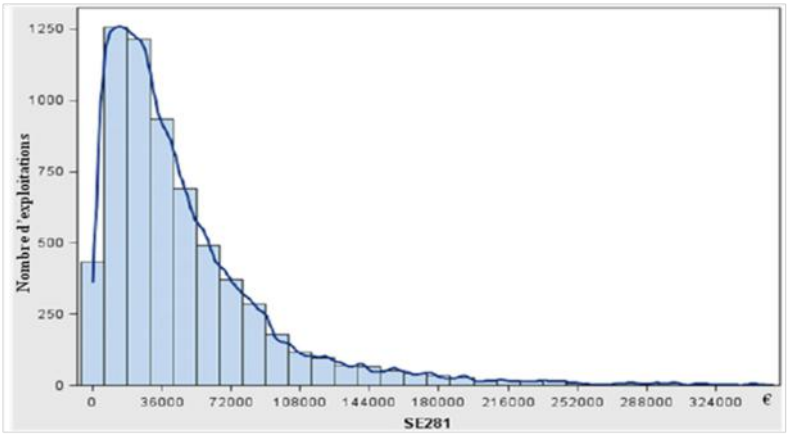
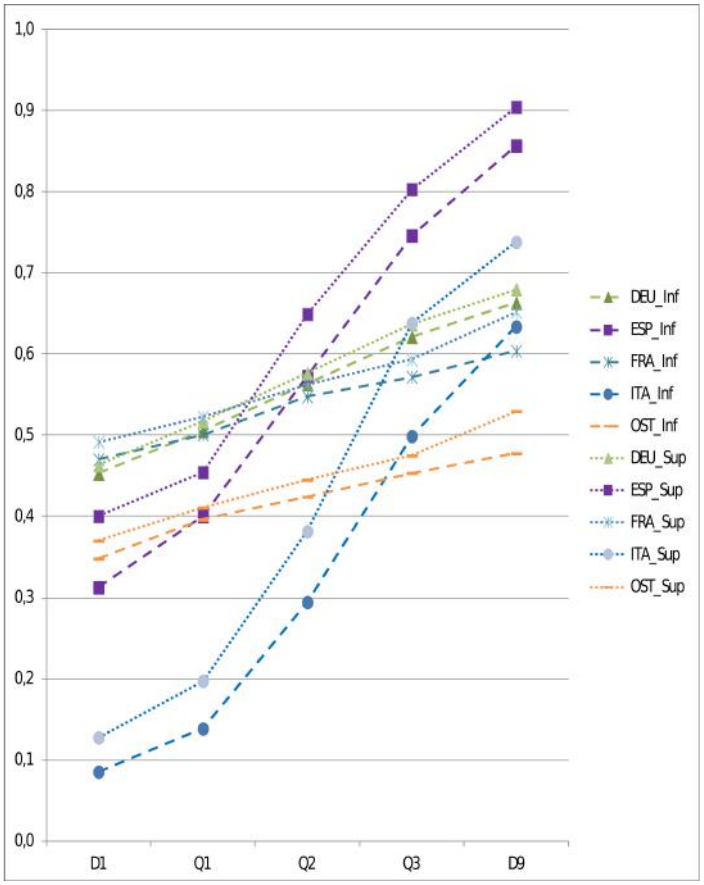
Introduction à la régression quantile



$$\underset{\beta}{Min}\left\{\sum_{i\, x_i \geq y_i'\beta} q\left|x_i - y_i'\beta\right| + \sum_{i\, x_i \leq y_i'\beta} (1-q)\left|x_i - y_i'\beta\right|\right\}$$



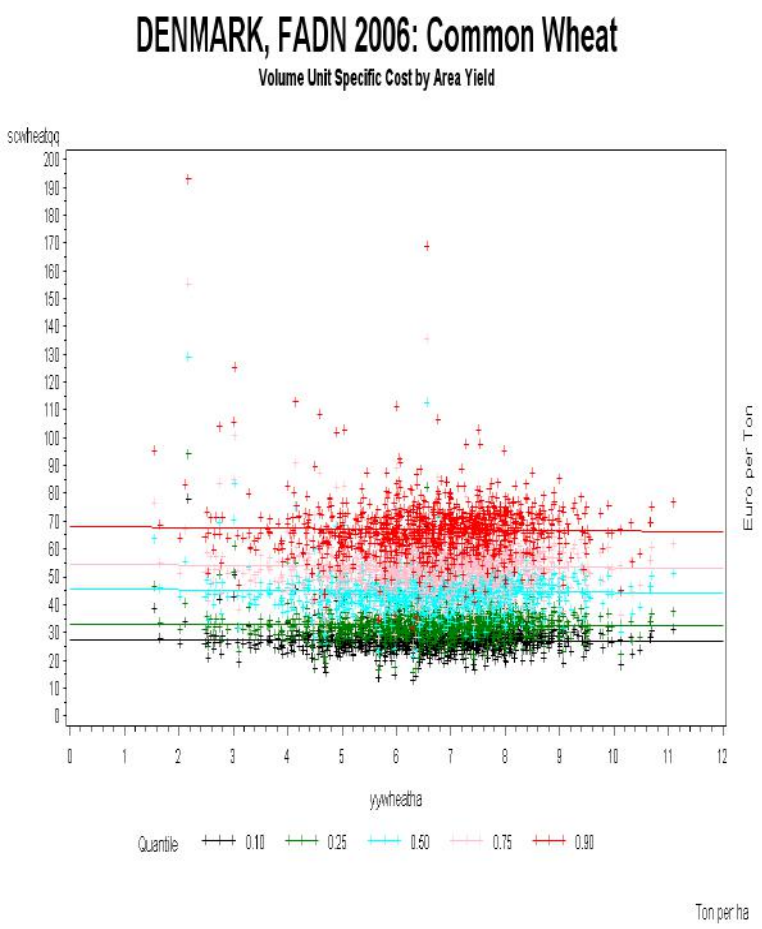
Dominique Desbois, UMR Economie publique, INRA-AgroParisTech

