# La régression quantile en pratique

Pauline Givord (INSEE-DMS)

Séminaire de Méthodologie Statistique

16 avril 2013

# Plan

# Plan

# Introduction

- Les régressions quantiles sont un outil dont l'usage s'est généralisé récemment
- Cette première présentation propose un aperçu de leur intérêt et un mode d'emploi pratique
- Elle s'appuie sur un document de méthodologie statistique (à paraître) rédigé avec Xavier D'Haultfœuille (CREST)

# Plan

# Enrichir le diagnostic sur certaines questions économiques.

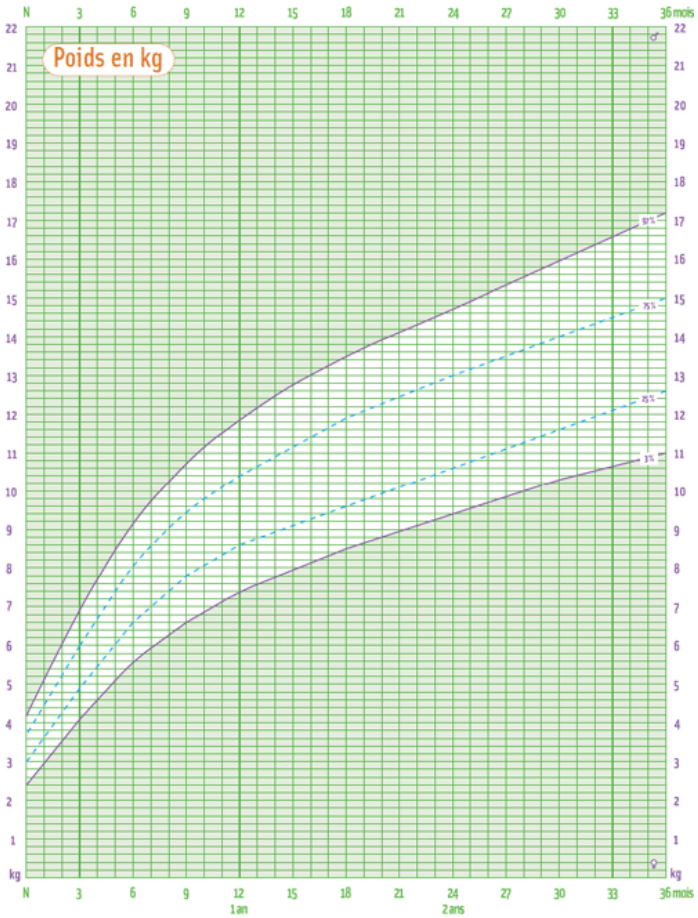
"Sortir de la dictature de la moyenne" : la plupart des études empiriques portent sur l'estimation d'effets moyens, mais la moyenne ne contient qu'une petite partie de l'information.

- Analyse des inégalités
  ex : stabilité du revenu moyen sur les dernières années, progression des derniers percentiles
- En terme d'évaluation des politiques publiques :
  Une mesure peut avoir un impact moyen nul mais être jugée "souhaitable" si elle affecte positivement *suffisamment* de personnes, ou *suffisamment certaines* personnes (exemples : échec scolaire, exclusion...)

# Aller au-delà de la moyenne

- La grande majorité des études empiriques s'intéressent à la moyenne de variable d'intérêt en fonction de déterminants observés : on modélise $E(Y|X)$
- Mais ces déterminants $X$ peuvent avoir un impact plus général sur la forme de la distribution
  Exemple : courbe de croissance (taille/poids en fonction de l'âge)

