# La régression quantile en pratique

*Xavier D'HAULTFŒUILLE(*), Pauline GIVORD(**)*

*(*) CREST*

*(**) INSEE, Direction de la méthodologie*

## Introduction

L'immense majorité des études empiriques économiques portent sur les effets moyens des variables auxquelles on s'intéresse. Cet état de fait s'explique par la diffusion d'un outil statistique, la régression linéaire, simple à utiliser et bien maitrisé. Cette limitation est de plus en plus contestée, le public ne se reconnaissant pas toujours dans cette construction virtuelle que serait l'individu moyen. L'une des préconisations du rapport Stiglitz-Sen-Fitoussi appelle ainsi à sortir de la "dictature de la moyenne" en présentant plus souvent des analyses sur la répartition des revenus. De fait, la moyenne apporte une information essentielle mais limitée sur la grandeur à laquelle on s'intéresse : le revenu moyen n'informe pas, par exemple, sur sa répartition plus ou moins inégale dans la population. Par exemple, le revenu moyen est resté stable aux Etats-Unis depuis les années soixante-dix, mais avec une progression plus importante du dernier décile [4]. Les rendements de l'éducation ont évolué différemment dans le haut et le bas de la distribution des revenus. Dit autrement, que l'augmentation du niveau de qualification a eu plus d'impact sur les hauts salaires que les bas salaires. En terme d'évaluation des politiques publiques, une mesure qui n'aurait d'impact positif que pour les plus défavorisés peut être considérée souhaitable même si son effet moyen est négligeable (par exemple, une mesure permettant d'améliorer le niveau scolaire des élèves les plus en difficultés, sans élever le niveau moyen).

Au-delà de l'objectif de description de l'ensemble de la distribution, la nature de la variable d'intérêt (ou de sa mesure) peut conduire à préférer d'autres outils à la régression classique. La moyenne est ainsi très sensible à la présence de valeurs extrêmes. Lorsque la variable d'intérêt a une distribution très étalée, travailler avec la médiane (par exemple) plutôt que la moyenne fournit des résultats plus robustes. Ainsi [10] utilisent ces régressions dites LAD (pour Least Absolute Deviation) pour l'estimation des courbes d'Engel (part de l'alimentation en fonction du revenu) à partir des enquêtes Budget de Famille. En présence de données censurées, l'estimation de la moyenne est également compromise. Supposons que nous nous intéressions à une variable d'intérêt $Y$, mais que nous n'observions cette variable que pour les valeurs supérieures à un certain seuil. Il n'est pas possible d'estimer la moyenne de variable censurée (sauf à faire des hypothèses paramétriques sur la distribution de cette variable en dessous du seuil). En revanche, au delà de ce seuil, les quantiles de la variable censurée coïncident avec ceux de la variable d'intérêt.

Les régressions de quantile sont des outils dont disposent l'économètre pour répondre à ce type de demande. Elles permettent en effet d'étudier l'impact de différents facteurs sur l'ensemble de la distribution de la variable d'intérêt. Si leur principe est ancien, elles ont connu récemment un regain d'intérêt [1]. Un ensemble de procédure préprogrammée en font aujourd'hui un outil simple d'utilisation. Ce document propose un guide d'utilisation pratique de ces méthodes. On en trouvera une présentation plus détaillée dans [16].

Après avoir défini le cadre dans lequel on se place et illustré par des exemples concrets l'avantage des régressions quantile dans la partie 1, on présente le principe de l'estimation par régression quantiles dans la partie 2, ainsi que les propriétés statistiques des estimateurs obtenus et les procédures disponibles dans les logiciels statistiques standard. La partie 3 recense plusieurs extensions utiles, dont en particulier la prise en compte de l'endogénéité. Enfin, la dernière partie présente deux applications à partir de données réelles.

# 1 L'intérêt des régressions quantiles

Pour poser les notations, nous nous intéressons à l'évolution d'une variable aléatoire $Y$, de fonction de répartition $F_Y$ ($F_Y(y) = P(Y \leq y)$). Rappelons que le $\tau^{\text{ième}}$ quantile est par définition $q_\tau(Y) = \inf\{y : F_Y(y) \geq \tau\}$ [2]. 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 $\tau = 0,75$).

Les régressions de quantiles tentent d'évaluer comment les quantiles conditionnels $q_{Y|X}(\tau)$ [3] de la variable d'intérêt se déforment en fonction de déterminants $X \in \mathbb{R}^p$ de cette variable d'intérêt. Il n'y a pas de raison en effet de supposer que l'impact d'une de ces caractéristiques $X$ soit le même aux différents quantiles de la distribution. On peut en trouver une illustration dans les classiques courbes de Quetelet 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 quantiles (traditionnellement les $3^{\text{ème}}$, $25^{\text{ème}}$, $75^{\text{ème}}$ et $97^{\text{ème}}$) de ces distributions conditionnelles à l'âge [4] (voir graphique 1). On peut constater que la distribution des poids conditionnelle n'évolue pas de manière simple avec l'âge. Ces poids sont par exemple plus dispersés lorsque les enfants sont plus âgés qu'à la naissance.

1. On en trouvera un exemple récent d'utilisation sur données françaises dans [6].

2. Si $F_Y$ est continue, on aura ainsi $F_Y(q_\tau(Y)) = \tau$ mais cette égalité est fausse en générale. Voir l'annexe pour des détails sur cette question et quelques propriétés des quantiles.

3. définis par $q_\tau(Y|X) = \inf\{y : F_{Y|X}(y) \geq \tau\}$. Notons qu'il n'existe pas pour l'instant de notation standard des quantiles conditionnels. $q_Y(\tau|X)$ ou encore $q_\tau(Y|X)$ sont également parfois adoptés.

4. On peut ainsi vérifier que la croissance d'un enfant est "normale" en le situant dans la distribution correspondant à son âge.

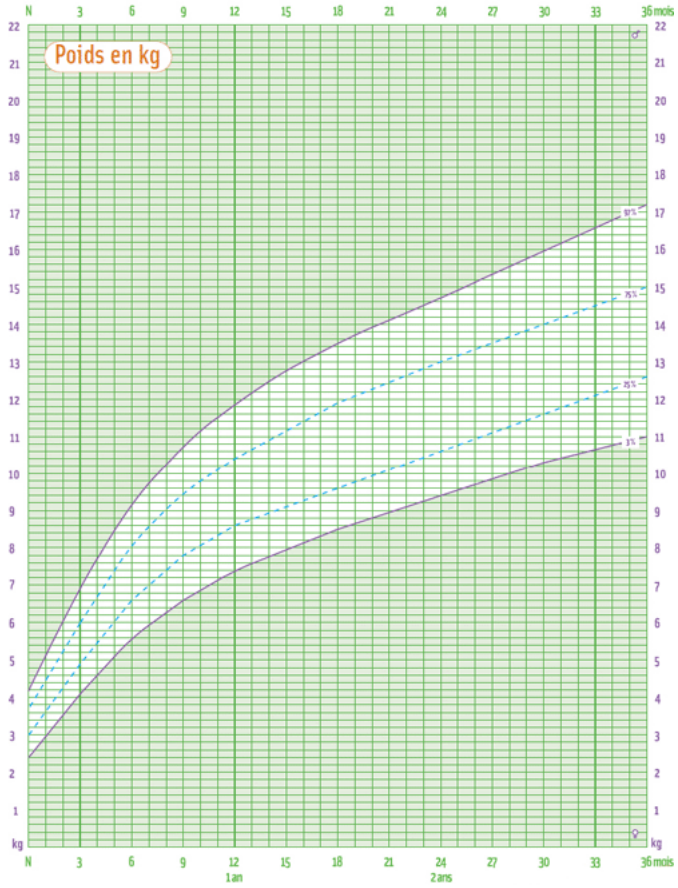



FIGURE 1 – Courbe de croissance.

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 de 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 observables $X$.

