

# Enhancing Research Capacity via Quantile Regression in an Inter-disciplinary Setting

**Edmore Ranganai**

**Department of Statistics**

**CSET, UNISA**

**Date: 5 Dec 2023**



Define tomorrow.

UNISA



# Appreciation

## Family

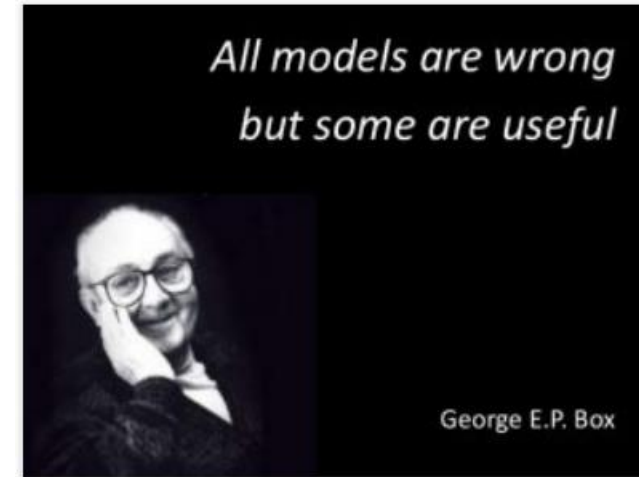
- My wife Christabell and two beautiful daughters
  - Trinity and Tallia
- My parents and Grandparents

## Supervisors, Collaborators & Funders

- Prof Tertius de Wet & the late Dr Johan van Vuuren
- Collaborators, UNISA and Funders
- Colleagues

# Introduction

- A model is a simplified representation of a (data generating) system.
- It follows simple rules (assumptions) formalized into
  - mathematical equations or
  - computational algorithm.
- All models are wrong at least in some of their details, but their overall direction is very likely to be correct.



Coined in 1976

# Introduction

- Inter-disciplinary scenarios call for a multiplicity of approaches depending on whether
  - It is of essence for the model to capture the overall direction of the data generating system (distribution)
  - Or at a local level, e.g. in the tails of the distribution.
  - Or all of these aspects.
- As coined in the Mosteller and Tukey (1977) concern.



Coined in 1976

# Introduction

- The average salary in South Africa for 2023 is R31,300 before taxes and other deductions, according to SalaryExplorer.  
How average is your salary? See how it compares to others in SA
- 75% of employees earn R41,100 ( $Q_3$ ) a month or less.
- 50% of employees earn R27,100 ( $Q_2$ ) a month or less
- 25% of employees earn less than R19,600 ( $Q_1$ ) per month.



Compare your salary to the average wages in South Africa this year. Picture: Karelien Kriel/Pixabay

Published Jul 22, 2023

# Introduction

- Characterization and quantification of the tail behaviour of rare events is important

Annals of Data Science (2023) 10(2):251–290  
<https://doi.org/10.1007/s40745-020-00294-w>



An Application of Extreme Value Theory for Measuring Financial Risk in BRICS Economies

Emmanuel Afuecheta<sup>1</sup> · Chigozie Utazi<sup>2</sup> · Edmore Ranganai<sup>3</sup> · Chibuzor Nnanatu<sup>4</sup>

RESEARCH

Open Access



Long memory mean and volatility models of platinum and palladium price return series under heavy tailed distributions

Edmore Ranganai<sup>\*</sup> and Sihle Basil Kubheka

J. Stat. Appl. Pro. 11, No. 1, 89-107 (2022)



89

Journal of Statistics Applications & Probability  
An International Journal

<http://dx.doi.org/10.18576/jsap/110107>

Value-at-Risk Estimation of Precious Metal Returns using Long Memory GARCH Models with Heavy-Tailed Distribution

Knowledge Chinhamu<sup>1,\*</sup>, Retius Chifurira<sup>1</sup> and Edmore Ranganai<sup>2</sup>

# Preliminaries

## The unconditional (univariate) case

- Denote the order statistics of the sample by  $Y_{(1)} \leq Y_{(2)} \leq \dots \leq Y_{(n)}$ , and the empirical distribution function (edf) by  $F_n(y) = n^{-1} \sum_{i=1}^n I(Y_i \leq y)$ .

- Since  $F_n$  is an estimator for  $F$ , a natural estimator for  $q_\theta$  is the  $\theta^{\text{th}}$  sample quantile,

$$\hat{q}_\theta \equiv F_n^{-1}(\theta) \equiv Y_{([\ln\theta])}$$

where  $[x]$  denotes the largest integer less than or equal to  $x$  and

- The sample the sample quartiles given as
  - First Quartile,  $Q_1 = \hat{q}_{0.25}$ ,
  - Second Quartile (sample median) =  $\hat{q}_{0.5}$  and
  - Third sample median =  $\hat{q}_{0.75}$ .