# La distribution des poids en fonction de l'âge



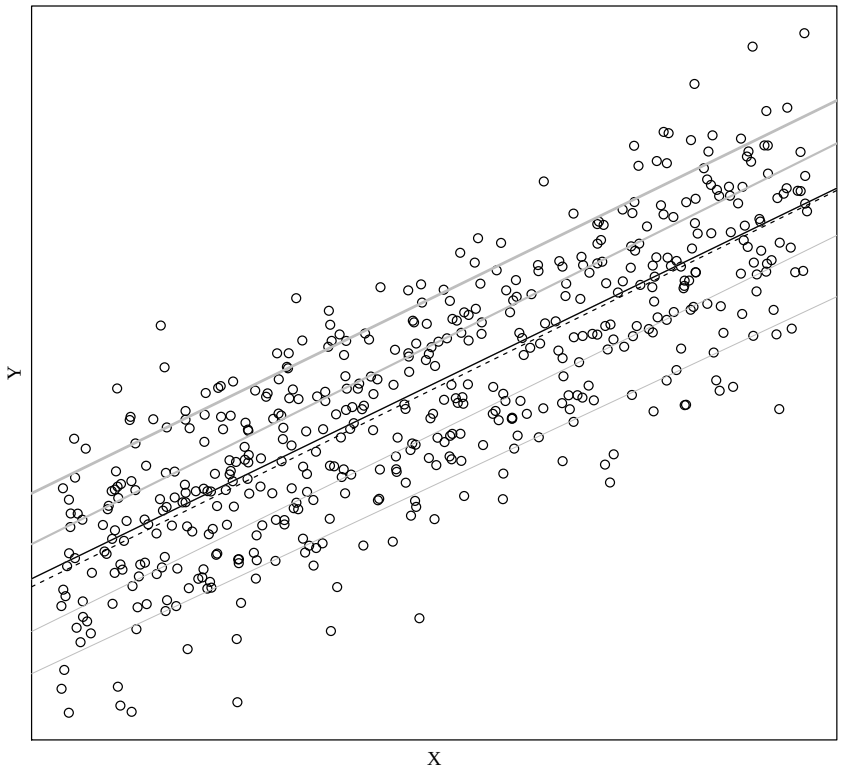

- Permet de vérifier que la croissance d'un enfant est "normale"
- Distribution des poids conditionnelle à l'âge
- Pour chaque âge (en abscisse), on représente les $3^{eme}$ et $97^{eme}$ percentiles, et premier et dernier quartiles.

# Modélisation des quantiles conditionnels

- Pour une v.a. $Y$ de distribution $F$ ($F(y) = P(Y < y)$), $\tau^{ieme}$ quantile : $q_\tau(Y) = inf\{y : F(y) \geq \tau\}$.
  soit si $F$ est continue $P(Y < q_\tau(Y)) = \tau$
- Remarque : il s'agit de la valeur telle que la probabilité d'observer une valeur de $Y$ inférieure dans la population soit $\tau$ et NON l'ensemble des personnes qui ont une valeur $Y$ inférieure à cette valeur
- On s'intéresse ici aux quantiles des distributions conditionnelles $F_{Y|X}$, notés $q_\tau(Y|X)$
- On utilise la modélisation :