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

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

Cette condition est à rapprocher de celle effectuée dans la régression linéaire standard, à savoir $E(Y|X) = X'\beta$. Une différence 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.

Le premier suppose que les observables n'ont d'impact que sur la moyenne de la variable d'intérêt. Il s'agit du modèle de translation linéaire :

$$Y = X'\beta + U \tag{1.2}$$

où $U$ est indépendant de $X$, de moyenne nulle. Sous cette hypothèse, les résidus sont en particulier homoscédastiques (i.e., $V(U|X) = \sigma^2$). Dans ce modèle, les distributions

conditionnelles $F_{Y|X=x}$ sont parfaitement parallèles lorsque $x$ varie. Ceci implique que tous les coefficients $\beta_{k,\tau}$ (où $\beta_\tau = (\beta_{1\tau}, ..., \beta_{p\tau})$, sauf celui correspondant à la constante, sont indépendants de $\tau$. On en trouvera une illustration dans la figure 2, dans le cas simple d'une régression univariée : dans ce cas, les droites correspondant aux régressions de quantiles sont des lignes parallèles (on parle d'*homogénéité des pentes*). Cela signifie également que ces coefficients $\beta_\tau$ (en dehors de la constante) sont les mêmes que ceux d'une régression linéaire. Cela semble limiter leur intérêt dans ce cas. Cependant, les estimateurs des régressions quantiles ne seront pas les mêmes que ceux obtenus par moindre carrés ordinaires. Ils possèdent en particulier des propriétés de robustesse qui les rendent intéressants dans certains cas de figure (cf. ci-dessous).

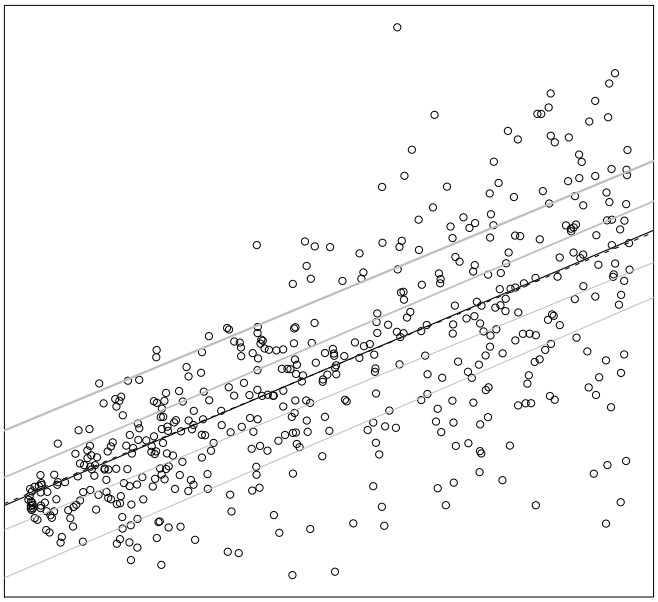


FIGURE 2 – Exemple de données distribuées selon un modèle de translation
Lecture : *Droites correspondant aux régressions quantile pour les déciles d'ordre 1,3,7 et 9 (en gris), la médiane (en noir) et à une régression linéaire classique (en noir pointillé).*

Le deuxième exemple est un peu plus général et suppose que les effets de ces déterminants ont à la fois un impact sur la moyenne mais aussi la variance de la variable d'intérêt. Ces modèles, appelés "translation-échelle", correspondent à une certaine forme d'hétéroscédasticité :

$$Y = X'\beta + (X'\gamma)U \tag{1.3}$$

avec encore une fois $U$ indépendant de $X$, de moyenne nulle et $X'\gamma > 0$. Dans un tel modèle, la dispersion de la variable dépendante 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 sans diplôme (cf. l'application section 4). Le modèle (1.3) implique que $q_\tau(Y|X) = X'(\beta + q_\tau(\varepsilon)\gamma)$. Ainsi, l'hypothèse (1.1) est bien vérifiée, avec $\beta_\tau = \beta + q_\tau(\varepsilon)\gamma$. L'impact des observables ne sera pas le même pour les différents quantiles, et il n'y a plus d'homogénéité des pentes (cf. graphique 3). Pour reprendre l'exemple des salaires, on s'attend ainsi à ce que l'effet du diplôme soit faible pour les premiers quantiles (car une partie non-négligeable de la population est au SMIC même pour des

diplômés du supérieur), mais fort pour les derniers quantiles. Cette information ne ressort pas d'une régression standard, qui se contente d'estimer $\beta$ (car dans le modèle (1.3), $E(Y|X) = X'\beta$).

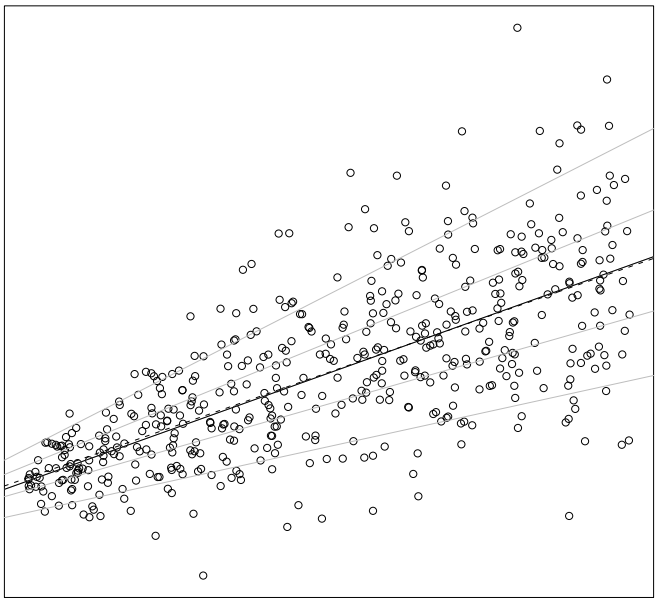


FIGURE 3 – Exemple de données distribués selon un modèle de translation échelle
Lecture : *Droites correspondant aux régressions quantile pour les déciles d'ordre 1,3,7 et 9 (en gris), la médiane (en noir) et à une régression linéaire classique (en noir pointillé).*

Plus généralement, une interprétation intéressante de l'hypothèse (1.1) est le modèle à coefficients aléatoires

$$Y = X'\beta_U, \quad U \text{ indépendant de } X \text{ et de loi uniforme sur } [0,1], \tag{1.4}$$

où la fonction $u \mapsto x'\beta_u$ est strictement croissante pour tout $x$ [5]. Dans ce modèle, $U$ peut s'interpréter comme une composante individuelle inobservée qui positionne l'individu dans la distribution de $Y$. Dans l'exemple des salaires, $U$ correspondrait à la productivité "intrinsèque" de l'individu, le salaire dépendant de cette productivité propre et des études. Ce modèle à coefficients aléatoires généralise les deux exemples précédents puisque dans le premier cas, $\beta_U = \beta_k$ (sauf pour la constante), tandis que dans le second, $\beta_U = \beta + U\gamma$. Il éclaire également l'interprétation du coefficient $\beta_\tau$ dans (1.1). Le modèle (1.4) implique en effet que si l'on modifie marginalement $X$ indépendamment de l'effet individuel $U$, l'effet sur $Y$ est égal à $\beta_U$. Ainsi, $\beta_\tau$ correspond à l'effet marginal de $X$ pour les individus au $\tau^{\text{ième}}$quantile de la distribution des caractéristiques inobservés. $\beta_\tau$ diffère donc en général du coefficient d'une régression linéaire, qui correspond à l'effet marginal moyen de $X$. Cette interprétation n'est valide qu'à la condition que $U$ reste inchangé lorsque $X$ varie. Cette

5. Ce modèle vérifie bien l'hypothèse (1.1) avec puisque $P(Y \leq x'\beta_\tau|X) = P(X'\beta_U \leq x'\beta_\tau|X) = P(U \leq \tau) = \tau$

condition forte est souvent appelée l'hypothèse d'*invariance de rang* dans la littérature, nous la discutons dans le cas de l'exemple présenté dans la partie 4.

Un autre intérêt des régressions quantiles est leur robustesse aux valeurs aberrantes ou à des erreurs très dispersées. Intuitivement, cette propriété est due au fait que les quantiles sont moins sensibles à la présence de valeurs très grandes que la moyenne. Considérons tout d'abord le cas des valeurs aberrantes. Supposons que la variable $Y^*$ vérifie le modèle de l'équation (1.2) mais que dans de très rares cas, nous observions non pas $Y^*$ mais une valeur très grande égale à $X'\alpha$. Formellement, on observe $Y = AX'\alpha + (1-A)Y^*$, où $A$ est une variable inobservée valant 1 lorsque $Y$ est aberrant, 0 sinon, avec $P(A = 1|X, \varepsilon) = p$ petit. Si l'on effectue une régression linéaire de $Y$ sur $X$, on obtient $\beta_{MCO} = \beta + p(\alpha - \beta)$. Si $\alpha$ est très différent de $\beta$, le terme de biais peut être important même si $p$ est petit. En revanche, on peut montrer que si $X'\alpha$ est très grand, $q_\tau(Y|X) = X'\widetilde{\beta}$, avec $\widetilde{\beta}_k = \beta_k$ sauf pour la constante. En d'autres termes, la présence de valeurs aberrantes n'affecte pas les résultat de la régression quantile, sauf ceux de la constante.

Dans un même ordre d'idée, dans le modèle (1.2) où les coefficients correspondant à une régression linéaire et à une régression quantile sont les mêmes, les résultats obtenus par régression quantile seront plus précis en général lorsque les résidus sont très dispersés. Un exemple extrême est celui où $\varepsilon$ n'a pas d'espérance, ce qui se produit lorsque $\varepsilon$ peut prendre des valeurs très grandes avec une probabilité importante [6]. Dans ce cas, l'estimateur des moindres carrés ordinaires n'est pas convergent : même pour des échantillons énormes, il pourra prendre des valeurs très différente du vrai paramètre $\beta$. 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 annexe). Cette propriété n'est bien sûr pas vérifiée par l'espérance. Ceci rend les restrictions de quantiles relativement naturelles et simples à utiliser dans des modèles non-linéaires comme les modèles binaires type logit ou les modèles à censure fixe type tobit.

Il est indispensable de souligner que si les régressions de quantiles ont été développées pour permettre une analyse plus complète de l'impact de déterminants $X$ que l'analyse de la moyenne fournie par les régressions linéaires classiques, elles ne résolvent en rien les problèmes d'endogénéité éventuels qui peuvent survenir. Des extensions au cadre des régressions de quantile ont donc été développés pour tenter d'y répondre. Nous en présentons certains dans la partie 3.

6. Cette situation n'est pas si rare en pratique. Certaines lois de Pareto, utilisées pour modéliser les hauts salaires ou les patrimoines, n'ont pas d'espérance.

# 2 Principes statistiques et mise en œuvre pratique

## 2.1 Définition de l'estimateur et propriétés statistiques

L'estimation des régressions quantiles part de l'observation cruciale que le quantile d'ordre $\tau$ est le résultat du programme de minisation (voir l'annexe pour une preuve) [7] :

$$q_\tau(Y) = \arg\min_a E[\rho_\tau(Y-a)],$$

où $\rho_\tau(.)$ est une "fonction test" définie par $\rho_\tau(u) = (\tau - \mathbb{1}\{u<0\})\,u$. Cette estimation peut sembler moins intuitive que l'approche directe, qui utilise la statistique d'ordre $Y_{(1)} < ... < Y_{(n)}$ en estimant $q_\tau(Y)$ par $\widehat{q}_\tau(Y) = Y_{(\lceil n\tau \rceil)}$, où $\lceil n\tau \rceil$ est le plus petit entier supérieur ou égal à $n\tau$. L'intérêt de cette approche est qu'elle s'étend facilement à un cadre conditionnel :

$$q_\tau(Y|X=x) = \arg\min_a E[\rho_\tau(Y-a)|X=x].$$

Ainsi, en intégrant sur la distribution de $X$ :

$$(x \mapsto q_\tau(Y|X=x)) = \arg\min_{h(.)} E[\rho_\tau(Y-h(X))].$$

Dans la régression quantile, on suppose que $q_\tau(Y|X) = X'\beta_\tau$. Donc :

$$\beta_\tau = \arg\min_\beta E[\rho_\tau(Y-X'\beta)]. \tag{2.1}$$

On peut noter l'analogie avec le modèle de régression linéaire classique, qui modélise l'espérance conditionnelle de $Y$ par une forme linéaire en $X$ : $E(X'\beta_0|X) = X'\beta_0$. Un estimateur de l'espérance d'une variable aléatoire pouvant être obtenu par la fonction de perte quadratique $\arg\min_\beta E[(Y-a)^2]$, on obtient un estimateur de $\beta_0$ par : où l'on a $\beta_0 = \arg\min_\beta E[(Y-X'\beta)^2]$. La fonction de perte quadratique est donc remplacée, dans la régression quantile, par la fonction test $\rho_\tau(.)$. Contrairement à la forme quadratique, seul le signe des écarts importe pour la maximimisation : elle pénalise donc moins les très grands écarts, ce qui explique la robustesse de la régression quantile aux valeurs extrêmes ou aberrantes.

L'estimation de $\beta_\tau$ s'appuie sur l'équation (2.1). Supposons que l'on observe un échantillon i.i.d. $(Y_i, X_i)_{i=1...n}$, nous considérons

$$\widehat{\beta}_\tau = \arg\min_\beta \frac{1}{n}\sum_{i=1}^{n} \rho_\tau(Y_i - X_i'\beta). \tag{2.2}$$

Quand $\tau = 1/2$, c'est-à-dire lorsque l'on s'intéresse à la médiane, $\widehat{\beta}_\tau$ minimise la somme des écarts absolus. On parle d'estimateur des moindres déviations (*least absolute deviation estimator*, ou encore LAD estimator en anglais).

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 [16]. Nous nous contentons ici du résultat principal sur la loi asymptotique de $\widehat{\beta}_\tau$.

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.

**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*

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

Cette variance asymptotique prend une forme particulièrement simple dans le cas du modèle de translation (1.2). Dans ce cas en effet, $\varepsilon_\tau = \varepsilon - q_\tau(\varepsilon)$ et la variance asymptotique $V_{\text{as}}$ s'écrit plus simplement

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

Cette formule 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 : ce sont les résidus autour de $q_\tau(\varepsilon)$ qui vont apporter de l'information sur la valeur du quantile conditionnel de $Y$. Ce résultat appuie également la remarque faite en partie 1 : pour certaines de distribution de $\varepsilon$, une estimation par régression quantile sera plus précise qu'une estimation par MCO. Ceci se produit lorsque $\tau(1-\tau)/f_\varepsilon(q_\tau(\varepsilon))^2 < \sigma^2$.

Le modèle de translation implique que les paramètres $\beta_\tau$ sont indépendants de $\tau$ (sauf le coefficient de la constante). Mais cette restriction n'est pas vérifiée en général, et en particulier dans le modèle de translation-échelle (1.3). Il peut donc être intéressant de faire un test, dit d'homogénéité des pentes, d'égalité des coefficients $\beta_{\tau_1}, ..., \beta_{\tau_m}$ correspondant à différents quantiles. Un tel test s'appuie sur leur distribution jointe asymptotique, qui est 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

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

Le théorème 2.1 est le point de départ à la construction de tests ou d'intervalles de confiance sur $\beta_\tau$. Les deux approches les plus courantes sont [8]

- la méthode directe, qui consiste à estimer directement 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. annexe pour plus de détails).
- le bootstrap, qui consiste, pour estimer la variabilité de $\widehat{\beta}_\tau$, à générer des échantillons "factices" par des tirages avec remises à partir de l'échantillon initial et à effectuer une régression quantile sur ces échantillons. L'inconvénient de cette méthode est qu'elle est souvent coûteuse en temps de calcul. Une solution récente ("Markov Chain Marginal Bootstrap", ou MCMB) a cependant récemment été proposée par [14] pour améliorer ce problème. Cette solution est moins générale que le bootstrap mais s'applique parfaitement au cas des régressions quantiles.

8. Il existe également une méthode basée sur les tests de rang (cf. [16]) et une méthode d'inférence à distance finie (cf. [8]). De par les difficultés qu'elles soulèvent, nous ne les présentons pas ici.

## 2.2 Algorithmes utilisés

Il n'existe pas de solution explicite à (2.2), si bien qu'il faut résoudre ce programme numériquement. Un problème est que la fonction objectif n'est ni différentiable (la fonction $\rho_\tau$ n'est pas dérivable en 0) ni strictement convexe. Les algorithmes standards tels que celui de Newton Raphson ne peuvent pas être utilisés ici. L'idée est alors de reformuler (2.2) 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 $n$. De tels problèmes linéaires peuvent être résolus efficacement par la méthode du simplexe (pour de petits échantillons) ou des méthodes de point intérieur (pour de grands échantillons). On en trouvera une présentation en annexe.

## 2.3 La régression quantile dans les logiciels

### 2.3.1 SAS

On utilise la `proc quantreg` (disponible à partir de la version 9.1) , dont la syntaxe est la suivante :

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

Par défaut, le calcul de $\widehat{\beta}_\tau$ est effectué par l'algorithme du simplexe. Pour éviter des temps de calcul trop long, il est nécessaire d'utiliser plutôt, dès que $n \geq 1000$, une méthode de point intérieur en utilisant l'option `algorithm=interior`.Pour calculer des intervalles de confiance, SAS utilise le bootstrap MCMB cité précédemment lorsque $n \geq 5000$ ou $p \geq 20$, et l'inversion des tests de rang sinon. Cette dernière méthode s'appuie cependant sur des hypothèses restrictives [9]. C'est pourquoi il est préférable d'indiquer l'option `ci=resampling`. L'inconvénient de cette méthode (comme celle basée sur les tests de rang) est qu'elle peut être coûteuse 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 (1.2).

### 2.3.2 Stata

On utilise la commande `sqreg` :

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

Cette commande sera préférée à `qreg` dans la mesure où (i) elle permet de faire des régressions sur plusieurs quantiles simultanément et (ii) les écarts-types sont calculés par bootstrap (le nombre de réplication est indiqué par l'option reps - 20 par défaut), et sont donc plus robustes que ceux calculés avec `qreg`,fondés sur (A.2) [10].

9. Ils ne sont en effet convergents que lorsque le vrai modèle est un modèle de translation-échelle, i.e. vérifie $Y = X'\beta + (X'\gamma)\varepsilon$, avec $\varepsilon$ indépendant de $X$.

10. Signalons que la commande bsqreg permet de faire de la régression quantile sur un seul quantile mais en calculant les écarts-types par bootstrap.

### 2.3.3 R

Un package R très complet a été développé par R. Koenker : quantreg.

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

Pour faire de l'inférence sur tous les quantiles on indiquera `tau=-1` (ou n'importe quelle nombre en dehors de $[0;1]$). 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, mais pas les écarts-types, statistiques de test ou intervalles de confiance correspondant. Pour les obtenir, 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" "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 statistique de test. Les option `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` (cette dernière s'appuyant sur l'approche de [22] détaillée en annexe) sont préférables. 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 [11].

# 3 Extensions

## 3.1 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

$$Y = h(X'\beta_0 + \varepsilon), \tag{3.1}$$

où $h$ est une fonction non-linéaire connue. Deux exemples importants sont le modèle binaire, pour lequel $h(x) = \mathbb{1}\{x > 0\}$, et le modèle Tobit, pour lequel $h(x) = \max(0, x)$ [12]. Dans ces modèles, il est difficile d'utiliser des restrictions de la forme $E(\varepsilon|X) = 0$ car en général, $E(Y|X) \neq h(X'\beta_0)$. L'approche standard consiste alors à imposer des hypothèses paramétriques beaucoup plus restrictives comme $\varepsilon|X \sim \mathcal{N}(0, \sigma^2)$.

Un approche alternative est de recourir à des restrictions sur les quantiles. Un premier intérêt est que cela évite des hypothèses de lois paramétriques difficiles à justifier. Par ailleurs, il est aisé d'étendre les retrictions sur les quantiles à un cadre non-linéaire, grâce à la propriété d'équivariance déjà présentée dans la partie **??** (cf. l'annexe pour une preuve) $g(q_\tau(U)) = q_\tau(g(U))$, valable pour toute variable aléatoire $U$ et toute fonction $g$

11. `http://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf`
12. Ce dernier modèle est particulièrement utilisé lorsque $Y$ est une consommation, et prend donc fréquemment des valeurs nulles. Pour plus de détails, cf. par exemple [24].

croissante et continue à gauche [13]. Ainsi, si l'on impose dans le modèle (3.1) 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'en suit que

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

Comme précédemment, on estime alors $\beta_0$ par

$$\widehat{\beta} = \arg\min_\beta \frac{1}{n}\sum_{i=1}^{n} \rho_\tau\left(Y_i - g(X_i'\beta)\right). \tag{3.2}$$

Cette idée a été appliquée en particulier dans le modèle binaire ($g(x) = \mathbb{1}\{x > 0\}$) avec $\tau = 1/2$. L'estimateur est alors appelé l'*estimateur du maximum de score* (cf. [18]), car il maximise le score [14]

$$\beta \mapsto \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}\{X_i'\beta > 0\} + (1 - Y_i)\mathbb{1}\{X_i'\beta \leq 0\}.$$

Un autre exemple standard est le modèle Tobit ($g(x) = \max(0, x)$) avec $\tau = 1/2$, qui correspond au "censored LAD estimator" développé par [20].

L'estimateur défini par (3.2) est très proche de celui de la régression quantile linéaire, la différence étant simplement l'ajout dans le programme de la fonction $g$. Cet ajout peut cependant conduire à des difficultés computationnelles et à des problèmes d'inférence délicats, en particulier dans le modèle binaire. L'estimateur du maximum du score converge ainsi en $n^{1/3}$ et non $\sqrt{n}$, et sa loi limite est non gaussienne (cf. [15]). Le calcul de l'estimateur est également délicat car la fonction objectif est constante par morceaux. Il n'existe pas à notre connaissance d'implémentation du maximum du score dans les logiciels standards. En revanche, $\widehat{\beta}$ est $\sqrt{n}-$convergent dans le modèle tobit, et peut être estimé simplement par une application itérative de régressions quantiles linéaires (cf. par exemple [4]). Cet algorithme est implémenté sous Stata (pour $\tau = 1/2$) avec la commande `clad`.

## 3.2 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 des rendements salariaux d'un dispositif de formation, 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 (2.2) n'est pas convergent.