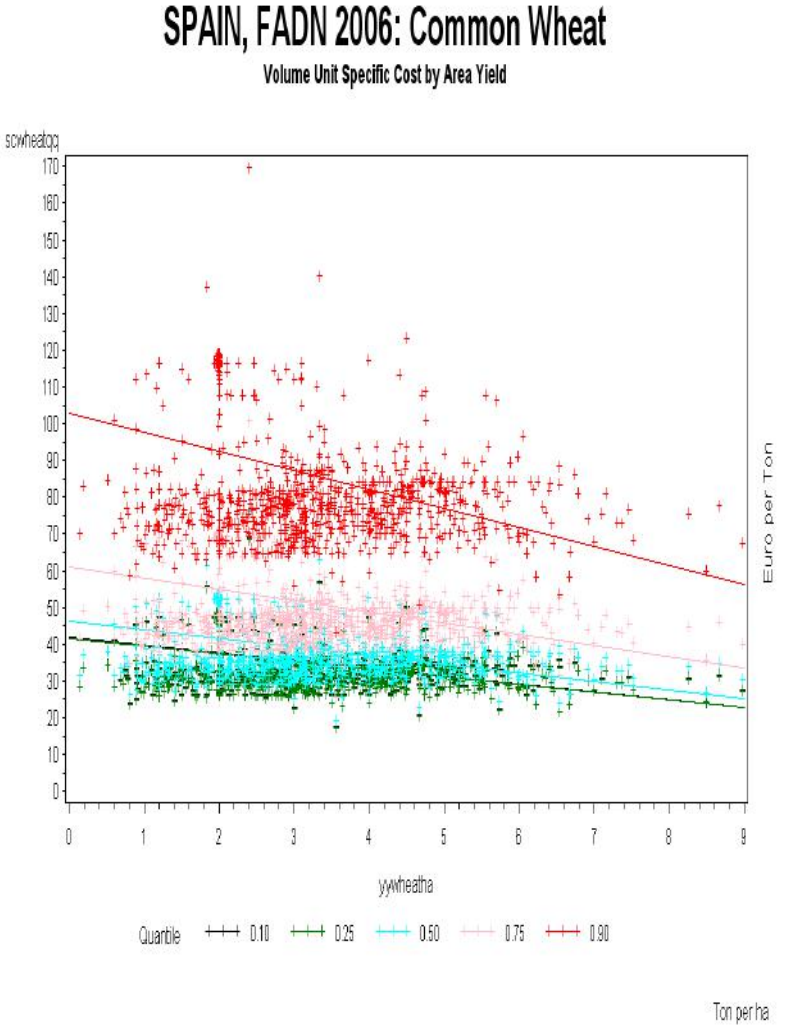
10es Journées de Recherches en Sciences Sociales
8 - 9 décembre 2016
PARIS, La Défense
JRSS 2016

Introduction à la Régression quantile : structure du tutoriel

- Problématique :
  - Exemples ;
- Méthodologie :
  - Concepts, modèles et propriétés ;
  - Estimation et inférence ;
  - Algorithmes et implantation ;
- Logiciels :
  - exemples sous SAS ;
  - exemple sous Stata ;
  - exemples sous R ;
- Résultats :
  - Exemples ;
- Développements
- Bibliographie

