Estimation and Testing in M-quantile Regression with application to small area estimation

A. Bianchi [1] E. Fabrizi[2] N. Salvati[3] N. Tzavidis[4]

[1]Università degli Studi di Bergamo

[2]Università Cattolica del S. Cuore

[3]Università di Pisa

[4]University of Southampton

July 15, 2016

- OLS regression models the conditional expectation of a r.v. $E(Y|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$

General background - M-quantile regression

- OLS regression models the conditional expectation of a r.v. $E(Y|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$
  Often it is not sufficient to do this because
  - the mean is not an adequate summary of the distribution
  - it is not outlier robust

General background - M-quantile regression

- OLS regression models the conditional expectation of a r.v. $E(Y|\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}$
  Often it is not sufficient to do this because
  - the mean is not an adequate summary of the distribution
  - it is not outlier robust
- Quantile regression models the quantiles of the conditional distribution $F_Y^{-1}(\tau|\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta}_\tau$

General background - M-quantile regression

- OLS regression models the conditional expectation of a r.v. $E(Y|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$
  Often it is not sufficient to do this because
  - the mean is not an adequate summary of the distribution
  - it is not outlier robust
- Quantile regression models the quantiles of the conditional distribution $F_Y^{-1}(\tau|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}_\tau$
  Solutions of estimating equations are not unique

General background - M-quantile regression

- OLS regression models the conditional expectation of a r.v. $E(Y|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}$
  Often it is not sufficient to do this because
  - the mean is not an adequate summary of the distribution
  - it is not outlier robust
- Quantile regression models the quantiles of the conditional distribution $F_Y^{-1}(\tau|\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}_\tau$
  Solutions of estimating equations are not unique
- More general approach is M-quantile regression (combines robustness properties of quantiles with the efficiency properties of mean regression)

General background - M-quantile regression

M-quantile of order $\tau \in (0, 1)$:

$$MQ_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_i^T \beta_\tau$$

General background - M-quantile regression

M-quantile of order $\tau \in (0,1)$:

$$MQ_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_i^T \beta_\tau$$

- β_τ is estimated by $\hat{\boldsymbol{\beta}}_\tau = \text{argmin} \sum_{i=1}^{n} \rho_\tau \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\hat{\sigma}_\tau} \right)$
- $\hat{\sigma}_\tau$ is a robust estimate of scale

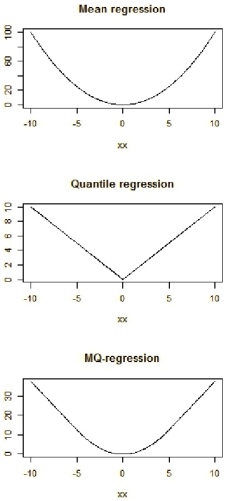
General background - M-quantile regression

M-quantile of order $\tau \in (0,1)$:

$$MQ_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_i^T \beta_\tau$$

- β_τ is estimated by $\hat{\boldsymbol{\beta}}_\tau = \text{argmin} \sum_{i=1}^n \rho_\tau \left(\frac{y_i - \mathbf{x}_i^T \boldsymbol{\beta}}{\hat{\sigma}_\tau} \right)$
- $\hat{\sigma}_\tau$ is a robust estimate of scale
- $\rho_\tau(u) = |\tau - I(u < 0)| \rho(u)$
- Huber loss function:

$$\rho(u) = 2 \begin{cases} (c|u| - c^2/2) & |u| > c \\ u^2/2 & |u| \leq c \end{cases}$$



General background

Chambers and Tzavidis (2006) apply MQ to Small Area Estimation problems

Since then a number of papers has been published

General background

Chambers and Tzavidis (2006) apply MQ to Small Area Estimation problems

Since then a number of papers has been published

Gap-Criticism: lack of model diagnostics and hypotheses testing tools for model selection

Objective

Expand available toolkit for inference in MQ regression

Objective

Expand available toolkit for inference in MQ regression

- Pseudo-R^2 goodness-of-fit measure

Objective

Expand available toolkit for inference in MQ regression

- Pseudo-R^2 goodness-of-fit measure
- LR-type test for linear hypotheses

Objective

Expand available toolkit for inference in MQ regression

- Pseudo-R^2 goodness-of-fit measure
- LR-type test for linear hypotheses
- Test for the presence of clustering