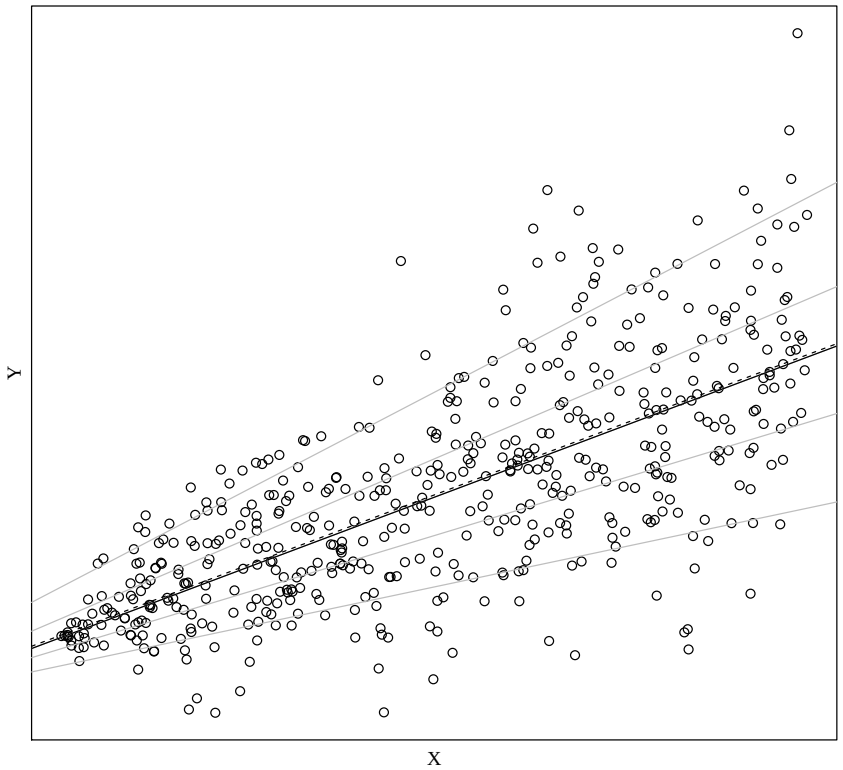
$$q_\tau(Y|X) = X'\beta_\tau$$

# Un exemple où la régression quantile apporte peu : modèle de translation




- On suppose que le modèle sous-jacent est simplement : $Y = X'\beta + U$
- En gris : droite de régression quantile pour les déciles d'ordre 1,3,5,7,9
  En pointillé : droite de régression MCO... de pente identique ($=\beta$)
- Dans ce cas, la régression quantile n'apporte pas grand chose... mais il s'agit d'un modèle très restrictif

# Un exemple où elle est plus utile : modèle de translation/échelle




- On ajoute un peu d'heteroscédasticité : $Y = X'\beta + (X'\gamma)U$
- En gris : droite de régression quantile pour les déciles d'ordre 1,3,5,7,9
  En pointillé : droite de régression MCO...
- Les régressions quantiles successives permettent de capter la dispersion croissance de $Y$ avec $X$ (les coefficients des régressions $\beta_{0.1}$,... $\beta_{0.9}$ augmentent).

# Répondre aux problèmes soulevés par la nature de certaines variables

- Moindre sensibilité que la moyenne à la présence de valeurs extrêmes.
- Données censurées, modèle Tobit... Une propriété intéressante des quantiles est l'équivariance par transformation monotone : si $h$ est une fonction croissante, $q_\tau(h(Y)|X) = h(q_\tau(h(Y)|X))$ (ce qui n'est pas le cas pour la moyenne !).

# Plan

# Principe

- Il est utile de voir les quantiles comme la solution d'un programme de minimisation.
- le quantile empirique $\widehat{q}_\tau(Y)$ satisfait :

$$\widehat{q}_\tau(Y) = \arg\min_b \sum_{i:Y_i \geq b} \tau|Y_i - b| + \sum_{i:Y_i < b} (1-\tau)|Y_i - b|$$

- Intuition : pour $\tau = 0.9$ par exemple, on pondère neuf fois plus les observations plus élevées que les plus faibles.
- ou encore : $\arg\min_b \sum_i \rho_\tau(Y_i - b)$ la fonction de pondération s'appelle la fonction de perte ("check function") :

$$\rho_\tau(u) = u(\tau - 1(u < 0))$$

# Estimation

- Principe de la régression quantiles : on cherche à modéliser le quantile

$$q_\tau(Y|X) = X'\beta_\tau$$

- On estime donc $\beta_\tau$ par :

$$\widehat{\beta}_\tau = \arg\min_\beta \sum \rho_\tau(Y_i - X_i'\beta)$$

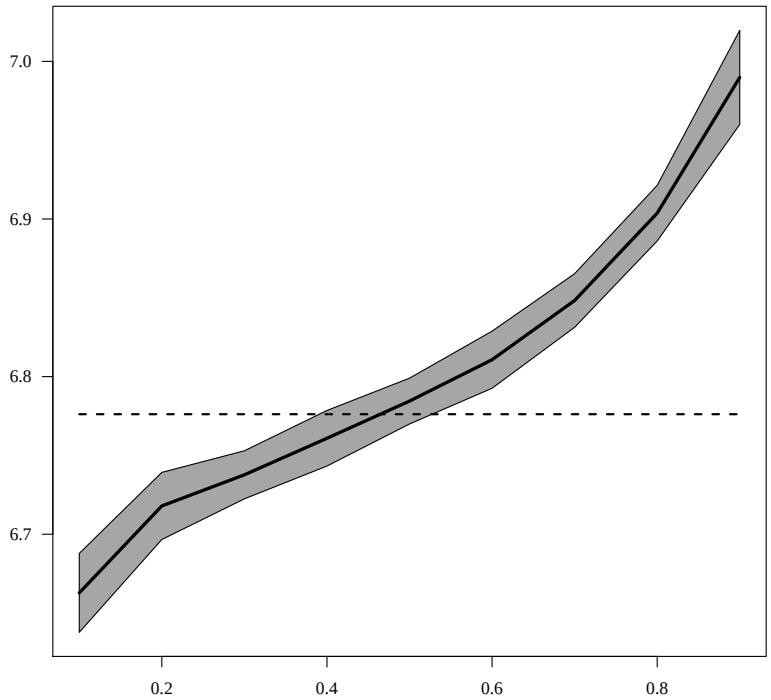
- Estimation maintenant standard sous SAS (quantreg),R (rq) et Stata (qreg, sqreg).