Problématique de la régression quantile : pertinence

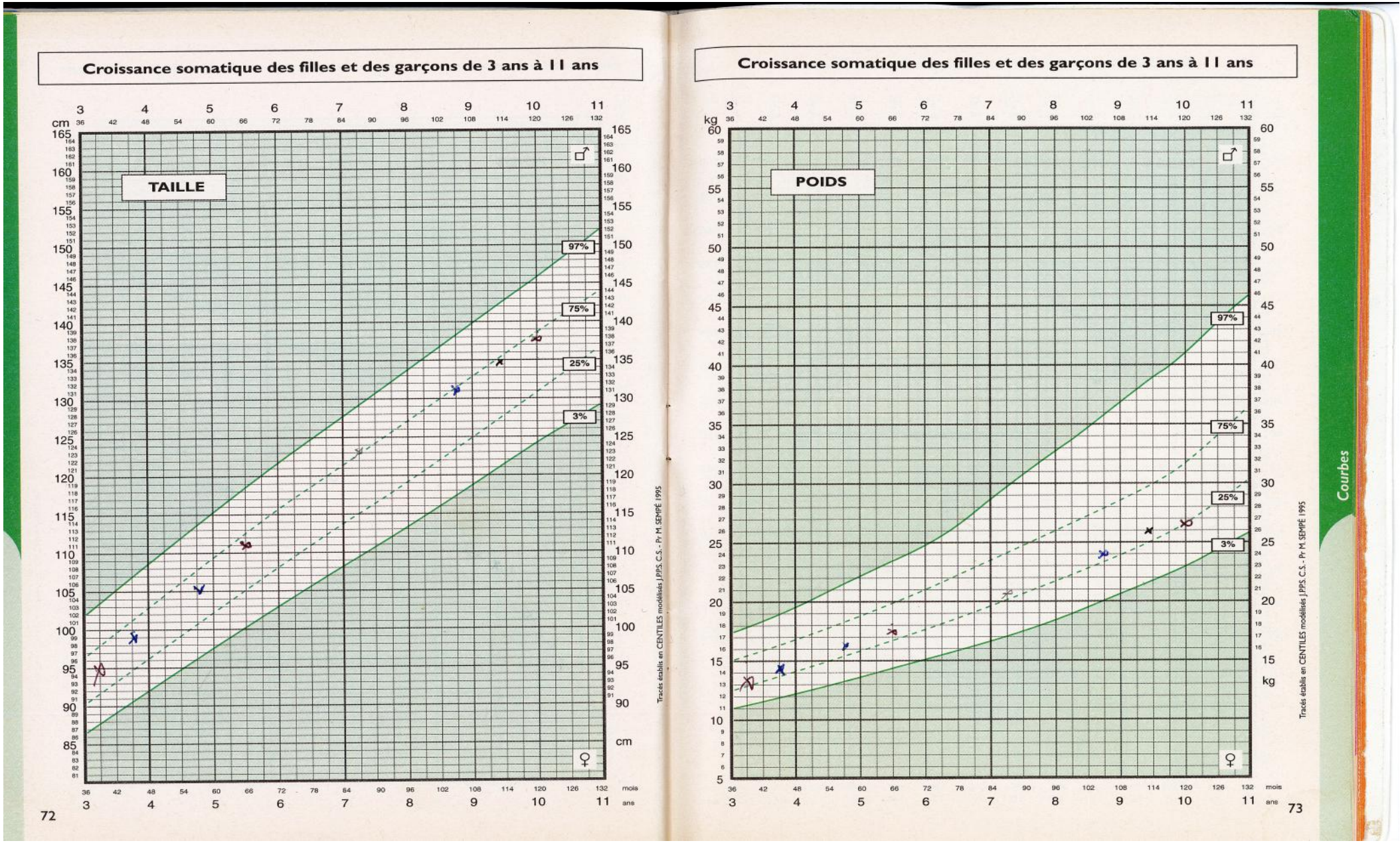




- La distribution de y est asymétrique relativement à la moyenne

- Les données sont de nature hétéroscédastique

Problématique en régression quantile : courbes de croissance



Les courbes de croissance peuvent être utilisées pour positionner des mesures individuelles par rapport à l'ensemble de la distribution balayée par les quantiles représentés.

Méthodologie : estimation

- Quantiles d'une distribution univariée:
  - Echantillon aléatoire :

$$y_1, y_2, \ldots, y_i, \ldots, y_n \ ;$$

  - Médiane :

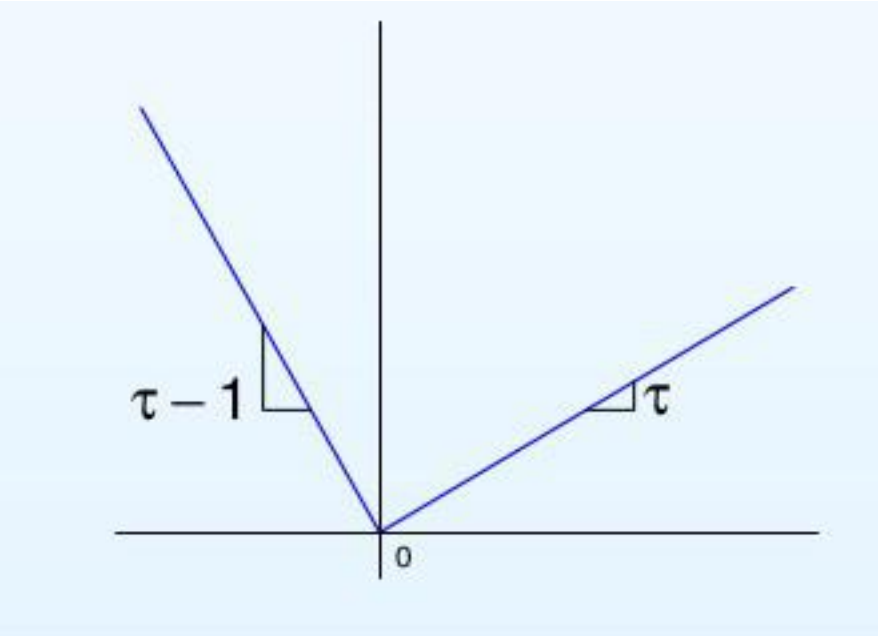
$$argmin_\theta \sum |y_i - \theta|$$

  - Fonction de perte :

$$\rho_\tau(\delta) = (\tau - 1_{\delta<0}) * \delta$$

  - Quantile d'ordre tau :

$$argmin_\theta \sum \rho_\tau(y_i - \theta)$$



Les quantiles peuvent être définis comme des optima minimisant une fonction de perte.

Méthodologie : quantile conditionnel d'ordre τ

- Modèle linéaire pour le quantile conditionnel d'ordre τ :

  - Modèle linéaire : $y_i = x_i^T \beta_\tau + e_i \qquad i = 1, \ldots, n$

    tel que le quantile univarié d'ordre τ des écarts soit nul

  - Quantile conditionnel multivarié :

    $$Q_\tau(Y / X) = X^T \beta_\tau$$

  - Estimateur quantile :

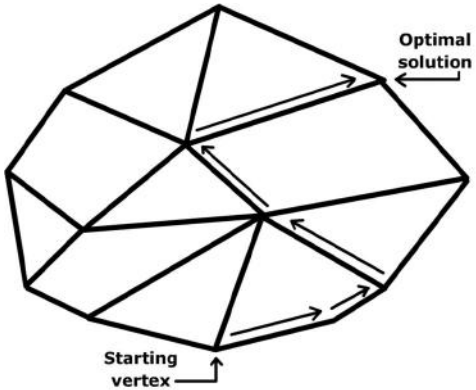
    $$\hat{\beta}_\tau = argmin_{\beta \in \mathbb{R}^p} \sum \rho_\tau(y_i - x_i^T \beta)$$

  - Régression quantile

    $$\hat{Q}_\tau(Y / X) = X^T \hat{\beta}_\tau$$

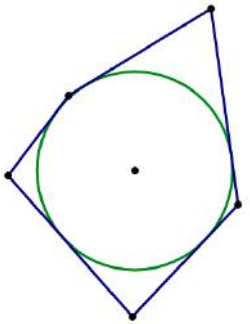
Méthodologie : Algorithmes de Calculs

• Pour de petits volumes de données : méthode du simplexe (Koenker et d'Orey, 1987 et 1993)



◦ initialisation avec un sommet aléatoire du polytope convexe et une recherche sur les arêtes d'un polygone circonscrit.

