

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Ian Meneghel Danilevicz

Robust linear mixed models, alternative methods to quantile regression for panel data, and adaptive LASSO quantile regression with fixed effects

Belo Horizonte
2022

Ian Meneghel Danilevich

Robust linear mixed models, alternative methods to quantile regression for panel data, and adaptive LASSO quantile regression with fixed effects

Final Version

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Computer Science.

Advisor: Valdério Anselmo Reisen
Co-Advisor: Pascal Bondon

Belo Horizonte
2022

Danilevicz, Ian Meneghel.

A283b Robust linear mixed models, alternative methods to quantile regression for panel data, and adaptive LASSO quantile regression with fixed effects [manuscrito] / Ian Meneghel Danilevicz. – 2022.

189 f. il.

Orientador: Valderio Anselmo Reisen.

Coorientador: Pascal Bondon.

Tese (doutorado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística.

Referências: f. 109-119

1. Estatística – Teses. 2. Dados longitudinais – Teses. 3. Regressão quantílica, – Teses. 4. M-estimation – Teses. 5. Observações atípicas – Teses I. Reisen, Valderio Anselmo. II. Bondon, Pascal. III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Estatística. IV. Título.

CDU 519.2(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais - ICEx



UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROGRAMA DE PÓS-GRADUAÇÃO EM ESTATÍSTICA



FOLHA DE APROVAÇÃO

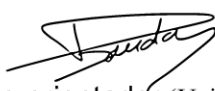
"Robust linear mixed models, alternative methods to quantile regression for panel data, and adaptive LASSO quantile regression with fixed effects"

IAN MENEGHEL DANILEVICZ

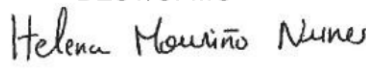
Tese submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em ESTATÍSTICA, como requisito para obtenção do grau de Doutor em ESTATÍSTICA, área de concentração ESTATÍSTICA E PROBABILIDADE.

Aprovada em 15 de dezembro de 2022, pela banca constituída pelos membros:


Prof(a). Valderio Anselmo Reisen - Orientador
DEST/UFES

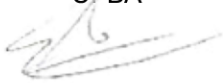

Pascal Bondon Co-orientador (Université Paris-Saclay)


Prof(a). Glaucia da Conceicao Franco
DEST/UFMG


Prof(a). Maria Helena Mouriño Silva Nunes
Universidade de Lisboa


Prof(a). Arthur Tenenhaus
CentraleSupélec


Prof(a). Paulo Canas Rodrigues
UFBA


Prof(a). Stephane Chretien
University of Lyon 2

Belo Horizonte, 15 de dezembro de 2022.

Acknowledgments

I would like to thank the committee that accepted to be part of my final evaluation, professors Maria Helena Mourino Silva Nunes, Glauro da Conceição Franco, Stéphane Chrétien, Paulo Cañas and Arthur Tennenhaus. I would also like to thank the professors who took part in the mid-term evaluation, professors Sara Taskien and Liudas Giratis.

I would like to thank the Universidade Federal de Minas Gerais workers who maintain this University, which I am proud to have been a part of for at least a brief period, Rogéria Ferreira Figueiredo (Secretary of the PPGE), Wilton Soares Ribeiro (IT), Samira Lott (DRI) and Pedro Fiuza (DRI). To the coordinators of the PPGE, who gave me great support to carry out a cotutelle, professors Flávio Bambirra Gonçalves and Rosângela Helena Loschi. To the professors who taught the courses I attended at UFMG, especially Luiz Henrique Duczmal and Sokol Ndreca. To my doctoral colleagues, especially my probability colleagues, Fernanda Gabriely Batista Mendes, Patrícia Viana da Silva and Walmir Reis Miranda. I would like to thank Ricardo Cunha Pedroso for the help in programming with RCPP, Ana Júlia Alves Câmara for reading the first version of my mid-term evaluation, and Uriel Moreira Silva for helping me with Probably.

I would like to thank the workers of Université Paris-Saclay and CentraleSupélec who helped me a lot during this period in France, Stéphanie Douesnard (HR), Anne Batalie (HR), Thomas Cuidu (IT), Christian David (IT). I would especially like to thank the Science Accueil team, it was vital in my adaptation, without it I would not even have been able to declare my income tax correctly. I would like to thank my colleagues, especially Higor Henrique Aranda Cotta, who helped me to understand how the French University system worked, and Stefano Fortunati, with whom I was lucky enough to share the worst coffees and the best conversations during the pandemic.

I would like to thank the workers of Universidade Federal do Espírito Santo, which hosted me informally for one year. I would like to thank my co-author Faradiba Sarquis Serpa, who kindly shared her research data. I would like to thank my friends who helped me in different ways, especially those who helped me get the French visa, such as Ricardo André Cecchin, Marco Pollo Almeida, and Hécio Martinho.

I would like to thank my supervisors who agreed to guide me, Valdério Anselmo Reisen and Pascal Bondon, who are for me as Plato and Aristotle. One imagines the impossible and the other returns to the real world, and he teaches me to be more precise and clear.

Finally, I would like to thank Paula.

“Etre un tout, c’est être une partie; le vrai voyage est le retour.”
(Laia Asieo Odo)

Resumo

A tese é composta por três capítulos. O primeiro enfoca a relação entre exposição à poluição do ar e doenças respiratórias em crianças e adolescentes. A coorte inclui 82 indivíduos observados mensalmente durante seis meses. Propomos um modelo linear misto robusto combinado com uma análise de componentes principais para gerenciar a multicolinearidade entre as covariáveis e o impacto de observações extremas nas estimativas. O segundo capítulo analisa dados em painel usando modelos de efeitos fixos e diferentes funções de perda. Para evitar que a dimensão paramétrica aumente com o tamanho da amostra, penalizamos cada método de regressão por LASSO. As propriedades assintóticas dessas novas técnicas são estabelecidas. Testamos o desempenho dos métodos com dados de painel econômico da OCDE. O objetivo almejado no terceiro capítulo é restringir as constantes de regressão individuais e as covariáveis explicativas simultaneamente. O LASSO adaptativo reduz a dimensionalidade garantindo assintoticamente a seleção correta do modelo. Testamos a precisão dos métodos propostos em dados de coorte de tamanho moderado.

Palavras-chave: dados longitudinais; regressão quantílica; estimação m; observações atípicas; seleção de variáveis; lasso

Abstract

The thesis consists of three chapters. The first focuses on the link between exposure to air pollution and respiratory diseases in children and adolescents. The cohort includes 82 individuals observed monthly for six months. We propose a robust linear mixed model combined with a principal component analysis to manage the multicollinearity between the covariates and the impact of extreme observations on the estimates. The second chapter analyzes panel data using fixed effects models and different loss functions. To avoid the parametric dimension increases with the sample size, we penalize each regression method by LASSO. The asymptotic properties of these new techniques are established. We test the performance of the methods with economic panel data from the OECD. The goal aspired in the third chapter is to constrain the individual regression constants and the explanatory covariates simultaneously. The adaptive LASSO reduces dimensionality by asymptotically ensuring the right model selection. We test the accuracy of the proposed methods on cohort data of moderate size.

Keywords: longitudinal data; quantile regression; m-estimation; outliers; variable selection; lasso

List of Figures

1.1	Squared loss function (red), Huber loss function (green) and Tukey loss function (blue).	18
1.2	Check function (red), M-quantile (green) and asymmetric squared function (blue) to $\tau = 0.3$	22
1.3	Funções perda quadrática (vermelho), Huber (verde) e Tukey (azul).	30
1.4	Funções <i>check</i> (vermelho), M-quantile (verde) e quadrática assimétrica (azul) para $\tau = 0.3$	34
1.5	Fonction de perte quadratique (rouge), fonction de perte de Huber (vert) et fonction de perte de Tukey (bleu).	43
1.6	Fonction <i>check</i> (rouge), fonction M-quantile (vert) et fonction quadratique asymétrique (bleu) par $\tau = 0.3$	47
2.1	sample mean (SM) and sample coverage probability (SCP) with 95% of confidence when $\xi = 0$, $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 1$. LS estimates (red square), Huber estimates with $\kappa_2 = 1.345$ (green circle) and Tukey estimates with $\kappa_3 = 4.685$ (blue triangle).	62
2.2	SM and SCP with 95% of confidence when $\xi = 0.5$, $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 1$. LS estimates (red square), Huber estimates with $\kappa_2 = 1.345$ (green circle) and Tukey estimates with $\kappa_3 = 4.685$ (blue triangle).	62
2.3	SM and SCP with 95% of confidence when $\xi = 0.9$, $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 1$. LS estimates (red square), Huber estimates with $\kappa_2 = 1.345$ (green circle) and Tukey estimates with $\kappa_3 = 4.685$ (blue triangle).	66
2.4	Estimated standardized residuals of Huber-robust principal component analysis (RPCA) model with AR(1) and independent errors.	68
3.1	Within each boxplot λ is estimated by quantile regression with fixed effects (QRFE) (red at the left), M -quantile regression with fixed effects (MQRFE) (green at the middle), expectile regression with fixed effects (ERFE) (blue at the left); and the plots follow LS (above) and LSS (below) models and Normal (on the left), t_3 (on the center) χ_3^2 (on the right) distributions.	79

3.2	Bias, RMSE and CP with 95% of confidence, LS model estimate by quantile regression (QR) (dotted line with cyan rhombus), QRFE (dashed line with red square), MQRFE (dashed line with green circle), ERFE (dashed line with blue triangle), penalized quantile regression with fixed effects (PQRFE) (solid line with red square), penalized M -quantile regression with fixed effects (PMQRFE) (solid line with green circle), penalized expectile regression with fixed effects (PERFE) (solid line with blue triangle).	81
3.3	Bias, RMSE and CP with 95% of confidence, LSS model estimate by QR (dotted line with cyan rhombus), QRFE (dashed line with red square), MQRFE (dashed line with green circle), ERFE (dashed line with blue triangle), PQRFE (solid line with red square), PMQRFE (solid line with green circle), PERFE (solid line with blue triangle).	82
3.4	exploratory data analysis (EDA) of high-technology exports (% of manufactured exports) (HTE), government expenditure on education (% of GDP) (EDU) and research and development expenditure (% of GDP) (R&D)	83
3.5	EDU and R&D coefficients and confidence intervals by QRFE, MQRFE and ERFE	84
3.6	$\alpha_i(\tau)$ coefficients and confidence intervals by MQRFE	85
4.1	Histogram for strengths and difficulties questionnaire - internalizing (SDQ-I) (left) and strengths and difficulties questionnaire - externalizing (SDQ-E) (right)	90
4.2	mean squared error (MSE) of estimates by QRFE (red square), LASSO quantile regression with fixed effects (LQRFE) (green circle) and adaptive LASSO quantile regression with fixed effects (ALQRFE) (blue triangle), to Normal, t_3 and χ_3^2 models with $d = 4, 40$	98
4.3	mean absolute error (MAE) of estimates by QRFE (red square), LQRFE (green circle) and ALQRFE (blue triangle), to Normal, t_3 and χ_3^2 models with $d = 4, 40$	99
A.1	The SM, sample mean squared error (SMSE) and SCP with 95% of confidence of estimates on scenario 1.1. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.	122
A.2	The SM, SMSE and SCP with 95% of confidence of estimates on scenario 1.2. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.	122
A.3	The SM, SMSE and SCP with 95% of confidence of estimates on scenario 1.3. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.	123
A.4	The SM, SMSE and SCP with 95% of confidence of estimates on scenario 1.4. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.	123

A.5	The SM, SMSE and SCP with 95% of confidence of estimates on scenario 1.5. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.	124
A.6	The SM, SMSE and SCP with 95% of confidence of estimates on scenario 1.6. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.	124
A.7	The SM, SMSE and SCP with 95% of confidence of estimates on scenario 2.1. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed. Time series with $m_i = 50$	125
A.8	The SM, SMSE and SCP with 95% of confidence of estimates on scenario 2.2. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.	126
A.9	The SM, SMSE and SCP with 95% of confidence of estimates on scenario 2.3. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.	127
A.10	The SM, SMSE and SCP with 95% of confidence of estimates on scenario 2.4. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.	183

List of Tables

1.1	Acronyms of regression models according to their characteristics and form of estimation	22
1.2	Acronyms of the regression models according to the type of penalty.	25
1.3	Acrônimos dos modelos de regressão conforme suas características e forma de estimação	34
1.4	Acrônimos dos modelos de regressão conforme o tipo de penalização.	37
1.5	Acronymes des modèles de régression en fonction de leurs caractéristiques et de leur forme d'estimation	47
1.6	Acronymes des modèles de régression selon le type de sanction.	50
2.1	SM, SMSE and SCP with 95% of confidence when $\xi = 0, 0.5$, $\delta = 0, 0.02$, $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 2$. linear mixed model (LMM) estimated by least squares (LS), Huber (H) and Tukey (T) with $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$, respectively.	63
2.2	SM, SMSE and SCP with 95% of confidence when $\xi = 0, 0.5$, $\delta = 0, 0.002$, $(n, m_i) = (40, 50)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 2$. LMM estimated by LS, Huber (H) and Tukey (T) with $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$, respectively.	64
2.3	SM, SMSE and SCP with 95% of confidence when $\xi = 0, 0.5$, $\delta = 0, 0.002$, $(n, m_i) = (40, 50)$, $(\phi_{1,1}, \phi_{2,2}) = (0.7, 0.6)$, $\mu = 2$. LMM estimated by LS, Huber (H) and Tukey (T) with $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$, respectively.	65
2.4	Correlations between environmental covariates.	66
2.5	Correlation between environmental covariates after replacing the 10% smallest and greatest values by the average of each covariate.	66
2.6	Loadings of the three first principal component (PC) of the classical principal component analysis (PCA) and RPCA, and cumulative proportion of the total variation (CPV).	67
2.7	Estimated parameters, standard errors (SE) and p-values in the fitted LMM LS-PCA, LS-RPCA, Huber-RPCA and Tukey-RPCA.	68
4.1	Estimated coefficient, SE and p-value of selected covariates for SDQ-E model by QRFE to $\tau = 0.1, 0.5, 0.9$	101
4.2	Estimated coefficient, SE and p-value of selected covariates for SDQ-E model by ALQRFE to $\tau = 0.1, 0.5, 0.9$	101

4.3	Estimated coefficient, SE and p-value of selected covariates for SDQ-I model by QRFE to $\tau = 0.1, 0.5, 0.9$	103
4.4	Estimated coefficient, SE and p-value of selected covariates for SDQ-I model by ALQRFE to $\tau = 0.1, 0.5, 0.9$	103
A.1	Number and label of the databases	127
A.2	Old random variables	184
A.3	New random variables	186
A.4	New random variables (continue).	187
A.5	Average time (in seconds) of each method for model XX in each Monte Carlo scenario with $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 1$, $d = 3$	187
A.6	Average time (in seconds) of each method for model XX in each Monte Carlo scenario with $(n, m_i) = (50, 5)$, $d = 1$	188
A.7	Average time (in seconds) of each method for model XX in each Monte Carlo scenario with $(n, m_i) = (50, 5)$, $d = 4, 40$	188

Contents

1	Introduction	15
1.1	English Introduction	15
1.2	Introdução em Português	27
1.3	Introduction en Français	40
2	A longitudinal study of the influence of air pollutants on respiratory health. A robust multivariate approach	53
2.1	Introduction	54
2.2	Linear mixed model	56
2.3	Monte Carlo simulations	58
2.4	Real data analysis	63
2.5	Conclusion	69
3	Alternative methods to quantile regression for panel data	70
3.1	Introduction	70
3.2	Model and estimation methods	72
3.2.1	The model	72
3.2.2	Estimation methods: quantile, expectile and M -quantile approaches	73
3.2.3	Asymptotic results for expectile regression with fixed effects	74
3.3	Monte Carlo study	75
3.3.1	Computational considerations	75
3.3.2	Monte Carlo design and measures of accuracy	77
3.3.3	Empirical results	78
3.4	Real data application	80
3.5	Conclusion	84
3.6	Appendix	86
4	Adaptive LASSO Quantile Regression with Fixed Effects with an application to the Millennium Cohort Study	87
4.1	Introduction	87
4.2	Data	89
4.3	Estimation methods	90
4.3.1	Quantile regression with fixed effects	91
4.3.2	Penalized quantile regression with fixed effects	92

4.3.3	Asymptotic results	92
4.4	Monte Carlo study	93
4.4.1	Computational considerations	93
4.4.2	Design and measures	96
4.4.3	Empirical results	97
4.5	Real data application	98
4.5.1	First outcome: externalizing difficulties	99
4.5.2	Second outcome: internalizing difficulties	100
4.6	Conclusion	104
5	Conclusion and perspectives	106
	References	108
	Appendix A	119
A.1	Correction factors	119
A.2	Weighted robust linear mixed models: additional simulation	120
A.2.1	Proposed simulation	120
A.2.2	Results	121
A.2.3	Conclusion of simulation	121
A.3	Millennium Cohort Study covariates	126
A.4	R packages	127
A.4.1	Robust linear mixed models	128
A.4.2	Penalized quantile regression with fixed effects	134
A.4.3	Adaptive Lasso Quantile Regression with Fixed Effects	161
A.5	Knight's identity	183
A.6	Computational performance	185

Chapter 1

Introduction

1.1 English Introduction

What is the effect of air pollutants on children and adolescents' health? This question is not fully answered in [102] considering robust methods. Because an aberrant or atypical observation can dramatically change the estimates of any linear model based on squared errors. Therefore, a new and more precise question must be done, taking into account that real data is subject to extreme fluctuations. In this sense, what is the average effect of air pollutants on the health of children and adolescents affected by atypical observations?

Before getting into the methodological question about robust alternatives and how much they protect against interference caused by additive outliers, I describe the data collected by [102] briefly. She researched a group of 82 children and adolescents in the city of Vitória Brazil. They come from low-income families, they live in the same neighborhood and study at the same school. Some of them suffer from asthma and some do not. The study was approved by the Ethics and Research Committee of the Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória-ES, CAAE 53176915.8.0000.5065, and by the Secretaria Municipal de Saúde de Vitória through the Escola Técnica do Sistema Único de Saúde (ETSUS). To determine whether the child is asthmatic or not, a clinical examination was performed, later the children are instructed to perform a self-test to measure their levels of respiratory capacity. Three measurements are performed each time and only the median is considered to lower the measurement error, because the self-test expected level of measurement errors is elevated. The measurements are repeated over six months, configuring a longitudinal study, in addition, socio-economic covariates of the sampled children are considered fixed as gender, age and breastfeeding, or over time as body mass index, if own a pet and passive smoking status. Observational climatological variables such as temperature, humidity and air pollutants such as CO_2 , SO_2 , $\text{PM}_{2.5}$, PM_{10} are also included. The pollutants are obtained by measurements from the Instituto do Meio Ambiente e Recursos Hídricos do Espírito Santo (IEMA-ES), which has 8 meteorological stations in the metropolitan region of Vitória, one of them close to the school where the young volunteers study. Each child performed the self-test on a different day of the month, thus, meteorological information from the previous day was

associated with each measurement of the child, as the effect of pollution tends to take a day to reach its maximum expression. Despite not being large, this sample presents a complex structure because there are individual and temporal dependences. The spatial dependence is restricted to a single neighborhood, but future works can explore this issue broadly. In addition to the multicollinearity present between air pollutants, as this is an observational study and not a controlled experiment, there is no way to perfectly separate the effect of each pollutant, since many of them come from similar sources such as traffic and industry.

Before proceeding, I would like to make a summary of the nomenclatures that I use in this thesis. The most natural and synthetic explanation which I know is available in [120, pages 1-2]. There he exposes the differences between similar terms with an example of a blood pressure test. If an individual has his systolic and diastolic blood pressure measured by a nurse we have a very simple example of multivariate data. If two families go to a health center and each of their members has their diastolic pressure measured, then we have information in clusters. If the nurse suspects that the tensiometer is defective, she can measure a patient twice to check, in which case we have repeated measurements. If these repeated measures are taken over time, let's say someone goes to the health center every Monday, then we have longitudinal data or panel data that we treat synonymously here and it is mostly used in econometrics. Understanding how the data hierarchy is structured helps the analyst to propose the model with the best qualities for the problem at hand. As for example, the most appropriate variance-covariance structure to be adopted for model residuals. Linear mixed models comprise a general classification of models that include a fixed component and a group of random parameters such as random intercepts or random coefficients. Repeated measures, longitudinal data or panel data can be modeled by LMM. Cluster analysis, factor analysis and principal components analysis, although they are ultimately multivariate data, do not present themselves as the best alternative to this type of data, in fact, they tend to have poor results in these cases [120], but there are exceptions such as I will show later.

[102] analyses her database without taking into account robust methods, but what is robustness? Robustness is a very broad term in statistics that usually means that a method is resistant to an assumption's violation. Let's say, even if some conditions for the application of the method are not achieved, the method can still be used and the results are expected to be similar to those in which the assumptions are fully achieved. For example, the analysis of variance (ANOVA) is said to be robust in relation to normality, which is an assumption that does not need to be fully achieved. In this thesis, the term robustness is mostly used as robustness for additive outliers. Observed meteorological observations have this characteristic, as they are observed every few minutes a day, there are times when the average rises extremely high and quickly drops to the average level.

There are books about LMM like [120, 90, 24]. There are also suitable books about

robust methods like [51, 82]. A fair introduction to robust LMM may be done by reading [42, 131]. To the best of my knowledge, there is no book that combines M-estimation and LMM, despite the title [57] is more related to generalized estimating equations (GEE). Robustness against outliers can be obtained by assuming a heavy-tailed distribution like the student's t distribution, or it is possible to let the data speak for itself assuming a loss function with adaptability like bisquared Tukey or winsorizing Huber, which ultimately does that. If the percentage of the contamination and the length of the outliers is not big, then heavy tail distributions or Huber result in similar estimations, but Huber function has interesting flexibility to deal with Gaussian and non-Gaussian data. [42, 97, 135] among many others work on mixed models with robust estimation and they point [53] as the father of the proposal. He offers a robust likelihood to repeated measures model with a known variance structure. To prove the consistency of his proposition, the author is extremely synthetic, he uses the Theorem 3.1 of [19]. However, this theorem is quite general and deals with several cases of estimation, but not necessarily with repeated measures or LMM. It seems to be an unfortunate choice because in [49] it is already said that if it exists a likelihood, then the estimator is consistent for sure. The likelihood assumptions and their consequences are exhaustively discussed in [50]. At the time Huggins published his work, Huber's theory was already very well explained in detail by [49, 48, 50, 51, 136]. So it seems like an active choice the author takes. [53] demands the existence of a second derivative of the loss function to guarantee the Taylor expansion as [51] to prove the convergence of simple linear regression. This second derivative requirement and the use of the Taylor expansion is also explored in the proof sketches proposed by [58, 105, 106]. However, the most complete proof of the consistency of the robust estimator for mixed model is in [104]. The later paper explores the Huber loss function and the weighting proposed by [15]. It is strange that a book dedicated to the asymptotic part of mixed models like [56] with a chapter on robust estimation does not cite either [53, 104].

In addition to the issue of robustness to outliers, there are points of influence and leverage. If the data does not come from a balanced designed experiment, but from an observational study, then the own matrix of covariates may have influent observations. This matrix of covariates should be investigated more deeply. [15] proposed to weight the observations to decrease the impact of influent or leverage points. These are a kind of multidimensional outliers. As the robust likelihood is constructed using the Huber function, outliers appear in the residuals, regardless of whether they come from a discrepant measure in the response variable or the design matrix. However, in observational studies where there is little control over the covariate matrix, this subtlety can make a difference. Because through the projection matrix, it is already possible to estimate which observations are too influential in the model and penalize them. Consequently, the observations have a more balanced weight in the final model. This proposition of [15] does not contradict the functions of [49], but they are easily harmonized in the same likelihood as

explored by [104, 71]. The asymptotic consistency is proved by [104].

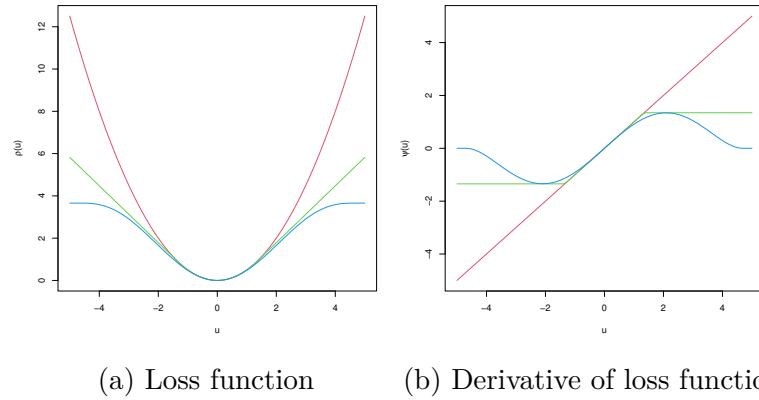


Figure 1.1: Squared loss function (red), Huber loss function (green) and Tukey loss function (blue).

Figure 1.1 shows the quadratic loss functions, Huber and Tukey on the left and their derivatives on the right. Clearly, the quadratic and Huber functions are convex and the Tukey function is not outside the range between $-c/\sqrt{5}$ and $c/\sqrt{5}$. The suspicion that non-convex functions do not generate consistent estimators is old, [49] suggests that additional restrictions should be imposed. [70, page 45] suggests for the case of the Tukey function that good starting points are enough for convergence. [78] is more radical and requires that most of the residuals need to fall into the convex region of the non-convex function and that it must be symmetric, which is the case for the Tukey function. Therefore the value of c can not be too small. [51, page 103] warns that using redescending functions like Andrews, Hampel and Tukey can be quite dangerous and that even manually removing “impossible” observations can be a more cautious alternative. This study of the loss function properties is vital for mixed model context and for the next subject of the thesis: quantile regression. For example, combining a check function, which is skewed, with a non-convex loss function like Tukey, might be very problematic from a convergence point of view. Apparently, this is a path that does not bring good results.

I do not follow a path in the thesis to optimize the tuning parameter automatically like [127, 8]. Because I could not find a quantity that was non-monotone or that do not cause more instability to the model. Thus I prefer to keep the traditional values that correspond to a relative efficiency of 95% compared to the average least square estimator [82, page 357]. Establishing a compact region with a maximum is a good starting point for anyone who desires to keep this investigation [31].

However, strong correlations between covariates are the main weakness of all these methods, as we saw in our simulations. And the environmental data from Vitória are quite complex and correlated, so we follow the previous authors’ proposal to reduce the data’s dimensionality by a robust technique of principal component analysis [17]. There

are several possible techniques for this, which explore different robust ways of estimating the variance of the data [18, 80, 52]

With the combination of the exposed methodologies, we are able to analyze [102]’s database and consider the effect of pollutants despite the discrepant observations. The model considers the characteristics inherent to each child, such as their allergic level and whether the child is a passive smoker. An obvious dummy variable as the presence/absence of asthma is not included in the model, because the asthmatic group uses the properly medicines. Therefore, asthmatics and non-asthmatics showed no significant difference in respiratory capacity measured over the months. In this way, we contribute to the public health debate with this study that combines a sampling in which children’s factors are controlled by an experiment design but climatological information is added, characteristic of observational studies. I emphasize that with these robust methodologies it was not necessary to delete any sample unit, which would be a problem if the sample size was smaller.

What is the effect of air pollutants on the health of the most vulnerable children and adolescents, for example, the most vulnerable 10% of children, or the most vulnerable 1%? This second question structures the second part of the thesis, in which the main issue is about quantile regression applied to longitudinal data.

To answer this question I took a tortuous path, which began with [107]. He proposed to unify quantile regression and copulas, giving an alternative interpretation to percentils. In which, the percentil is no longer a value established by the analyst to fit a hyperplane and becomes a hyperparameter of the model. This method is very sensitive as it depends on the estimation and transformation of residuals. Reading articles that combined quantile regression and repeated measures or longitudinal data, I realized that there is a consensus: all articles cited [65]. Just as [66] is a seminal paper for quantile regression, [65] is the master paper for quantile regression for longitudinal data. However, reading this article is not easy because it is difficult to understand why the author treats individual intercepts as fixed effects.

There are works that used the asymmetric Laplace distribution (ALD) like [61, 39, 40, 33]. The closed expression of a likelihood looks solid, and the analyst chooses the percentile, resulting in expressions whose asymmetry increases as the percentile approaches the extremes: zero or one. The estimation procedure is the canonical Expectation and maximization (EM). There is an available algorithm in R language ready to be downloaded and modified. There are many possibilities for improvements including Bayesian priors. Later, I read the work of [118], who proposed an M-quantile for mixed models or an M-quantile for random effects and promised to extend the method to mixed models in future works. [118] cited works like [53, 97, 106]. They proposed a change in Huggins’ robust likelihood following the type II correction proposed by [131] for robust estimation of the mean. This quantile plugin instead of an average struck me as weird and unlikely.

Another valuable citation from this work is the study developed by [10]. This is the inaugural article on the M-quantile method for simple regression. This article is not cited more because one of the main appeals of M-regression for applied work is its robustness and quantile regression is already robust. The other advantages listed in [68, 11, 62] are not so obvious or trivial. However, the same issue still remains. Is it possible to use an M-quantile function as a plugin on a built-to-mean likelihood without degenerating that likelihood?

There is a literature review on quantile regression for independent and longitudinal data, as well as pointing out the difference between authors who prefer marginal or conditional estimation [81]. Quantile regression is in vogue and there are a lot of recent works after 2015. However, the map sketched by Marino is still a great starting point. They describe the progress of quantile regression from independent data to dependent data. Thus they highlighted the difference between marginal and conditional estimation in longitudinal data. I explored some of the marginal methods, for which a fundamental reference is [32]. As he uses a working variance-covariance matrix [128, 20, 94]. But Marino's articles also pointed to works on the conditional estimation that continued Koenker's paper as [73, 37].

Returning to [65], the author explains that minimizing the BLUP of a model with random intercepts is equivalent to minimizing a simple regression model with many intercepts where the intercepts suffer a restriction: the sum of their squares cannot exceed a certain value. This Ridge-like regression construction solves two problems. First, the model's estimation procedure considers the sample units' individual effects on the longitudinal data. Second, there is a restriction on the number of parameters. This answers why Ridge and BLUP are efficient techniques. The author cites [98] who credits the idea to a paper that apparently only the abstract [46] is left. Koenker proposes to perform a quantile regression for a model with multiple intercepts that are penalized by least absolute shrinkage and selection operator (LASSO). Finally, it is clear why Koenker does not use the term LMM. However, so many fixed effects are still a bit of a confusing issue for me. Fortunately, there is an easy explanation for the analysis of longitudinal data like [137]. Repeated measures might be divided into two types of analysis: random effects and fixed effects. The former are generally assumed to have zero expectation and to be independent of the other explanatory variables. Due to these properties, the first class may omit the location effect of random effects and only include them in the variance structure. The second class, relax those assumptions and treated the individual effect as dummy variables. However, the number of parameters grows along with the sample size and the estimation becomes inefficient [2, 47, 38, 137]. Henderson's insight is to be flexible in assumptions and assume fixed effects, but to be practical in estimation and penalize them by LASSO which amounts to random effects. In Koenker's words, this demonstration of equivalence and balance opens the doors to other penalties and methodological proposals.

It represents the dawn of the quantile regression with fixed effects. A natural extension is proposing an M-quantile for fixed effects instead of random like [118].

So far I have written about expectile and quantile regression without going into details, but each method has its history. Quantile regression becomes famous after the article of [66] that generalizes the non-parametric concept of univariate quantile to a conditional relationship between at least two random variables or a design matrix and a response variable. The authors point out that, unlike the univariate case in which the quantile has a closed formula, the bivariate case was open to discussion, it could be estimated by the check function, but could also be estimated by ranks, among other possibilities. [86] proposes a return to the average, more specifically the expectiles which are a form of expected quantiles. Using instead of the check function an asymmetric quadratic function. His method has good properties such as lower risk of quantile crossing and better estimation of extreme quantiles. However, this method is sensitive to the probability distribution of errors, being evidently not robust to outliers. Quantile crossing is the phenomenon in which a predicted quantile does not increase monotonously for a fixed point of $\mathbf{X}$. For example, suppose we estimate the coefficients of a regressor for two percentiles: 0.3 and 0.6. The value predicted by the 0.3 coefficient times a fixed value of $\mathbf{X}$ must be less or equal to the value predicted by the 0.6 percentile. If not, the assumption that the cumulative probability function is increasing monotonously is not satisfied. Often the empirical estimation of quantile regression can result in quantile crossing. The expectile regression is also not free of this problem, but it is less frequent. One year later, we have the work of [10], who proposes a balance between the quantile and the expectile: the M-quantile. Just as Huber weighs between the mean and the median with different loss functions, Breckling's work weighs between the quantile and the expectile. The authors also mention that the estimate is not unique, it can be obtained by the loss or its derivative, which can result in small differences depending on the type of problem. Finally, we have an article that proves that the three proposals, although with different names, are equally and true quantiles [59].

Figure 1.2 displays the check, M-quantile and asymmetric quadratic functions. On the left are the losses and on the right are their derivatives. Clearly, all functions are convex, making estimation and consistency proofs much easier to achieve. The derivative of the check function has a discontinuity point at the root which can cause numerical instability depending on the type of problem to be minimized. In this example, the percentile chosen is 0.3, but more extreme values result in stronger skewness. Values greater than 0.5 invert the asymmetry relationship, e.g., 0.7 would mirror this graphic horizontally.

So far I have written about three types of optimization: quantile, M-quantile and expectile. In addition, we have four possible models: simple regression, random effects, fixed effects, and fixed effects with LASSO penalty. The combination of these models with

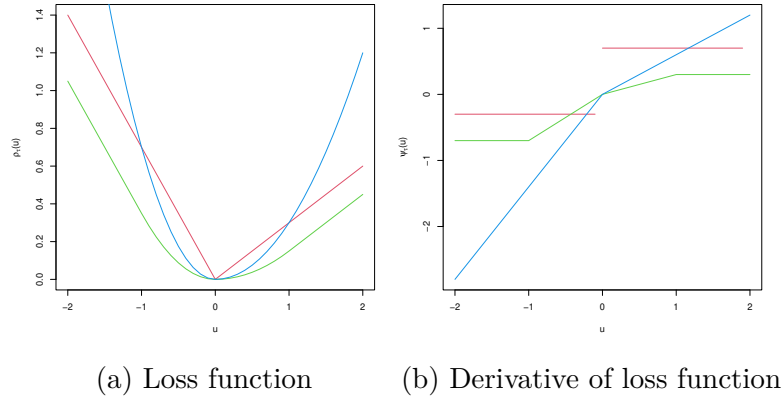


Figure 1.2: Check function (red), M-quantile (green) and asymmetric squared function (blue) to $\tau = 0.3$.

Table 1.1: Acronyms of regression models according to their characteristics and form of estimation

Intercept(s)	Quantile	M-quantile	Expectile
Single	QR	MQR	ER
Random	QRRE/LQMM	MQRRE	ERRE
Fixed	QRFE	MQRFE	ERFE
Penalized	PQRFE	MQRFE	PERFE

those functions results in 12 methods without considering the details which each author adds to her or his method such as splines, EM, etc. Comparing so many methods results in several acronyms and the reading becomes hard and tiresome. Therefore, in Table 1.1 I prepare a summary. About the intercepts characteristics: if there is a single intercept, then it is a simple regression, if they are assumed random, then they are random effects or mixed models, if they are interpreted as fixed, then they form a set of dummy variables, if they are constricted, then they are penalized fixed effects. A brief mention of the authors who proposed the respective methods follows:

- QR was proposed by [66],
- LQMM was proposed by [39],
- QRFE was proposed by [65],
- PQRFE was proposed by [65],
- MQR was proposed by [10],
- MQRRE was proposed by [118],
- MQRFE is proposed by us in this thesis,
- PMQRFE is proposed by us in this thesis,

- ER was proposed by [86],
- ERRE was proposed by [7, 123],
- ERFE is proposed by us in this thesis,
- PERFE is proposed by us in this thesis.

A path that we did not follow in the thesis is to optimize the tuning parameter of the M-quantile automatically, as some authors did for the M-estimation of the average [127]. Even an analytical choice of the tuning constant is an open problem. To choose a static constant, some authors compare the variance of the M-estimation with the least square estimator under the normality of errors, where the least square has the lowest variability [82, page 276]. However, most works about M-quantile compare their results with quantile, which is the generalization of the median, not the mean. If you choose $c = 1.345$, then your M-estimation has 95% of the efficiency of the least square under Gaussianity, but 95% is completely meaningless information about the median. Furthermore for other quantiles. Works like [5, 88] do not conclude which value of c is optimal in a quantile framework. Thus, recycling values like $c = 1.345$ seems like an inappropriate copy and paste, but repeated in the literature [100, 9, 118]. One of the few works that analyze the impact of different values of constant tuning is [62]. For this reason, we ended up choosing $c = 1$ according to [69, 62].

The asymptotic consistency and normality of the estimators are proved by [65] for the QRFE method. The author left how to estimate the constraint parameter as an open problem, but [73] proves the consistency of the tuning parameter. Koenker uses the convergences explained by [64] and an identity from [63]. This identity is only valid for the check function and cannot be applied to M-quantile or expectile, unfortunately. Even though, the asymptotic results of the new models are partially resolved. We have already proved the consistency and normality of the expectile estimators, that is, for the ERFE. But the same has not yet been done for the M-quantile: MQRFE. The quadratic form of the expectile helps with proofs. We presume that M-quantile for fixed effects is consistent and that the proof is analytically possible. Because M-regression is consistent [51], M-regression with LASSO has estimators which are asymptotic Gaussian [108], and M general linear mixed models are consistent [104, 106].

Without definitive mathematical proof, we have good results from Monte Carlo simulation and application to real data that seem to prove our claim. To illustrate the method with real data, I looked for panel data in which I could be very confident that there were no errors and whose relationship between the covariates was already a consensus in the specific literature. Although I started this part of the text by asking a question about the effect of air pollutants on the most vulnerable children, alluding to issues related to public health, the database I chose for this second part is of an economic nature. In

this way, the method proves to be useful for another applied science such as economics. Furthermore, the source is the World Bank, and in order to be in such a repository, the database underwent extensive revisions, [click here to explore the data](#).¹ After all, it concerns the indexes of the member countries of the Organization for Economic Co-operation and Development (OECD). The economic literature has already detailed the relationship between the percentage of high technology in exports and predictive variables such as investment in education and in science and development.

Another challenging database structure is the Millennium Cohort Study (MCS), analyzed by [118, 41, 4]. The MCS constitutes a landmark in longitudinal research in the United Kingdom. Almost 19,000 children were sampled to participate in the study that began in 2001 and continues to the current days, [click here to explore the data](#).² More than 3,000 questions are asked each biennium to parents, teachers, family members, and the child itself. Such an amount of information can be used to study public health issues, physical and psychological development, socioeconomic relationships, among many others. [118] studied only two response variables and elected 11 predictor covariates, this represents the tip of the iceberg of what there is to analyze in the MCS. Large databases lead us to the third structuring question of the thesis.

What is the effect of air pollutants on the health of the most vulnerable children and adolescents, for example, the 5% most vulnerable children, if the number of pollutants and predictor variables is very large and I do not have a priori information to select them? This question structures the third part of the thesis, in which the main topic is quantile regression applied to high-dimensional longitudinal data.

To answer this question I took a more direct route, as there is a recently published book that provided me with an updated map of the options available so far for this type of problem [13]. There are techniques such as LASSO [114], adaptive LASSO [140], Elastic-net [141] for regression with quadratic errors, M-estimation and quantile regression.

The solution I have proposed is a small change to the original model by [65]. In which exploratory variables are penalized as well as the intercepts by LASSO. This seems natural, but it has some consequences, such as how to estimate the constraint parameter. We no longer have the guarantee shown by [73] that the standard deviation of the residuals divided by the standard deviation of the intercepts converges to the true value of λ . Because now the tuning parameter extends to the predictors as well. It seems that we would have to resort to the strategy that most analysts use: cross-validation (CV). However, I preferred to use the Bayesian Information Criterion (BIC) for two reasons. First, quantile regression has a more explanatory character than a predictive finality. For sure, the prediction for the 10% percentile will certainly have a large quadratic error or even absolute errors. Second, longitudinal data have an intrinsic dependence relationship

¹Source visited at May 26 2022

²Source visited at June 13 2022

among observations. Of course, it is possible to get around this limitation and resampling longitudinal data, but if it is possible to avoid yet another source of errors, I prefer to follow this way.

Table 1.2: Acronyms of the regression models according to the type of penalty.

Method	Intercepts	LASSO on α	LASSO on β	Adaptive
QR	No	No	No	No
QRFE	Yes	No	No	No
PQRFE/LQRFE	Yes	Yes	No	No
DLQRFE	Yes	Yes	Yes	No
ALQRFE	Yes	Yes	Yes	Yes

This time, the number of methods is not so extensive. Even so, a table can help better understand each author's contribution and nomenclature. In Table 1.2 we have the acronyms of simple quantile regression (QR) proposed by [66], quantile regression with fixed effects (QRFE) proposed by [65], quantile regression with fixed effects and penalty on intercepts by LASSO (PQRFE) proposed by [65], we also call this method as quantile regression with fixed effects and penalty on intercepts and exploratory variables by LASSO (LQRFE), quantile regression with fixed effects and penalty on intercepts and exploratory variables by LASSO in two steps (DPQRFE) proposed by [77], which is a sub-case of Group LASSO see [138], and quantile regression with fixed effects and penalty on intercepts and exploratory variables by adaptive LASSO (ALQRFE) proposed by us. Simple quantile regression with LASSO as proposed by [142, 134] is a particular case.

- QR was proposed by [66],
- QRFE was proposed by [65],
- PQRFE was proposed by [65],
- LQRFE is proposed by us in this thesis,
- DPQRFE was proposed by [77],
- ALQRFE is proposed by us in this thesis.

Just as [118, 4] restricted their MCS studies to England (2004, 2006 e 2008), I restricted the analysis to Northern Ireland on the same years. The reason is that Ireland is the country with the smallest sample size, just over 650 children each biennium considering that the missing data has already been removed. The simplest fixed model of all already requires 702 intercepts to be estimated, if we restrict the database to a little more than 1200 covariates, we have already achieved an example of high dimension, as there are more parameters than observations. However, we don't have to go so far

as to observe confounding problems and need select the covariates. In our example, I considered approximately 100 covariates. In this case, I observed a low correlation between the exploratory variables. However, if the correlation is moderate or high, as commonly observed, LASSO restriction would be unwise and it would be more appropriate to proceed with an adaptive LASSO. I propose this general variant of LASSO to help future applied researchers. There are several similar longitudinal databases with high value for epidemiological and social studies, such as the 1982 Birth Cohort, developed by researchers from the Universidade Federal de Pelotas (UFPel) Brazil with an initial sample of almost 6,000 children, [click here to see the project](#) and Étude Longitudinale Française depuis l'Enfance (ELFE) France with an initial sample of over 18,000 children, [click here to see the project](#). Just to cite two examples, one Brazilian and the other French.

The two response variables chosen by [118] are external and internal scoring the strengths and difficulties questionnaire (SDQ). Those inform the degree of social maladjustment and internal suffering of children and adolescents. By adding a large number of exploratory variables, I found new predictors in addition to those that the specialized literature usually establishes. An interesting phenomenon occurred, when intimate issues of each family were included, general variables such as neighborhood conditions, economic situation, or ethnic status were no longer relevant.

The asymptotic normality of the estimators for the new model is obtained following the path of [63, 64, 65]. Furthermore, I proceeded with Monte Carlo simulation studies. These simulations show that adaptive LASSO dominates the other methods, presenting much better results than the others.

Although I have always opened the pieces of this introduction using the example of air pollutants and children's respiratory capacity, the real questions I answer in each chapter of this thesis change according to the real data analyzed. Thus, the real question for the second part would be: what is the effect of investment in education and science on the technological qualification of exports from the most advanced, medium and least advanced OECD countries? Likewise, the third part answers the question: what are the most relevant variables to impact the internal and external SDQ scores of the most vulnerable children from Northern Ireland?

Computational methods are briefly described in each chapter. All are developed in R and RCPP³ [95, 27]. The optimization is performed by variations of the Newton-Raphson method, where the core of the function is written in RCPP and the auxiliary functions in R to speed up the process. These packages are available on my personal page at the GitHub [22]. In addition, two of these packages are already in the official R repository [21, 23]. As soon as the documentation for the last package is ready, it will be submitted to the R CRAN. Also, vignettes and small tutorials are in my future plans.

³The RCPP package offers a seamless integration of R and C++

This thesis is the result of a joint agreement between the Universidade Federal de Minas Gerais (UFMG) and the Université Paris-Saclay (UPS) and therefore follows some peculiarities such as having an introduction in three languages. In addition, it does not have great eccentricities, it follows a format of a collection of articles in the chronological order in which they were written, but it also has a theoretical-methodological logic. I could not implement computational methods for a quantile regression for longitudinal data if I had not first written codes for a robust LMM and tested this realization by Monte Carlo simulations. This necessary step back resulted in the first manuscript which is also the material defended in the mid-term evaluation.

The remainder of this thesis follows the following structure. Chapter 2 is mainly devoted to discussing robust LMM models. The Vitória database on the effect of air pollutants on children with and without asthma is the motivating problem. The main results are gathered in the manuscript that was submitted on June 13, 2022 to a scientific journal of statistics. After the mid-term evaluation, a simulation study was added with weights to minimize the impact of the most influential observations, since in this context they did not contribute to an intense improvement of the technique they were not added to the manuscript, only cited, but are present in the thesis. In Chapter 3, I discuss quantile regression methods for panel data. I emphasize the treatment by fixed effects and the new methods we present to the scientific community that use alternative loss functions to the check function. This manuscript has not yet been submitted to any scientific journal. Chapter 4 is motivated by extending quantile regression to longitudinal data and simultaneously proceeding with the selection of variables. This manuscript has not yet been submitted to any scientific journal. Chapter 5 is reserved for the general conclusion of the thesis. There is a smorgasbord of information in the Appendix, in Section A.1 are the correction factors for the robust LMM presented in Chapter 3. In Section A.2 are simulations using Cantoni's weights. In Section A.3 is a list of covariates and how to extract them from the MCS digital repository. In Section A.4 are the help of the packages I wrote and a brief comment to install them. In Section A.5 is the proof of Knight's identity, which is presented in [63] without proof. Section A.6 presents the computational performance of the methods in terms of estimation time.

1.2 Introdução em Português

Qual o efeito dos poluentes do ar na saúde de crianças e adolescentes? Essa pergunta não tinha sido completamente respondida na tese de [102], na qual não levou em conta métodos robustos, uma vez que, uma observação aberrante ou atípica, pode modificar dramaticamente as estimativas de qualquer modelo linear baseado em erros quadráticos. Portanto, uma nova pergunta mais precisa é necessária para levar em consideração os dados reais que estão sujeitos a oscilações extremas. Nesse sentido, o efeito médio dos

poluentes do ar na saúde de crianças e adolescentes é afetado por observações atípicas?

Antes de entrar na questão metodológica sobre alternativas robustas e o quanto elas protegem contra a interferência causada por *additive outliers*, os dados levantados por [102] são descritos rapidamente. Ela pesquisou em um grupo de 82 crianças e adolescentes na cidade de Vitória Brasil, sendo que parte dessas crianças sofria de asma e parte não, mas todas eram oriundas de famílias de baixo poder aquisitivo e estudavam todas na mesma escola. O estudo foi aprovado pelo Comitê de Ética em Pesquisa (CEP) da Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória-ES, CAAE 53176915.8.0000.5065, e pela Secretaria Municipal de Saúde de Vitória por meio da Escola Técnica do Sistema Único de Saúde (ETSUS). Para determinar se a criança era asmática ou não, foi realizado um exame clínico. Posteriormente as crianças foram instruídas a realizar um autoteste para medir seus níveis de capacidade respiratória. Três medidas foram realizadas por vez e somente a mediana foi considerada para baixar o nível de erro de medição por tratar-se de um autoteste. Tais medidas foram repetidas ao longo de seis meses, configurando um estudo longitudinal. Além disso foram consideradas covariáveis como: condições socioeconômicas das crianças fixas ao longo do tempo, gênero, IMC, se fumante passivo, tipo de parto e forma de amamentação. Também foram incluídas variáveis observacionais climatológicas como: temperatura, umidade e poluentes do ar como CO_2 , SO_2 , PM_{10} , $\text{PM}_{2.5}$. Os poluentes foram obtidos pelas medições do IEMA-ES, que possui 8 estações meteorológicas na região metropolitana de Vitória, uma delas nas proximidades da escola em que as crianças voluntárias estudavam.

Cada criança realizou o autoteste em um dia diferente do mês, sendo assim, foi conectado, a cada medida da criança, as informações meteorológicas do dia anterior, visto que o efeito da poluição tende a demorar 24 horas para ter sua expressão máxima. Essas amostras, mesmo não sendo grandes, apresentam uma estrutura complexa, dependendo do indivíduo, os causadores das diferenças serial temporal ou espacial, sendo que essa última não foi explorada e constitui uma questão em aberto a futuros trabalhos. Além da multicolinearidade presente entre os poluentes do ar, sendo esse um estudo observacional e não um experimento controlado, não há como separar perfeitamente o efeito de cada poluente visto que, muitos deles advêm de fontes similares como o trânsito e a indústria.

Antes de prosseguir, gostaria de fazer um apanhado de nomenclaturas que serão usadas nesta tese, a explicação mais natural e sintética que conheço está disponível em [120, página 1]. Em tal obra, ele expõe as diferenças entre termos semelhantes com um exemplo de exame de pressão arterial, ou seja, se um indivíduo tem a sua pressão arterial sistólica e diastólica medidas por uma enfermeira temos, um exemplo muito simples de dado multivariado. Se duas famílias vão a um posto de saúde e cada um dos seus membros tem a sua pressão diastólica aferida, então temos informação em clusters. Se a enfermeira desconfia que o tensiômetro está com defeito, ela pode medir o paciente mais de uma vez para checar. Em tal caso, temos medidas repetidas. Se essas medidas repetidas são

retiradas ao longo do tempo, digamos que alguém vai ao posto toda segunda-feira, então, temos dados longitudinais ou dados de painel, que aqui tratamos como sinônimo e que é mais usado na econometria. Entender a forma como a hierarquia dos dados se estrutura ajuda o analista a propor o modelo com melhores qualidades para o problema em questão. Sendo uma questão chave, a estrutura de variância-covariância que deve ser adotada aos resíduos do modelo. Modelos lineares mistos compõem uma classificação geral de modelos que incluem um componente fixo e um grupo de parâmetros aleatórios como interceptos aleatórios ou coeficientes aleatórios. Medidas repetidas, dados longitudinais ou dados de painel podem ser modelados por LMM. Análise de cluster, análise fatorial e análise de componentes principais, embora tratem de dados em última instância multivariados, não se apresentam como a melhor alternativa a esse tipo de dado, na verdade tendem a ter um resultado pífio nesses casos [120], mas há exceções como será mostrado adiante.

Como foi exposto, o banco de dados de [102] foi analisado sem se levar em conta métodos robustos, mas o que vem a ser robustez? Robustez é um termo bastante amplo em estatística. Ela costuma querer dizer que um método é resistente a uma quebra de pressuposto, ou seja, mesmo que exista algumas suposições para a aplicação do método que não são atendidas, ainda assim o método pode ser usado e os resultados serão esperados por serem semelhantes aos que as suposições eram plenamente atendidas. Por exemplo, o método de análise de variância (ANOVA) é dito robusto em relação à normalidade pois é um pressuposto, mas não precisa ser plenamente atendido. Nesta tese, o termo robustez é usado majoritariamente como robustez para additive outliers. As observações meteorológicas observadas têm essa característica, como observadas a cada poucos minutos por dia, há momentos nos quais a média sobe de forma extremamente elevada e cai rapidamente ao nível em que estava.

Existe livros sobre LMM como [120, 90, 24], também há livros sobre métodos robustos como [51, 82]. Não conheço um livro que junte esses dois assuntos, mas o artigo de [42] e o capítulo de [131] são boas introduções. Robustez a outliers pode ser obtida ao se assumir uma distribuição de cauda pesada como a distribuição-t de Student, ou ainda é possível deixar os dados falarem por si só como em funções de perda com adaptabilidade como a função bisquared de Tukey, na qual ao final das contas não deixa de ser equivalente a assumir que tais erros são uma mistura de uma distribuição Gaussiana e outliers de outra distribuição de cauda pesada, porém, nesse caso, não se determina o percentual de contaminação, ou seja, essa parte é deixada livre para que os próprios dados falem por si.

[42, 97, 135] entre muitos outros trabalhos sobre modelos mistos com estimação robusta apontam para [53] como o pai da proposta. Embora no momento em que Huggins escreveu seu artigo a teoria de Huber já estivesse muito bem exposta em detalhes por [49, 48, 50, 51], ainda assim na hora de provar sua proposição ele é extremamente sintético e faz referência a um teorema no trabalho de [19], um trabalho que é bem geral trata de diversos casos de estimação. Parece ser uma escolha infeliz pois em [49] já está dito que no

caso de verossimilhança a convergência é esperada e em [50] os casos de verossimilhança são exaustivamente discutidos. Portanto, [53] propõem uma robustificação da verossimilhança de um modelo de medida repetida com estrutura de variância conhecida e não tira proveito disso. Ainda faz exigências sobre a existência da segunda derivada da função perda para garantir a expansão de Taylor como fez [51, página 167] para provar a convergência da regressão simples. Essa exigência de segunda derivada e o uso da expansão de Taylor também é explorada nos esboços de prova propostos por [58, 105, 106]. No entanto, a prova de consistência do estimador robusto para modelo misto mais completa está em [104], em que o autor explora a função perda de Huber e a ponderação proposta por [15]. É estranho que um livro dedicado a parte assintótica de modelos mistos como [56] com um capítulo sobre estimação robusta não cita nem [53] nem [104].

Além da questão da robustez a outliers, precisa-se levar em conta ainda os pontos de influência e de alavancagem. [15] propôs que se investigasse mais profundamente a matriz de covariáveis ou matriz de desenho, pois, a partir dela, algumas observações poderiam estar causando excesso de influência no modelo. Da forma como a verossimilhança robusta é construída usando a função de Huber os outliers aparecem nos resíduos, aparentemente pouco importando se eles são provenientes de uma medida discrepante na variável resposta ou na matriz de desenho. Porém, em estudos observacionais em que se tem pouco controle sobre a matriz de covariáveis, essa sutileza pode fazer diferença. Pois pela matriz de projeção já se pode estimar quais observações são demasiado influentes no modelo e penalizá-las para que as observações tenham um peso mais equilibrado no modelo final. Essa proposta de [15] não compete com as funções de [49], mas sim harmonizam facilmente em uma mesma verossimilhança como explorou [71] e já tinha demonstrado e provado a consistência assintótica [104].

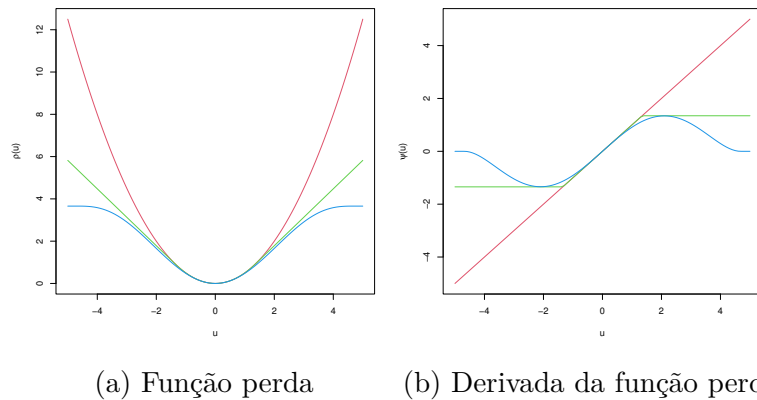


Figure 1.3: Funções perda quadrática (vermelho), Huber (verde) e Tukey (azul).

Na Figura 1.3 estão as funções de perda quadrática, Huber e Tukey, a esquerda e a direita, suas derivadas. Claramente, as funções quadrática e Huber são convexas e a função Tukey não o é. Fora do intervalo entre $-c/\sqrt{5}$ e $c/\sqrt{5}$. A desconfiança de

que funções não convexas não geram estimadores consistentes é antiga [49] sugere que restrições adicionais devem ser impostas. [70, página 45] sugere, para o caso da função Tukey, que bons pontos iniciais são o suficiente para a convergência. Já [78] é mais radical e exige que a maioria dos resíduos cai na região convexa da função não convexa e que essa seja simétrica, o que é o caso da função Tukey. Portanto o valor de c não pode ser muito pequeno. [51, página 103] alerta que usar funções rescendentes como a Andrews, Hampel e Tukey pode ser bastante perigoso e que, inclusive, remover manualmente as observações impossíveis pode ser uma alternativa mais cautelosa. Esse estudo das propriedades das funções foi vital para não propor uma combinação de função *check* para estimar quantis com funções não convexas, aparentemente esse é um caminho que não traz bons frutos.

Um caminho que não seguimos na tese consiste em otimizar o parâmetro tuning automaticamente [127, 8], mas não foi possível achar uma quantidade que fosse não monótona ou que não causasse mais instabilidade ao modelo. Por isso, prefere-se manter valores tradicionais que correspondem a uma eficiência relativa de 95% comparando com a estimativa média por mínimos quadrados [82, página 357]. Estabelecer uma região compacta com um máximo é um bom ponto de partida para quem seguir essa problemática [31].

No entanto, como vimos nas nossas próprias simulações, correlações fortes entre as covariáveis são a principal fragilidade para todos esses métodos. E os dados ambientais de Vitória são bastante complexos e correlacionados, portanto preferiu-se seguir a proposta de autores anteriores de reduzir a dimensionalidade dos dados por uma técnica robusta de análise de componente principal [17]. Há várias técnicas possíveis para isso, que exploram formas de robustas diferentes de estimar a variância dos dados [18, 80, 52].

Com a combinação da metodologia exposta, conseguimos analisar o banco de dados [102] e considerar o efeito dos poluentes, apesar das observações discrepantes. Também foram levados em consideração fatores próprios de cada criança como: o seu nível alérgico, se a criança era ou não fumante passiva. Não foi incluída uma variável dicotômica de que se apresenta ou não asma, porque os asmáticos do grupo visto que os asmáticos do grupo apresentam um quadro controlado e fazem uso de medicamento quando necessário. Portanto, asmáticos e não asmáticos não apresentaram diferença significativa quanto a capacidade respiratória medida ao longo dos meses. De tal forma pode-se contribuir para o debate de saúde pública com tal estudo que combina uma amostragem de fatores de crianças que são controlados por um desenho de experimento, mas são adicionadas informações climatológicas com característica de estudos observacionais. Ressalta-se ainda que com essas metodologias robustas não foi necessário deletar nenhuma unidade amostral, o que seria um problema em um tamanho de amostra relativamente pequeno como este.

Qual o efeito dos poluentes do ar na saúde de crianças e adolescentes mais vulneráveis? Por exemplo, nas 10% crianças mais vulneráveis? Ou no 1% mais vulnerável? Tais perguntas estruturam a segunda parte da tese, na qual a principal questão é a re-

gressão quantílica aplicada a dados longitudinais.

Para responder essa pergunta percorreu-sei um caminho tortuoso, que começou com [107], que propunha unificar regressão quantílica e cópulas, dando uma interpretação alternativa aos percentis, na qual, o percentil deixa de ser um valor estabelecido pelo analista para ajustar um hiperplano e passa a ser um hiperparâmetro do modelo. Tal método é muito sensível, pois depende de uma estimação e transformação de resíduos. Lendo artigos que uniam regressão quantílica e medidas repetidas ou dados longitudinais foi possível perceber que havia um consenso: todos os artigos citavam [65]. Assim sendi, como [66] é um artigo seminal para regressão quantílica, [65] é o artigo fundamental para regressão quantílica para dados longitudinais. No entanto, a leitura desse artigo não é fácil porque é difícil compreender visto que o autor trata os interceptos individuais como efeitos fixos.

Há trabalhos em que usam a distribuição assimétrica de Laplace como [61, 39, 40, 33]. A expressão fechada parecia sólida, possuía uma verossimilhança, e o percentil é escolhido pelo analista resultando em expressões cuja assimetria aumenta conforme o percentil se aproxima dos extremos zero ou um. O método de estimação através de EM já estava disponível em R e poderia ser modificado, inclusive prioris bayesianas poderiam ser facilmente incluídas no método. Posteriormente, o trabalho de [118] foi lido, percebeu-se que ele propunha um M-quantile para modelos mistos ou melhor, um M-quantile para efeitos aleatórios e prometia estender o método para modelos mistos nos trabalhos futuros. [118] citava trabalhos como [53, 97, 106]. Ele Ppropunha uma alteração na verossimilhança robusta de Huggins seguindo a correção de tipo II proposta por [131] para estimação robusta da média. Esse plugin de quantil no lugar de uma média parece estranho e improvável. Outra citação de valor de tal trabalho é o estudo de [10]. Nesse artigo inaugural do método M-quantile para regressão simplesele não só é mais citado porque um dos principais apelos da M-regressão para trabalhos aplicados é a sua robustez e a regressão quantílica já é robusta. Outras vantagens enumeradas em [68? , 62] não são tão chamativas. No entanto, a mesma questão se mantinha: é possível usar uma função M-quantil como plugin numa verossimilhança construída para o efeito médio sem degenerar essa verossimilhança?

Retornando a [65], o autor explica que minimizar o BLUP de um modelo com interceptos aleatórios equivale a minimizar um modelo de regressão simples com muitos interceptos em que os eles sofrem uma restrição: a soma de seus módulos não podem ultrapassar uma certa quantidade. Essa construção do método LASSO permite resolver duas questões: estimar os coeficientes das variáveis preditoras levando em conta o efeito individual das unidades amostrais em um modelo longitudinal e não permitir que o número de parâmetros cresça em demasia. Isso responde porque o LASSO e o BLUP são técnicas eficientes. O autor cita [98], que credita a ideia a um artigo que aparentemente só resta o resumo [46]. A proposta de [65] é realizar uma regressão quantílica para um modelo

com vários interceptos que são penalizados por LASSO. Logo, fica claro porquê [65] não usa o termo LMM. No entanto, tantos efeitos fixos ainda eram uma questão um pouco confusa. Faltava um pesquisador como [137] separar os métodos com maestria. Dados longitudinais podem ser divididos em dois tipos de análise: efeitos aleatórios e efeitos fixos: os primeiros geralmente são assumidos por terem esperança zero e serem independentes das demais variáveis preditoras. Em razão dessas propriedades eles podem ser omitidos do processo de estimação em termos de locação e apenas incluídos na estrutura de variância. Sobre os segundos, nenhuma dessas duas suposições são necessárias, sendo tratados como variáveis dummies. Porém o número de parâmetros cresce junto com o tamanho amostral e a estimação se torna pouco eficiente [2, 47, 38, 137]. O insight de [46] é ser flexível nas suposições e assumir efeitos fixos, mas ser prático na estimação e penalizá-los por LASSO o que equivale a efeitos aleatórios. Nas palavras de Koenker, essa demonstração de equivalência e equilíbrio abre as portas para outras penalizações e propostas metodológicas. Nascia aí a regressão quantílica com efeitos fixos e penalização por LASSO [65]. A extensão natural seria propor uma M-quantile para efeitos fixos ao invés de efeitos aleatórios como [118].

Até aqui foi discutido sobre regressão quantílica sem entrar em detalhes, mas cada método tem a sua história. A regressão quantílica fica famosa a partir do artigo de [66] que generaliza o conceito não paramétrico de quantil univariado para uma relação condicional entre ao menos duas variáveis aleatórias ou uma matriz de desenho e uma variável resposta. [66] ressalta que diferentemente do caso univariado em que o quantil tem uma fórmula fechada, o caso bivariado estava aberto à discussão, poderia ser estimado pela função *check*, mas também poderia ser estimado por *ranks*, entre outras possibilidades que ainda podem ser propostas. [86] propõe o retorno à média, mais especificamente os expectiles que são uma forma de quantil médio. Usando ao invés da função *check* uma função quadrática assimétrica. Seu método tem boas propriedades como menor risco de *quantile crossing* e melhor estimação dos quantis extremos. No entanto, esse método é sensível à distribuição de probabilidade dos erros, não sendo robusto a outliers evidentemente. *Quantile crossing* é como chamamos quando a estimativa do coeficiente quantile de uma regressão não é monotonamente crescente para um $\mathbf{X}$ fixo. Por exemplo, suponha-se que estimamos os coeficientes de um regressor para dois percentis: 0.3 e 0.6. O valor predito pelo coeficiente 0.3 vezes um valor fixo de $\mathbf{X}$ tem de ser menor ou igual ao valor predito por um percentil 0.6. Se isso não ocorrer, a suposição de a função acumulada de probabilidade ser monotonamente crescente não é satisfeita. Muitas vezes a estimativa empírica da regressão quantílica pode resultar em *quantile crossing*. A regressão expectile também não está livre desse problema, mas ele é menos frequente. Logo depois temos o trabalho de [10], que propõe um equilíbrio entre a quantile e a expectile: a M-quantile. Assim como o trabalho de [49] pondera entre a média e a mediana com diferentes funções de perda, o trabalho de [10] pondera entre o quantil e o expectile. Os autores também

mencionam que a estimativa não é única, podendo ser obtida pela função de perda ou pela sua derivada, o que pode resultar em pequenas diferenças a depender do tipo de problema. Finalmente, temos um artigo que prova que as três propostas, embora com nomes diferentes, são igualmente quantis [59].

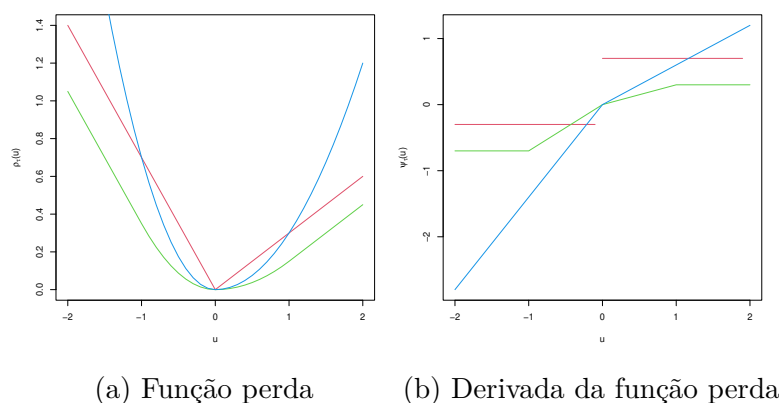


Figure 1.4: Funções *check* (vermelho), M-quantile (verde) e quadrática assimétrica (azul) para $\tau = 0.3$.

Na Figura 1.4 estão as funções *check*, M-quantile e quadrática assimétrica. A esquerda estão as funções perda e a direita suas derivadas. Claramente todas as funções são convexas o que facilita muito a estimação e as provas de consistência. A derivada da função *check* tem um ponto de descontinuidade bem na raiz o que pode causar instabilidade numérica a depender do tipo de problema a ser minimizado. Neste exemplo o percentil escolhido é 0.3, mas valores mais extremos resultam em assimetria mais forte. Valores maiores do que 0.5 invertem a relação de assimetria, e.g., 0.7 espelharia horizontalmente este gráfico.

Table 1.3: Acrônimos dos modelos de regressão conforme suas características e forma de estimação

Intercepto(s)	Quantile	M-quantile	Expectile
Único	QR	MQR	ER
Aleatórios	QRRE/LQMM	MQRRE	ERRE
Fixos	QRFE	MQRFE	ERFE
Penalizados	PQRFE	MQRFE	PERFE

Até o momento discorre-se sobre três tipos de otimização: quantile, M-quantile e expectile. Além disso, temos quatro modelos possíveis: regressão simples, efeitos aleatórios, efeitos fixos e efeitos fixos com penalização tipo LASSO. O cruzamento dessas formas de utilização pelos tipos de modelo resulta em 12 métodos possíveis, além das particularidades e restrições que cada autor agrega ao seu método como *splines*, EM, etc. A comparação de tantos métodos resulta em diversos acrônimos e a leitura se torna cansativa. Portanto na Tabela 1.3 foi elaborado um resumo. Sobre as características dos

interceptos: se há um único intercepto temos uma regressão simples, se eles são aleatórios temos efeitos aleatórios ou modelos mistos, se eles são fixos temos uma regressão com muitas variáveis *dummies*, por fim os interceptos penalizados são fixos, porém com a restrição é do tipo LASSO. Segue uma breve menção dos autores que propuseram os respectivos métodos:

- QR foi proposto por [66],
- LQMM foi proposto por [39],
- QRFE foi proposto por [65],
- PQRFE foi proposto por [65],
- MQR foi proposto por [10],
- MQRE foi proposto por [118],
- MQFE é proposto por nós nesta tese,
- PMQRFE é proposto por nós nesta tese,
- ER foi proposto por [86],
- ERRE foi proposto por [123],
- ERFE é proposto por nós nesta tese,
- PERFE é proposto por nós nesta tese.

Um caminho que não foi seguido na tese consiste em otimizar o parâmetro tuning da M-quantile automaticamente como alguns autores fizeram para a M-estimation da média [127]. Mesmo uma escolha analítica da constante tuning é um problema em aberto. Visto que a maioria dos trabalhos compara a M-quantile com a *check* e não com a quadrática assimétrica, então a relação clássica de eficiência deve ser atualizada [82, página 357]. Dessa forma, reciclar valores como $c = 1.345$ parece um “cópia e cola” inapropriado, porém repetido na literatura [100, 9, 118]. Um dos poucos trabalhos em que o impacto de diversos valores da constante tuning é analisado é o de [62]. Por esse motivo foi elegido $c = 1$ conforme [69, 62].

Os resultados assintóticos de consistência e normalidade dos estimadores são provados por [65] para os modelos QRFE e PQRFE. Estimar o parâmetro de restrição é deixado como um problema em aberto pelo autor, mas [73] prova a consistência do parâmetro tuning. Koenker usa as convergências explicitadas por [64] e uma identidade de [63]. Infelizmente essa identidade só vale para a função *check* e não pode ser aplicada na M-quantile

ou expectile. Mesmo assim, os resultados assintóticos dos novos modelos foram parcialmente resolvidos. Já provamos a consistência e normalidade dos estimadores do tipo expectile, ou seja, para o ERFE e PERFE, mas ele ainda não foi feito para o M-quantile: MQRFE e PMQRFE. A forma quadrática do expectile ajuda nas provas. Imagina-se que a consistência da M-quantile pode ser provada, pois casos especiais como a função Huber já foram exaustivamente investigados. Inclusive para casos com penalização LASSO [108], e modelos com efeitos aleatórios [104, 106].

Na falta de uma prova matemática definitiva, temos bons resultados por simulação de Monte Carlo e aplicação em dados reais que parecem comprovar as suposições. Para a ilustração do método com dados reais foram procurados dados de painel nos quais pudesse ter muita confiança de não haver erros e cuja relação entre as covariáveis já fosse um consenso na literatura específica. Embora tenha-se começado essa parte do texto fazendo uma pergunta sobre o efeito dos poluentes do ar nas crianças mais vulneráveis fazendo alusão a questões relacionadas à saúde pública, o banco de dados escolhido para essa segunda parte é de natureza econômica. Dessa forma o método se mostra útil para outra ciência aplicada como a Economia. Além disso, a fonte é o Banco Mundial, e para estar em tal repositório o banco de dados passou por exaustivas revisões, afinal diz respeito aos índices dos países membros da Organização para a Cooperação e Desenvolvimento Econômico (OCDE). A literatura econômica já explanou a relação entre o percentual da alta tecnologia nas exportações e variáveis preditoras como o investimento em educação e em ciência e desenvolvimento.

Outro banco de dados, cuja estrutura é desafiante, é o Millennium Cohort Study (MCS), analisado por [118, 41, 4]. O MCS constitui um marco nas pesquisas longitudinais no Reino Unido, no qual quase 19 mil crianças foram feitas parte de amostras para participar do estudo que começou em 2001 e segue ativo até os dias de hoje. Mais de 3 mil perguntas são feitas a cada biênio para os pais, professores, familiares e para a própria criança. Tamanha quantidade de informação pode ser usada para estudar questões de saúde pública, desenvolvimento físico e psicológico, relações socioeconômicas, entre muitas outras áreas. [118] estudou somente duas variáveis respostas e 11 covariáveis preditoras, isso representa apenas a ponta do iceberg do que há para analisar no MCS. Grandes bancos de dados nos levam à terceira pergunta estruturante da tese.

Qual o efeito dos poluentes do ar na saúde de crianças e adolescentes mais vulneráveis nos quais 5% são vulneráveis aos está vulnerável aos poluentes e variáveis preditoras ? 5% é uma quantidade muito grande e não tem-se informação a priori para selecioná-las. Tal pergunta estrutura a terceira parte da tese, na qual a principal questão é a regressão quantílica aplicada a dados longitudinais de alta dimensão.

Para responder essa pergunta, um caminho mais direto foi percorrido, pois existe um livro recém publicado que forneceu um mapa atualizado das opções até aqui disponíveis para esse tipo de problema [13]. No livro estão presentes técnicas como LASSO

[114], LASSO adaptativo [140], Elastic-net [141] para regressão com erros quadráticos, M-estimation [49], entre outros.

A solução que foi proposta é uma pequena alteração no modelo original de [65]. Em que as variáveis exploratórias são penalizadas assim como os interceptos por LASSO. Isso parece natural, porém traz algumas consequências tais quais: como estimar o parâmetro de restrição. Não temos mais a garantia mostrada por [73] de que o desvio padrão dos resíduos dividido pelo desvio padrão dos interceptos converge para o verdadeiro valor de λ . Pois agora o parâmetro tuning se estende às preditoras também. Parece que teríamos que recorrer à estratégia que a maioria dos analistas utiliza: cross-validation (CV). Porém, prefere-se usar o critério Bayesian Information Criteria (BIC) por dois motivos: porque a regressão quantílica tem um viés mais explicativo do que preditivo. Com certeza a predição para o percentil 10% terá grandes erros quadráticos ou mesmo absolutos; e dados longitudinais possuem uma relação de dependência intrínseca. Fica claro que é possível contornar essa limitação e reamostrar dados longitudinais, mas se é possível evitar mais uma fonte de erros prefiro seguir esse caminho.

Table 1.4: Acrônimos dos modelos de regressão conforme o tipo de penalização.

Método	Interceptos	Lasso em α	Lasso em β	Adaptativo
QR	Não	Não	Não	Não
QRFE	Sim	Não	Não	Não
PQRFE	Sim	Sim	Não	Não
LQRFE	Sim	Sim	Sim	Não
DLQRFE	Sim	Sim	Sim	Não
ALQRFE	Sim	Sim	Sim	Sim

Agora, o número de métodos não é tão extenso, mesmo assim julga-se que um quadro pode ajudar a melhor compreender a contribuição de cada autor e a nomenclatura. Na Tabela 1.4 temos os acrônimos da regressão quantílica simples (QR) proposta por [66], regressão quantílica com efeitos fixos (QRFE) proposta por [65], regressão quantílica com efeitos fixos e penalização nos interceptos por LASSO (PQRFE) proposta por [65], regressão quantílica com efeitos fixos e penalização nos interceptos e nas variáveis exploratórias por LASSO (LQRFE) proposta por nós, regressão quantílica com efeitos fixos e penalização nos interceptos e nas variáveis exploratórias por LASSO em duas etapas (DLQRFE) proposta por [77], que não deixa de ser um subcaso de Group LASSO ver [138], e regressão quantílica com efeitos fixos e penalização nos interceptos e nas variáveis exploratórias por LASSO adaptativo (ALQRFE) proposta por nós. A regressão quantílica simples com LASSO como a proposta por [134] é um caso particular.

- QR foi proposto por [66],
- QRFE foi proposto por [65],

- PQRFE foi proposto por [65],
- LQRFE é proposto por nós nesta tese,
- DLQRFE foi proposto por [77],
- ALQRFE é proposto por nós nesta tese.

Assim como [118, 4] restringiram seus estudos do MCS à Inglaterra, a análise feita restringiu-se à Irlanda do Norte, pois a Irlanda é o país com menor tamanho amostral, pouco mais de 600 crianças a cada biênio considerando que os dados faltantes já foram removidos. O modelo fixo mais simples de todos já requer 600 interceptos para serem estimados, se restringirmos o banco de dados a um pouco mais de 1200 covariáveis já teremos alta dimensão, pois são mais parâmetros do que observações. No entanto, não precisamos ir tão longe para termos problemas de confundimento e ser necessário selecionar as covariáveis. Em nosso exemplo, foram consideradas aproximadamente 100 covariáveis do MCS. Nesse caso, observou-se que a correlação entre as variáveis exploratórias era baixa. No entanto, caso ela fosse moderada ou alta, como é mais comum, a restrição por LASSO seria desaconselhável, e seria mais apropriado proceder com um LASSO adaptativo. Essa variante foi escrita de forma mais geral do LASSO para ajudar futuros pesquisadores aplicados, pois há diversos bancos de dados longitudinais semelhantes com alto valor para estudos epidemiológicos e sociais somente para citar dois exemplos, um brasileiro e outro francês: o 1982 Birth Cohort, desenvolvido por pesquisadores da Universidade Federal de Pelotas (UFPel) Brasil com uma amostra inicial de quase 6 mil crianças, e Étude Longitudinale Française depuis l'Enfance (ELFE) França com uma amostra inicial de mais de 18 mil crianças.

As duas variáveis respostas escolhidas por [118] são: strengths and difficulties questionnaire (SDQ), externa e interna. Que informam o grau de desajuste social e sofrimento interno de crianças e adolescentes. Ao acrescentar muitos variáveis exploratórias encontramos novas preditoras além daquelas que literatura especializada normalmente estabelece. Um fenômeno interessante ocorreu: ao serem incluídas questões íntimas de cada família as variáveis gerais como as condições do bairro, conjuntura econômicas ou status étnico deixaram de ser importantes.

A normalidade assintótica dos estimadores para o novo modelo foi fácil de ser obtida, pois a parte mais trabalhosa já havia sido feita por [63, 64, 65]. Além disso, procederam-se estudos de simulação de Monte Carlo, que mostram variáveis independentes o LASSO domina o LASSO adaptativo e o modelo fixo sem restrições.

Embora tenha sempre aberto novas questões usando os poluentes do ar e capacidade respiratória de crianças, na verdade as verdadeiras perguntas que responde-se nos capítulos dessa tese mudam de acordo com os dados reais analisados. Desse modo, a

verdadeira pergunta para a segunda parte seria: qual o efeito do investimento em educação e ciência na qualificação tecnológica das exportações dos países mais avançados, dos medianos e menos avançados da OCDE? Igualmente, a terceira parte responde à pergunta: quais as variáveis mais relevantes no escore SDQ interno e externo das crianças mais vulneráveis a traumas familiares ou sociais?

Os métodos computacionais foram brevemente descritos em cada capítulo. Todos foram desenvolvidos em R e RCPP⁴ [95, 27]. A otimização foi realizada por variações do método Newton-Raphson em que o cerne da função foi escrito em RCPP e as funções auxiliares em R para acelerar o processo. Aos poucos sobem os códigos a minha página pessoal no GitHub onde ficarão disponíveis a quem se interessar. A documentação das funções e uma vinheta ou breve tutorial estão em planos futuros.

Essa tese é fruto de um acordo de cotutela entre a Universidade Federal de Minas Gerais (UFMG) e a Universidade de Paris-Saclay (UPS) e, portanto, segue algumas peculiaridades como ter uma introdução em três idiomas. Além disso, ela não possui grandes excentricidades, segue um formato de coletânea de artigos na ordem cronológica em que foram escritos, mas que também possui uma lógica teórico metodológica. Eu não conseguia implementar métodos computacionais para uma regressão quantílica para dados longitudinais se não tivesse primeiramente escrito códigos para um LMM robusto e testado essa realização por simulações de Monte Carlo. Esse necessário passo atrás resultou no primeiro manuscrito que também foi o material defendido na prova de qualificação.

O restante desta tese segue a seguinte estrutura: o Capítulo 2 é dedicado, principalmente, a discutir modelos LMM robustos. O banco de dados de Vitória sobre o efeito dos poluentes do ar em crianças com e sem asma foi o problema motivador. Sendo os principais resultados reunidos no manuscrito que foi submetido no dia 13 de junho de 2022 a uma revista científica de estatística. Após o exame de qualificação foi acrescentado um estudo de simulação com ponderações para minimizar o impacto das observações mais influentes, visto que, nesse contexto, elas não contribuíram para um intenso aperfeiçoamento da técnica. eElas não foram adicionadas ao manuscrito, somente citadas, mas estão presentes na tese. No Capítulo 3 discorro sobre métodos de regressão quantílica para dados de painel. Enfatizo o tratamento por efeitos fixos e os novos métodos que apresentam-seos a comunidade científica que utilizam funções perda alternativas a função *check*. Esse manuscrito ainda não foi submetido a nenhuma revista científica. O Capítulo 4 é motivado por estender a regressão quantílica para dados longitudinais no caso de alta dimensão. Esse manuscrito ainda não foi submetido a nenhuma revista científica. O Capítulo 5 é reservado à conclusão geral da tese. Há uma miscelânea de informações no Apêndice, na Seção A.1 estão os fatores de correção para o LMM robusto apresentado no Capítulo 3. Na Seção A.2 estão as simulações usando os pesos de Cantoni. Na Seção A.3 está uma lista de covariáveis e como extraí-las do repositório digital MCS. Na Seção

⁴O pacote RCPP oferece uma integração direta e rápida entre R e C++

A.4 está a ajuda dos pacotes que escrevi e um breve comentário para instalá-los. Na Seção A.5 está a prova da identidade de Knight, que é apresentada em [63] sem prova. A Seção A.6 apresenta a performance computacional dos métodos em termos de tempo para estimação.

1.3 Introduction en Français

Quel est l'effet des polluants dans l'air à la santé des enfants et des adolescents ? Cette question n'a pas été entièrement répondue dans la thèse de [102], dans laquelle il n'a pas été considéré les méthodes robustes, car une observation aberrant ou atypique, elle peut modifier radicalement les estimations de tout modèle linéaire basé sur des erreurs quadratiques. Par conséquent, une nouvelle question plus précise est nécessaire pour prendre en compte les données réelles qui sont sujettes à des fluctuations extrêmes. Dans ce cas, l'effet moyen des polluants de l'air sur la santé des enfants et des adolescents est-il affecté par des observations atypiques ?

Avant d'entrer dans la question méthodologique des alternatives robustes et de leur protection contre les interférences causées par textitadditive outliers, les données soulevées par [102] sont décrites rapidement. Elle a fait des recherches sur un groupe de 82 enfants et adolescents dans la ville de Vitoria au Brésil, dont une partie souffrait d'asthme et l'autre partie non, mais tous provenaient de familles à faible pouvoir d'achat et étudiaient tous dans la même école. L'étude a été approuvée par le comité d'éthique et de recherche de Escola Superior de Ciências da Santa Casa de Misericórdia de Vitória-ES, CAAE 53176915.8.0000.5065, et par le Secretaria Municipal de Saúde de Vitória par le Escola Técnica do Sistema Único de Saúde (ETSUS). Pour déterminer si l'enfant était asthmatique ou pas, un examen clinique a été effectué. Par la suite, les enfants ont reçu l'instruction d'effectuer un autotest pour mesurer leurs niveaux de capacité respiratoire. Trois mesures ont été effectuées à la fois et seule la médiane a été prise en compte pour baisser le niveau d'erreur de mesure car il s'agit d'un autotest. Telles mesures ont été répétées sur six mois, ce qui constitue une étude longitudinale. En outre, ils ont été considérés comme variables : les conditions socio-économiques des enfants fixes au fil du temps, le sexe, l'IMC, si fumeur passif, le type d'accouchement et la forme d'allaitement. Aussi ont été considérées des variables d'observation climatologiques telles que la température, l'humidité et les polluants atmosphériques tels que CO₂, SO₂, PM₁₀, PM_{2.5}. Les polluants ont été obtenus par les mesures de l'aciema, qui possède 8 stations météorologiques dans la région métropolitaine de Vitoria, dont l'une près de l'école où les enfants volontaires étudiaient. Chaque enfant a effectué l'autotest dans un jour différent du mois, de sorte qu'il a été connecté, à chaque mesure de l'enfant, les informations météorologiques de la veille, car l'effet de la pollution a tendance à prendre 24 heures pour avoir son expression maximale. Ces exemplaires, même s'ils ne sont pas grands, présentent une structure

complexe, en fonction de l'individu, qui est à l'origine des différences série temporelles ou spatiales, cette dernière n'ayant pas été exploitée et reste ouverte aux travaux futurs. En plus de la multicolinéarité présente parmi les polluants de l'air, cette étude étant une étude observationnelle et non une expérience contrôlée, il n'y a pas de séparation parfaite entre l'effet de chaque polluant puisque beaucoup d'entre eux proviennent de sources similaires comme le trafic et l'industrie.

Avant de continuer, je voudrais faire un aperçu des nomenclatures qui seront utilisées dans cette thèse, l'explication la plus naturelle et synthétique que je connais est disponible en [120, page 1]. Dans tel travail, il expose les différences entre les termes similaires avec un exemple d'examen de la pression artérielle, c'est-à-dire si un individu a sa pression artérielle systolique et diastolique mesuré par une infirmière nous avons un exemple très simple de données multivariées. Si deux familles se rendent dans un centre de santé et que chacun de leurs membres ont sa pression diastolique mesurée, alors nous avons des informations en clusters. Si l'infirmière pense que le tensiomètre est défectueux, elle peut mesurer le patient plus d'une fois pour être sûre. Dans tel cas, nous avons des mesures répétées. Si ces mesures répétées sont retirées au fil du temps, disons que quelqu'un va au poste tous les lundis, alors nous avons des données longitudinales ou des données de panel, que nous traitons ici comme synonyme et qui est le plus utilisé en économétrie. Comprendre la structure de la hiérarchie des données aide l'analyste à proposer le modèle le mieux adapté au problème étant une question clé, la structure de variance-covariance qui doit être adoptée aux résidus du modèle. Les modèles linéaires mixtes constituent une classification générale des modèles comprenant un composant fixe et un groupe de paramètres aléatoires tels que des interceptions aléatoires ou des coefficients aléatoires. Les mesures répétées, les données longitudinales et les données de Dashboard peuvent être modélisées par LMM. L'analyse de cluster, l'analyse factorielle et l'analyse des composants principaux, bien qu'ils traitent des données en dernière instance multivariées, ne se présentent pas comme la meilleure alternative à ce type de données, ils ont en fait tendance à avoir un résultat python dans ces cas [120], mais il y a des exceptions comme cela sera montré plus loin.

Comme indiqué, la base de données [102] a été analysée sans tenir compte des méthodes robustes, mais qu'est-ce qui est robuste ? La robustesse est un terme assez large en statistique. Il veut souvent dire qu'une méthode est résistante à une rupture d'hypothèse, c'est-à-dire, même s'il y a des hypothèses pour l'application de la méthode qui ne sont pas respectées, cependant, la méthode peut être utilisée et les résultats seront attendus car ils sont similaires à ceux auxquels les hypothèses ont été pleinement respectées. Par exemple, la méthode d'analyse de variance (ANOVA) est dite robuste par rapport à la normalité car il s'agit d'une hypothèse, mais elle n'a pas besoin d'être pleinement satisfaite. Dans cette thèse, le terme de robustesse est utilisé principalement comme robustesse pour additive outliers. Les observations météorologiques observées ont

cette caractéristique, comme observées toutes les quelques minutes par jour, il y a des moments où la moyenne monte extrêmement haute et tombe rapidement au niveau où elle était.

Il existe des livres sur LMM comme [120, 90, 24], il y a également des livres sur les méthodes robustes comme [51, 82]. Je ne connais aucun livre qui rassemble ces deux sujets, mais l'article de [42] et le chapitre de [131] sont de bonnes introductions. La robustesse aux valeurs aberrantes peut être obtenue en supposant une distribution de queue lourde comme celle de la distribution t de Student, ou encore il est possible de laisser les données parler d'elles-mêmes, comme dans les fonctions de perte avec l'adaptabilité qui fait comme la fonction bisquared de Tukey, dans laquelle, elle n'équivaut pas à supposer que ces erreurs sont un mélange d'une distribution gaussienne et d'une distribution aberrante d'une autre distribution à queue lourde, mais dans ce cas, le pourcentage de contamination n'est pas déterminé, c'est-à-dire que cette partie est laissée libre pour que les données elles-mêmes parlent d'elles-mêmes. [42, 97, 135] parmi beaucoup d'autres travaux sur les modèles mixtes avec l'estimation robuste indiquent [53] comme le créateur de la proposition. Bien qu'au moment où Huggins a écrit son article, la théorie de Huber était déjà très bien exposée en détail par [49, 48, 50, 51] Pourtant, à l'heure de prouver sa proposition, il est extrêmement synthétique et fait référence à un théorème dans le travail de [19], un travail qui est assez général et qui traite de plusieurs cas de compagnie. