'estimateur des moindres carrés, il n'existe pas de forme explicite pour $\widehat{\beta}_\tau$. Pour plus de détails, on se référera par exemple à l'ouvrage de Koenker (2005). Nous nous contentons ici du résultat principal sur la loi asymptotique de $\widehat{\beta}_\tau$.

Théorème 2.1. *Supposons que $\varepsilon_\tau = Y - X'\beta_\tau$ admette, conditionnellement à X, une densité en 0 $f_{\varepsilon_\tau|X}(0|X)$ et que $J_\tau = E\left[f_{\varepsilon_\tau|X}(0|X)XX'\right]$ soit inversible. Alors*

$$\sqrt{n}\left(\widehat{\beta}_\tau - \beta_\tau\right) \xrightarrow{d} \mathcal{N}\left(0, \tau(1-\tau)J_\tau^{-1}E\left[XX'\right]J_\tau^{-1}\right) \tag{2.3}$$

8. On peut montrer que pour un échantillon de taille n, le nombre de quantiles pour lesquels les coefficients estimés par une régression quantiles seront différents est quasi linéaire en n (il est plus précisément en $O(nln(n))$).

Dans le cas du modèle de translation simple (1.3), la variance asymptotique prend une forme particulièrement simple. On a en effet, $\varepsilon_\tau = \varepsilon - q_\tau(\varepsilon)$ et la variance asymptotique V_{as} s'écrit plus simplement

$$V_{\text{as}} = \frac{\tau(1-\tau)}{f_\varepsilon(q_\tau(\varepsilon))^2} E\left[XX'\right]^{-1}.$$

Cette variance est très proche de celle des MCO avec résidus homoscédastiques, si ce n'est que $\sigma^2 = V(\varepsilon)$ est remplacé par $\tau(1-\tau)/f_\varepsilon(q_\tau(\varepsilon))^2$. Le terme de densité est logique : seuls les résidus autour de $q_\tau(\varepsilon)$ vont apporter de l'information sur la valeur du quantile conditionnel de Y. Ce résultat explique que même dans le cas de translation simple, il peut être parfois préférable d'utiliser une régression quantile pour certaines distributions des termes inobservés ε. L'estimation par régression quantile sera plus précise qu'une estimation par MCO lorsque $\tau(1-\tau)/f_\varepsilon(q_\tau(\varepsilon))^2 < \sigma^2$.

En dehors du modèle restrictif de translation simple, la variance asymptotique est plus complexe à estimer que dans le cadre d'un modèle de régression linéaire simple. Plusieurs méthodes d'inférence ont été proposées pour construire des tests ou des intervalles de confiance sur β_τ, et il n'existe pas à l'heure actuelle de consensus sur la méthode à utiliser. On trouvera dans Kocherginsky et al. (2005) une présentation générale de ces méthodes et une discussion pratique des cas où certaines sont plus ou moins indiquées. Le choix dépend des hypothèses plus ou moins restrictives qu'on accepte de faire sur le modèle sous-jacent (modèle de translation...), de la taille de l'échantillon ou du nombre de variables du modèle.

Certaines méthode s'appuient sur une estimation directe de la variance asymptotique en partant de la formule 2.3. La difficulté principale de cette approche est la présence de la densité conditionnelle $f_{\varepsilon_\tau|X}(0|X)$, qui est délicate à estimer (cf. l'annexe A.2.1 pour plus de détails). Dans le cadre restrictif d'un modèle de translation-échelle, une méthode basée sur les tests de rang est parfois utilisée (cf. Koenker, 2005). Il est surtout courant de s'appuyer sur des méthodes de bootstrap. Elles consistent à générer des échantillons« factices » par des tirages avec remise à partir de l'échantillon initial, et à effectuer une régression quantile sur ces échantillons (cf. l'annexe A.2.2). L'inconvénient de ces méthodes est qu'elles sont aussi coûteuses en temps de calcul. Ce dernier augmente à la fois avec la taille de l'échantillon et le nombre de variables explicatives. Une solution récente (« Markov Chain Marginal Bootstrap » , ou MCMB) a été proposée par He & Hu (2002) pour résoudre en partie ce problème quand le nombre de variables explicatives est important. Ces méthodes ne sont pas toujours performantes sur de petits échantillons.

Enfin, l'un des intérêts de la régression quantile étant de ne pas supposer *a priori* que les variables explicatives ont un effet homogène sur l'ensemble de la distribution de la variable d'intérêt, il est tout à fait possible de tester cette hypothèse à partir des estimations obtenues. Par exemple, l'homogénéité de l'effet de l'une des variables X_k correspond à l'égalité des coefficients $\beta_{k,\tau_1}, ..., \beta_{k,\tau_m}$ (où $(\tau_1, ...\tau_m)$ peuvent être par exemple l'ensemble des déciles), ce qui peut se tester simplement. Un tel test s'appuie sur la distribution jointe asymptotique de $(\widehat{\beta}_{\tau_1}, ..., \widehat{\beta}_{\tau_m})$, donnée par le résultat suivant :

$$\sqrt{n}\left(\widehat{\beta}_{\tau_k} - \beta_{\tau_k}\right)_{k=1}^{m} \xrightarrow{d} \mathcal{N}\left(0, V\right), \tag{2.4}$$

où V est une matrice par bloc dont le bloc $V_{k,l}$ vérifie

$$V_{k,l} = \left[\tau_k \wedge \tau_l - \tau_k\tau_l\right] J_{\tau_k}^{-1} E\left[XX'\right] J_{\tau_l}^{-1}.$$

Ce résultat est une généralisation du théorème 2.1 à plusieurs quantiles.

2.4 La régression quantile dans les logiciels

2.4.1 SAS

On utilise la `proc quantreg`, disponible à partir de la version 9.1 et dont la syntaxe est la suivante [9] :

```
proc quantreg data=(table) algorithm=(choix de l'algorithme) ci= (méthode de
    calcul des intervalles de confiance);
    class (variables qualitatives);
    model (y) = (x) /quantile = (liste des quantiles);
run;
```

Par défaut, le calcul de $\widehat{\beta}_\tau$ est effectué par l'algorithme du simplexe. Pour éviter des temps de calcul trop longs, il est recommandé d'utiliser plutôt, dès que $n \geq 1000$, une méthode de point intérieur via l'option `algorithm=interior`. Pour calculer des intervalles de confiance, il est préférable d'utiliser l'option `ci=resampling` qui utilise le bootstrap MCMB cité précédemment. Le nombre de réplications peut être indiqué en ajoutant `(NREP=)` (il est de 200 par défaut). Si l'option `ci` n'est pas utilisée, SAS utilise la méthode de MCMB lorsque $n \geq 5000$ ou $p \geq 20$, et l'inversion des tests de rang sinon. Cette dernière méthode s'appuie cependant sur l'hypothèse restrictive que les données vérifient un modèle de translation-échelle. L'inconvénient du bootstrap est qu'il peut être coûteux en temps. L'option `ci=sparsity` permet alors de gagner du temps, mais la variance asymptotique est alors estimée sous l'hypothèse très restrictive que le vrai modèle est un modèle de translation défini par (1.3). On peut écrire l'ensemble des quantiles souhaités, ou de manière plus condensée par exemple pour l'ensemble des percentiles : `quantile= 0.01 to 0.99 by 0.01` [10]. Il est possible de faire des tests de contrainte linéaire en utilisant l'option `test`.

2.4.2 Stata

Plusieurs commandes sont disponibles sous Stata. La commande `sqreg` permet de faire des régressions sur plusieurs quantiles simultanément. Sa syntaxe est :

```
sqreg y x, quantiles(choix des quantiles) reps(nombre de réplication du bootstrap)
```

D'autres commandes sont également disponibles, par exemple `qreg` et `bsqreg`, mais elles ne sont pas conseillées. D'une part, elles ne permettent pas de faire des régressions quantiles sur plusieurs quantiles. D'autre part, il n'est pas possible d'estimer via `qreg` la précision des estimateurs par bootstrap. Dans cette procédure, l'estimation des écarts-types est fondée sur l'hypothèse restrictive d'un modèle de translation.

2.4.3 R

Un package R très complet a été développé par R. Koenker : quantreg. La syntaxe principale s'écrit :

9. On trouvera une description plus précise de l'utilisation de cette procédure par Colin Chen "An Introduction to Quantile Regression and the QUANTREG Procedure", disponible à l'adresse : `http://www2.sas.com/proceedings/sugi30/213-30.pdf`

10. L'option `quantile= ALL` permet d'estimer des régressions quantiles pour toutes les valeurs de τ comprises entre $[0, 1]$ telles que les estimations de β_τ sont différentes. Cette estimation n'est possible qu'en utilisant l'algorithme du simplex (`algorithm= simplex`. Voir par exemple Koenker & D'Orey (1987) pour une description d'un algorithme possible.

```
library(quantreg)
rq(y ~ x1 + x2, tau = (vecteur de quantiles),  data=(table),
    method=("br" ou "fn"))
```

La méthode "br" correspond au simplexe (par défaut), tandis que "fn" sélectionne une méthode de point intérieur. Par défaut, R n'indique que les paramètres estimés de la régression quantile [11]. Pour obtenir également les écarts-types, statistiques de test ou intervalles de confiance correspondants, il faut utiliser la commande suivante :

```
fit1 <- rq(y ~ x1 + x2, tau = (vecteur de quantiles),  data=(table),
    method=("br" ou "fn"))
summary(fit1, se="iid" ou "nid" ou "ker" ou "boot")
```

Par défaut, si l'on ne précise pas l'option `se`, R fournit simplement des intervalles de confiance, par inversion de tests de rang. L'option `se` permet d'obtenir des écarts-types et statistiques de test. Les options `iid`, `nid` et `ker` sont des variantes de la méthode directe (cf. annexe). L'option `iid` s'appuyant sur l'hypothèse restrictive que le vrai modèle est un modèle de translation, les options `nid` ou `ker` sont préférables. L'option `ker` s'appuie sur une estimation de la matrice J_τ proche de $\widehat{J}_\tau$ définie par (A.1) en annexe A.2.1, l'indicatrice étant simplement remplacée par un noyau. Enfin, l'option `boot` estime les écarts-types par bootstrap. Plusieurs méthodes de bootstrap sont proposées. On se référera à l'aide R ou à un tutorial disponible sur la page web de Roger Koenker pour plus de détails [12].

3 Extensions*

3.1 Les régressions quantiles instrumentales

Comme en régression linéaire, il arrive fréquemment que certaines composantes des variables X soient a priori endogènes. Par exemple, dans une étude sur l'impact d'un dispositif de formation sur le salaire, le fait de participer à ce dispositif peut être lié à des caractéristiques inobservées qui influent également le salaire. Dans ce cas, l'estimateur $\widehat{\beta}_\tau$ défini par (2.1) ne mesure pas l'effet causal du dispositif de formation.

En revanche, on peut disposer d'instruments affectant ces variables mais pas directement les composantes inobservées de la variable d'intérêt (représentées par le résidu ε_τ). Plus précisément, si l'on se place dans le cadre de la régression quantile précédente,

$$Y = X'\beta_\tau + \varepsilon_\tau,$$

on suppose qu'il existe des variables Z corrélées à X et telles que

$$q_\tau(\varepsilon_\tau|Z) = 0. \tag{3.1}$$

Cette hypothèse est l'équivalent de l'hypothèse $E(\varepsilon|Z) = 0$ en régression linéaire instrumentale.