# Preliminaries

- The mean

$$\mu = \arg \min_{\xi \in \mathcal{R}} E(Y - \xi)^2$$

- The median

$$\text{Med} = \arg \min_{\xi \in \mathcal{R}} E |Y - \xi|$$

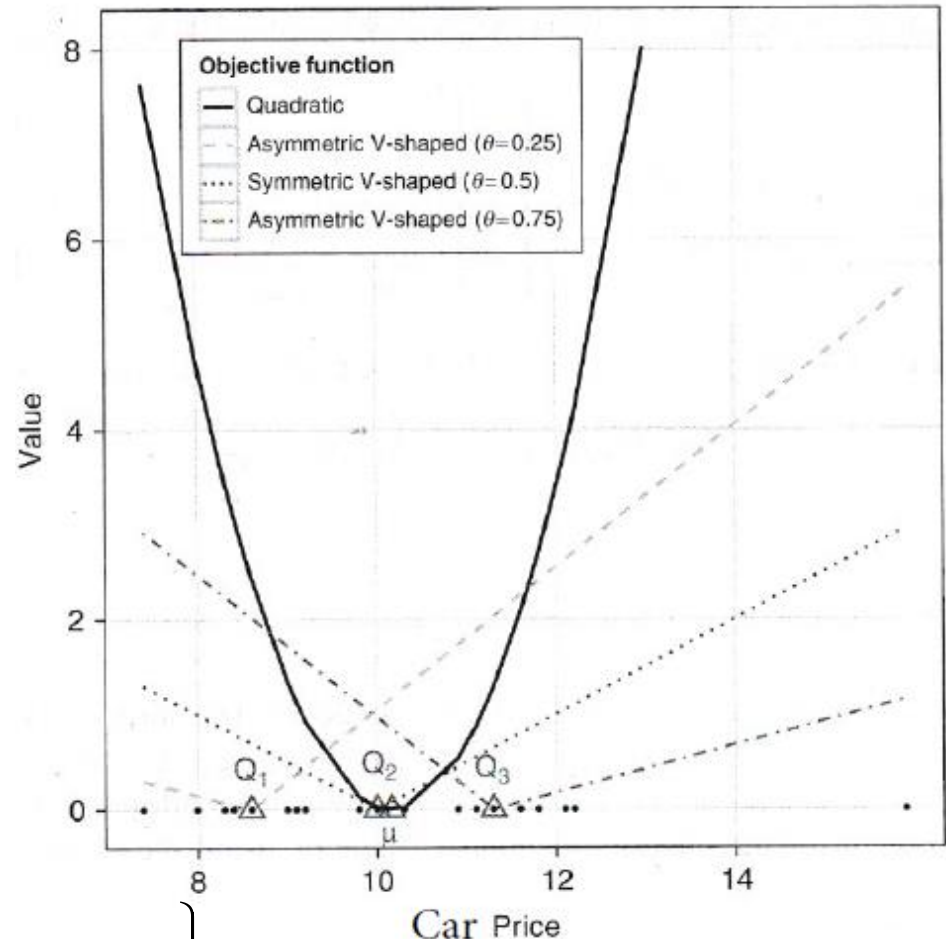
- All quantiles:

$$\begin{aligned} \rho_{\theta}(\xi) &= \xi [\tau - I(\xi < 0)] \\ &\equiv \xi [\theta I(\xi \geq 0) + (1 - \theta) I(\xi < 0)]. \end{aligned}$$

$$\hat{q}_{\theta} = \arg \min_{\xi \in \mathcal{R}} \sum_{i=1}^n (\rho_{\theta}(Y_i - \xi)), \quad 0 < \theta < 1,$$

$$\equiv \min_{\xi \in \mathcal{R}} \left\{ \sum_{i:(Y_i - \xi) \geq 0} \theta |Y_i - \xi| + \sum_{i:(Y_i - \xi) < 0} (1 - \theta) |Y_i - \xi| \right\}$$

A VISUAL INTRODUCTION TO QUANTILE REGRESSION



Source: [www.wiley.com/go/quantile\\_regression](http://www.wiley.com/go/quantile_regression)



