La régression quantile en pratique

Xavier D'haultfœuille (CREST) et Pauline Givord (INSEE-DM)

Journées de Méthodologie Statistique

24 Janvier 2012

Plan

Plan

Introduction

- Les régressions quantiles sont un outil dont l'usage s'est généralisé récemment

Introduction

- Les régressions quantiles sont un outil dont l'usage s'est généralisé récemment
- Ce document présente un mode d'emploi pratique, à destination des chargés d'études

Introduction

- Les régressions quantiles sont un outil dont l'usage s'est généralisé récemment
- Ce document présente un mode d'emploi pratique, à destination des chargés d'études
- Beaucoup de développement récent, donc ne vise pas à l'exhaustivité !

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 aux US mais forte progression du dernier décile (Buchinsky, 1998,...)

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 aux US mais forte progression du dernier décile (Buchinsky, 1998,...)
- 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...)

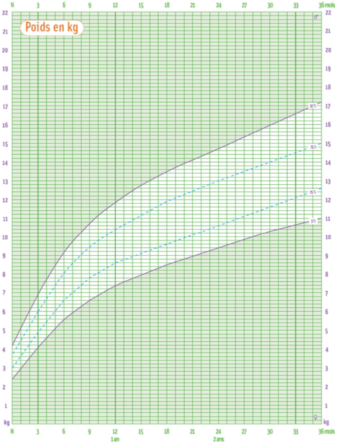
Aller au-delà de la moyenne

- La grande majorité des études empiriques s'intéressent à la moyenne de variable d'intérêt en fonction de déterminants observés : on modélise $E(Y|X)$

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 Quetelet (taille/poids en fonction de l'âge)

La distribution des poids en fonction de l'âge



Modélisation des quantiles conditionnels

- Pour une v.a. Y de distribution F ($F(y) = P(Y < y)$), τ^{ieme} quantile : $Q_\tau(Y) = inf\ \{y : F(y) \geq \tau\}$.

Modélisation des quantiles conditionnels

- Pour une v.a. Y de distribution F ($F(y) = P(Y < y)$), τ^{ieme} quantile : $Q_\tau(Y) = inf\,\{y : F(y) \geq \tau\}$.
- On s'intéresse ici aux quantiles des distributions conditionnelles $F_{Y|X}$, notés $q_\tau(Y|X)$

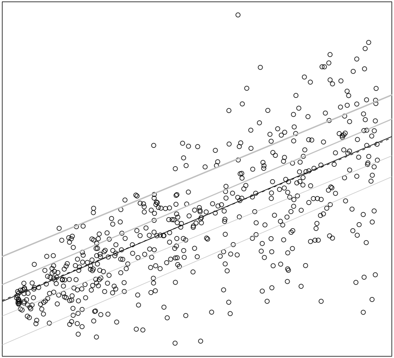
Modélisation des quantiles conditionnels

- Pour une v.a. Y de distribution F $(F(y) = P(Y < y))$, τ^{ieme} quantile : $Q_\tau(Y) = inf\ \{y : F(y) \geq \tau\}$.
- On s'intéresse ici aux quantiles des distributions conditionnelles $F_{Y|X}$, notés $q_\tau(Y|X)$
- on utilise la modélisation :

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

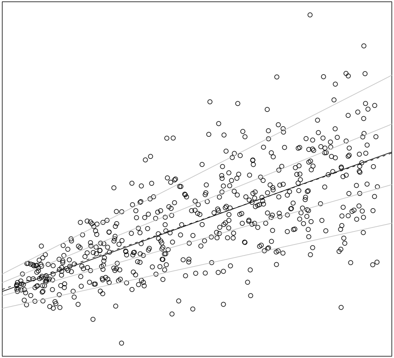
Exemple : modéle de translation

On suppose que le modèle sous-jacent est simplement :
$Y = X'\beta + U$



Exemples : modéle de translation/échelle

Si on ajoute un peu d'homoscédasticité : $Y = X'\beta + (X'\gamma)U$



Répondre aux problèmes soulevés par la nature de certaines variables

- Moindre sensibilité que la moyenne à la présence de valeurs extrêmes.

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.

Principe

- Il est utile de voir les quantiles comme la solution d'un programme de minimisation.
- Cas général :

$$argmin_b \sum_{i:Y_i \geq b} \tau |Y_i - b| + \sum_{i:Y_i < b} (1-\tau)|Y_i - b|$$

Principe

- Il est utile de voir les quantiles comme la solution d'un programme de minimisation.
- Cas général :

$$argmin_b \sum_{i:Y_i \geq b} \tau |Y_i - b| + \sum_{i:Y_i < b} (1-\tau)|Y_i - b|$$

- Intuition : pour $\tau = 0.9$ par exemple, on pondère neuf fois plus les observations plus élevées que les plus faibles.