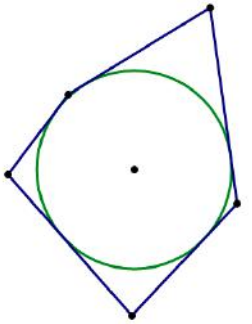
• Pour de gros volumes de données : méthode du point intérieur (Portnoy et Koenker, 1997) à complexité de cacul polynomiale



◦ utilise l'algorithme de Karmakar (1984) pour l'optimisation linéaire en passant par l'intérieur de l'ensemble convexe des solutions réalisables.

Méthodologie : Algorithmes utilisés

• Pour de très gros volumes de données (n >100 000) : méthode du point intérieur avec prétraitement (Portnoy et Koenker, 1997)



• recherche de l'optimum par lissage de la fonction de perte (Chen, 2004)

$$\rho_\tau(.)$$

Inférence en régression quantile

• L'inférence en régression quantile peut s'effectuer à partir des trois familles de méthodes suivantes :

◦ l'estimation directe de la matrice de variance-covariance

L'estimation directe converge asymptotiquement ; elle est de faible complexité algorithmique mais elle est sensible aux paramètres du lissage ;

◦ les scores de rang

La méthode des scores de rang ne converge asymptotiquement que pour certains modèles (*iid*) ; elle est de complexité algorithmique plus élevée mais elle est robuste par rapport aux hypothèses courantes ;

◦ le rééchantillonnage :

Le rééchantillonnage converge asymptotiquement ; sa complexité algorithmique est la plus élevée, cependant il est algorithmiquement efficace en grande dimension ;

Inférence en régression quantile : la pratique

• Les recommandations pratiques en matière d'inférence sont :

◦ utiliser la méthode des scores pour des problèmes d'estimation de relativement faible dimension
(*n* < 1000 et *p* < 10) ;

◦ utiliser la méthode MCMB pour des problèmes d'estimation d'assez grande dimension
(10 000 < *np* < 2 000 000) ;

◦ utiliser les méthodes d'estimation directe pour les problèmes non *iid* de très grande dimension
(He et Hu, 2002) ;

Ces recommandations pratiques (Kocherginsky, He et Mu, 2005) sont les options par défaut de la proc QUANTREG de SAS

Logiciel : implantation sous SAS (Windows)

- Package SAS/STAT, procédure QUANTREG, version SAS 9.2

◦ documentation :
https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#qreg_toc.htm

- Syntaxe élémentaire :

```
PROC QUANTREG DATA = table_sas <options>;
BY variables;
CLASS variables;
MODEL var_expliquee = covariants </ options>;
RUN ;
```

Logiciel : implantation sous SAS (Windows)

- Pour spécifier le niveau du quantile :
  - utiliser l'option QUANTILE de l'instruction MODEL :

    MODEL Y = X / QUANTILE = < number list | ALL > ;

  - exemples

Pour un seul quantile (premier quartile) : QUANTILE = 0.25 ;
Pour plusieurs quantiles (les trois quartiles) : QUANTILE = 0.25 0.5 0.75 ;
Pour l'ensemble du processus quantile : QUANTILE = ALL ;

  - par défaut, estimation médiane : QUANTILE = 0.5 ;

Logiciel : inférence en régression quantile sous SAS

• Par défaut, la méthode des scores de rang est utilisée pour les petits échantillons ($n < 5\,000$ et $p < 20$), sinon la méthode utilisée est le ré-échantillonnage :

◦ spécification de la proc QUANTREG

PROC QUANTREG CI = <NONE | RANK |...> ALPHA = valeur ;

◦ méthode estimation directe

Modèlel iid CI = SPARCITY / IID

Modèle non iid CI = SPARCITY ;

◦ méthode des rangs CI = RANK ;

◦ méthode de rééchantillonnage CI = RESAMPLING ;

Implantation sous SAS (Windows)

- Pour spécifier l'algorithme :

◦ utiliser l'option ALGORITHM de l'instruction PROC QUANTREG :

PROC QUANTREG DATA = table_sas ALGORITHM=<option> ;

◦ options

Algorithme du simplexe :	ALGORITHM = SIMPLEX ;
Algorithme du point intérieur :	ALGORITHM = INTERIOR ;
Algorithme du point intérieur avec prétraitement :	ALGORITHM = INTERIOR PP ;
Algorithme de lissage :	ALGORITHM = SMOOTHING ;