Objective

Expand available toolkit for inference in MQ regression

- Pseudo-R^2 goodness-of-fit measure
- LR-type test for linear hypotheses
- Test for the presence of clustering

- Data-driven tuning parameter selection in Huber influence function

Goodness-of-fit measure

Partitioned MQ model:

$$MQ_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_{i1}^T\boldsymbol{\beta}_{\tau 1} + \mathbf{x}_{i2}^T\boldsymbol{\beta}_{\tau 2}$$

- $\boldsymbol{\beta}_\tau = (\boldsymbol{\beta}_{\tau 1}^T, \boldsymbol{\beta}_{\tau 2}^T)^T$, $\boldsymbol{\beta}_{\tau 1}$ is $(p-k) \times 1$, $\boldsymbol{\beta}_{\tau 2}$ is $k \times 1$ $(0 < k < p)$

Goodness-of-fit measure

Partitioned MQ model:

$$MQ_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_{i1}^T\boldsymbol{\beta}_{\tau 1} + \mathbf{x}_{i2}^T\boldsymbol{\beta}_{\tau 2}$$

- $\boldsymbol{\beta}_\tau = (\boldsymbol{\beta}_{\tau 1}^T, \boldsymbol{\beta}_{\tau 2}^T)^T$, $\boldsymbol{\beta}_{\tau 1}$ is $(p-k)\times 1$, $\boldsymbol{\beta}_{\tau 2}$ is $k \times 1$ $(0 < k < p)$
- $\hat{\boldsymbol{\beta}}_\tau$ MQ estimator of the full model
- $\tilde{\boldsymbol{\beta}}_\tau = (\tilde{\boldsymbol{\beta}}_{\tau 1}^T, \mathbf{0}^T)^T$ MQ estimator under the restriction $\boldsymbol{\beta}_{\tau 2} = \mathbf{0}$

Goodness-of-fit measure

- Relative goodness-of-fit measure comparing the unrestricted to the restricted MQ regression model

$$R^2_\rho(\tau) = 1 - \frac{\sum_{i=1}^{n} \rho_\tau\left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\tau}{\hat{\sigma}_\tau}\right)}{\sum_{i=1}^{n} \rho_\tau\left(\frac{y_i - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_\tau}{\hat{\sigma}_\tau}\right)}$$

Goodness-of-fit measure

- Relative goodness-of-fit measure comparing the unrestricted to the restricted MQ regression model

$$R_\rho^2(\tau) = 1 - \frac{\sum_{i=1}^n \rho_\tau \left(\frac{y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_\tau}{\hat{\sigma}_\tau} \right)}{\sum_{i=1}^n \rho_\tau \left(\frac{y_i - \mathbf{x}_i^T \tilde{\boldsymbol{\beta}}_\tau}{\hat{\sigma}_\tau} \right)}$$

- When the restricted model includes only the intercept, this measure is the natural analog of the usual R^2 goodness-of-fit measure used in mean regression
- It varies between 0 and 1 and it represents a measure of goodness-of-fit for a specified τ

Testing linear hypotheses

Model:

$$MQ_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_{i1}^T\boldsymbol{\beta}_{\tau1} + \mathbf{x}_{i2}^T\boldsymbol{\beta}_{\tau2}$$

Hypotheses:

$$H_0 : \boldsymbol{\beta}_{\tau2} = \mathbf{0}$$
$$H_A : \boldsymbol{\beta}_{\tau2} \neq \mathbf{0}$$

Testing linear hypotheses

Model:

$$MQ_\tau(y_i|\mathbf{x}_i) = \mathbf{x}_{i1}^T\boldsymbol{\beta}_{\tau 1} + \mathbf{x}_{i2}^T\boldsymbol{\beta}_{\tau 2}$$

Hypotheses:

$$H_0 : \boldsymbol{\beta}_{\tau 2} = \mathbf{0}$$
$$H_A : \boldsymbol{\beta}_{\tau 2} \neq \mathbf{0}$$