Principe

- Il est utile de voir les quantiles comme la solution d'un programme de minimisation.
- Cas général :

$$argmin_b \sum_{i:Y_i \geq b} \tau |Y_i - b| + \sum_{i:Y_i < b} (1-\tau)|Y_i - b|$$

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

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

Principe

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

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

Principe

- Principe de la régression quantiles : on cherche à modéliser le quantile
$$q_\tau(Y|X) = X'\beta_\tau$$
- On remplace donc dans le programme :

$$\beta_\tau = \arg\, min_\beta \sum \rho_\tau(Y_i - X_i\beta)$$

Principe

- Principe de la régression quantiles : on cherche à modéliser le quantile
$$q_\tau(Y|X) = X'\beta_\tau$$
- On remplace donc dans le programme :
$$\beta_\tau = \arg min_\beta \sum \rho_\tau(Y_i - X_i\beta)$$
- Remarque : équivalent à la démarche MCO, qui modélise l'espérance conditionnelle $E(Y|X)$ à partir de la fonction de perte quadratique :
$$\beta = argmin_\beta E[(Y - X'\beta)^2]$$

Estimation

- La fonction objectif non différentiable, donc la procédure du gradient classique non utilisable

Estimation

- La fonction objectif non différentiable, donc la procédure du gradient classique non utilisable
- On peut l'écrire comme solution d'un modèle de programmation linéaire (Koenker et Bassett, 1978).

Estimation

- La fonction objectif non différentiable, donc la procédure du gradient classique non utilisable
- On peut l'écrire comme solution d'un modèle de programmation linéaire (Koenker et Bassett, 1978).
- Implémentation maintenant standard sous stata (qreg, sqreg) ou R (rq), sas (quantreg).

Inférence

- Il n'existe pas de forme explicite de l'estimateur

Inférence

- Il n'existe pas de forme explicite de l'estimateur
- On peut montrer que la variance asymptotique s'écrit :

$$\sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \to N(0, \Lambda_\tau)$$

avec
$$\Lambda_\tau = \tau(1-\tau)(E[f_{u_\tau}(0|X)X'X])^{-1}E[X'X]E[f_{u_\tau}(0|X)X'X]^{-1}$$

Inférence

- Il n'existe pas de forme explicite de l'estimateur
- On peut montrer que la variance asymptotique s'écrit :

$$\sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \to N(0, \Lambda_\tau)$$

avec
$\Lambda_\tau = \tau(1-\tau)(E[f_{u_\tau}(0|X)X'X])^{-1}E[X'X]E[f_{u_\tau}(0|X)X'X]^{-1}$

- Deux solutions :

Inférence

- Il n'existe pas de forme explicite de l'estimateur
- On peut montrer que la variance asymptotique s'écrit :

$$\sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \rightarrow N(0, \Lambda_\tau)$$

avec
$\Lambda_\tau = \tau(1-\tau)(E[f_{u_\tau}(0|X)X'X])^{-1}E[X'X]E[f_{u_\tau}(0|X)X'X]^{-1}$

- Deux solutions :
  1. Simplification si les termes d'erreurs u_τ sont indépendants de x :

$$\Lambda_\tau = \frac{\tau(1-\tau)}{f_{u_\tau}^2(0)} E[X'X]^{-1}$$

Inférence

- Il n'existe pas de forme explicite de l'estimateur
- On peut montrer que la variance asymptotique s'écrit :

$$\sqrt{n}(\hat{\beta}_\tau - \beta_\tau) \to N(0, \Lambda_\tau)$$

  avec
  $\Lambda_\tau = \tau(1-\tau)(E[f_{u_\tau}(0|X)X'X])^{-1}E[X'X]E[f_{u_\tau}(0|X)X'X]^{-1}$
- Deux solutions :
  1. Simplification si les termes d'erreurs u_τ sont indépendants de x :
  $$\Lambda_\tau = \frac{\tau(1-\tau)}{f^2_{u_\tau}(0)} E[X'X]^{-1}$$
  2. Bootstrap : plus général mais plus coûteux en temps.

Interprétation des résultats

- Cas général : β_j mesure $\frac{\partial EQ_\tau(Y|X)}{\partial X_j}$, soit le changement marginal du τ^{ieme} quantile suite à un changement marginal de X_j.

Interprétation des résultats

- Cas général : β_j mesure $\frac{\partial EQ_\tau(Y|X)}{\partial X_j}$, soit le changement marginal du τ^{ieme} quantile suite à un changement marginal de X_j.
- Variable binaire : β_j mesure l'écart entre les deux distributions (conditionnelles aux autres observables).

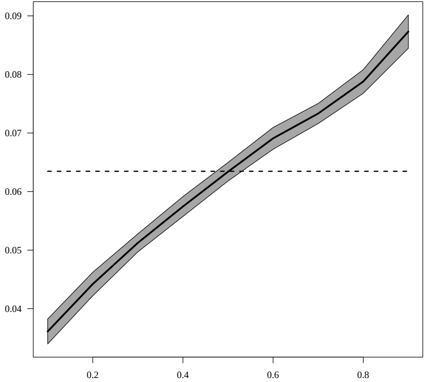
Illustration

- Estimation d'une équation de salaires à partir de l'enquête Emploi en continu 2008

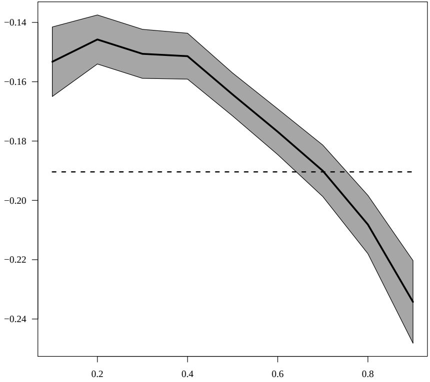
Illustration

- 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'étude, de l'expérience potentielle, du sexe, de la nationalité

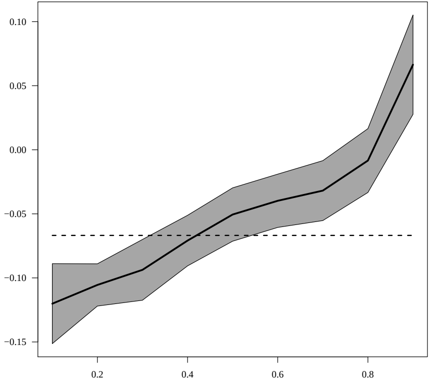
Effet du nombre d'année d'études



Effet du fait d'être une femme



Effet du fait d'être une femme



Remarque 1 : le problème d'endogénéité

- La régression quantiles permet d'avoir un diagnostic plus complet sur l'influence d'une (des) observables,

Remarque 1 : le problème d'endogénéité

- La régression quantiles permet d'avoir un diagnostic plus complet sur l'influence d'une (des) observables,
- Elle ne règle aucun des problèmes d'endogénéité éventuels qui peuvent survenir

Remarque 1 : le problème d'endogénéité

- La régression quantiles permet d'avoir un diagnostic plus complet sur l'influence d'une (des) observables,
- Elle ne règle aucun des problèmes d'endogénéité éventuels qui peuvent survenir
- Plusieurs extensions ont été proposées. Nous présentons dans le document une méthode de régressions de quantile instrumentées proposée par Chernozhukov et Hansen (2009), avec une application.

Remarque 2 : pas d'interprétation individuelle

- Les estimations comparent les quantiles des distributions conditionnelles entre eux

Remarque 2 : pas d'interprétation individuelle

- Les estimations comparent les quantiles des distributions conditionnelles entre eux
- Par exemple, pour une variable binaire, on estime la différence des quantiles des distributions conditionnelles, mais pas la distribution de la différence.

Remarque 2 : pas d'interprétation individuelle

- Les estimations comparent les quantiles des distributions conditionnelles entre eux
- Par exemple, pour une variable binaire, on estime la différence des quantiles des distributions conditionnelles, mais pas la distribution de la différence.
- A priori, pas d'interprétation individuelle : une personne dont le revenu se situe dans le quantile τ de la distribution conditionnelle correspondant à $X = x$ pourrait avoir un revenu qui se situe à un autre niveau de la distribution conditionnelle à $X = x'$.

Remarque 2 : pas d'interprétation individuelle

- Les estimations comparent les quantiles des distributions conditionnelles entre eux
- Par exemple, pour une variable binaire, on estime la différence des quantiles des distributions conditionnelles, mais pas la distribution de la différence.
- A priori, pas d'interprétation individuelle : une personne dont le revenu se situe dans le quantile τ de la distribution conditionnelle correspondant à $X = x$ pourrait avoir un revenu qui se situe à un autre niveau de la distribution conditionnelle à $X = x'$.
- Pour passer à l'interprétation individuelle, il faut faire une hypothèse d'invariance des rangs

Remarque 3 : revenu conditionnel

- Les estimations comparent les quantiles des distributions conditionnelles entre eux

Remarque 3 : revenu conditionnel

- Les estimations comparent les quantiles des distributions conditionnelles entre eux
- Mais les quantiles n'ont pas de propriétés de linéarité :

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

Pas d'interprétation directe sur la distribution de la variable d'intérêt d'un changement dans la distribution des covariates

Remarque 3 : revenu conditionnel

- Les estimations comparent les quantiles des distributions conditionnelles entre eux
- Mais les quantiles n'ont pas de propriétés de linéarité :

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

  Pas d'interprétation directe sur la distribution de la variable d'intérêt d'un changement dans la distribution des covariates
- Il sera plus complexe d'obtenir de telles distributions contrefactuelles. On peut préférer mobiliser des méthodes spécifiques.