◦ par défaut, algorithme du simplexe

Implantation sous SAS : code

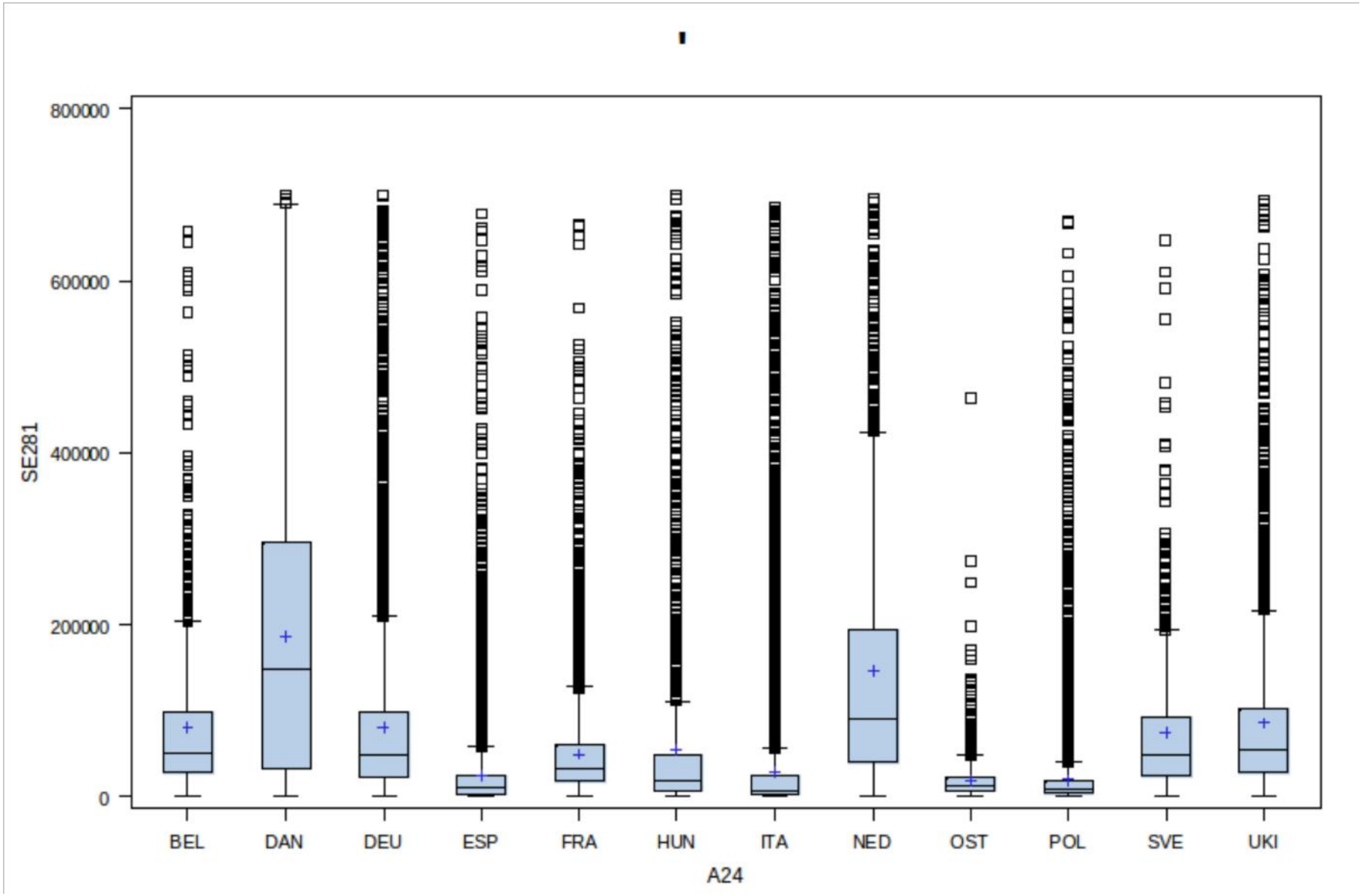
- **application coûts spécifiques :**

```
* regression quantile ;
ods html ;
ods graphics on ;

proc quantreg  data=fra2006t alpha= 0.05 ci=resampling algorithm=simplex outest=oemargb ;
model SE281 =  WHEAT OTHCER DRYPU INDCROP OILSEED HORTIC ALLFRT  WINE
                    OTHCROP CATTL PIG__ EGGPOUL CMILK OTHLST  OTHACT
 / quantile= 0.01 0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09 0.10  0.11 0.12 0.13 0.14 0.15 0.16 0.17
   0.18 0.19 0.20 0.21 0.22 0.23 0.24 0.25 0.26 0.27 0.28 0.29 0.30 0.31 0.32 0.33 0.34 0.35 0.36
   0.37 0.38 0.39 0.40 0.41 0.42 0.43 0.44 0.45 0.46 0.47 0.48 0.49 0.50 0.51 0.52 0.53 0.54 0.55
   0.56 0.57 0.58 0.59 0.60 0.61 0.62 0.63 0.64 0.65 0.66 0.67 0.68 0.69 0.70 0.71 0.72 0.73 0.74
   0.75 0.76 0.77 0.78 0.79 0.80 0.81 0.82 0.83 0.84 0.85 0.86 0.87 0.88 0.89 0.90 0.91 0.92 0.93
   0.94 0.95 0.96 0.97 0.98 0.99
seed=12345  noint ;
weight sys02;
output out=omargb p=pmargb res=rmargb / columnwise ;
run;

ods graphics off ;
ods html close;
```

Application coûts spécifiques : asymétrie et hétérogénéité



La comparaison selon les pays met en évidence l'hétérogénéité de la distribution des coûts spécifiques et l'asymétrie des distributions nationales

Coûts spécifiques : liste des produits et des pays

Produits	
Blé	Bovins
Autres céréales	Porcins
Cultures industrielles	Volailles
Protéagineux	Lait de vache
Oléagineux	Autres productions animales
Productions horticoles	Autres produits animaux
Fruits	Autres produits
Vins	
Autres production végétales et forestières	

Pays sélectionnés	
Dénomination	*Codification*
Allemagne	DEU
Autriche	OST
Belgique	BEL
Danemark	DAN
Espagne	ESP
France	FRA
Hongrie	HUN
Italie	ITA
Pays-Bas	NED
Pologne	POL
Royaume-Uni	UKI
Suède	SVE

La décomposition du produit brut s'effectue selon 16 catégories de produits
Les douze pays européens sélectionnés représentent une part déterminante de la production européenne en blé, lait de vache et porc.
(Desbois, Butault et Surry, 2012 et 2015)