Test statistic (under H_0, provided assumptions verified):

$$-2\frac{(n-p)^{-1}\sum_{i=1}^n \hat{\psi}'_{\tau i}}{n^{-1}\sum_{i=1}^n \hat{\psi}^2_{\tau i}}\left[\sum_{i=1}^n \rho_\tau\left(\frac{y_i-\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_\tau}{\hat{\sigma}_\tau}\right) - \sum_{i=1}^n \rho_\tau\left(\frac{y_i-\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_\tau}{\hat{\sigma}_\tau}\right)\right] \xrightarrow{d} \chi^2_k$$

with $\hat{\psi}'_{\tau i} := \psi'_\tau(\hat{\varepsilon}_{i\tau}/\hat{\sigma}_\tau)$, $\hat{\psi}_{\tau i} = \psi_\tau(\hat{\varepsilon}_{i\tau}/\hat{\sigma}_\tau)$

Testing linear hypotheses

Model:

$$MQ_{\tau}(y_i|\mathbf{x}_i) = \mathbf{x}_{i1}^T\boldsymbol{\beta}_{\tau 1} + \mathbf{x}_{i2}^T\boldsymbol{\beta}_{\tau 2}$$

Hypotheses:

$$H_0 : \boldsymbol{\beta}_{\tau 2} = \mathbf{0}$$
$$H_A : \boldsymbol{\beta}_{\tau 2} \neq \mathbf{0}$$

Test statistic (under H_0, provided assumptions verified):

$$-2\frac{(n-p)^{-1}\sum_{i=1}^{n}\hat{\psi}'_{\tau i}}{n^{-1}\sum_{i=1}^{n}\hat{\psi}^2_{\tau i}}\left[\sum_{i=1}^{n}\rho_{\tau}\left(\frac{y_i-\mathbf{x}_i^T\hat{\boldsymbol{\beta}}_{\tau}}{\hat{\sigma}_{\tau}}\right)-\sum_{i=1}^{n}\rho_{\tau}\left(\frac{y_i-\mathbf{x}_i^T\tilde{\boldsymbol{\beta}}_{\tau}}{\hat{\sigma}_{\tau}}\right)\right]\xrightarrow{d}\chi^2_k$$

with $\hat{\psi}'_{\tau i} := \psi'_{\tau}(\hat{\varepsilon}_{i\tau}/\hat{\sigma}_{\tau})$, $\hat{\psi}_{\tau i} = \psi_{\tau}(\hat{\varepsilon}_{i\tau}/\hat{\sigma}_{\tau})$

$\rightarrow$ LR type test

Test for the presence of clustering

Suppose we want to test the presence of group effects

Test for the presence of clustering

A. Bianchi , E. Fabrizi, N. Salvati, N. Tzavidis

Suppose we want to test the presence of group effects

Idea

Characterize heterogeneity in the data using group specific MQ-coefficients

Test for the presence of clustering

Suppose we want to test the presence of group effects

Idea

Characterize heterogeneity in the data using group specific MQ-coefficients

- $j = 1, \ldots, d$ group, $i = 1, \ldots, n_j$ unit
- Variable of interest y_{ij}, Unit level covariates $\mathbf{x}_{ij}$
- For each group j, τ_j=group specific MQ-coefficient

$$\min_{\tau} E\left[\rho\left(\frac{y_{ij} - \mathbf{x}_{ij}^T \boldsymbol{\beta}_{\tau}}{\sigma}\right) |j\right]$$

It is the τ value that corresponds to the MQ plane which is closer to observations from group j, according to metric ρ

Test for the presence of clustering

Group MQ-coefficient original definition (Chambers and Tzavidis, 2006):

Test for the presence of clustering

Group MQ-coefficient original definition (Chambers and Tzavidis, 2006):

1. Define MQ-coefficient for each unit, τ_{ij}, s.t.

$$MQ_{\tau_{ij}}(y_{ij}|\mathbf{x}_{ij}) = y_{ij}$$

Test for the presence of clustering

Group MQ-coefficient original definition (Chambers and Tzavidis, 2006):

1. Define MQ-coefficient for each unit, τ_{ij}, s.t.

$$MQ_{\tau_{ij}}(y_{ij}|\mathbf{x}_{ij}) = y_{ij}$$

2. Characterize each group j by the average of the units MQ-coefficients that belong to that group.

Test for the presence of clustering

A. Bianchi , E. Fabrizi, N. Salvati, N. Tzavidis

Hypotheses:

$$H_0 : \tau_j = 0.5 \; \forall j = 1, \ldots, d$$
$$H_A : \tau_j \neq 0.5 \text{ for at least one } j$$