# Preliminaries

## Formulation of the linear programming (LP) problem

- Naturally we can define the sample quantiles,  $\hat{q}_\theta$  as the solution to the corresponding minimization problem based on the sample, viz.,

$$\hat{q}_\theta = \arg \min_{\xi \in \mathbb{R}} \sum_{i=1}^n (\rho_\theta(Y - \xi)).$$

- This minimization problem may be reformulated as

$$\min \left[ \theta \mathbf{1}'_n \mathbf{u}^+ + (1 - \theta) \mathbf{1}'_n \mathbf{u}^- \right]$$

$$\text{subject to } \mathbf{Y} = \mathbf{1}'_n \xi + \mathbf{1}'_n \mathbf{u}^+ - \mathbf{1}'_n \mathbf{u}^-,$$

$$\mathbf{u}^+, \mathbf{u}^- \geq \mathbf{0}$$

where  $\mathbf{1}_n$  is the vector of ones and  $\{u_i^+, u_i^- : i = 1, \dots, n\}$  represent the positive and the negative residuals respectively.

- In this formulation it is clearly a linear programming (LP) problem to which the available LP tools could be applied (see *e.g.* Koenker, 2005).

# Motivation: In the Regression Case

- Quantile Regression (QR) addresses the Mosteller and Tukey (1977) concern in their influential remark:

“What the Ordinary Least Squares (OLS) regression curve does is give a grand summary for the averages of the distributions corresponding to the set of covariates. We could go further and **compute several different regression curves** corresponding to **the various percentage points of the distributions** and thus get **a more complete picture of the set**. Ordinarily this is not done, and so regression often gives a rather **incomplete picture**. **Just as the mean gives an incomplete picture of a single distribution, so the regression gives a correspondingly incomplete picture for a set of distributions.**”

# Motivation: In the Regression Case

- The Koenker and Basset (1978) quantile regression (QR) falls under the domain of robust statistical methodologies.
- QR like other robust statistical methodologies is able to detect outliers by first fitting the majority of the data and then flagging data points that deviate from it; filling a void left by classical statistical methods such as ordinary least squares (OLS) based methods

# Motivation: In the Regression Case

- While the OLS needs the assumption of data Normality (homoscedasticity of the error term) for both mathematical tractability and to produce the best possible coefficient estimates, QR performs well on data drawn from a wide range of probability distributions.
- QR has been applied to a multiplicity of interdisciplinary areas which include electricity demand, medical reference charts, survival analysis, financial economics, environmental modelling and the detection of heteroscedasticity and high leverage points, etc to minimal extent.

# Motivation: In the Regression Case

In a nutshell:

- QR is the robust method of choice since unlike other robust procedures QR is not only supplementary to the OLS procedure but also complementary (and alternate) to it due to its versatility QR procedure can detect heterogeneous effects of covariates at different quantiles).
- Unequal variation implies that there is more than a single slope (rate of change) describing the relationship between a response variable and predictor variables measured on a subset of factors.

# Motivation: In the Regression Case

In a nutshell:

- Quantile regression estimates multiple rates of change (slopes) from the minimum to maximum response, providing a more complete picture of the relationships between variables missed by other regression methods.
- In some areas research often focus on rates of change in quantiles near the maximum/minimum response, where a much smaller subset of limiting factors are measured, e.g., ecology, extreme electricity demand; and near the minimum, e.g., value at risk.

# Questions

- Why is QR still playing a second fiddle role to the OLS despite its inherent advantages?
- Are statistical practitioners employing appropriate methodologies/technologies?
- If not, what are their reasons?
- What needs to be done to popularise the use of QR?

# What do we know or Challenges?

- Robust statistics is now about some 40 years old:
  - Tukey (1960), Huber (1964), and Hampel (1968) laid the foundations of modern robust statistics'
  - The Koenker and Basset (1978) quantile regression (QR) falls under the domain of robust statistical methodologies.
- A common understanding
  - required to routinely use both OLS and robust estimators and only examine the data more closely in case of "large" discrepancies-whatever this means?



# What do we know or Challenges?

- at the interface of statistics and its applications there are non-statisticians who find it insurmountable
  - to deal with this vague idea of “large” discrepancies and the necessary choices of types of estimators
  - and tuning constants involved in the robust statistical methodology.

## Possible Answers

- Encouraging statistical practitioners to adopt the recommendation to use OLS and QR as a robust procedure of choice due to the latter's versatility.
- To make the robust estimators more appealing to statistical practitioners, an endeavor to studentize robust estimators has been undertaken by some researchers (McKean and Sheather 1991; Yohai et al. 1991).
- I have demonstrated that studentization can be achieved via relating QR to the OLS.
- Actually, many useful statistics derive from this relationship.