Il est utile de distinguer, parmi les variables explicatives X, les variables *a priori* endogènes, notées $X_1 \in \mathbb{R}^q$ (c'est à dire telles que $q_\tau(\varepsilon_\tau|X_1) \neq 0$), et les variables exogènes

11. Comme pour la procédure quantreg de Sas, l'utilisation de l'algorithme du simplex permet d'estimer un ensemble de valeurs sur $[0;1]$ (en nombre proportionnel à $nln(n)$) tels que les paramètres estimés soient différents. Pour cela on indiquera n'importe quel nombre en dehors de $[0;1]$ (par exemple `tau = -1`).

12. `http://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf`.

X_2. Supposons qu'on dispose de $Z_1 \in \mathbb{R}^r$ (avec $r \geq q$) variables supplémentaires aux explicatives, telles que $q_\tau(\varepsilon_\tau|Z) = 0$, avec $Z = (Z_1, X_2)$. Cette condition implique que :

$$q_\tau(Y - X_1'\beta_{1\tau}|Z) = X_2'\beta_{2\tau}. \tag{3.2}$$

Cette propriété est à la base d'une méthode proposée récemment par Chernozhukov & Hansen (2008). L'équation (3.2) signifie que dans une régression quantile de $Y - X_1'\beta_{1\tau}$ sur Z_1 et X_2, le coefficient de Z_1 est égal à 0. L'idée de Chernozhukov et Hansen est alors d' « inverser » la régression quantile, en estimant $\beta_{1\tau}$ par le paramètre $\widehat{\beta}_{1\tau}$ permettant d'obtenir, dans la régression quantile de $Y - X_1'\widehat{\beta}_{1\tau}$ sur Z un coefficient égal à 0 pour Z_1. En pratique, les auteurs proposent l'algorithme suivant :

1. Définir une grille sur $\beta_{1\tau}$, $\{b_1, ..., b_J\}$.
2. Pour $j = 1$ à J :
   - Calculer les estimateurs de régression quantile de $Y - X_1'b_j$ sur (Z_1, X_2). Soit $(\widehat{\gamma}(b_j), \widehat{\beta}_{2\tau}(b_j))$ les estimateurs correspondants.
   - Calculer la statistique de Wald correspondant au test de $\gamma(b_j) = 0$:

$$W_n(b_j) = n\widehat{\gamma}(b_j)'\widehat{V}_{\text{as}}^{-1}(\widehat{\gamma}(b_j))\widehat{\gamma}(b_j).$$

3. Définir l'estimateur de β_τ par

$$\widehat{\beta}_{1\tau} = \arg\min_{j=1...J} W_n(b_j),\ \widehat{\beta}_{2\tau} = \widehat{\beta}_{2\tau}(\widehat{\beta}_{1\tau}).$$

L'intérêt de cet algorithme est qu'il ne s'appuie que sur des régressions quantiles classiques. Il peut donc être mis en œuvre simplement avec des logiciels standards. La commande `Stata ivqreg` a d'ailleurs été introduite récemment. En pratique, la grille doit être suffisamment fine pour ne pas altérer les propriétés asymptotiques de l'estimateur (cf. Chernozhukov & Hansen, 2008 pour plus de détails). Pour que le temps de calcul reste raisonnable, le nombre de variables endogènes doit donc être petit ($q = 1$ ou 2).

Notons que d'autres solutions existent pour estimer des régressions quantiles instrumentales. Abadie et al. (2002) proposent de recourir à une approche par régression quantile pondérée dans le cas où la variable endogène X_1 et l'instrument Z_1 sont binaires. On peut estimer directement les coefficients en s'appuyant sur l'équation (3.2) et la méthode des moments généralisés. Comme toujours, la difficulté est évidemment de trouver un instrument valide, c'est-à-dire vérifiant (3.1). Nous proposons dans la partie 4.2 un exemple d'instrument utilisant une expérimentation sociale.

3.2 Les régressions quantiles avec des données de panel

L'utilisation de données de panel, c'est-à-dire de données répétées pour les mêmes unités, peut être également une manière de traiter la présence d'hétérogénéité individuelle inobservée. Ces données sont en effet plus souvent disponibles que des variables instrumentales valides. Sous l'hypothèse, certes restrictive, que l'hétérogénéité individuelle est fixe dans le temps et indépendante des termes résiduels, une simple différenciation permet dans le cadre des régressions linéaires de se débarrasser de ces effets fixes classiques. Cependant, ces méthodes ne s'appliquent plus directement dans le cas des quantiles. Ces derniers n'ont pas de propriété de linéarité comme la moyenne. Les quantiles des variables différenciées ne correspondent pas directement aux quantiles d'intérêt.

Sur la période très récente, de nombreux estimateurs permettant de tenir compte de la présence d'effets fixes dans des régressions quantiles ont été proposés. L'estimateur de Canay (2011) a l'avantage de s'appuyer sur des techniques standards utilisant des procédures couramment disponibles dans les logiciels statistiques. Pour en comprendre le principe, il est utile d'utiliser les notations du modèle à coefficients aléatoires présenté en introduction, dans lequel on prend en compte également un effet fixe individuel α_i :

$$Y_{it} = X_{it}\beta_{U_{it}} + \alpha_i, \tag{3.3}$$

où α_i et U_{it} sont inobservables, U_{it} suivant une distribution uniforme sur $[0, 1]$. X_{it} représentent les variables explicatives ainsi que la constante. En introduisant $e_{it} = X_{it}(\beta_{U_{it}} - \beta_\tau)$, on peut réécrire le modèle comme :

$$Y_{it} = X_{it}\beta_\tau + \alpha_i + e_{it}, \text{ avec } \; q_\tau(e_{it}|X_i) = 0. \tag{3.4}$$

Le terme inobservé U_{it} est indépendant des caractéristiques individuelles observées et inobservées (X_{it}, α_i). On suppose également que les effets fixes individuels α_i ont un effet de translation simple sur la distribution. Sous des hypothèses techniques détaillées par Canay (2011), on peut montrer que les termes β_τ sont identifiés et proposer un estimateur convergent lorsque T tend vers l'infini. Plus précisément, Canay (2011) propose un estimateur basé sur les deux étapes suivantes :

1. Estimation par un estimateur within classique de la régression linéaire

   $$Y_{it} = X_{it}\beta_\mu + \alpha_i + u_{it}, \text{ avec } E(u_{it}|X_{it}, \alpha_i) = 0.$$

   A partir de cette estimation de β_μ, on peut obtenir des estimations des effets fixes individuels : $\hat{\alpha}_i = \frac{1}{T}\sum_{t=1}^{T}(Y_{it} - X_{it}\hat{\beta}_\mu)$.

2. Régression quantile classique sur la variable transformée $\hat{Y}_{it} = Y_{it} - \hat{\alpha}_i$ sur les régresseurs X_{it}.

Canay montre également que ces estimateurs sont asymptotiquement normaux. Cependant, la matrice de variance-covariance ne correspond pas à celle produite par défaut dans la deuxième étape. Canay (2011) fournit un estimateur de la matrice de variance-covariance, mais propose également d'utiliser une procédure de bootstrap compte tenu de sa complexité. Dans cette procédure, il s'agit de répliquer, pour chacun des échantillons bootstrap, les deux étapes précédentes de l'estimation. Il n'existe pas encore dans les logiciels statistiques standards de procédure ou de package utilisant cette méthode. On trouvera cependant sur le site d'Ivan Canay, un programme `R` permettant d'utiliser cette procédure en deux étapes [13].

Canay n'établit la convergence de cet estimateur que lorsque le nombre de périodes T tend vers l'infini. Comme il le montre sur des données simulées, l'estimateur n'est pas toujours performant sur des panels contenant un petit nombre de périodes. D'autres méthodes ont également été proposées, toujours sous l'hypothèse (3.3) que les effets fixes individuels ont un simple effet de translation. Koenker (2004) propose d'estimer simultanément l'ensemble des effets fixes, en utilisant un terme de pénalisation pour éviter la

13. L'adresse actuelle du site est `http://faculty.wcas.northwestern.edu/~iac879/research.htm`. Le programme, nommé QRPanel.R, contient la fonction `pqr.estimator(dataframe,eqformula,tau)` qui estime β_τ pour $\tau =$ `tau`. `eqformula` permet de spécifier Y et les X, et `dataframe` correspond au fichier de données. Une estimation de la précision est ensuite fournie par la fonction `pqr.se`

trop grande dispersion des nombreux termes (voir aussi Lamarche, 2010). L'estimation correspond alors à l'estimation simultanée de q quantiles $(\beta_{\tau_k})_{k=1...q}$ et des effets fixes individuels $(\alpha_i)_{i=1...n}$:

$$((\widehat{\beta}_{\tau_k})_{k=1...q}, (\widehat{\alpha}_i)_{i=1...n}) = \arg \min_{\substack{(\alpha_i)_{i=1...n} \\ (\beta_{\tau_k})_{k=1...q}}} \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \sum_{k=1}^{q} \rho_{\tau_k}(Y_{it} - X'_{it}\beta_{\tau_k} - \alpha_i) + \lambda \sum_{i=1}^{n} |\alpha_i|.$$

Le package R `rqpd` permet de faire cette estimation. Cette procédure est cependant très coûteuse en temps de calcul lorsque le nombre d'individus est élevé, et soulève la question du choix du coefficient de pénalisation λ. Comme celui de Canay, cet estimateur n'est convergent qu'asymptotiquement avec T. Notons que l'hypothèse commune à ces deux estimateurs est que les effets fixes n'ont qu'un effet de translation sur la distribution d'intérêt, ce qui constitue une restriction importante. Des articles récents proposent l'utilisation d'estimateurs non-paramétriques, mais ces méthodes sont difficiles à mettre en œuvre et peu adaptées lorsque le nombre de variables explicatives est important. L'utilisation des données de panel pour des régressions quantiles est aujourd'hui encore un champ de recherche actif, et il est probable que d'autres estimateurs seront proposés dans les prochaines années.