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 $\varepsilon_\tau$). Plus précisément, si l'on se place dans le cadre de la régression quantile précédente,

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

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

14. On parle de score en marketing pour désigner la proportion d'individus bien classés par le modèle.

on suppose qu'il existe des variables $Z$ telles que

$$q_\tau(\varepsilon_\tau|Z) = 0.$$

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 $X$ entre les variables endogènes, notés $X_1$ (c'est à dire telles que $q_\tau(\varepsilon_\tau|X_1) \neq 0$), et les variables exogènes $X_2$ (telles que cette même condition soit vérifiée). Supposons qu'on dispose de $Z_2 \in \mathbb{R}^r$ (avec $r \geq q$) variables supplémentaires aux explicatives, qui vérifient cette condition d'indépendance,on note alors $Z = (X_1, Z_2)$. La condition d'indépendance implique que :

$$q_\tau(Y - X_1'\beta_{1\tau} - X_2'\beta_{2\tau}|Z) = 0. \tag{3.3}$$

Cette propriété est à la base d'une méthode proposée récemment par [7] [15]. L'équation (3.3) signifie que dans une régression quantile de $Y - X_2'\beta_{2\tau}$ sur $Z$, le coefficient de $X_1$ (resp. de $Z_2$) est égal à $\beta_{1\tau}$ (resp. à 0). L'idée de Chernozhukov et Hansen est alors d' "inverser" la régression quantile, en estimant $\beta_{2\tau}$ par le paramètre $\widehat{\beta}_{2\tau}$ qui permet d'obtenir, dans la régressions quantile de $Y - X_2'\widehat{\beta}_{2\tau}$ sur $Z$ un coefficient égal à 0 pour $Z_2$. En pratique, les auteurs proposent l'algorithme suivant :

1. Définir une grille sur $\beta_{2\tau}$, $\{b_1, ..., b_J\}$.
2. Pour $j = 1$ à $J$ :
   - Calculer les estimateurs de régression quantile de $Y - X_2'b_j$ sur $(X_1, Z_2)$. Soit $(\widehat{\beta}_\tau^1(b_j), \widehat{\gamma}(b_j))$ 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 $\beta_\tau$ par

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

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. En pratique, la grille doit être suffisamment fine pour ne pas altérer les propriétés asymptotiques de l'estimateur (cf. [7] pour plus de détails). Ceci implique que le nombre de variables endogènes ne doit pas être trop grand ($q = 1$ ou 2).

En pratique, la difficulté est évidemment de trouver un instrument valide. Nous proposons dans la partie 5 un exemple, fondée sur l'utilisation de données issues d'une expérimentation sociale.

15. Notons que d'autres solutions existent. On peut estimer directement les coefficients en s'appuyant sur l'équation (3.3) et la méthode des moments généralisées. On peut également recourir à une approche par régression quantile pondérée dans le cas où où $X_2$ et $Z_2$ sont binaires (cf. [1]).

## 3.3 Les "Quantile Treatment Effects"

Souvent, le chargé d'études est intéressé par l'effet non pas de l'ensemble des covariables, mais plus spécifiquement l'une d'entre elles, à savoir une indicatrice $T$ d'être "traité" par une politique publique [16]. Dans ce cas, il est utile de se placer dans le cadre des méthodes issues de la littérature sur l'évaluation empirique de politique publique. Dans ce cadre, 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). Il faut donc considérer a priori deux distributions de revenu, $F_{Y_0}$ qui correspond au revenu en l'absence du programme et $F_{Y_1}$ à celle du revenu avec le programme.

On peut alors définir le $\tau^{\text{ième}}$quantile treatment effect (QTE) comme la "distance" horizontale entre les deux distributions (Lehmann, 1974 et Doksum, 1974) :

$$\delta_\tau = q_{Y_1}(\tau) - q_{Y_0}(\tau)$$

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

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

Pour une personne donnée, on n'observe cependant dans la réalité qu'un seul de ces revenus potentiels (le revenu potentiel sans traitement $Y_0$ si il n'a pas bénéficié du programme et le revenu potentiel avec traitement $Y_1$ sinon). On pourrait être tenté d'estimer simplement l'effet de la politique $T$ sur la distribution du revenu $Y$ pourrait être simplement obtenu par une régression quantile de la variable d'intérêt $Y$ sur le traitement $T$.
En général cependant, cette estimation peut fournir des résultats qui ne correspondent pas au paramètre d'intérêt. Tout d'abord, très classiquement, pour les problèmes d'endogénéité déjà discutés : en général, s'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 qui ce programme a un effet attendu positif), on n'estimera pas l'effet causal de la politique par cette simple comparaison. En effet, dans ce cas, la distribution des revenus observées parmi les bénéficiaires $F_{Y|T=1}$ n'est pas représentatives 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 (voir [13] pour une présentation plus détaillée). Des extensions de ces méthodes à l'analyse des quantiles ont été proposées récemment. Elles se heurtent cependant à des difficultés supplémentaires, dont l'une est spécifique à la régression quantiles. Pour le comprendre, il est utile de se placer dans le cadre le plus favorable, dans lequel on peut considérer qu'on dispose de suffisamment d'informations dans nos données pour que conditionnellement à des caractéristiques observables, le fait de bénéficier du programme ou pas n'est pas lié au gain escompté. Cette hypothèse d'indépendance conditionnelle (Conditional Independance Assumption, ou CIA) peut s'écrire :

$$Y_0, Y_1 \perp\!\!\!\perp T|X \tag{3.4}$$

où $Y_0$ représente la variable d'intérêt lorsque $T = 0$ et $Y_1$ représente la variable d'intérêt lorsque $T = 1$. Cette hypothèse correspond, dans le cadre d'une régression quantile,

16. Le vocabulaire employé ("effet de traitement", "treatment effect" en anglais) peut surprendre. Il a une raison historique : le cadre méthodologique sous-jacent a d'abord été développé pour étudier l'efficacité de traitements médicaux.

à l'exogénéité conditionnelle de $T$. On pourrait donc envisager d'estimer l'impact du programme $T$ par une régression quantile en "contrôlant" de l'effet des observables $X$. Ceci a deux limites. Le premier est l'hypothèse de linéarité de l'effet des $X$ qui est faite pour la régression de quantile : cette hypothèse est forte. C'est ce qui conduit souvent à préférer utiliser des méthodes d'appariement (matching) plutôt qu'une simple régression linéaire quand on s'intéresse aux seuls effets moyens du traitement.
Un autre problème est cette fois spécifique aux quantiles. En effet, dès lors qu'on inclue des variables de contrôle supplémentaires $X$ la régression quantile n'estime pas de quantile treatment effects mais le paramètre $\widetilde{\delta}_\tau = q_{Y_1|X=x}(\tau) - q_{Y_0|X=x}(\tau)$. Même si l'hypothèse (1.1) implique que $\widetilde{\delta}_\tau$ est indépendant de $x$, $\widetilde{\delta}_\tau$ est différent de $\delta_\tau$ ou de $\delta_{\tau|T=1}$ en général, du fait de la non-linéarité des quantiles.

On peut cependant proposer une méthode pour résoudre ce deux problèmes qui est le pendant des méthodes d'appariement ou matching (qui permettent d'estimer l'average treatment effect $E(Y_1 - Y_0)$) à l'estimation des quantile treatment effects (cf. [11]), sous l'hypothèse 3.4. On fait également une hypothèse de support commun, également classique dans les méthodes d'appariement :

$$p(X) = P(T = 1|X) \in ]0, 1[ \tag{3.5}$$

Cette hypothèse signifie que pour chaque bénéficiaire du programme, on peut trouver une personne qui n'en a pas bénéficié et qui présente les mêmes caractéristiques observables. 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 $(Y, T, X)$. On a en effet que [17]

$$\tau = E\left[\frac{T\mathbb{1}_{Y\leq q_\tau(Y_1)}}{p(X)}\right] \tag{3.6}$$

Cette expression signifie qu'on peut estimer à partir des données que l'on observe le quantile de la distribution de la variable d'intérêt conditionnellement au fait de bénéficier du programme, $q_\tau(Y_1)$, et bien que l'on n'observe cette variable dans cette situation que pour une population particulièrement des personnes qui ont choisi d'en bénéficier. Il faut cependant estimer la probabilité de bénéficier du programme conditionnelle aux observables (qu'on appelle classiquement le score) $p(X) = P(T = 1|X)$. En pratique, Firpo propose une procédure en deux étapes pour estimer $\delta_\tau$ :

1. estimer le score $p(X)$ ;
2. estimer les quantiles de revenus en adaptant la méthode de régression de quantiles classiques : on estime $\hat{q}_\tau(Y_t)$ $(t = 0, 1)$ par $\arg\min_b \sum \hat{\omega}_{t,i|T=1}\rho_\tau(Y_i - b)$, avec $\hat{\omega}_{1,i} = \frac{T_i}{N\hat{p}(X_i)}$ et $\hat{\omega}_{0,i} = \frac{(1-T_i)}{N(1-\hat{p}(X_i))}$.