➤ Workshop at  
the University of  
Venda

# Possible Answers

IEEE Access<sup>®</sup>

Multidisciplinary | Rapid Review | Open Access Journal

Received August 31, 2020, accepted September 13, 2020, date of publication September 18, 2020,  
date of current version September 30, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024661

## Capturing Long-Range Dependence and Harmonic Phenomena in 24-Hour Solar Irradiance Forecasting: A Quantile Regression Robustification via Forecasts Combination Approach

EDMORE RANGANAI<sup>1</sup> AND CASTON SIGAUKE<sup>2</sup>, (Member, IEEE)

<sup>1</sup>Department of Statistics, University of South Africa, Florida Campus, Johannesburg 1709, South Africa

<sup>2</sup>Department of Statistics, University of Venda, Thohoyandou 0950, South Africa



energies



Article

## Prediction of Extreme Conditional Quantiles of Electricity Demand: An Application Using South African Data

Norman Maswanganyi<sup>1,†</sup>, Caston Sigauke<sup>2,\*,†</sup> and Edmore Ranganai<sup>3,†</sup>

# Quantile Regression

- Consider the linear model in the usual notation:

$$\mathbf{Y} = \mathbf{1}\beta_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \text{with } \varepsilon_i \sim F, \text{ (say) .}$$

- Unlike the univariate case, in the regression case (structured) the data cannot be ordered.
- As Analogues to sample quantiles the  $\theta^{\text{th}}$  regression quantile (RQ) based on the sample  $(Y_i, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , is

$$\hat{\boldsymbol{\beta}}(\theta) = \arg \min_{\beta_0, \boldsymbol{\beta}} \sum_{i=1}^n \rho_{\theta}(Y_i - (\beta_0 + \mathbf{x}_i' \boldsymbol{\beta})),$$

where  $\mathbf{x}_i'$  is the  $i^{\text{th}}$  row of the design matrix  $\mathbf{X}$  **without** the constant covariate,

$\beta_0$  is the intercept term,  $\boldsymbol{\beta}$  is the slope coefficient and  $\rho_{\theta}(u)$  as defined earlier.

# Quantile Regression

- Let  $Q_{Y|\mathbf{x}}(\theta)$  denote the conditional quantile function of  $Y$  given the covariate  $\mathbf{x}$ .
  - Since we have the linear shift model,

$$\begin{aligned}\hat{q}_{Y|\mathbf{x}}(\theta) &= F^{-1}(\theta) + \hat{\beta}_0 + \mathbf{x}'\hat{\beta}_1(\theta) \\ &= \hat{\beta}_0(\theta) + \mathbf{x}'\hat{\beta}_1(\theta).\end{aligned}$$

# Quantile Regression

- This can be written as

$$\hat{q}_{Y|X}(\theta) = (1 \quad \mathbf{x}') \hat{\boldsymbol{\beta}}(\theta),$$

with

$$\hat{\boldsymbol{\beta}}(\theta) = \begin{pmatrix} F^{-1}(\theta) + \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_1(\theta) \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0(\theta) \\ \hat{\boldsymbol{\beta}}_1(\theta) \end{pmatrix}.$$

- Clearly  $\hat{\boldsymbol{\beta}}(\theta)$  estimates  $\boldsymbol{\beta}(\theta)$ . The former is the  $\theta^{\text{th}}$  population regression quantile estimator.
- Note that  $\hat{\boldsymbol{\beta}}(\theta)$  is an M-estimator (see *e.g.* Huber, 1981) with check function  $\rho_\theta(\cdot)$ .

# Quantile Regression

- RQs are fairly robust to  $Y$ -space outliers since their influence functions are bounded in the  $Y$ -space.
- Also, for  $\theta = 0.5$  we obtain the usual  $L_1$  (median) regression estimator.
- Unlike the explicit solution giving the ordinary least squares (OLS) estimator

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\boldsymbol{\beta}}_1 \end{pmatrix}.$$

# Quantile Regression

- The minimization problem; consider the vector of residuals

$$\mathbf{r}(\mathbf{b}) = \mathbf{Y} - (\mathbf{1}_n \quad \mathbf{X})\mathbf{b} \equiv \mathbf{r}^+(\mathbf{b}) - \mathbf{r}^-(\mathbf{b}),$$

then we can write

$$\sum_{i=1}^n \rho_{\theta}(r_i(\mathbf{b})) = \theta \sum_{i=1}^n r_i^+(\mathbf{b}) + (1 - \theta) \sum_{i=1}^n r_i^-(\mathbf{b}).$$

- Hence in vector-matrix notation the minimization problem becomes

$$\min \left[ \theta \mathbf{1}'_n \mathbf{r}^+(\mathbf{b}) + (1 - \theta) \mathbf{1}'_n \mathbf{r}^-(\mathbf{b}) \right]$$

$$\text{subject to } \mathbf{Y} = (\mathbf{1}_n \quad \mathbf{X})\mathbf{b} + \mathbf{r}^+(\mathbf{b}) - \mathbf{r}^-(\mathbf{b}),$$

$$\mathbf{r}^+(\mathbf{b}), \mathbf{r}^-(\mathbf{b}) \geq \mathbf{0}.$$



# Quantile Regression Link to OLS

- The optimal LP basic solution gives a RQ corresponding to the  $p$  (equal to the number of covariates) points of the data set.
- The LP problem optimal solution  $\hat{\beta}(\theta)$  coefficient coincide with the OLS coefficient

$$\hat{\beta}_{J_\theta} = \left( \mathbf{X}'_{J_\theta} \mathbf{X}_{J_\theta} \right)^{-1} \mathbf{X}'_{J_\theta} \mathbf{Y}_{J_\theta} = \mathbf{X}_{J_\theta}^{-1} \mathbf{Y}_{J_\theta},$$

for non-singular  $\mathbf{X}_{J_\theta}$  where the subset  $J_\theta$  corresponds to the set of subscripts

$\{h_1, \dots, h_p\}$  such that  $(\mathbf{x}'_{h_i}, \mathbf{y}_{h_i})$ ,  $i = 1, \dots, p$ , is the the  $i^{\text{th}}$  case of elemental set (ES)  $J_\theta$ ;

in other words, an elemental regression (ER)  $J_\theta$ .

- Using simple example with a historical perspective the connection between RQs and ESs is easily illustrated.

# Quantile Regression Link to OLS

## Example: Boscovitch's ellipticity of the earth

- Arc length is measured as the excess over 56 700 toise per degree where one toise  $\approx$  6.39 feet or 1.95 meters.

Table 1: Boscovich data set

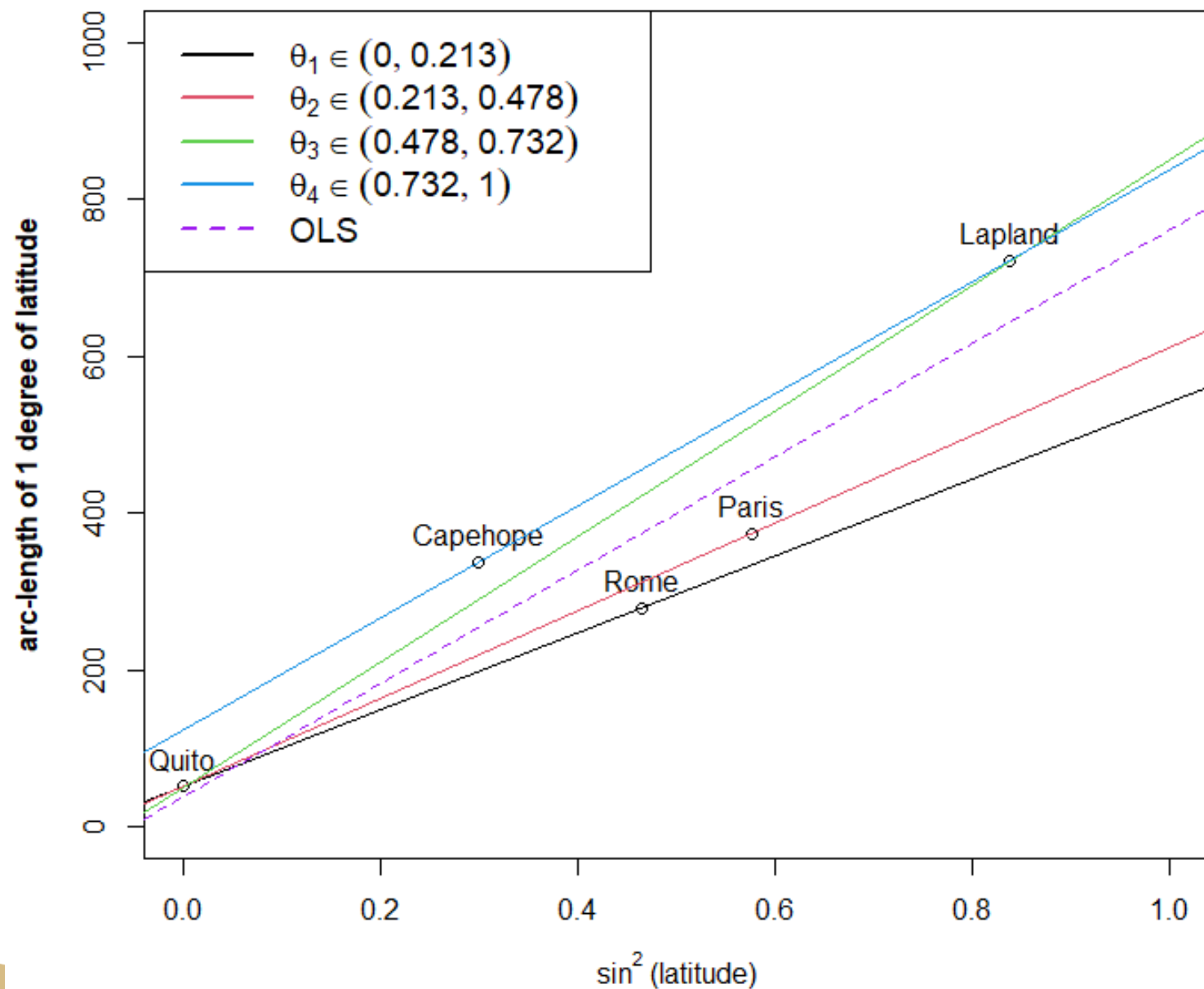
	Latitude	$\sin^2(\text{Latitude})$	Arc Length
Quito	0° 0'	0	51
Cape of Good Hope	33° 18'	0.2987	337
Rome	42° 59'	0.4648	279
Paris	49° 23'	0.5762	374
Lapland	66° 19'	0.8386	722

- Here RQs are illustrated using a very simple bivariate data set by considering the Boscovich (1755)'s approximation for short arcs by

$$y = a + b \sin^2 \lambda$$

in Table 1.

## Boscovitch Ellipticity of the Earth Example



# Quantile Regression Link to OLS

- The total number of ESs (ERs) is  $K = \binom{n}{p} = \frac{n!}{(n-p)!p!}$ .

- So in this case we have  $\binom{5}{2} = \frac{5!}{3!2!} = 10 = K$  ESs.

- Now, let

$$J(\theta) = \{\hat{\beta}(\theta_1), \hat{\beta}(\theta_2), \dots, \hat{\beta}(\theta_{n'})\}, \text{ for } 0 < \theta_1 < \theta_2 < \dots < \theta_{n'-1} < \theta_{n'} < 1,$$

be the complete set of solutions to the LP problem giving  $n'=4$  RQs,

where  $n'$  is approximately equal to  $n < K$  as  $n$  increases.

# Quantile Regression Link to OLS

- The solutions to the LP problem (2.2) do not change over the intervals  $[\theta_{k-1}, \theta_k)$ , for  $k = 2, \dots, n'$ .
- Thus, the  $n'$  RQs corresponding to  $n'$  specific ERs are efficiently computed from LP problem drastically reducing the computational load since computing all the  $K > n'$  ERs is avoided.
- The LP problem optimal solution  $\hat{\beta}(\theta)$  coefficient coincide with the OLS coefficient  $\hat{\beta}_{J_\theta} = (\mathbf{X}'_{J_\theta} \mathbf{X}_{J_\theta})^{-1} \mathbf{X}'_{J_\theta} \mathbf{Y}_{J_\theta} = \mathbf{X}_{J_\theta}^{-1} \mathbf{Y}_{J_\theta}$ ,  
for nonsingular  $\mathbf{X}_{J_\theta}$

# Quantile Regression Link to OLS

- Thus we have

$$J(\tau) = \{\widehat{\beta}(\theta_1), \widehat{\beta}(\theta_2), \widehat{\beta}(\theta_3), \dots, \widehat{\beta}(\theta_4)\}$$

such that  $(\mathbf{x}'_{h_i}, \mathbf{y}_{h_i})$ ,  $i = 1, \dots, p$ , is the  $i^{\text{th}}$  case of elemental set (ES)  $J_\theta$ , i.e.,

- $\{(\mathbf{x}'_{h_i}, \mathbf{y}_{h_i}), i : \text{Quito, Rome}\}$  corresponds to  $\widehat{\beta}(\theta_1)$  for  $\theta_1 \in (0, 0.213)$ .
  - $\{(\mathbf{x}'_{h_i}, \mathbf{y}_{h_i}), i : \text{Quito, Paris}\}$  corresponds to  $\widehat{\beta}(\theta_2)$  for  $\theta_2 \in (0.213, 0.478)$ .
  - $\{(\mathbf{x}'_{h_i}, \mathbf{y}_{h_i}), i : \text{Quito, Lapland}\}$  corresponds to  $\widehat{\beta}(\theta_3)$  for  $\theta_3 \in (0.478, 0.732)$ .
  - $\{(\mathbf{x}'_{h_i}, \mathbf{y}_{h_i}), i : \text{Capehope, Lapland}\}$  corresponds to  $\widehat{\beta}(\theta_4)$  for  $\theta_4 \in (0.732, 1)$ .
- Here we have only  $K = 10$  Ess (ERs) but for a data set with  $n = 50$  and  $p = 8$  we have  $K = 536\,878\,650$ .

# Quantile Regression Link to OLS

- The Linear Model can be expressed as

$$\begin{pmatrix} \mathbf{Y}_J \\ \mathbf{Y}_I \end{pmatrix} = \begin{pmatrix} \mathbf{X}_J \\ \mathbf{X}_I \end{pmatrix} \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where  $\mathbf{X}_J : p \times p$ , is non-singular,  $\mathbf{X}_I : (n-p) \times p$ ,  $\boldsymbol{\varepsilon} : n \times 1$ ,  $\boldsymbol{\beta} : p \times 1$ .

- $J^{th}$  elemental regression is

$$\hat{\boldsymbol{\beta}}_J = (\mathbf{X}'_J \mathbf{X}_J)^{-1} \mathbf{X}'_J \mathbf{Y}_J = \mathbf{X}^{-1}_J \mathbf{Y}_J,$$

where  $\mathbf{X}_J$  is square and assumed to be nonsingular.

# Quantile Regression Link to OLS

- ES Residuals (Exact fit property):

$$e_i = \begin{cases} 0, & i \in J \\ Y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_J, & i \in I \end{cases}, \quad i = 1, 2, \dots, n.$$

- Thus ESs are based only on the minimum number of observations to estimate the parameters of the model (see e.g. Hawkins *et. al.* 1984).



# Relationship between OLS and ESs

- The properties of ESs (ERs) can be developed and used study RQs under design space data aberrations.
- Also, ordinary least squares (OLS) statistics can be expressed as weighted averages of ES statistics *e.g.*

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{OLS} &= \sum_J \omega_J \hat{\boldsymbol{\beta}}_J \\ &= \sum_{J \in B(\theta)} \omega_J \hat{\boldsymbol{\beta}}_J + \sum_{J \notin B(\theta)} \omega_J \hat{\boldsymbol{\beta}}_J.\end{aligned}$$

where  $\omega_J = |\mathbf{X}'_J \mathbf{X}_J| / |\mathbf{X}' \mathbf{X}|$  is the elemental regression weight such that  $0 \leq \omega_J \leq 1$  and  $\sum_J \omega_J = 1$ .

- Furthermore, the three-tier relationship amongst RQs, ESs and OLS procedures can be used to address problems of interest in the RQ scenario.

# Relationship between OLS and ESs

- RQs are affected more adversely by design space aberrations *viz.* leverage and collinearity. This is so since RQs were designed for dealing with outliers and therefore they are fairly robust to them.
- In 1755, half a century before the advent of OLS due to Legendre's work, Boscovich used ER procedure when he was attempting to find the length of the median arc near Rome.
- Today the OLS are the standard tools of statistical analysis due to their mathematical tractability under Normality Assumptions, hence they are part of standard statistical software.
- However, the OLS are amenable to deviations from the Normality (Classical) Assumptions.

# Remarks

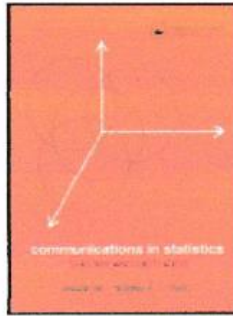
- Today the OLS are the standard tools of statistical analysis due to their mathematical tractability under Normality Assumptions, hence they are part of standard statistical software.
- OLS are amenable to deviations from the Normality (Classical) Assumptions (outliers).
- However, due the perceived complexity of the robust statistical methodology QR is still playing a second fiddle role to the OLS estimator like their robust counterparts despite,

# Remarks

- Giving a more comprehensive view of the relationship between covariates and the response variable (including extremes of the conditional distribution of response), i.e., its versatility.
- Being robust to response variable outliers.
- The inherent three-tier relationship amongst ESs (ERs), QR and the OLS which can be exploited fruitfully in model development.
- Thus, QR has to be viewed as both an alternative and complementary approach in interdisciplinary settings.

# Model Diagnostics and Inference

Stellenbosch University and Manchester University Collaborations



*Communications in Statistics - Theory and Methods*



ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: <http://www.tandfonline.com/loi/lsta20>

## Multiple Case High Leverage Diagnosis in Regression Quantiles

Edmore Ranganai, Johan O. Van Vuuren & Tertius De Wet



*Communications in Statistics - Simulation and Computation*



ISSN: 0361-0918 (Print) 1532-4141 (Online) Journal homepage: <http://www.tandfonline.com/loi/lssp20>

## A predictive leverage statistic for quantile regression with measurement errors

Edmore Ranganai & Saralees Nadarajah

# Model Diagnostics and Inference

Some individually authored articles



Contents lists available at [ScienceDirect](#)

## Statistics and Probability Letters

journal homepage: [www.elsevier.com/locate/stapro](http://www.elsevier.com/locate/stapro)



## Quality of fit measurement in regression quantiles: An elemental set method approach

Edmore Ranganai

*University of South Africa, Department of Statistics, College of Science, Engineering and Technology, Private Bag X6, Florida 1710, Roodepoort, Johannesburg, South Africa*

Ranganai *SpringerPlus* (2016)5:1231  
DOI 10.1186/s40064-016-2898-6



 SpringerPlus

**METHODOLOGY**

**Open Access**

## On studentized residuals in the quantile regression framework

Edmore Ranganai\*



# Variable Selection and Regularization

- Unlike the OLS which is susceptible to both outliers and predictor space data aberrations (collinearity and high leverage points), QR (a robust procedure) is only susceptible to predictor space data aberrations.
- QR penalized with the RIDGE ( $l_2$ -squared) penalty (Hoerl and Kennard, 1970) denoted by QR-RIDGE is given by the minimization problem

$$\arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \sum_{i=1}^n \rho_{\tau}(Y_i - \beta_0 - \mathbf{x}'_i \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p-1} \beta_j^2 \right\}, \quad i = 1, 2, \dots, n \text{ and } \lambda > 0 \text{ (tuning parameter)}.$$

- QR penalized with the LASSO ( $l_1$ ) penalty (Hoerl and Kennard, 1970) denoted by QR-LASSO is given by the minimization problem

$$\arg \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \sum_{i=1}^n \rho_{\tau}(Y_i - \beta_0 - \mathbf{x}'_i \boldsymbol{\beta}) + n\lambda \sum_{j=1}^{p-1} |\beta_j| \right\}, \quad i = 1, 2, \dots, n \text{ and } \lambda > 0.$$

- While QR-RIDGE does not shrink any coefficients to zero (fails to select any variables) QR-LASSO tend to be too “greedy”.

## Variable Selection and Regularization

- Therefore a compromised version of the approaches is a combination of the two penalties via QR penalized with the elastic NET penalty (QR-E-NET) given by

$$\arg \min_{\beta \in R^p} \left\{ \sum_{i=1}^n \rho_{\tau}(Y_i - \beta_0 - \mathbf{x}_i' \boldsymbol{\beta}) + \alpha \lambda \sum_{j=1}^{p-1} |\beta_j| + (1-\alpha) \lambda \sum_{j=1}^{p-1} \beta_j^2 \right\}, \quad i = 1, 2, \dots, n \text{ and } \lambda \geq 0,$$

where  $\alpha \in [0,1]$  is the mixing parameter between RIDGE ( $\alpha = 0$ ) and LASSO ( $\alpha = 1$ ).

- Adaptive versions of penalized QR where the tuning parameter  $\lambda_j = \lambda \omega_j$  for  $i = p-1$ , tend to perform better than their non-adaptive counterparts.





# Variable Selection and Regularization

Work with student



*Article*

## Variable Selection and Regularization in Quantile Regression via Minimum Covariance Determinant Based Weights

Edmore Ranganai <sup>1,\*</sup>  and Innocent Mudhombo <sup>2</sup> 



*Article*

## Robust Variable Selection and Regularization in Quantile Regression Based on Adaptive-LASSO and Adaptive E-NET

Innocent Mudhombo  and Edmore Ranganai <sup>\*</sup> 

## Take home message

- Critical review on postgraduate curriculum.
- Developing robust statistics that parallels the OLS.
- Sensitize young students to research in/ using QR.
- More workshops.
- More interdisciplinary collaboration using QR.



# Thank you

Define tomorrow.

UNISA