3.3 Les « Quantile Treatment Effects »

En pratique, on est souvent intéressé par l'effet non pas de l'ensemble des variables explicatives, mais plus spécifiquement de l'une d'entre elles. On peut par exemple s'intéresser à l'effet d'avoir suivi une formation professionnelle sur les revenus, d'une politique éducative sur la réussite scolaire... L'objectif est alors d'évaluer l'effet causal du fait d'avoir bénéficié du programme évalué, qu'on peut noter par une indicatrice binaire T. Pour disposer d'un cadre pour traiter de ces questions, on peut utiliser le formalisme des méthodes issues de la littérature sur l'évaluation empirique de politique publique. Dans ce cadre, on suppose que chaque personne a deux revenus « potentiels » , Y_0 (celui qu'il peut espérer en l'absence du programme) et Y_1 (celui qu'il peut espérer avec le programme). A ces revenus potentiels sont associées les deux distributions F_{Y_0} et F_{Y_1}. On peut alors définir le $\tau^{\text{ième}}$« quantile treatment effect » (QTE) comme la « distance » horizontale entre les deux distributions (Doksum (1974)) :

$$\delta_\tau = q_\tau(Y_1) - q_\tau(Y_0).$$

De même, on peut définir sa restriction aux personnes qui ont effectivement bénéficié du programme(*quantile treatment effect on the treated*, QTET) :

$$\delta_{\tau|T=1} = q_\tau(Y_1|T=1) - q_\tau(Y_0|T=1).$$

On peut ici s'attarder sur l'interprétation de ces paramètres. L'effet du traitement sur le quantile pour les traités (traduction très littérale du quantile treatment effect on the treated) correspond à la différence entre le τ^{eme} quantile de la distribution de la variable d'intérêt parmi les personnes qui ont bénéficié du programme T, et le quantile équivalent de la distribution de cette variable parmi ces mêmes personnes, si elles n'avaient pas bénéficié de ce programme. L'effet du traitement sur le quantile (quantile treatment effect)

est plus général, puisqu'il correspond à la différence entre les quantiles des distributions qu'on s'attend à observer dans la population respectivement si le programme est généralisé à tous ou au contraire en son absence. Comme on l'a déjà souligné, ces paramètres ne correspondent pas en général à l'effet de ce programme pour les personnes qui se trouvent au niveau du $\tau^{\text{ième}}$quantile de la distribution de Y en l'absence de ce programme. Ce paramètre ne permet donc en principe pas de dire si ce sont les personnes initialement les plus (dé)favorisées qui bénéficieraient du programme qu'on évalue. Pour passer à cette interprétation, il est ici encore nécessaire de faire une hypothèse d'invariance des rangs (les personnes seraient « ordonnées » de la même manière dans la distribution de la variable d'intérêt Y en l'absence ou en présence du programme). Enfin, du fait de la non linéarité des quantiles, ces paramètres ne correspondent pas à la distribution de l'effet du programme ($q_\tau(Y_1) - q_\tau(Y_0) \neq q_\tau(Y_1 - Y_0)$). On trouvera dans Clements et al. (1997) une discussion des conditions sous lesquelles il est possible de borner cette distribution des effets.

Au-delà de ces questions d'interprétation, la difficulté pour estimer ces effets du programme sur les quantiles tient à ce que pour une personne donnée, on n'observe en fait que $Y = TY_1 + (1 - T)Y_0$, c'est-à-dire le revenu potentiel avec traitement (Y_1) si elle a bénéficié du programme et le revenu potentiel sans traitement sinon. On pourrait être tenté d'estimer simplement δ_τ (ou $\delta_{\tau|T=1}$) par la différence de quantiles conditionnels des revenus observés, $q_\tau(Y|T = 1) - q_\tau(Y|T = 0)$. En général cependant, cette différence ne correspond pas au paramètre d'intérêt. En effet, dès qu'il existe une (auto)sélection dans l'entrée dans le programme (par exemple, lorsque les personnes qui ont choisi d'en bénéficier sont celles pour lesquelles il sera le plus efficace), la distribution des revenus observés parmi les bénéficiaires $F_{Y|T=1}$ n'est pas représentative de la distribution du revenu potentiel avec le programme de l'ensemble de la population F_{Y_1}.

Plusieurs méthodes ont été proposées pour identifier les effets moyens d'un programme en présence d'effets de sélection (on en trouvera une description par exemple dans Givord, 2010). Des extensions de ces méthodes à l'analyse des quantiles ont été proposées récemment. Nous ne présentons ici qu'une possibilité, qui est facilement implémentable. Elle repose sur l'hypothèse d'indépendance conditionnelle (Conditional Independance Assumption, ou CIA) suivante :

$$(Y_0, Y_1) \perp\!\!\!\perp T|X. \tag{3.5}$$

Cette hypothèse correspond à l'exogénéité conditionnelle de T (on parle aussi de sélection sur observables, c'est-à-dire que toute la sélection dans le programme peut être expliquée par les variables observées X). Elle est à la base des méthodes d'appariement (matching) ou simplement des régressions linéaires lorsque l'on s'intéresse à la moyenne. On pourrait donc envisager d'estimer l'impact du programme T par une régression quantile en « contrôlant » de l'effet des observables X. Cependant, une telle régression quantile ne permet pas d'estimer directement le paramètre d'intérêt. Lorsque l'on inclut des variables de contrôles supplémentaires X, la régression quantile estime en effet le paramètre $\widetilde{\delta}_\tau = q_\tau(Y_1|X = x) - q_\tau(Y_0|X = x)$. Du fait de la non-linéarité des quantiles, ce paramètre ne correspond pas à δ_τ ni même à $\delta_{\tau|T=1}$ en général. Les quantiles de la distribution des revenus potentiels Y_0 et Y_1 ne sont pas les mêmes que ceux des distributions de ces revenus potentiels conditionnelles aux observables.

Firpo (2007) propose une méthode pour résoudre ces deux problèmes. Celle-ci s'apparente aux méthodes d'appariement (ou matching) utilisées pour estimer l'effet moyen du traitement $E(Y_1 - Y_0)$ sous l'hypothèse 3.5 d'indépendance conditionnelle. On fait

tout d'abord une hypothèse de support commun, nécessaire également dans les méthodes d'appariement :

$$p(X) = P(T = 1|X) \in]0, 1[\tag{3.6}$$

Cette hypothèse signifie que pour chaque bénéficiaire du programme, il existe une personne qui n'en a pas bénéficié et présente les mêmes caractéristiques observables (et inversement). Firpo montre que sous les hypothèses ci-dessus il est possible d'identifier les deux quantiles $q_\tau(Y_1)$ et $q_\tau(Y_0)$, à partir des seules données observées(Y, T, X). Il utilise pour cela les relations [14] :

$$\tau = E\left[\frac{T\mathbb{1}\{Y \leq q_\tau(Y_1)\}}{p(X)}\right] = E\left[\frac{(1-T)\mathbb{1}\{Y \leq q_\tau(Y_0)\}}{1-p(X)}\right] \tag{3.7}$$

Comme on observe (T, Y, X) et que l'on peut identifier le score de propension $p(X)$, ces relations permettent d'estimer $q_\tau(Y_1)$ et $q_\tau(Y_0)$. En pratique, Firpo montre que l'on peut estimer $\delta_\tau = q_\tau(Y_1) - q_\tau(Y_0)$ par une procédure en deux étapes :

1. estimer le score $p(X)$. Notons $\widehat{p}(X)$ un tel estimateur ;
2. estimer $q_\tau(Y_0)$ puis $q_\tau(Y_1)$ en utilisant une régression quantile sur la seule constante : $\widehat{q}_\tau(Y_t) = \arg\min_b \sum \widehat{\omega}_{t,i}\rho_\tau(Y_i - b)$ $(t = 0, 1)$, avec des pondérations $\widehat{\omega}_{1,i} = \frac{T_i}{\widehat{p}(X_i)}$ et $\widehat{\omega}_{0,i} = \frac{1-T_i}{1-\widehat{p}(X_i)}$.

Il s'agit donc d'une régression quantile simple, mais pondérée afin de tenir compte des effets de sélection. Intuitivement, on pondère ainsi parmi les personnes non traitées celles qui ont néanmoins une probabilité plus grande de l'être. Firpo propose également un jeu de pondérations pour estimer l'effet du traitement sur les seuls traités $\delta_{\tau|T=1}$: dans ce cas on utilise comme pondération respectivement $\hat{\omega}_{0,i|T=1} = \frac{\frac{\hat{p}(X_i)}{1-\hat{p}(X_i)}(1-T_i)}{\sum_1^l T_l}$ pour estimer $q_\tau(Y_0|T = 1)$ et $\hat{\omega}_{1,i|T=1} = \frac{T_i}{\sum_1^l T_l}$ pour estimer $q_\tau(Y_1|T = 1)$.

Cette méthode peut donc être implémentée simplement en utilisant des régressions quantiles standards, en pondérant les observations par le poids estimé correspondant au score de propension (option `weight` pour la procédure `quantreg` de Sas et `weights` pour la procédure du même nom de `R` ; il semble que l'option ne soit pas possible pour la procédure `sqreg` de `Stata`).

3.4 Les régressions quantiles dans les modèles non linéaires

Nous considérons ici des extensions de la régression linéaire quantile aux modèles non-linéaires de la forme

$$Y = g(X'\beta_0 + \varepsilon), \tag{3.8}$$

où g est une fonction non-linéaire connue. Deux exemples importants sont le modèle binaire, pour lequel $g(x) = \mathbb{1}\{x > 0\}$, et le modèle à censure fixe, pour lequel $g(x) = \max(s, x)$ (ou $g(x) = \min(s, x)$) avec s une constante connue. Ce dernier modèle est

14. Ce résultat se montre comme suit (ici pour Y_1) :

$$\begin{aligned} E\left[\frac{T\mathbb{1}\{Y \leq q_\tau(Y_1)\}}{p(X)}\right] &= E\left[\frac{\mathbb{1}\{Y_1 \leq q_\tau(Y_1)\}}{p(X)}E(T|Y_1, X)\right] = E\left[\frac{\mathbb{1}\{Y_1 \leq q_\tau(Y_1)\}}{p(X)}E(T|X)\right] \\ &= E\left[\mathbb{1}\{Y_1 \leq q_\tau(Y_1)\}\right] = \tau. \end{aligned}$$

souvent utilisé pour modéliser la consommation d'un bien, qui prend la valeur nulle quand il n'est pas consommé [15]. Ceci peut être rationalisé par l'existence d'une valuation implicite du bien c^* par les consommateurs éventuels, qui ne consomment ce bien que lorsque cette valuation est strictement positive. On observe donc la consommation $c = max(0, c^*)$. Dans ces modèles, il est difficile d'utiliser des restrictions de la forme $E(\varepsilon|X) = 0$ car en général, $E(Y|X) \neq g(X'\beta_0)$. L'approche standard consiste alors à imposer des hypothèses paramétriques sur la distribution des résidus. Par exemple, il est fréquent de supposer l'indépendance entre X et ε et la normalité de ces derniers (on parle alors de modèle probit lorsque $g(x) = \mathbb{1}\{x > 0\}$ et de modèle tobit lorsque $g(x) = \max(0, x)$). Ces hypothèses sont cependant restrictives et souvent difficiles à justifier.

Une approche alternative à ces hypothèses paramétriques est de recourir à des restrictions sur les quantiles. En effet, on peut facilement étendre les restrictions sur les quantiles des termes de perturbations ϵ à une transformation non linéaire, grâce à la propriété d'invariance évoquée dans la section 1 :

$$g(q_\tau(U)) = q_\tau(g(U)),$$

valable pour toute variable aléatoire U et toute fonction g croissante et continue à gauche [16]. Ainsi, si l'on impose dans le modèle non linéaire (3.8) la restriction $q_\tau(\varepsilon|X) = 0$ et que g est croissante continue à gauche, on obtient

$$q_\tau(Y|X) = g(q_\tau(X'\beta_0 + \varepsilon|X)) = g(X'\beta_0).$$

Par le même argument que celui développé dans la section 3, il s'ensuit que

$$\beta_0 \in \arg\min_\beta E\left[\rho_\tau\left(Y - g(X'\beta)\right)\right].$$

Comme précédemment, on estime alors β_0 par

$$\widehat{\beta} = \arg\min_\beta \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(Y_i - g(X_i'\beta)\right). \tag{3.9}$$

L'estimateur défini par (3.9) est très proche de celui de la régression quantile linéaire, la différence étant simplement l'ajout dans le programme de la fonction g.