# Plan

# Exemple

- Estimation d'une équation de salaires à partir de l'enquête Emploi en continu 2008
- Modélisation des différents déciles du log du salaire en fonction du nombre d'années d'études, de l'expérience potentielle, du sexe, de la nationalité
- Remarque : on a un jeu de cœfficients par décile... on choisit une présentation graphique

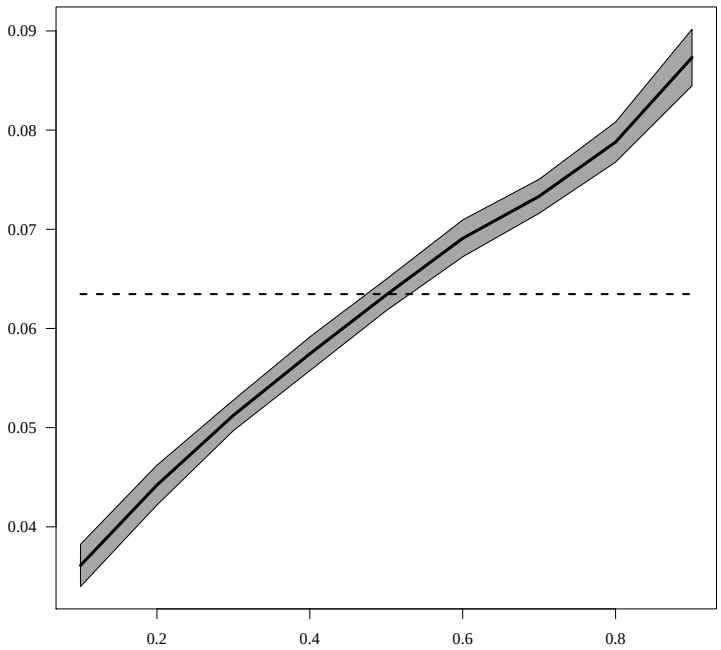
# Résultats : constante




La zone grisée correspond à l'intervalle de confiance à 95% ; la courbe pointillée à l'estimation obtenue par les moindres carrés ordinaires (moyenne)

# Cas d'une variable continue : Nombre d'années d'études

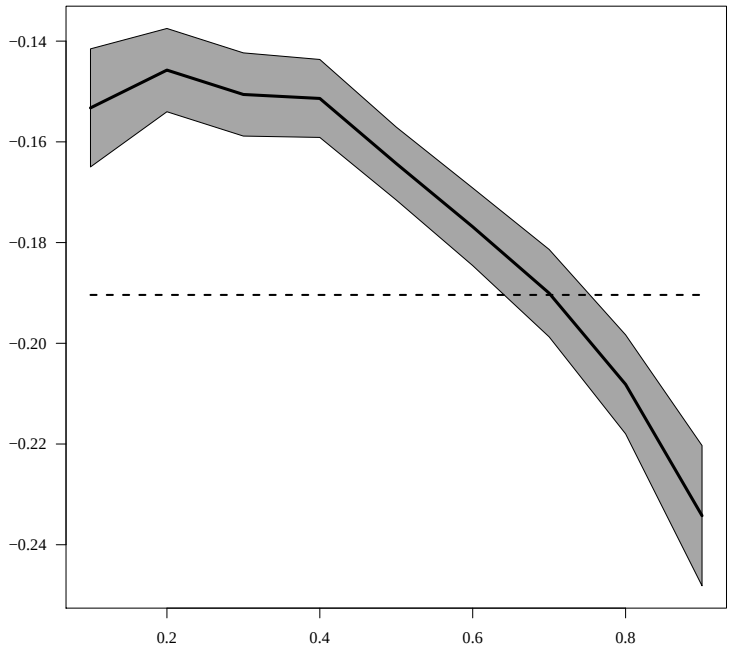
Le coefficient estimé pour un décile correspond à l'augmentation marginale de ce décile liée à une augmentation marginale du nombre d'années d'études