Test for the presence of clustering

Hypotheses:

$$H_0 : \tau_j = 0.5 \ \forall j = 1, \dots, d$$
$$H_A : \tau_j \neq 0.5 \text{ for at least one } j$$

Test statistic (under H_0, provided assumptions verified):

$$-2\frac{(n-p)^{-1}\sum_{ij}\hat{\psi}'_{ij}}{n^{-1}\sum_{ij}\hat{\psi}^2_{ij}}\left[\sum_{j=1}^{d}\sum_{i=1}^{n_j}\rho\left(\frac{y_{ij}-\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{\hat{\tau}_j}}{\hat{\sigma}}\right)-\sum_{j=1}^{d}\sum_{i=1}^{n_j}\rho\left(\frac{y_{ij}-\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{0.5}}{\hat{\sigma}}\right)\right]\xrightarrow{d}\chi^2_{d-1}$$

Test for the presence of clustering

Hypotheses:

$$H_0 : \tau_j = 0.5 \ \forall j = 1, \ldots, d$$
$$H_A : \tau_j \neq 0.5 \text{ for at least one } j$$

Test statistic (under H_0, provided assumptions verified):

$$-2\frac{(n-p)^{-1}\sum_{ij}\hat{\psi}'_{ij}}{n^{-1}\sum_{ij}\hat{\psi}^2_{ij}}\left[\sum_{j=1}^{d}\sum_{i=1}^{n_j}\rho\left(\frac{y_{ij}-\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{\hat{\tau}_j}}{\hat{\sigma}}\right)-\sum_{j=1}^{d}\sum_{i=1}^{n_j}\rho\left(\frac{y_{ij}-\mathbf{x}_{ij}^T\hat{\boldsymbol{\beta}}_{0.5}}{\hat{\sigma}}\right)\right]\xrightarrow{d}\chi^2_{d-1}$$

Usefulness in Small Area context: if no area effect, a more efficient estimator may be used (synthetic estimator)

Data driven tuning constant selection

- Huber loss function

$$\rho_\tau(u) = 2\begin{cases} (c|u| - c^2/2)|\tau - I(u \leq 0)| & |u| > c \\ u^2/2|\tau - I(u \leq 0)| & |u| \leqslant c \end{cases}$$

- c determines robustness level (tradeoff robustness-efficiency)
- In general $c = 1.345$

Data driven tuning constant selection

- Huber loss function

$$\rho_\tau(u) = 2 \begin{cases} (c|u| - c^2/2)|\tau - I(u \leq 0)| & |u| > c \\ u^2/2|\tau - I(u \leq 0)| & |u| \leqslant c \end{cases}$$

- c determines robustness level (tradeoff robustness-efficiency)
- In general $c = 1.345$
- Interpret c as a parameter of a proper density

Data driven tuning constant selection

A. Bianchi , E. Fabrizi, N. Salvati, N. Tzavidis

- Asymmetric Least Informative distribution

$$f_\tau(y;\mu_\tau,\sigma_\tau)=\frac{1}{\sigma_\tau B_\tau} exp\left\{-\rho_\tau\left(\frac{y-\mu_\tau}{\sigma_\tau}\right)\right\}, -\infty < y < +\infty$$

Data driven tuning constant selection

- Asymmetric Least Informative distribution

$$f_\tau(y;\mu_\tau,\sigma_\tau)=\frac{1}{\sigma_\tau B_\tau} exp\left\{-\rho_\tau\left(\frac{y-\mu_\tau}{\sigma_\tau}\right)\right\}, -\infty < y < +\infty$$

- Choose c maximizing log-likelihood function based on $f_\tau(y;\mu_\tau,\sigma_\tau)$

Data driven tuning constant selection

- Asymmetric Least Informative distribution

$$f_\tau(y;\mu_\tau,\sigma_\tau)=\frac{1}{\sigma_\tau B_\tau} exp\left\{-\rho_\tau\left(\frac{y-\mu_\tau}{\sigma_\tau}\right)\right\}, -\infty < y < +\infty$$

- Choose c maximizing log-likelihood function based on $f_\tau(y;\mu_\tau,\sigma_\tau)$
- Simulation results: effective in reflecting different levels of contamination in the data

Simulation studies

Null distributions for the test statistics have been derived asymptotically only. We explored finite sample properties by means of simulations.

Three sets of simulations:

- LR-type test
- Test for the presence of clusters
- Choice of the tuning constant

Data generated under mixed (random) effects model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + u_i + \varepsilon_{ij},$$

where $i = 1, \ldots, n_j$, $j = 1, \ldots, d$,

- $\beta_0 = 0,\ \beta_1 = 0.5,\ n_j = 5$
- $(\beta_2, \beta_3) \in \{(0,0), (0.25, 0.25), (0.5, 0.5), (1,1)\}$
- $d = 20,\ 100$
- u_i and ε_{ij}: $N(0,1)$, t_3, and $\chi^2(2)$

Null hypothesis: $H_0 : \beta_{\tau 2} = \beta_{\tau 3} = 0$

Simulation studies - LR-type test

d		N, $c = 100$			t_3, $c = 1.345$		
	τ	0.50	0.75	0.90	0.50	0.75	0.90
	α	$(\beta_2, \beta_3) = (0, 0)$					
	0.10	0.110	0.114	0.133	0.103	0.114	0.147
20	0.05	0.059	0.062	0.075	0.050	0.063	0.089
	0.01	0.012	0.015	0.021	0.012	0.016	0.030
	0.10	0.101	0.105	0.109	0.102	0.108	0.122
100	0.05	0.052	0.058	0.058	0.052	0.053	0.063
	0.01	0.010	0.011	0.012	0.013	0.012	0.017

- Type I error very close to the nominal level
- small deviations in the case of $\tau = 0.9$ in the t_3 (and $\chi^2(2)$) with $d = 20$ ($n = 100$) where the test turns out to be slightly conservative

Simulation studies - LR-type test

Power analysis:

- Gaussian scenario: the power of the test tends to 1 as soon as the values of β_2 and β_3 increase for both sample sizes.
- t_3 scenario: the value of the power of the test tends to 1 at $\tau = 0.5$ and 0.75 once the $\beta_2, \beta_3 = 0.25$ especially for $d = 100$. At $\tau = 0.9$ the likelihood ratio type test performs well as regression coefficients increase (as soon as $\beta_2, \beta_3 = 0.5$).

Data generated under mixed (random) effects model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + \varepsilon_{ij},$$

where $i = 1, \ldots, n_j$, $j = 1, \ldots, d$,

- $\beta_0 = 1$, $\beta_1 = 2$, $x \sim U(0,5)$, $\varepsilon_{ij} \sim N(0,5)$
- $u_i \sim N(0, \sigma_u^2)$, with $\sigma_u^2 = 0,\ 1,\ 2.5,\ 7.5$
- $n_j = 5, 20, 50$
- $d = 20,\ 100$

$$\begin{aligned} \textit{Hypotheses}: \quad & H_0 : \tau_j = 0.5\ \forall j = 1, \ldots, d \\ & H_A : \tau_j \neq 0.5 \text{ for at least one } j \end{aligned}$$

Simulation studies - Test for clusters

- Under the null, Type I error is very close to the nominal value
- As the value of σ_u^2 increases the power of the test increases too
- Power increases more sharply for larger within cluster sample sizes.

Simulation studies - Tuning constant

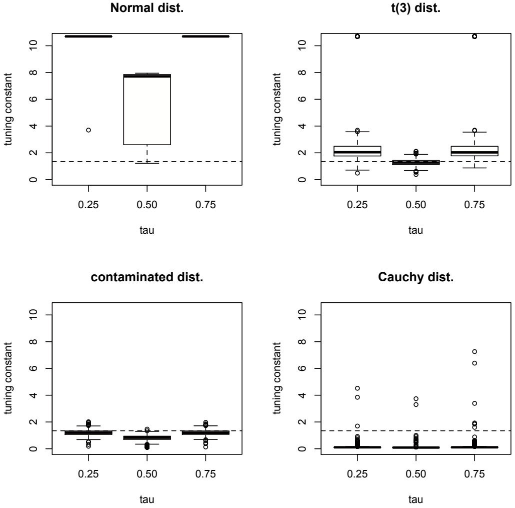
Data generated under mixed (random) effects model:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_i + \varepsilon_{ij},$$

where $i = 1, \dots, n_j$, $j = 1, \dots, d$,

- $\beta_0 = 1$, $\beta_1 = 2$, $x \sim U(0,5)$, $d = 100$, $n_j = 5$
- $u_i \sim N(0,1)$
- ε_{ij}: $N(0,1)$, t_3, $(1-\gamma)N(0,1) + \gamma N(0,25)$ with $Pr(\gamma = 1) = 0.1$, Cauchy with location 0 and scale 1

Simulation studies - Tuning constant



Real data application

- Battese et al. (1988) data $\rightarrow$ predict the hectares of corn and soybean by county (small area)

Real data application

- Battese et al. (1988) data $\rightarrow$ predict the hectares of corn and soybean by county (small area)
- y_1= n. of hectares of corn, y_2= n. of hectares of soybean
- x_1, x_2= number of pixels classified as corn / soybean by satellite
- cluster= 12 counties in Iowa

Real data application

- Battese et al. (1988) data → predict the hectares of corn and soybean by county (small area)
- y_1= n. of hectares of corn, y_2= n. of hectares of soybean
- x_1, x_2= number of pixels classified as corn / soybean by satellite
- cluster= 12 counties in Iowa
- Battese et al. (1988) use these data to predict the hectares of corn and soybean by county (using linear mixed random effects model)

Real data application

Estimated values for the tuning constant (at $\tau = 0.5$):

- Corn (y_1): $\hat{c} = 1.94 \rightarrow$ relatively low value, consistent with the presence of outlier identified in Sinha and Rao (2009)
- Soybean (y_2): $\hat{c} = 7.85 \rightarrow$ no issue with contamination.

Real data application

LR-type test

- joint significance of both covariates for modeling y_1 and y_2
- Corn (y_1): after controlling for the n. of pixels classified as corn (x_1), the n. of pixels classified as soybean (x_2) not significant
- Soybean (y_2): after controlling for the n. of pixels classified as soybean (x_2), the n. of pixels classified as corn (x_1) not significant

Real data application

LR-type test

- joint significance of both covariates for modeling y_1 and y_2
- Corn (y_1): after controlling for the n. of pixels classified as corn (x_1), the n. of pixels classified as soybean (x_2) not significant
- Soybean (y_2): after controlling for the n. of pixels classified as soybean (x_2), the n. of pixels classified as corn (x_1) not significant

To validate results, we run the model proposed by Battese et al. and got similar results

Real data application

Clustering test

- Corn: p-value=0.103 $\rightarrow$ leading to synthetic small area estimation and more efficient estimators
- Soybean: p-value=0.005 $\rightarrow$ presence of clustering

Real data application

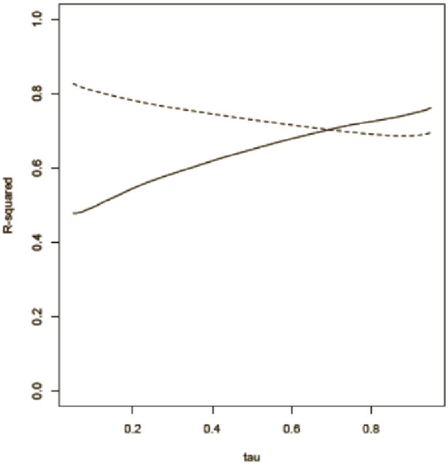
A. Bianchi , E. Fabrizi, N. Salvati, N. Tzavidis

Clustering test

- Corn: p-value=0.103 → leading to synthetic small area estimation and more efficient estimators
- Soybean: p-value=0.005 → presence of clustering

Test for the presence of clustering effect in linear mixed model framework confirms results.

Real data application



- Corn (solid line): R^2 increases as τ increases
- Soybean (dashed line): almost constant high value for R^2 at all values of τ

Concluding remarks

- We have expanded the available toolkit for inference in MQ-regression

- The theory can be applied in small area estimation framework to validate MQ models and improve small area estimates

Thank you for your attention!