Pour le modèle de censure fixe pour lequel $g(x) = \max(s, x)$, $\widehat{\beta}$ est $\sqrt{n}-$convergent, et peut être estimé par une application itérative de régressions quantiles linéaires (cf. par exemple Buchinsky, 1994). Pour la médiane ($\tau = 1/2$), l'estimateur proposé par Powell (1984) (« censored LAD estimator », i.e. l'estimateur des moindres valeurs absolues censuré) est implémenté sous Stata via la commande `clad`. L'estimation de la médiane n'est pas suffisante si la censure se trouve très haut par rapport à la distribution. Hong & Chernozhukov (2002) proposent donc un estimateur en trois étapes qui peut être utilisé pour l'ensemble des quantiles. Cette méthode est décrite dans Fack & Landais (2009) qui l'appliquent pour modéliser l'impact des incitations fiscales sur les dons aux œuvres. Ces dons sont en effet minoritaires, puisque environ 15% des ménages déclarent un tel don, et une modélisation de l'effet moyen ne permettrait pas de mettre en évidence un effet.

15. Pour plus de détails, cf. par exemple Wooldridge (2001).

16. Rappelons qu'une fonction g est continue à gauche si pour tout x, $\lim_{u\uparrow x} g(u) = g(x)$. Les fonctions $g(x) = \mathbb{1}\{x > 0\}$ et $g(x) = \max(0, x)$ sont donc continues à gauche.

4 Exemples d'application

4.1 Comment lire les résultats d'une régression quantile ?

A titre d'illustration, on estime une équation de salaire classique à partir de l'enquête Emploi 2008. Cet exercice n'a d'autre prétention que d'illustrer les résultats issus d'une régression quantile sur un cas pratique. Pour une étude plus complète de la question des rendements salariaux de l'expérience et de l'éducation, et de leur évolution en France, on se reportera à Charnoz et al. (2011).

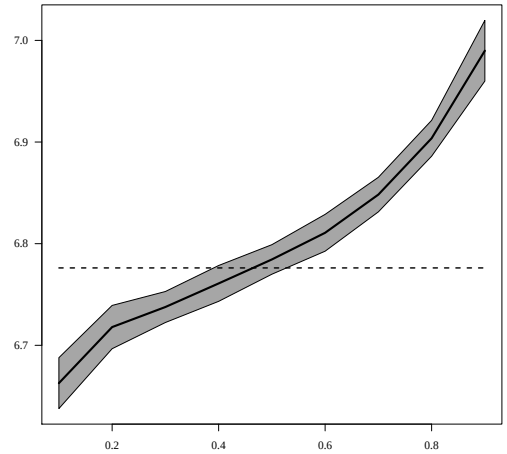
La variable d'intérêt est ici le salaire (exprimé en log), et les variables explicatives sont les caractéristiques observables du salarié, à savoir le nombre d'années d'études, le sexe, sa nationalité, le nombre d'années d'expérience potentielle ainsi que le carré de celui-ci. Les estimations ont été faites pour chaque décile de la distribution conditionnelle du logarithme du salaire. On modélise donc, pour chaque décile (à titre d'illustration, le code sas correspondant à cette estimation est fourni en annexe) :

$$\text{décile}_j(\ln(\text{salaire}|X)) = X'\beta_j$$

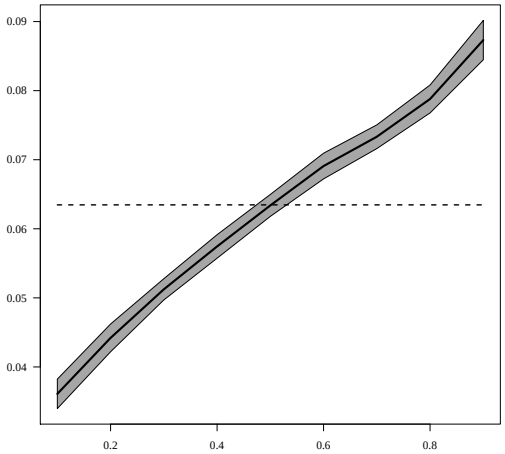
Les régressions quantiles permettent de déterminer comment varie chaque décile en fonction des déterminants auxquels on s'intéresse. Par exemple, le paramètre β_{kj} dans la régression $\text{décile}_j(\ln(\text{salaire})|X) = X'\beta_j$ correspond, à $\text{décile}_j(\ln(\text{salaire})|X_{-k}, X_k = x_k) - \text{décile}_j(\ln(\text{salaire})|X_{-k}, X_k = x_k + 1)$. Il s'agit du changement du $j^{\text{ième}}$ décile de la distribution conditionnelle de salaire suite à une augmentation d'une unité de X_k, par exemple une augmentation d'une année d'études, toutes choses égales par ailleurs (les autres variables X_{-k} restent constantes). Dans le cas d'une variable explicative binaire, comme le fait d'être un homme pour un salarié, β_{kj} mesure simplement l'écart entre le $j^{\text{ième}}$ décile de la distribution des salaires des hommes (conditionnelle à l'ensemble des autres variables explicatives X_{-k}) et le $j^{\text{ième}}$ décile de la distribution des salaires des femmes (également conditionnelle à X_{-k}).

En termes de présentation, on notera qu'on a un jeu de coefficients estimés pour chaque quantile auquel on s'intéresse. Les résultats sont donc plus lourds à présenter. Dans la littérature, on les trouve présentés sous forme d'un tableau regroupant l'ensemble des coefficients, ou de manière peut être plus parlante sous forme de graphiques. C'est la solution que nous avons retenue ici (figure 4).

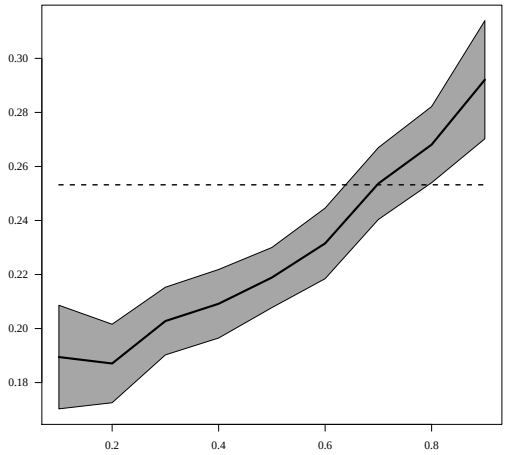
Nous avons choisi de représenter les estimations des coefficients pour les différents déciles, avec l'intervalle de confiance à 95% (zone grisée), ainsi, à titre de comparaison, que la valeur du coefficient des moindres carrés ordinaires (en pointillé).



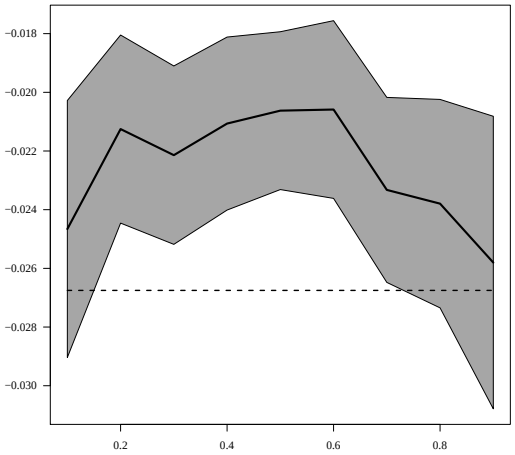
Constante



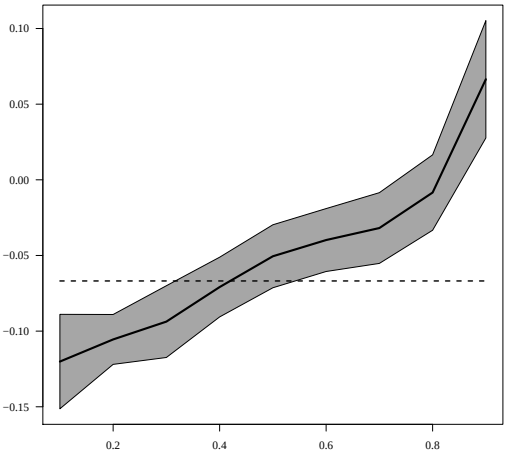
Nombre d'années d'étude



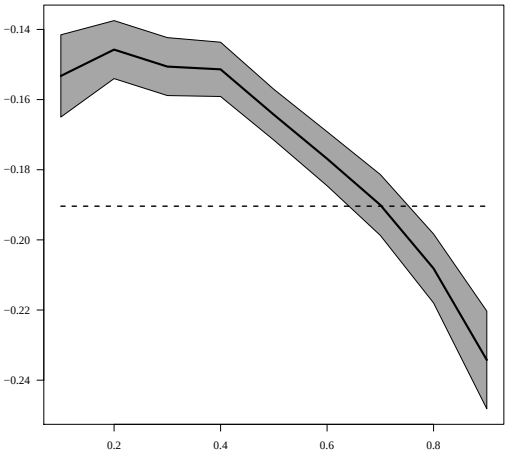
Expérience potentielle



Expérience potentielle au carré



Etranger



Femme

FIGURE 4 – Estimation des coefficients par régressions quantiles (en pointillé : estimation par les moindres carrés ordinaires).

Le coefficient correspondant à la constante peut être considéré comme le décile des salariés ayant les modalités de référence (ici, le fait d'être un homme salarié avec la nationalité française, sans expérience potentielle et dont le niveau d'étude est minimal). Il est sans surprise croissant avec le décile (premier graphique en haut à gauche). On passe ainsi de $6,5$ pour le premier décile à $7,0$ pour le neuvième décile. Le coefficient estimé par les moindres carrés ordinaires (qui correspond pour sa part donc au salaire moyen des salariés aux modalités de référence, sans expérience potentielle et de niveau d'étude minimal), est plus proche des premiers déciles (autour de $6,7$), ce qui exprime bien que la distribution du logarithme des salaires est asymétrique.

Le coefficient correspondant au nombre d'années d'étude est toujours positif, ce qui traduit que le niveau d'étude décale globalement la distribution des salaires vers le haut. Son effet est très nettement croissant avec le décile. Ainsi, l'augmentation du premier décile de la distribution des salaires conditionnelle suite à une augmentation d'une année d'études est, une fois contrôlé des autres variables observées, de 0,04, contre 0,08 pour le dernier décile. Une autre manière de présenter ce résultat est d'observer que la dispersion des salaires augmente avec le nombre d'années d'études, ou encore que les distributions de salaires des plus diplômés sont plus inégales que celles des moins diplômés. C'est également le cas pour l'expérience potentielle [17].

Les salaires des femmes sont systématiquement inférieurs à ceux des hommes, mais ces différences sont d'autant plus fortes que l'on s'élève dans la distribution : conditionnellement aux autres caractéristiques observables, le neuvième décile de la distribution des salaires des femmes est ainsi inférieur de 24% au neuvième décile de la distribution des salaires des hommes, tandis que cette différence n'est « que » de 15% pour le premier décile (ces différences peuvent s'expliquer par exemple par la présence de « plafonds de verre »). A l'inverse, le fait de ne pas disposer de la nationalité française a un impact négatif pour le bas de la distribution, mais les deux distributions conditionnelles se rapprochent ensuite : le coefficient augmente avec le décile, il n'est plus significatif au niveau des septième et huitième déciles et même positif au niveau du neuvième décile.