Coûts spécifiques : informations sur le modèle

Dependent Variable	SE281
Weight Variable	SYS02
Number of Independent Variables	15
Number of Observations	7740
Optimization Algorithm	Simplex
Method for Confidence Limits	Resampling

Number of Observations Read	7740
Number of Observations Used	7740
Sum of Weights	652147,8

Quantile and Objective Function	
Quantile	0,5
Objective Function	1100943897,9
Predicted Value at Mean	8978,0428

Les informations sur le modèle fournissent : i) l'identification de la variable d'intérêt et de la pondération, le nombre de régresseurs et d'observations, la méthode d'optimisation et de calcul des intervalles de confiance ; ii) le nombre d'observations lues et utilisées, et la somme des poids ; iii) les valeurs du quantile, de la fonction objectif et de la prédiction en moyenne.

Sorties SAS : résumé de statistiques descriptives

Summary Statistics						
Variable	**Q1**	**Median**	**Q3**	**Mean**	**Standard Deviation**	**MAD**
WHEAT	0	0	0	1164,4	4505,4	0
othcer	0	0	6513	5990,3	14096	0
DRYPU	0	0	0	150,6	1371,3	0
INDCROP	0	0	0	2615,9	17484,7	0
OILSEED	0	0	0	120	1174,3	0
HORTIC	0	0	0	12036,6	105479	0
ALLFRT	0	0	0	2354,3	13479,3	0
WINE	0	0	0	4941,9	15989,9	0
OTHCROP	0	0	7603	7324,5	19192,9	0
CATTL	0	0	0	4210,6	15804,6	0
PIG__	0	0	0	8573,8	85091,8	0
EGGPOUL	0	0	0	968,7	17443,3	0
CMILK	0	0	0	12107,5	45054,4	0
OTHLST	0	0	0	6266,6	23265,4	0
OTHACT	0	0	0	1036,3	8412,4	0
SE281	3488	9004,5	22605	23039,1	55710,7	9993,5

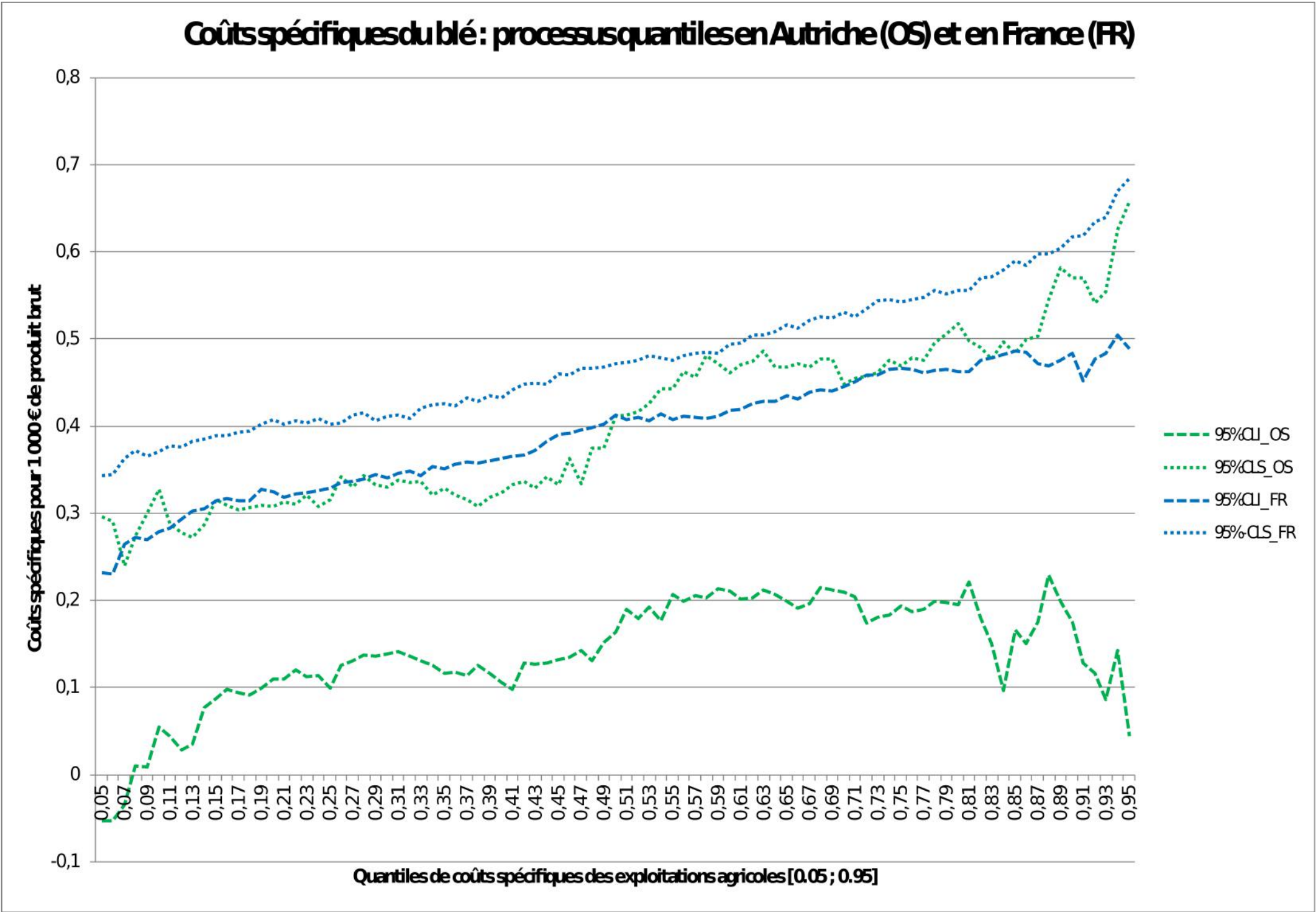
Le résumé des statistiques descriptives fournit les trois quartiles, la moyenne, l'écart-type et l'écart médian absolu pour la variable à expliquer (SE281) et l'ensemble des régresseurs