Par rapport à la régression de quantiles classique, les poids sont modifiés pour tenir compte des effets de sélection. Firpo propose également un jeu de pondération pour estimer l'effet du traitement sur les seules traités $\delta_{\tau|T=1}$. Cette méthode peut donc être implémentée simplement en utilisant des régressions de quantile standard, 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`).

17. Ce résultat se montre comme suit :

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

# 4 Exemples d'application

## 4.1 Comment lire les résultats d'une régression quantile ?

A titre d'illustration, on a estimé 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 [6].

La variable d'intérêt est ici le salaire (exprimé en log), et les variables explicatives sont les caractéristiques observables du salariés, à savoir le nombre d'années d'étude, le sexe, sa nationalité, le nombre d'années d'expérience potentielle ainsi que le carré de celle-ci. Les estimations ont été faites pour chaque décile de la distribution du logarithme du salaire conditionnelle. On modélise donc, pour chaque décile :

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

Les régressions quantiles permettent de déterminer comment varie chaque décile en fonction des déterminants auxquels on s'intéresse. Par exemple, le paramètre $\beta_j$ dans la régression de $\text{décile}_j(\ln(\text{salaire}|X) = X'\beta_j$ correspond à $\frac{\partial \text{décile}_j(\ln(\text{salaire}|X)}{\partial X_j}$, c'est-à-dire le changement marginal du $\text{j}^{\text{ième}}$ décile de la distribution de revenu conditionnelle suite à un changement marginal de $X_j$ (par exemple, une augmentation du nombre d'années d'étude). Cette notion est simple à interpréter dans le cas d'une variable binaire (le fait d'être un homme pour un salarié). Dans ce cas, on compare le $j^{\text{ième}}$ décile de la distribution des salaires des hommes (conditionnelle à l'ensemble des autres variables observables auxquelles on s'intéresse) au $j^{\text{ième}}$ décile de la distribution des salaires des femmes (également conditionnelle à l'ensemble des autres variables observables auxquelles on s'intéresse).

En terme 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).

Le coefficient correspondant à la constante peut être considéré comme un niveau moyen de chaque décile (pour les modalités de référence (ici, le fait d'être un homme salarié avec la nationalité française), et il est sans surprise croissant avec le décile (premier graphique en haut à chaque). 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 donc à un niveau moyen), est plus proche des premiers décile (autour de $6,7$), ce qui exprime bien que la distribution de salaire à une queue de distribution assez étalée.

Le coefficient correspondant au nombre d'années d'étude est toujours positif. Son effet est très nettement croissant avec le décile. Une interprétation est que le rendement des études est plus fort pour les salaires les plus élevés que pour les salaires plus faibles.

On retrouve ce résultat pour l'expérience potentielle [18]. 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 0.24 au neuvième décile de la distribution des salaires des hommes, tandis que cette différence n'est "que" de 0.15 pour le premier décile (ce qui peut s'expliquer par exemple par des mécanismes type "plafond de verre"). Á 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 décile et même positif au niveau du neuvième décile.

Au moins trois remarques doivent être faites.

Tout d'abord, la régression quantile est un outil pour estimer les effets de covariables sur l'ensemble de la distribution d'une variable d'intérêt, et non plus seulement sur sa moyenne. Elles ne règlent pas les problèmes d'endogénéité potentielle de certains déterminants du revenu. Par exemple, le nombre d'années d'étude peut être lié à des caractéristiques individuelles inobservées qui ont également un effet sur le salaire. Le fait que le neuvième décile de la distribution conditionnelle des salaires des étrangers soit supérieur à celui des salariés français peut traduire un effet de sélection plus qu'un effet de discrimination positive à ces niveaux de salaires. Exactement les mêmes précautions d'interprétation que dans le cas d'une régression linéaire s'imposent. Comme discuté plus haut, il est nécessaire pour une interprétation causale de mobiliser des techniques spécifiques, comme par exemple la méthode de variable instrumentale décrite dans la partie 3 (mais il n'est évidemement pas toujours possible de disposer d'un tel instrument).

D'autre part, les estimations obtenues correspondent à l'effet des variables sur les distributions des revenus conditionnelles à ces variables (et non la distribution de revenu sur l'ensemble de la distribution). Pour prendre un cas extrêmement simplifié dans lequel on s'intéresse à l'impact du seul fait d'être de nationalité française, $\beta_\tau$ mesurerait $q_\tau(W|\text{Nat=français}) - q_\tau(W|\text{Nat=étranger})$ (écart entre le quantile d'ordre $\tau$ de la distribution de salaire des salariés de nationalité française et le quantile d'ordre $\tau$ de la distribution de salaire des salariés étrangers). En revanche, cette estimation ne permet pas de répondre à une autre question, qui serait celle de l'impact d'une variation de la proportion de salariés ayant la nationalité française dans la population sur le quantile d'ordre $\tau$ de la distribution de salaires sur l'ensemble de la population [19]. Cette limite vient du fait que les quantiles n'ont pas de propriété de linéarité. De ce fait, $E_X(q_\tau(Y|X)) \neq q_\tau(Y)$. A noter, ce problème ne se pose pas quand on s'intéresse aux moyennes : par la loi des espérances itérées on a $E_X(E(Y|X)) = E(Y)$. Sur les années très récentes, de nombreuses études ont

18. Du fait du terme quadratique, on peut interpréter l'augmentation marginale de l'expérience potentielle sur un décile de la distribution de salaire comme $\beta_1 + \beta_2 exppot$, où $\beta_1$ correspond à la variable en niveau et $\beta_2$ à son carré.

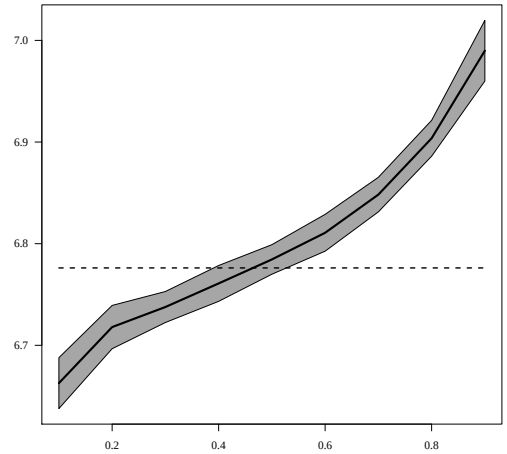
19. L'impact sur le quantile général d'une modification de la proportion $p$ de salariés français s'écrit :

$$\frac{\partial q_\tau(W)}{\partial p} = \frac{f(q_\tau)}{F(q_\tau|\text{Nat=étranger}) - F(q_\tau|\text{Nat=français})}$$

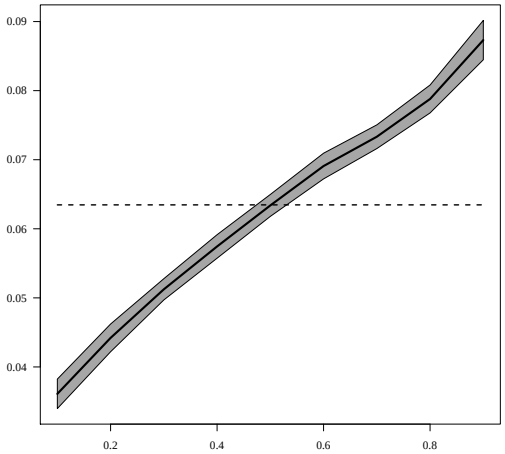
. En effet, on a $\tau = F(q_\tau) = pF(q_\tau|\text{Nat=français}) + (1-p)F(q_\tau|\text{Nat=étranger})$ donc en dérivant, on a par le théorème des fonctions implicites que $F(q_\tau|\text{Nat=étranger}) - F(q_{tau}|\text{Nat=français}) = \frac{\partial q_\tau(W)}{\partial p} \underbrace{(pf(q_{tau}|\text{Nat=étranger}) + (1-p)f(q_\tau|\text{Nat=français}))}_{f(q_\tau)}$.

donc proposées des méthodes permettant de traiter spécifiquement cette question (voir par exemple [12] ou [23]).

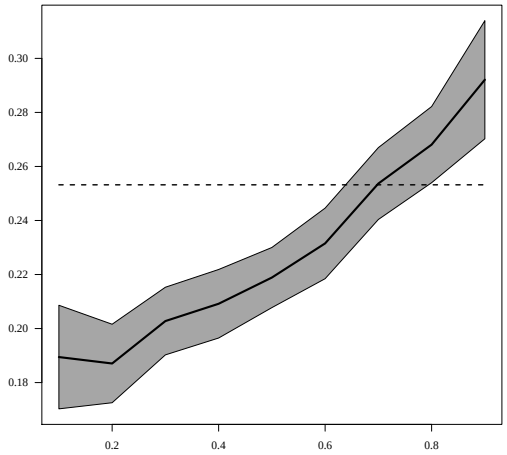
Enfin, les régressions de quantile permettent de comparer des distributions conditionnelles entre elles, mais elle ne permettent pas a priori de déterminer la distribution de l'écart entre ces distributions sauf à faire une hypothèse forte. Une régression de quantile des salaires sur le fait d'être français par exemple permet d'évaluer l'écart entre le quantile d'ordre $\tau$ de la distribution de salaire des salariés de nationalité françaises et le quantile d'ordre $\tau$ de la distribution de salaire des salariés n'ayant pas la nationalité française. Elle ne permet pas de donner directement une information sur la distribution de l'impact individuel d'obtenir la nationalité française ou non, sauf à faire des hypothèses plus forte sur les distributions jointes. Ce serait le cas si l'ensemble des salariés seraient ordonnés de manière exactement identique en terme de salaires, qu'il soit cadre ou non (on parle d'"invariance des rangs"). Cette hypothèse est évidemment très forte (voir [9] pour une discussion).



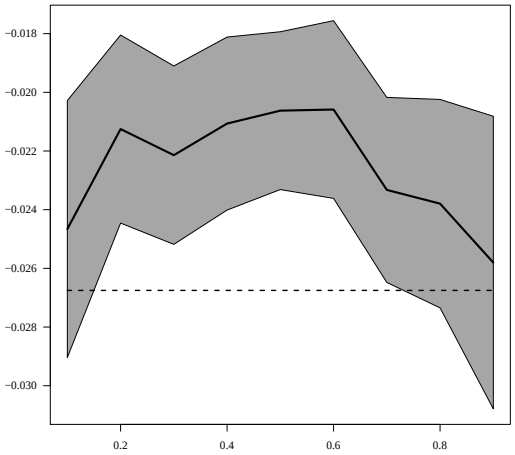

Constante



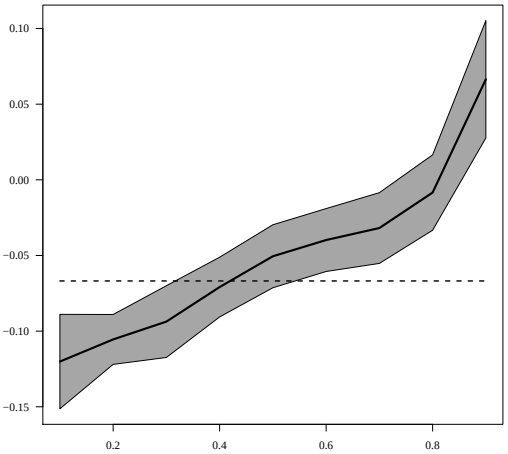

Nombre d'années d'étude



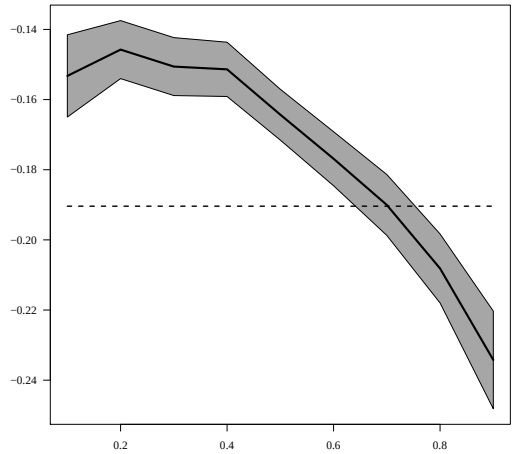

Expérience potentielle




Expérience potentielle au carré




Etranger




Femme

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

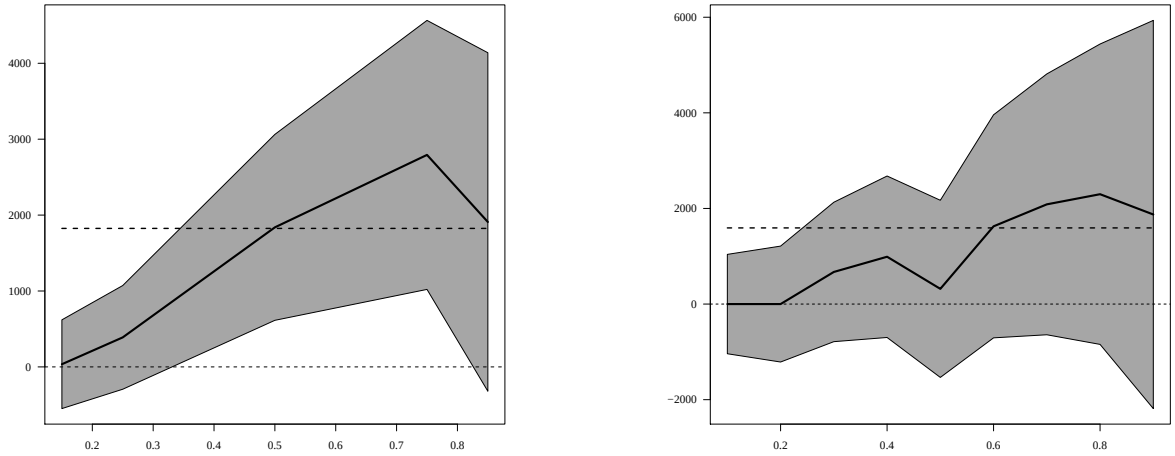
# 5 Un exemple de régression quantile instrumentale

Ce deuxième exemple permet d'illustrer l'utilisation des méthodes de régression quantiles instrumentales développée dans la section **??**. 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 États-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 rendu 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 ont été proposés à seulement deux tiers de ces jeunes.

La particularité de cette évaluation est que les jeunes auxquels le programme est proposé sont désignés au hasard, pour éviter des effets de sélection. Des données ont été collectées sur l'insertion professionnelle de l'ensemble de l'échantillon : celle-ci est mesurée par la séquence de revenu pendant au moins trente mois après l'entrée dans le dispositif. Cependant, il ne s'agit que d'une proposition, et la participation au dispositif n'est pas contraignante : les personnes tirées au sort peuvent choisir de ne pas en profiter, tandis qu'à l'inverse des personnes qui n'avaient pas été tirées au sort peuvent en bénéficier. Dans le cas du JTPA, seulement 60% des personnes en ont profité. Inversement, une (très faible) proportion, 2%, de jeunes qui faisaient partie du groupe de contrôle (à qui le dispositif n'avait pas été proposé) ont finalement suivi les programmes de formation. De ce fait, il n'est pas possible de comparer directement le groupe des bénéficiaires des programmes et les autres pour évaluer l'efficacité de ces programmes. Dans ce cas, il existe à nouveau de la sélection, et une comparaison simple risque de fournir un résultat biaisé. On dispose néanmoins d'un instrument, l'affectation au dispositif : issue d'un tirage au sort, elle n'est pas corrélée aux déterminants inobservés du revenu. En revanche, on peut supposer que les personnes à qui ont a proposé le dispositif ont été plus incitées à en profiter. Cette variable doit donc être corrélée avec le fait d'en avoir bénéficié *in fine*. Nous appliquons donc la méthode de [7] décrite dans la partie 3.2 pour évaluer l'impact de ce programme sur l'ensemble des jeunes [20].

Les estimations sont conduites séparément pour les hommes et les femmes. Les résultats sont représentés dans les graphiques suivants. Les régressions instrumentées fournissent des résultats près de deux fois plus faibles que les simples régressions linéaires, ce qui traduit bien qu'il y a une autosélection 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 est de 1 825 dollars, mais de seulement 390 dollars pour le premier quartile et 2 800 dollars pour le dernier (voir figure 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. [1] proposent une méthode alternative pour des régressions de quantile instrumentées, en utilisant simplement des repondérations. Il est rassurant de constater que les résultats obtenus par l'une et l'autre méthode sont très proches, bien que peu précis dans les deux cas.

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




Femmes Hommes

FIGURE 5 – Estimation de l'impact du programme de formation, régression quantile instrumentée (en pointillé : estimation par les doubles moindres carrés).

# A Annexe

## 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 $q_\tau(U) = F_U^{-1}(\tau)$. La figure 6 illustre la définition des quantiles dans le cas général.

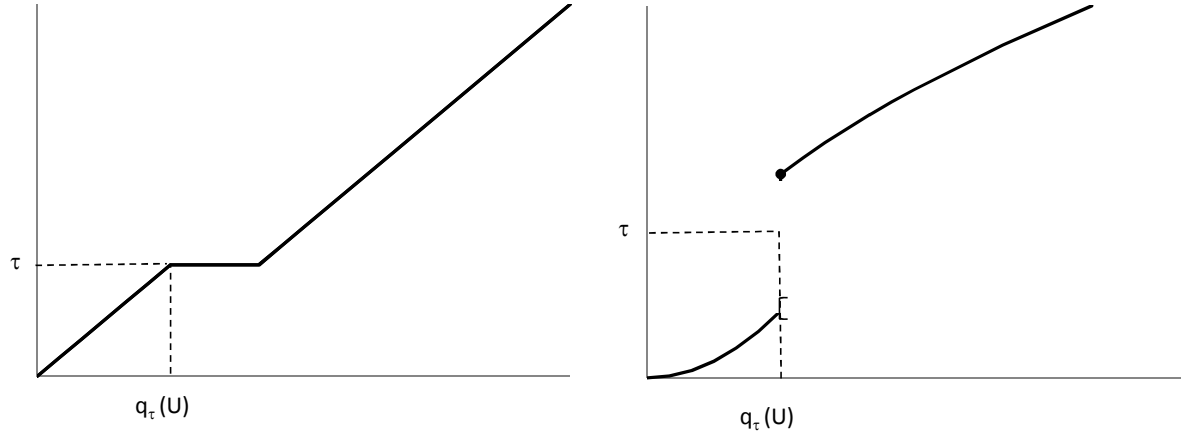



FIGURE 6 – Quantile d'une variable dans le cas général.

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'équivariance 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 $q_\tau(\max(0,U)) = \max(0, q_\tau(U))$, 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

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

Cette fonction est dérivable, et

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

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. figure 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, [22] 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 $J_\tau$ 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

21. On a en effet

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

de la variance asymptotique vaut alors :

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

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 $V_{\text{as}}$, l'inférence sur $\beta_\tau$ est aisée. Un intervalle de confiance de niveau $1-\alpha$ sur $\beta_\tau$ 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 $\chi^2_p$ sous l'hypothèse nulle.

### 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 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.

## A.3 Les méthodes du simplexe et du point intérieur

La méthode du simplexe permet de résoudre des programmes linéaires de la forme

$$\min_{x\in\mathbb{R}^n} c'x \quad \text{s.t. } x \in S = \{u/Au \geq b, Bu = c\}, \tag{A.3}$$

où $c \in \mathbb{R}^n$, $A$ et $B$ sont deux matrices et "$\geq$" doit être considéré élément par élément. On peut alors montrer que (i) $S$ est un polyèdre convexe et (ii) que si des solutions existent, alors elles sont sommets de $S$. La méthode du simplexe consiste alors à aller d'un sommet à l'autre, en choisissant à chaque fois l'arête correspondant à la pente la plus forte.

Les méthodes de points intérieurs, quant à elles, s'appuient sur l'idée de modifier légèrement (A.3) pour en faire un programme standard, facilement résoluble. Considéront (A.3) avec $A = I_n$ et $b = 0$, on résout par exemple

$$\min_{x \in \mathbb{R}^n} c'x - \mu \sum_{k=1}^{n} \ln x_k \quad \text{s.t. } B\,x = c. \tag{A.4}$$

(A.4) peut être facilement résolu par une méthode de Newton. Il suffit alors de faire tendre $\mu$ vers 0.

# Références


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

[2] Athey, S., and Imbens, G. W. Identification and inference in nonlinear difference-in-differences models. *Econometrica 74*, 2 (03 2006), 431–497.

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

[4] Buchinsky, M. Changes in the u.s. wage structure 1963-1987 : Application of quantile regression. *Econometrica 62* (1994), 405–458.

[5] Buchinsky, M., and Hahn, J. An alternative estimator for the censored quantile regression model. *Econometrica 66*, 3 (1998), 653–672.

[6] Charnoz, P., Coudin, ., and Gaini, M. Wage inequalities in france 1976-2004 : a quantile regression analysis. Tech. rep., 2011.

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

[8] Chernozhukov, V., Hansen, C., and Jansson, M. Finite sample inference for quantile regression models. *Journal of Econometrics 152* (2005), 93–103.

[9] Clements, N., Heckman, J., and Smith, J. Making the most out of social experiments : Reducing the intrinsic uncertainty in evidence from randomized trials with an application to the jtpa exp. NBER Technical Working Papers 0149, National Bureau of Economic Research, Inc, Jan. 1994.

[10] Coudin, E., and Clerc, M.-E. L'ipc, miroir de l'évolution du coÃ»t de la vie en france ? ce qu'apporte l'analyse des courbes d'engel. *Economie et Statistique 433*, 1 (2010), 77–99.

[11] Firpo, S. Efficient semiparametric estimation of quantile treatment effects. *Econometrica 75*, 1 (01 2007), 259–276.

[12] Firpo, S., Fortin, N. M., and Lemieux, T. Unconditional quantile regressions. *Econometrica 77*, 3 (05 2009), 953–973.

[13] Givord, P. 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, 2010.

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

[15] Kim, J., and Pollard, D. Cube root asymptotics. *Annals of Statistics 18* (1990), 191–219.

[16] Koenker, R. *Quantile Regression*. Econometric Society Monograph Series, Cambridge University Press, 2005.

[17] Koenker, R., and Hallock, K. F. Quantile regression. *Journal of Economic Perspectives 15*, 4 (Fall 2001), 143–156.

[18] Manski, C. F. Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics 3* (1975), 205–228.

[19] Manski, C. F. Identification of binary response models. *Journal of the American Statistical Association 83* (1988), 729–738.

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

[21] Powell, J. L. Censored regression quantiles. *Journal of Econometrics 32*, 1 (June 1986), 143–155.

[22] Powell, J. L. *Estimation of monotonic regression models under quantile restrictions*. Cambridge : Cambridge University Press, 1991.

[23] Rothe, C. Nonparametric estimation of distributional policy effects. *Journal of Econometrics 155*, 1 (March 2010), 56–70.

[24] Wooldridge, J. W. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2001.