Ces résultats sont à interpréter avec la même prudence que ceux que donnerait la modélisation du salaire conditionnel moyen. La régression quantile est un outil qui permet d'estimer les effets de variables explicatives sur l'ensemble de la distribution d'une variable d'intérêt, mais elle ne règle aucun des éventuels problèmes d'endogénéité de certaines de ces variables. Par exemple, le fait d'avoir fait des études longues peut être lié à des compétences spécifiques, des réseaux familiaux qui ont également un effet positif sur le salaire. Ces éléments ne sont pas observés dans l'enquête. Dans ce cas, le coefficient des années d'études refléterait également l'effet positif de ces caractéristiques inobservées, et pas uniquement l'effet causal d'une augmentation d'une année d'étude. De même, l'interprétation du coefficient correspondant au fait d'être étranger est délicate. Les coefficients négatifs puis positifs ne peuvent s'interpréter comme la présence d'une discrimination négative puis positive envers les étrangers, sauf à considérer que la distribution des caractéristiques ayant un effet sur le salaire de l'ensemble des salariés étrangers travaillant en France est identique à cette même distribution pour les salariés français. Cependant, les étrangers décidant de travailler en France ont sans doute des profils professionnels particu-

17. Du fait du terme quadratique, on peut interpréter l'augmentation marginale de l'expérience potentielle (exppot) sur le décile j de la distribution de salaire comme $\beta_{1j} + 2\beta_{2j}$ exppot, où β_{1j} est le coefficient relatif à exppot et β_{2j} le coefficient du carré d'exppot.

liers, qui peuvent expliquer ces coefficients [18]. Au final, exactement les mêmes précautions d'interprétation que dans le cas d'une régression linéaire s'imposent. En cas d'endogénéité de certaines variables explicatives, il est nécessaire, pour obtenir une interprétation causale des coefficients, de mobiliser par exemple la méthode des variables instrumentales décrite dans la partie 3 (même s'il n'est pas toujours évident d'obtenir de tels instruments).

D'autre part, les estimations obtenues correspondent à l'effet des variables explicatives sur les distributions de la variable d'intérêt conditionnelles à ces variables. Elles renseignent sur les écarts entre les quantiles d'ordre τ de la distribution de salaire des salariés conditionnelle à ces variables. Elle ne permet pas d'évaluer directement comment se modifierait le quantile d'ordre τ de la distribution de salaires de *l'ensemble de la population* si la distribution de ces variables explicatives était différente. Ceci vient de la propriété de non linéarité des quantiles évoquée en section 1. Ceci a des conséquences importantes quand on souhaite analyser les déterminants des inégalités. Supposons par exemple que l'on souhaite interpréter l'évolution des inégalités de salaires sur les dernières décennies. Celles-ci peuvent être liées à de multiples facteurs, par exemple le niveau du salaire minimum ou la structure de taxation du travail. Des effets de composition sont aussi susceptibles de jouer, comme l'arrivée de générations plus diplômées que les précédentes. Sur notre exemple, on observe ainsi que les distributions de salaires des plus diplômés sont plus inégales que celles des moins diplômés. On peut donc souhaiter quantifier l'effet de l'augmentation du niveau de qualification sur l'évolution des inégalités. Lorsque l'on s'intéresse aux seules évolutions du salaire moyen, il est simple de contrôler de ces effets de composition. Par exemple, la variation du salaire moyen sur les vingt dernières années liée à l'augmentation du niveau d'études s'estime simplement en multipliant le rendement d'une année d'étude, estimé par régression linéaire, par l'augmentation moyenne du nombre d'années d'étude sur la période. Il n'est pas possible d'effectuer directement cette opération pour les quantiles. Autrement dit, on ne peut pas estimer l'effet de l'augmentation de la qualification sur, par exemple, le dernier décile, à partir de la régression quantile correspondante et de l'évolution de la qualification sur la période. Dans ce cas en effet, on cherche à modéliser le décile de la distribution des salaires sur l'ensemble de la population (i.e., le décile inconditionnel), qui ne s'obtient pas simplement en intégrant les quantiles conditionnels qui sont modélisés dans les régressions quantiles. On trouvera dans Fortin et al. (2011) une discussion approfondie des différentes méthodes qui peuvent être utilisées pour répondre à ce type de question.

Enfin, comme expliqué dans la section 1.1.4, les résultats quantiles, tout comme ceux obtenus par une régression linéaire, n'ont pas directement d'interprétation individuelle. Le principe des régressions quantiles est *stricto sensu* de comparer des distributions conditionnelles entre elles. Elles permettent par exemple de dire que le premier décile des salaires des salariés étrangers est inférieur de 12% à celui des salariés ayant la nationalité française, toutes choses égales par ailleurs. En revanche, elles ne permettent pas *a priori* de dire que le salaire d'un salarié étranger qui se trouve au niveau du premier décile de la distribution de salaire de cette population augmenterait d'autant s'il acquérait la nationalité française. Pour pouvoir ainsi interpréter les résultats obtenus, il faut supposer que ce salarié occupe le même rang dans les deux distributions (correspondant respectivement aux salaires des salariés de nationalité française et aux salaires des salariés ne disposant pas de la nationalité française). Comme expliqué plus haut, cette hypothèse d'invariance des rangs, qui a le mérite de fournir une interprétation simple des coefficients

18. Cette difficulté explique le recours au *testing* pour mettre en évidence la discrimination.

des régressions quantiles, est restrictive.

Cette mise en garde dans l'interprétation des résultats est de même nature que celle qu'on peut avoir pour les résultats de la modélisation de la moyenne du logarithme des salaires par une régression linéaire. Les estimations ainsi obtenues montrent que la moyenne du logarithme du salaire des salariés qui n'ont pas la nationalité française est supérieur de 0,7 à celui des salariés français. Cela ne signifie pas qu'un salarié étranger dont le salaire se trouve au niveau au salaire moyen verrait son salaire augmenter d'autant s'il acquérait la nationalité française.

4.2 Un exemple de régression quantile instrumentale

Ce deuxième exemple, qui reprend l'application d'Abadie et al. (2002), permet d'illustrer l'utilisation des méthodes de régressions quantiles instrumentales développées dans la section 3.1. Il utilise les données issues de l'expérimentation de l'efficacité d'un programme de formation de chômeur, le « Job Training Partnership Act (JTPA) » , mis en place à partir de 1983 aux Etats-Unis. Il s'agit d'un ensemble de programmes de formation et d'assistance destinés aux jeunes défavorisés. L'évaluation de l'efficacité de ce genre de programme est souvent rendue difficile par les effets d'auto-sélection : en général, ce sont les personnes qui peuvent en retirer le plus grand bénéfice qui choisissent de rentrer dans le dispositif.

Une expérimentation a cependant été mise en place entre 1987 et 1989 dans 16 structures locales auprès d'un échantillon initial de 20 000 jeunes environ. Les programmes de formation correspondant au JTPA n'ont été proposés qu'à deux tiers de ces jeunes, tirés aléatoirement. Des données ont ensuite été collectées sur la séquence de revenus de l'ensemble des jeunes de l'échantillon initial.

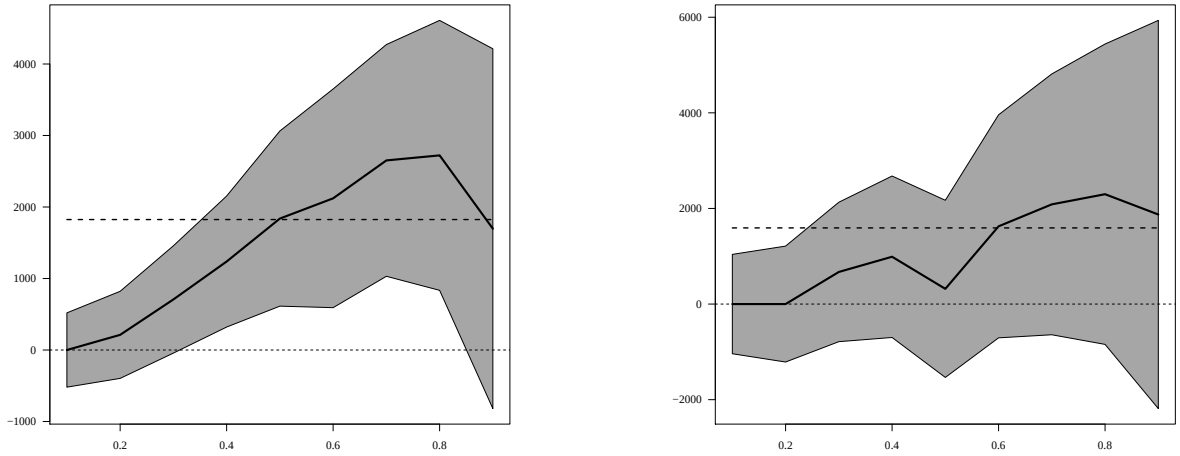
Pour évaluer l'efficacité du JTPA sur les revenus futurs, on ne peut cependant pas comparer directement les jeunes participant au programme de formation et ceux ne participant pas. En effet, malgré l'affectation aléatoire initialement créée par le dispositif expérimental, les personnes qui, par le tirage au sort, étaient affectées au programme pouvaient choisir de ne pas en profiter. Ainsi, seules 60% des personnes tirées au sort ont profité des programmes de formation. Inversement, l'entrée dans le programme n'était pas complètement fermée aux jeunes qui n'avaient pas été sélectionnés par le tirage au sort puisque 2% des jeunes non tirés au sort les ont néanmoins suivis. Ainsi, malgré le tirage au sort, il y a une part d'auto-sélection dans la participation effective au programme. On dispose ici cependant d'un instrument, l'affectation aléatoire au dispositif. Issue d'un tirage au sort, elle n'est évidemment pas corrélée aux déterminants inobservés du revenu. En revanche, elle explique fortement le fait d'avoir bénéficié ou non du programme. Nous appliquons donc la méthode de Chernozhukov & Hansen (2008) décrite dans la partie 3.1 pour évaluer l'impact de ce programme sur l'ensemble des jeunes [19]. Le code sas de la macro est fournie en annexe de ce document [20].

Les estimations sont conduites séparément pour les hommes et les femmes. Du fait du nombre limité de données, on a estimé les trois quartiles, le premier et le dernier décile.

19. Les données sont disponibles à l'adresse `http://econ-www.mit.edu/faculty/angrist/data1/data/abangim02`.

20. Une macro Stata, ivqte, permet également de faire des régressions quantiles instrumentées. Elle s'appuie sur la méthode alternative proposée par Abadie et al. (2002) On trouvera des détails à l'adresse `http://www.econ.brown.edu/fac/Blaise_Melly/code_ivqte.html`.