Sorties SAS : estimation des paramètres

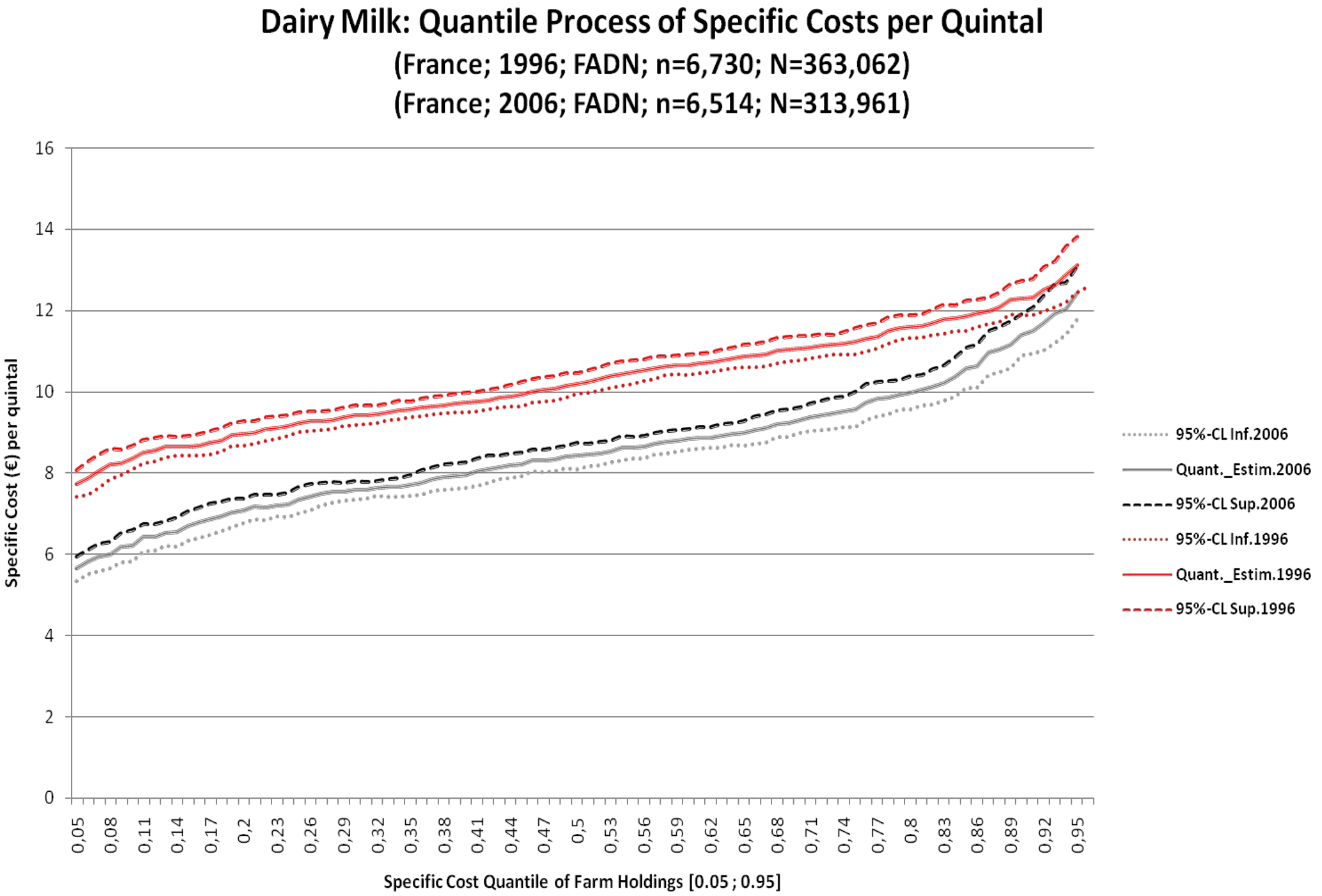
Parameter	DF	Estimate	Standard Error	95% Limits	Confidence	t Value	Pr > |t|
Intercept	0	0	0	0	0	,	,
WHEAT	1	0,2521	0,0314	0,1907	0,3136	8,04	<,0001
othcer	1	0,2336	0,0108	0,2123	0,2548	21,53	<,0001
DRYPU	1	0,2211	0,0604	0,1026	0,3395	3,66	0,0003
INDCROP	1	0,2686	0,043	0,1843	0,353	6,24	<,0001
OILSEED	1	0,7227	0,0898	0,5466	0,8988	8,05	<,0001
HORTIC	1	0,165	0,0099	0,1457	0,1843	16,74	<,0001
ALLFRT	1	0,1398	0,0152	0,1101	0,1695	9,22	<,0001
WINE	1	0,0683	0,0105	0,0476	0,0889	6,47	<,0001
OTHCROP	1	0,0913	0,0062	0,0792	0,1034	14,77	<,0001
CATTL	1	0,5393	0,035	0,4706	0,608	15,39	<,0001
PIG__	1	0,6102	0,0432	0,5256	0,6948	14,14	<,0001
EGGPOUL	1	0,7439	0,0305	0,684	0,8037	24,35	<,0001
CMILK	1	0,4372	0,0151	0,4075	0,4669	28,87	<,0001
OTHLST	1	0,4021	0,0053	0,3917	0,4125	75,69	<,0001
OTHACT	1	0,0049	0,0164	-0,0274	0,0371	0,3	0,7675

Le tableau des estimations donne également l'erreur-type, les limites inférieure et supérieure de l'intervalle de confiance à 95 %, la valeur test et la probabilité associée.