- les cœfficients sont toujours positifs : le niveau d'étude décale la distribution des salaires vers le haut...
- cet écart augmente avec le décile : la dispersion des salaires augmente avec le nombre d'années d'études

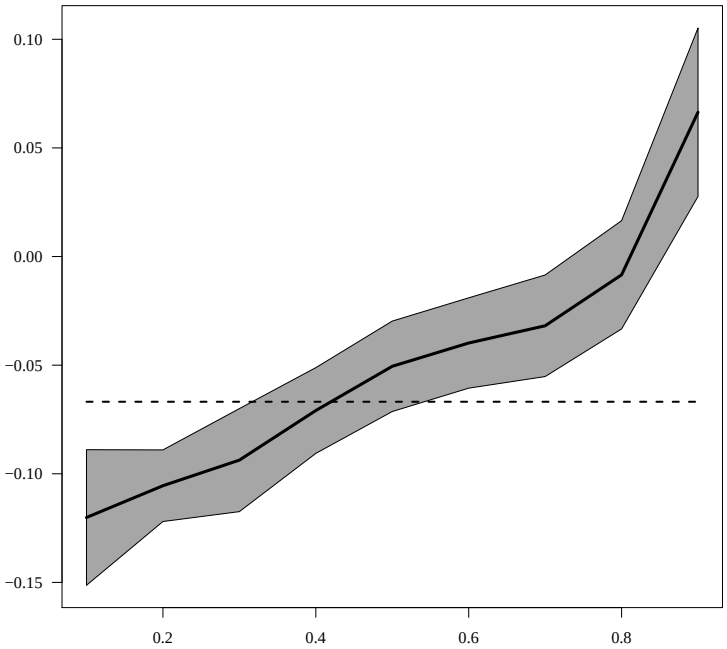
# Cas d'une variable discrète : Genre (= femme)

Le coefficient estimé correspond à l'écart entre le décile de la distribution de salaires conditionnelle des hommes et le décile de celle des femmes




- les salaires des femmes sont toujours inférieurs à ceux des hommes
- cet écart augmente dans le haut de la distribution

# Cas d'une variable discrète : Nationalité ($\neq$ française)




- les premiers déciles des distributions de salaires des salariés n'ayant pas la nationalité française sont nettement inférieurs à ceux correspondant aux salariés ayant la nationalité française
- mais les deux distributions se rapprochent ensuite

# Remarque 1 : Endogénéité

- Les régressions quantiles permettent d'analyser l'ensemble de la distribution de $Y$ en fonction de certaines variables explicatives
- Ne règle aucun des problèmes d'endogénéité éventuelle de ces variables
- Estimer l'effet causal de celles-ci demande d'utiliser des méthodes spécifiques
- Extensions des méthodes classiques (présentées dans le document méthodo) : variable instrumentale, effets fixes (données de panel), contrôle des observables...

# Remarque 2 : pas d'interprétation individuelle

1. Attention, le premier décile est strictement la valeur du salaire telle que 10% de la population a une valeur inférieure.
2. En s'en tenant à la stricte définition statistique, il ne s'agit pas des 10% de la population avec le salaire le moins élevé
3. Les régressions quantiles fournissent une description de la manière dont le décile se modifie en fonction de certaines variables explicatives...
4. ... pas sur l'impact de ces variables explicatives sur les personnnes qui sont situées dans ces quantiles.

# Remarque 3 : Distribution conditionnelle vs inconditionnelle

- On modèlise la distribution conditionnelle aux variables explicatives...
- N'indique pas directement comment la distribution de $Y$ se modifie lorsque la distribution de certains $X$ évolue
  Exemple : déterminer ce qui relève de l'évolution de la qualification dans les inégalités
- du fait de la non linéarité des quantiles ($\neq$ espérance) :

$$E_X(q_\tau(Y|X)) \neq q_\tau(Y)$$

- Mais on peut adapter les régressions

# En conclusion

- L'utilisation des régressions quantiles se diffuse très rapidement
- Elles sont relativement simples d'utilisation
- Elles proposent une description plus riche, et donc aussi plus complexe à analyser
- Elles ne permettent pas de répondre à toutes les questions qu'on pourrait avoir sur l'analyse des distributions, donc attention à l'interprétation
- Domaine en pleine expansion...