Les résultats sont présentés sur le graphique 5. Les régressions instrumentales fournissent des résultats près de deux fois plus faibles que les simples régressions linéaires (droites pointillées en gras), ce qui traduit bien l'auto-sélection dans le dispositif. Les régressions quantiles indiquent également que la moyenne masque de grandes disparités dans l'effet du programme. Pour les femmes, l'effet moyen du traitement, c'est-à-dire la différence entre la moyenne espérée des revenus en l'absence du programme et celle des revenus avec le programme est de 1 825 dollars. Mais il n'augmente en fait que de 390 dollars le premier quartile de la distribution de revenus, alors que l'augmentation atteint 2 800 dollars pour le dernier quartile (voir le graphique 5). Les différences sont également très marquées pour les hommes, mais les estimations sont bien plus imprécises et ne permettent jamais d'exclure leur nullité aux seuils ordinaires de significativité. En termes d'interprétation, on rappelle que ces valeurs correspondent à la différence des quantiles des distributions de revenus qu'on s'attend à observer respectivement en l'absence ou en présence du programme de formation. Cela ne signifie pas *a priori* que les femmes dont le revenu serait au niveau du premier quartile en l'absence de la formation vont bénéficier d'une augmentation de 390 dollars grâce à celle-ci, sauf à faire l'hypothèse que la formation ne modifie pas l'ordre relatif des revenus des personnes. Cela pourrait ne pas être le cas si le programme bénéficie beaucoup plus à des personnes en bas de l'échelle des revenus qu'à ceux au dessus et que les revenus potentiels se croisent. Par ailleurs, si on peut interpréter le résultat obtenu par les doubles moindres carrés comme l'effet moyen de la formation (et pas seulement comme la différence des moyennes des revenus avec et sans la formation), il n'est pas en général possible d'interpréter des résultats obtenus par la régression quantile (ici instrumentée) comme ceux de la distribution des effets. Autrement dit, il n'est pas possible de dire qu'un quart des jeunes femmes va bénéficier d'au moins 390 dollars grâce à la formation, car la différence des premiers quartiles des distributions de revenus avec et sans formation ne correspond pas *a priori* avec le premier quartile de la différence des revenus avec et sans formation.



Femmes Hommes

Lecture : les paramètres estimés sont la moyenne (en pointillé, estimation par les doubles moindres carrés), et l'ensemble des déciles. Les zones grisées correspondent à l'intervalle de confiance à 95%, estimé par une méthode de bootstrap. L'effet moyen de la formation sur les revenus des jeunes femmes est de 1 825 dollars. L'impact de cette formation sur le premier décile de la distribution des revenus des jeunes femmes est proche de zéro.

FIGURE 5 – Estimation de l'impact du programme de formation, régression quantile instrumentée.

5 Pour conclure

Au final, la régression quantile est un outil, facile d'utilisation, qui permet d'enrichir la description quantitative des phénomènes économiques et sociaux. Ce document en présente les principes, et permet également de cadrer l'utilisation qui peut en être fait. Il est ainsi nécessaire de rappeler que les régressions quantiles « classiques » ne fournissent qu'une description des différences observées dans les distributions d'une variable d'intérêt conditionnelles à des covariables, qui ne peuvent pas toujours s'interpréter comme un effet causal de ces variables. Pour cela, il est nécessaire de mobiliser des outils d'identification spécifiques détaillés dans la partie présentant les extensions (à condition de disposer des instruments ou des données nécessaires). Par ailleurs, les régressions quantiles permettent de décrire comment les quantiles conditionnels se modifient en fonction de ces déterminants. Par exemple, elle permettra de dire qu'une politique de réduction de la taille des classes permet d'augmenter le niveau du premier décile de x points. Elles ne permettent pas d'étudier la distribution de l'effet de l'un de ces déterminants sur la variable d'intérêt. L'exemple précédent ne permettrait pas de dire que grâce à une réduction de la taille des classes, 10% des élèves verraient leur niveau augmenter de x points. Elles peuvent répondre à une question un peu plus précise, qui serait de savoir comment les personnes qui étaient dans tel ou tel niveau de la distribution initiale ont vu leur situation se modifier, sauf à faire une hypothèse simple mais restrictive. Dans notre exemple, on ne peut pas dire que les élèves dont le niveau initial se situe en dessous du premier décile verrait leur niveau augmenter de x points, grâce à une réduction de la taille des classes. Cette interprétation n'est valide qu'à condition de faire l'hypothèse que quelle que soit la taille des classes observées, les élèves ont des niveaux relatifs identiques. Enfin, les régressions quantiles peuvent être mobilisées même dans des cas où l'objet d'intérêt n'est pas l'ensemble de la distribution, car elles possèdent des propriétés statistiques intéressantes comme la robustesse ou la propriété d'équivariance.

A Annexe A : propriétés des quantiles et détails sur l'inférence

A.1 Quelques propriétés des quantiles

Le quantile d'ordre $\tau \in (0,1)$ d'une variable aléatoire réelle U est défini par

$$q_\tau(U) = \inf\{x | F_U(x) \geq \tau\},$$

F_U étant la fonction de répartition de U. Dans le cas où F_U est continue et strictement croissante, on a simplement $q_\tau(U) = F_U^{-1}(\tau)$. Le graphique 6 illustre la définition des quantiles dans le cas général.

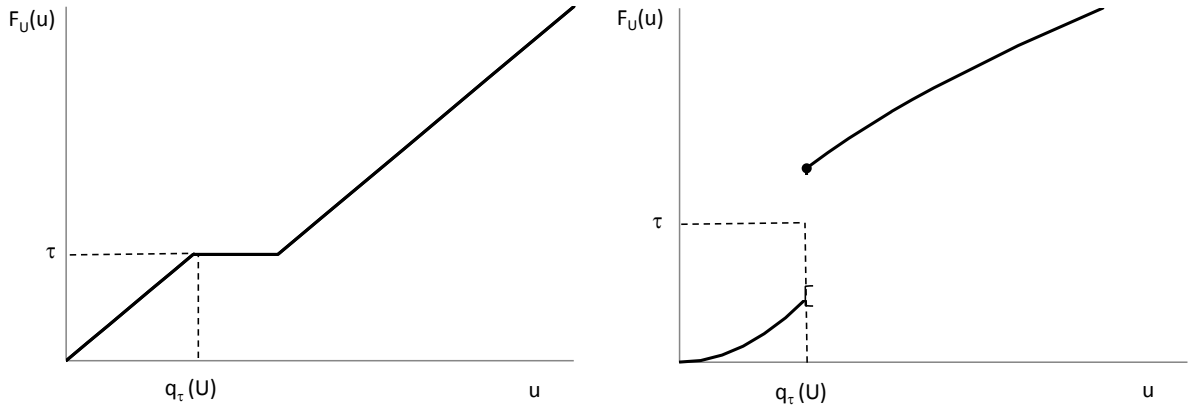


FIGURE 6 – Quantile d'une variable dans le cas général.

Pour deux variables aléatoires U et V, le quantile conditionnel $q_\tau(U|V)$ est défini de manière similaire par :

$$q_\tau(U|V) = \inf\{x | F_{U|V}(x) \geq \tau\},$$

où $F_{U|V}$ est la fonction de répartition de U conditionnelle à V. Les quantiles satisfont l'importante propriété d'invariance suivante.

Proposition A.1. *Soit g une fonction croissante et continue à gauche. Alors :*

$$g(q_\tau(U)) = q_\tau(g(U)).$$

Preuve : Grâce à la monotonicité de g on a $P(U \leq q_\tau(U)) = P(g(U) \leq g(q_\tau(U)))$ et par définition de $q_\tau(U)$: $\tau \leq P(U \leq q_\tau(U))$. Or par définition on a aussi $q_\tau(g(U)) = \inf\{x \in \mathbb{R} | F_{g(U)}(x) \geq \tau\}$, donc $g(q_\tau(U)) \geq q_\tau(g(U))$. Réciproquement, en définissant $g^-(v) = \sup\{x | g(x) \leq v\}$, on a :

$$P(g(U) \leq q_\tau(g(U))) \leq P(U \leq g^-(q_\tau(g(U)))).$$

Par définition de $q_\tau(g(U))$ et $q_\tau(U) = \inf\{x \in \mathbb{R} | F_U(x) \geq \tau\}$ on en déduit que : $g^-(q_\tau(g(U))) \geq q_\tau(U)$. De la continuité de g à gauche, on a aussi que $g(g^-(q_\tau(g(U)))) \leq q_\tau(g(U))$. Donc $q_\tau(g(U)) \geq g(q_\tau(U))$, ce qui conclut la preuve. □

Ce résultat implique notamment que $q_\tau(aU+b) = aq_\tau(U)+b$, ou, de même, $q_\tau(a(X)U+b(X)|X) = a(X)U+b(X)$. Mais il implique également que $q_\tau(\max(s,U)) = \max(s, q_\tau(U))$,

ou que $q_\tau(\mathbb{1}\{U > 0\} = \mathbb{1}\{q_\tau(U) > 0\}$. En revanche, et contrairement à l'espérance, la fonction quantile n'est pas linéaire : on a en général $q_\tau(U_1 + U_2) \neq q_\tau(U_1) + q_\tau(U_2)$.

La propriété suivante est cruciale pour l'estimation.

Proposition A.2. *Supposons F_U dérivable et strictement croissante, et soit $\rho_\tau(u) = (\tau - \mathbb{1}\{u < 0\})\, u$. On a :*

$$q_\tau(U) \in \arg\min_a E[\rho_\tau(U - a)].$$

Preuve : soit $f_U = F_U'$, on a

$$E\left[\rho_\tau(U-a)\right] = \tau(E(U) - a) - \int_{-\infty}^{a} (u-a) f_U(u) du.$$

Cette fonction est dérivable, et

$$\frac{\partial E\left[\rho_\tau(U-a)\right]}{\partial a} = -\tau - (a-a) f_U(a) + \int_{-\infty}^{a} f_U(u) du = F_U(a) - \tau.$$

Cette fonction est croissante, par conséquent $a \mapsto E\left[\rho_\tau(U-a)\right]$ est convexe et atteint son minimum en $q_\tau(U)$ □

Lorsqu'on omet les conditions de régularité sur F_U, le minimum de $a \mapsto E[\rho_\tau(U-a)]$ n'est pas unique en général. Ceci provient du fait que l'équation $F_U(a) = \tau$ peut ne pas avoir de solution, ou en avoir plusieurs (cf. le graphique 6). On peut cependant montrer que $q_\tau(U)$ est toujours l'un des minimum de $E[\rho_\tau(U-a)]$.

A.2 Détails sur les méthodes d'inférence

A.2.1 Estimation directe

Cette approche consiste à estimer directement la variance asymptotique en partant de la formule 2.3. Dans le cas général, la difficulté principale est d'estimer $J_\tau = E(f_{\varepsilon_\tau|X}(0|X)XX')$. Pour ce faire, Powell (1991) propose de s'appuyer sur l'idée suivante :

$$J_\tau = \lim_{h \to 0} E\left[\frac{\mathbb{1}\{|\varepsilon_\tau| \leq h\}}{2h} XX'\right].$$

On estime alors J_τ par

$$\widehat{J}_\tau = \frac{1}{2nh_n} \sum_{i=1}^{n} \mathbb{1}\{|\widehat{\varepsilon}_{i\tau}| \leq h_n\} X_i X_i'. \tag{A.1}$$

où $h_n \to 0$ et $\sqrt{n} h_n \to \infty$.

Cette formule est plus simple dans le cas du modèle de translation, puisque seule l'estimation de $1/f_\varepsilon(q_\tau(\varepsilon)$ est problématique. Soit $\widehat{\varepsilon}_{i\tau} = Y_i - X_i'\widehat{\beta}_\tau$, on peut alors estimer $1/f_\varepsilon(q_\tau(\varepsilon)$ [21] par $(\widehat{\varepsilon}_{(\lceil n(\tau+h_n)\rceil)\tau} - \widehat{\varepsilon}_{(\lceil n(\tau-h_n)\rceil)\tau})/2h_n$. L'estimateur de la variance asymptotique vaut alors :

$$\widehat{V}_{\text{as}} = \tau(1-\tau) \left(\frac{\widehat{\varepsilon}_{(\lceil n(\tau+h_n)\rceil)\tau} - \widehat{\varepsilon}_{(\lceil n(\tau-h_n)\rceil)\tau}}{2h_n}\right)^2 \left[\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right]^{-1}. \tag{A.2}$$

21. On a en effet $\frac{1}{f_\varepsilon(q_\tau(\varepsilon))} = \frac{1}{f_\varepsilon(F_\varepsilon^{-1}(\tau))} = \frac{\partial F_\varepsilon^{-1}}{\partial \tau}(\tau) = \lim_{h\to 0} \frac{F_\varepsilon^{-1}(\tau+h) - F_\varepsilon^{-1}(\tau-h)}{2h}$.