Processus quantile : blé, comparaison France-Autriche

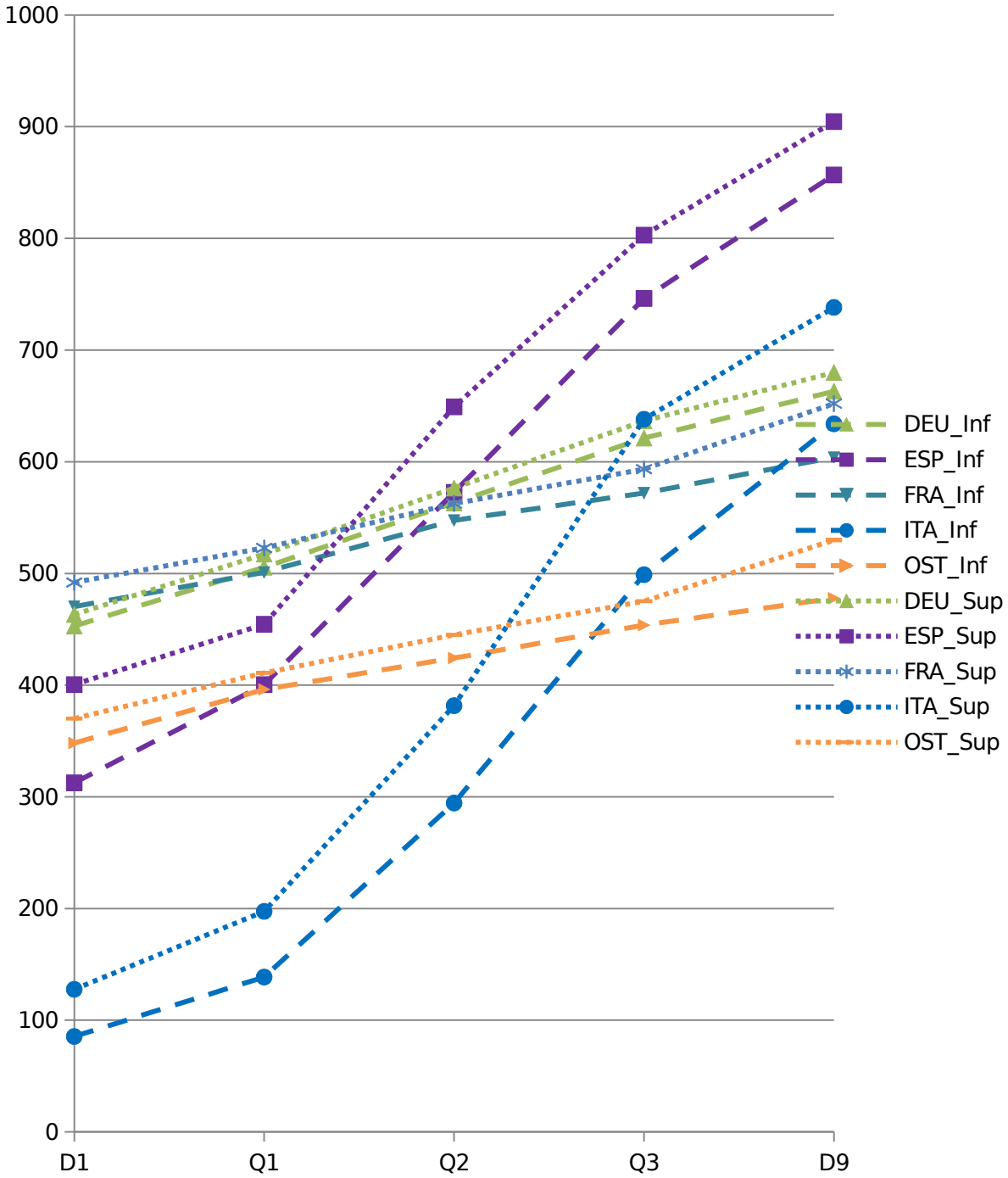


Processus quantile : lait, comparaison France, 1996 versus 2006



Coûts spécifiques : porc, quantiles conditionnels par pays

L'Espagne et l'Italie présentent des formes de distributions plus hétérogènes que l'Allemagne, la France ou l'Autriche.
Parmi les distributions homogènes, l'Autriche se distingue de l'Allemagne et de la France par un
niveau global plus faible



Références bibliographiques

- He X. et Hu F. (2002), Markov Chain Marginal Bootstrap. *Journal of the American Statistical Association*, Vol. 97, 783-795.

- Koenker, R. and Bassett, G. J. (1978), Regression quantiles. *Econometrica*, Vol. 46, 33-50.

- Koenker R. (2005), *Quantile Regression*, Econometric Society Monographies, Cambridge University Press, 349 p.

- Kocherginsky M., He X. et Mu Y. (2005), Practical Confidence Intervals for Regression Quantiles. *Journal of Computational and Graphical Statistics*, Vol. 14, 41-55.

- Surry Y., Desbois D., et Butault J.-P. (2012) Quantile Estimation of Specific Costs of Production. *FACEPA*, D8.2, 49 p.

- Desbois D., Butault J.-P., et Surry Y. (2013) Estimation des coûts de production en phytosanitaires pour les grandes cultures. Une approche par la régression quantile. *Economie Rurale*, n° 333, 27-49.

- Desbois D., Butault J.-P., et Surry, Y. (2015) Distribution des coûts spécifiques de production dans l'agriculture de l'Union européenne : une approche reposant sur la méthode de régression quantile. *9ièmes Journées de Recherches en Sciences Sociales*, n° 333, 27-49.