Cet estimateur est parfois proposé par défaut dans des logiciels standard. Il faut cependant garder à l'esprit qu'il n'est convergent que dans le très restrictif modèle de translation.

Une fois obtenu un estimateur convergent de V_{as}, l'inférence sur β_τ est aisée. Un intervalle de confiance de niveau $1-\alpha$ sur β_τ s'écrit ainsi :

$$IC_\alpha = \left[\widehat{\beta}_\tau - z_{1-\alpha/2}\sqrt{\widehat{V}_{\text{as}}}, \widehat{\beta}_\tau + z_{1-\alpha/2}\sqrt{\widehat{V}_{\text{as}}}\right],$$

où $z_{1-\alpha/2}$ est le quantile d'ordre $1-\alpha/2$ d'une loi $\mathcal{N}(0,1)$. De même, la statistique de Wald T du test $\beta_\tau = 0$ s'écrit $T = n\widehat{\beta}'_\tau \widehat{V}_{\text{as}}^{-1}\widehat{\beta}_\tau$, avec T qui tend vers un χ^2_p sous l'hypothèse nulle, où p est le nombre de variables explicatives.

A.2.2 Bootstrap

Une autre possibilité pour faire de l'inférence est de recourir au bootstrap. Rappelons que le principe de bootstrap est de générer des échantillons "factices" par des tirages avec remises à partir de l'échantillon initial. Dans le cas du bootstrap standard, on applique l'algorithme suivant.

De $b = 1$ à B :
- Tirer aver remise un échantillon de taille n à partir de l'échantillon initial $(Y_i, X_i)_{i=1...n}$. Soit $(k^*_{b1}, ..., k^*_{bn})$ les indices correspondants aux observations tirées ;
- Calculer $\widehat{\beta}^*_{\tau b} = \arg\min_\beta \sum_{j=1}^n \rho_\tau(Y_{k^*_{bj}} - X'_{k^*_{bj}}\beta)$.

On peut alors estimer la variance asymptotique par

$$V^*_{\text{as}} = \frac{1}{B}\sum_{b=1}^{B}(\widehat{\beta}^*_{\tau b} - \widehat{\beta})^2.$$

Des intervalles de confiance ou tests peuvent être alors construits comme précédemment, en utilisant l'approximation normale. Pour construire des intervalles de confiance, on peut également s'appuyer sur le *percentile bootstrap*. Soit q^*_u le quantile empirique d'ordre u de $(\widehat{\beta}^*_{\tau 1}, ..., \widehat{\beta}^*_{\tau B})$, on construit simplement l'intervalle de confiance par

$$IC_{1-\alpha} = [q^*_{\alpha/2}, q^*_{1-\alpha/2}].$$

Par rapport à l'estimateur (A.2), les méthodes de bootstrap ont l'avantage de ne pas supposer que le vrai modèle est un modèle de translation. Elles évitent également de devoir choisir le paramètre de lissage h_n, sachant que les résultats peuvent être sensibles à ce choix.

B Annexe B : Codes Sas correspondant aux exemples

B.1 Exemple de la partie 4.1

```
proc quantreg data=emploi algorithm=interior;
model lnsal = francais femme exppot exppotsq etudes /quantile= 0.1 to 0.9 by 0.1 ;
run;
```

Ce code permet simplement d'estimer sur la table emploi les régressions quantiles correspondants aux neufs premiers déciles. Les variables explicatives sont la nationalité **francais**, le genre **femme** l'expérience potentielle (âge moins âge de fin d'études) **exppot** et son carré **exppotsq** et l'âge de fin d'études **etudes**.

B.2 Exemple de la section 4.2

```
%quantile_iv(table=male_jtpa, Y=Sal, D=Trait, Z=Aff,
X=hsorged black hispanic married wkless13 age2225
age2629 age3035 age3644 age4554 class_tr ojt_jsa f2sms,
min_beta_D=0, max_beta_D=3500, nb_pts=100,
liste_tau=0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90);
```

La macrovariable `table` indique la table de travail (ici `male_jtpa`, qu'on peut obtenir à l'adresse `http://econ-www.mit.edu/faculty/angrist/data1/data/abangim02`). Les macrovariables `Y, D, Z et X` correspondent respectivement à la variable d'intérêt (ici le salaire), la variable endogène (ici le fait d'avoir suivi la formation), l'instrument (ici l'affectation aléatoire) et les variables explicatives supplémentaires (ici `hsorged black hispanic married wkless13 age2225 age2629 age3035 age3644 age4554 class_tr ojt_jsa f2sms`). Les macrovariables `min_beta_D` et `max_beta_D` correspondent aux limites du support qui bornent donc la grille (ici 0 et 3500). La macrovariable `nb_pts` correspond au nombre de points utilisés pour la grille (ici 100). Enfin, on indique les quantiles pour lesquels on souhaite faire l'estimation à travers la macrovariable `liste_tau` (ici `0.10 0.20 0.30 0.40 0.50 0.60 0.70 0.80 0.90`).

Le code de la macro `quantile_iv` (écrite par Xavier D'Haultfœuille) est le suivant :

```
%macro quantile_iv(table=, Y=, D=, Z=, X=, min_beta_D= , max_beta_D=, nb_pts=50,
liste_tau=);
/* Construction de la variable d'intérêt modifiée */
data temp_&table.;
set &table.(keep = &Y. &D. &Z. &X.);
%do i=1 %to &nb_pts.;
Ytemp&i.=&Y.-(&min_beta_D.+(&i-1)*(&max_beta_D.-&min_beta_D.)/(&nb_pts.-1)) *&D.;
%end;
run;
/* Nombre de quantiles */
%let nb_tau=0;
%do %while(%qscan(&liste_tau.,&nb_tau.+1,%str( )) ne %str());
%let nb_tau = %eval(&nb_tau.+1);
%end;
/* Number of exogenous Xs */
%let nb_x=0;
%do %while(%qscan(&X.,&nb_x.+1,%str( )) ne %str());
%let nb_x = %eval(&nb_x.+1);
%end;
%do j=1 %to &nb_tau. ;
%let wald_min_&j. = 1000000;
%end;

/* Estimation de la régression quantile pour la variable d'intérêt sur la grille */
%do i=1 %to &nb_pts.;
ods listing close;
ods output ParameterEstimates=pr_Wald;
proc quantreg data=temp_&table. algorithm=interior(tolerance=1.5e-4);
```

```
model Ytemp&i. = &Z. &X.  /quantile=&liste_tau.;
run;
quit;
ods output close;

data pr_Wald;
set pr_Wald;
indic_instr = (parameter="&Z.");
run;

proc sort data=pr_Wald;
by quantile descending indic_instr;
run;

/* Enregistrement des paramètres estimés dans des macro variables */
data pr_Wald;
set pr_Wald;
retain good_alpha 0 quantile_prec 0 j 0;
if quantile_prec ne quantile then do ;
good_alpha = 0;
j = j +1;
end;
new_coeff = &min_beta_D. + (&i-1) * (&max_beta_D. - &min_beta_D.) /(&nb_pts.-1);
if abs(tValue) < abs(symget(compress("wald_min_"||j)))
and parameter="&Z." then do;
good_alpha = 1;
call symput(compress("wald_min_"||j),tValue);
call symput(compress("coeff_alpha_"||j),new_coeff);
end;
if good_alpha = 1 and parameter ne "&Z." then do;
call symput(compress("coeff_"||parameter||"_"||j),Estimate);
end;
quantile_prec = quantile ;
run;

%end;

ods listing ;

/* Estimation des écarts types. La matrice de variance asymptotique
s'écrit V = J^{-1} S J'^{-1}, avec
S = tau(1-tau) E((Z X)' (Z X)) et J = E[f_eps_tau(0|X,Z,D) (Z X)' (D X)]. */

/* Pour obtenir E((Z X)' (Z X)) */
proc reg data=&table. outsscp=for_S;
var &Z. &X.;
run;quit;
```

```
proc sql noprint;
select Intercept into :n_S from for_S where _TYPE_ = "N";
quit ;

data for_S;
set for_S;
where _TYPE_ ne "N";
keep Intercept &Z. &X.;
run;

/* Pour le calcul de J. On estime d'abord les résidus correspondant à chaque tau*/

data for_J;
set temp_&table.;
where &Y. ne . and &Z. ne . and &D. ne .
%do k=1 %to &nb_x.; and %scan(&X.,&k.) ne . %end; ;
%do j=1 %to &nb_tau. ;
residual&j. = &Y. - &&coeff_Intercept_&j. - &D. * &&coeff_alpha_&j.
%do k=1 %to &nb_x.;
- %scan(&X.,&k.) * symget("coeff_%scan(&X.,&k.)_&j.") %end; ;
%end;
run;

/* Estimation de la taille de la fenêtre c_n pour estimer f_eps_tau(0|X,Z,D) dans J */
/* On utilise c_n = std(eps) * [Phi^{-1}(tau+h_n)-Phi^{-1}(tau-hn)]
où hn est donnée par Hall and Shether (1988), voir Koenker, p.140. */

proc sql noprint;
select count(*) into :n from for_J ;
quit;

data pipo;
%do j=1 %to &nb_tau. ;
tau = %scan(&liste_tau.,&j., %str( ));
/* Here 1.79 = 1.5^(1/3) * (1.96)^{2/3}, see Koenker p.140 */
hn = (&n.**(-1/3)) * 1.79 *
(pdf('NORMAL',probit(tau))**2/(2*(probit(tau)**2)+1))**(1/3);
call symput("hn&j.",hn);
%end;
run;

proc sql ;
select count(*) into :n from for_J ;
select 0 %do j=1 %to &nb_tau. ; , std(residual&j.) *
(probit(min(%scan(&liste_tau.,&j., %str( )) + &&hn&j., 0.9999)) -
 probit(max(%scan(&liste_tau.,&j., %str( )) - &&hn&j., 0.0001))) %end;
into :pipo %do j=1 %to &nb_tau.; , :c&j. %end;
from for_J ;
```

```
quit;

/* Table utilisée pour estimer la matrice J qui
intervient dans la variance asymptotique*/

data for_J;
set for_J;
one = 1;
%do j=1 %to &nb_tau. ;
weight&j. = (abs(residual&j.)<=&&c&j.)/(2*&&c&j.);
Z_weighted&j. = weight&j. * &Z.;
   %do k=1 %to &nb_x.;
X&k._weighted&j. = weight&j. * %scan(&X.,&k.);
%end;
drop residual&j. ;
%end;
run;

%do j=1 %to &nb_tau. ;
%let tau = %scan(&liste_tau.,&j., %str( ));
proc iml;
edit for_S;
read all into S;
S = S/&n_S.;
edit for_J;
read all var {one &D. &X.} into DX;
read all var {weight&j. Z_weighted&j.
%do k=1 %to &nb_x.; X&k._weighted&j. %end;} into mat;
invJ = inv(mat‘ * DX / &n.) ;
V = &tau. * (1- &tau.) * invJ * S * (invJ‘) ;
et = sqrt(vecdiag(V/&n.));
create et from et[colname={et}];
append from et;
quit;
title "Quantile IV estimated coefficients for tau= &tau.";
/* On enregistre les  macro variables dans une table sas */
data table_res;
Variable = "Intercept";
Estimate = &&coeff_Intercept_&j.;
output;
Variable = "&D.";
Estimate = &&coeff_alpha_&j. ;
output;
%do k=1 %to &nb_x.;
Variable = "%scan(&X.,&k.)";
Estimate = symget(compress("coeff_"||Variable||"_&j."));
output;
%end;
```

```
run;
data table_res;
set table_res;
set et;
label et = "Standard errors";
run;
proc print data=table_res label noobs;
run;
%end;
title ;
%mend;
```

Références

Abadie, A., Angrist, J. & Imbens, G. (2002), 'Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings', *Econometrica* **70**(1), 91–117.

Bilias, Y. & Koenker, R. (2001), 'Quantile regression for duration data : A reappraisal of the pennsylvania reemployment bonus experiments', *Empirical Economics* **26**(1), 199–220.

Buchinsky, M. (1994), 'Changes in the U.S. wage structure 1963-1987 : Application of quantile regression', *Econometrica* **62**, 405–458.

Cade, B. S. & Noon, B. R. (2003), 'A Gentle Introduction to Quantile Regression for Ecologists', *Frontiers in Ecology and The Environment* **1**, 412–420.

Canay, I. A. (2011), 'A simple approach to quantile regression for panel data', *The Econometrics Journal* **14**(3), 368–386.

Charnoz, P., Coudin, . & Gaini, M. (2011), Wage inequalities in france 1976-2004 : a quantile regression analysis, Technical report.

Chernozhukov, V. & Hansen, C. (2008), 'Instrumental variable quantile regression : A robust inference approach', *Journal of Econometrics* **142**, 379–398.

Clements, N., Heckman, J. & Smith, J. (1997), 'Making the most out of programme evaluations and social experiments : Accounting for heterogeneity in programme impacts', *Review of Economic Studies* **64**(4), 487–535.

Cornec, M. (2010), Constructing a conditional gdp fan chart with an application to french business survey data. mimeo Insee.

Doksum, K. (1974), 'Empirical probability plots and statistical inference for nonlinear models in the two-sample case', *The Annals of Statistics* **2**(2), pp. 267–277.

Fack, G. & Landais, C. (2009), 'Les incitations fiscales aux dons sont-elles efficaces ?', *Économie et Statistique* **427**(1), 101–121.

Firpo, S. (2007), 'Efficient semiparametric estimation of quantile treatment effects', *Econometrica* **75**(1), 259–276.

Fitzenberger, B. & Wilke, R. A. (2005), Using quantile regression for duration analysis, ZEW Discussion Papers 05-65, ZEW - Zentrum für Europäische Wirtschaftsforschung / Center for European Economic Research.

Fortin, N., Lemieux, T. & Firpo, S. (2011), *Decomposition Methods in Economics*, Vol. 4 of *Handbook of Labor Economics*, Elsevier, chapter 1, pp. 1–102.

Givord, P. (2010), Méthode économétrique pour l'évaluation des politiques publiques, Documents de Travail de la DESE - Working Papers of the DESE g2010-08, Institut National de la Statistique et des Etudes Economiques, DESE.

He, X. & Hu, F. (2002), 'Markov chain marginal bootstrap', *Journal of the American Statistical Association* **97**(459), pp. 783–795.

Hong, H. & Chernozhukov, V. (2002), 'Three-step censored quantile regression and extramarital affairs', *Journal of the American Statistical Association* **97**, 872–882.

Kocherginsky, M., He, X. & Mu, Y. (2005), 'Practical confidence intervals for regression quantiles', *Journal of Computational and Graphical Statistics* **14**(1), 41–55.

Koenker, R. (2004), 'Quantile regression for longitudinal data', *Journal of Multivariate Analysis* **91**(1), 74–89.

Koenker, R. (2005), *Quantile Regression*, Econometric Society Monograph Series, Cambridge University Press.

Koenker, R. & Hallock, K. F. (2001), 'Quantile regression', *Journal of Economic Perspectives* **15**(4), 143–156.

Koenker, R. W. & D'Orey, V. (1987), 'Algorithm as 229 : Computing regression quantiles', *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **36**(3), pp. 383–393.

Lamarche, C. (2010), 'Robust penalized quantile regression estimation for panel data', *Journal of Econometrics* **157**(2), 396–408.

Landais, C. (2007), Les hauts revenus en France (1998-2006) : Une explosion des inégalités ? mimeo Paris School Economics.

Machado, J. A. & Silva, J. M. C. S. (2005), 'Quantiles for counts', *Journal of the American Statistical Association* **100**, 1226–1237.

Portnoy, S. & Koenker, R. (1997), 'The gaussian hare and the laplacian tortoise : Computability of squared- error versus absolute-error estimators', *Statistical Science* **12**(4), pp. 279–296.

Powell, J. (1984), 'Least absolute deviations estimation for the censored regression model', *Journal of Econometrics* **25**, 303–325.

Powell, J. L. (1991), *Estimation of monotonic regression models under quantile restrictions*, Cambridge : Cambridge University Press.

Solard, J. (2010), 'Les très hauts revenus : des différences de plus en plus marquées entre 2004 et 2007', *Insee Références* **Editions 2010**, 45–64.

Wooldridge, J. W. (2001), *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

Série des Documents de Travail
« Méthodologie Statistique »

9601 : Une méthode synthétique, robuste et efficace pour réaliser des estimations locales de population.
G. DECAUDIN, J.-C. LABAT

9602 : Estimation de la précision d'un solde dans les enquêtes de conjoncture auprès des entreprises.
N. CARON, P. RAVALET, O. SAUTORY

9603 : La procédure FREQ de SAS - Tests d'indépendance et mesures d'association dans un tableau de contingence.
J. CONFAIS, Y. GRELET, M. LE GUEN

9604 : Les principales techniques de correction de la non-réponse et les modèles associés.
N. CARON

9605 : L'estimation du taux d'évolution des dépenses d'équipement dans l'enquête de conjoncture : analyse et voies d'amélioration.
P. RAVALET

9606 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT).
S. LOLLIVIER, M. MARPSAT, D. VERGER

9607 : Enquêtes régionales sur les déplacements des ménages : l'expérience de Rhône-Alpes.
N. CARON, D. LE BLANC

9701 : Une bonne petite enquête vaut-elle mieux qu'un mauvais recensement ?
J.-C. DEVILLE

9702 : Modèles univariés et modèles de durée sur données individuelles.
S. LOLLIVIER

9703 : Comparaison de deux estimateurs par le ratio stratifiés et application aux enquêtes auprès des entreprises.
N. CARON, J.-C. DEVILLE

9704 : La faisabilité d'une enquête auprès des ménages.
1. au mois d'août.
2. à un rythme hebdomadaire

C. LAGARENNE, C. THIESSET

9705 : Méthodologie de l'enquête sur les déplacements dans l'agglomération toulousaine.
P. GIRARD.

9801 : Les logiciels de désaisonnalisation TRAMO & SEATS : philosophie, principes et mise en œuvre sous SAS.
K. ATTAL-TOUBERT, D. LADIRAY

9802 : Estimation de variance pour des statistiques complexes : technique des résidus et de linéarisation.
J.-C. DEVILLE

9803 : Pour essayer d'en finir avec l'individu Kish.
J.-C. DEVILLE

9804 : Une nouvelle (encore une !) méthode de tirage à probabilités inégales.
J.-C. DEVILLE

9805 : Variance et estimation de variance en cas d'erreurs de mesure non corrélées ou de l'intrusion d'un individu Kish.
J.-C. DEVILLE

9806 : Estimation de précision de données issues d'enquêtes : document méthodologique sur le logiciel POULPE.
N. CARON, J.-C. DEVILLE, O. SAUTORY

9807 : Estimation de données régionales à l'aide de techniques d'analyse multidimentionnelle.
K. ATTAL-TOUBERT, O. SAUTORY

9808 : Matrices de mobilité et calcul de la précision associée.
N. CARON, C. CHAMBAZ

9809 : Échantillonnage et stratification : une étude empirique des gains de précision.
J. LE GUENNEC

9810 : Le Kish : les problèmes de réalisation du tirage et de son extrapolation.
C. BERTHIER, N. CARON, B. NEROS

9811 : Vocabulaire statistique Français - Chinois - Anglais.
LIU Xiaoyue, CUI Bin

9901 : Perte de précision liée au tirage d'un ou plusieurs individus Kish.
N. CARON

9902 : Estimation de variance en présence de données imputées : un exemple à partir de l'enquête Panel Européen.
N. CARON

0001 : L'économétrie et l'étude des comportements. Présentation et mise en œuvre de modèles de régression qualitatifs. Les modèles univariés à résidus logistiques ou normaux (LOGIT, PROBIT) (version actualisée).
S. LOLLIVIER, M. MARPSAT, D. VERGER

0002 : Modèles structurels et variables explicatives endogènes.
J.-M. ROBIN

0003 : L'enquête 1997-1998 sur le devenir des personnes sorties du RMI - Une présentation de son déroulement.
D. ENEAU, D. GUILLEMOT

0004 : Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables.
O. GODECHOT

0005 : Estimation dans les enquêtes répétées : application à l'Enquête Emploi en Continu.
N. CARON, P. RAVALET

0006 : Non-parametric approach to the cost-of-living index.
F. MAGNIEN, J. POUGNARD

0101 : Diverses macros SAS : Analyse exploratoire des données, Analyse des séries temporelles.
D. LADIRAY

0102 : Économétrie linéaire des panels : une introduction.
T. MAGNAC

0201 : Application des méthodes de calages à l'enquête EAE-Commerce.
N. CARON

C 0201 : Comportement face au risque et à l'avenir et accumulation patrimoniale - Bilan d'une expérimentation.
L. ARRONDEL, A. MASSON, D. VERGER

C 0202 : Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002.
J.-A. VALLET, G. BONNET, J.-C. EMIN, J. LEVASSEUR, T. ROCHER, P. VRIGNAUD, X. D'HAULTFOEUILLE, F. MURAT, D. VERGER, P. ZAMORA

0203 : General principles for data editing in business surveys and how to optimise it.
P. RIVIERE

0301 : Les modèles logit polytomiques non ordonnés : théories et applications.
C. AFSA ESSAFI

0401 : Enquête sur le patrimoine des ménages - Synthèse des entretiens monographiques.
V. COHEN, C. DEMMER

0402 : La macro SAS CUBE d'échantillonnage équilibré
S. ROUSSEAU, F. TARDIEU

M0501 : Correction de la non-réponse et calage de l'enquêtes Santé 2002
N. CARON, S. ROUSSEAU

M0502 : Correction de la non-réponse par répondération et par imputation
N. Caron

M503 : Introduction à la pratique des indices statistiques - notes de cours
J-P BERTHIER

M0601 : La difficile mesure des pratiques dans le domaine du sport et de la culture - bilan d'une opération méthodologique
C. LANDRE, D. VERGER

M2013/01 : La régression quantile en pratique
P. GIVORD, X. D'HAULTFOEUILLE

M 0801 : Rapport du groupe de réflexion sur la qualité des enquêtes auprès des ménages
D. VERGER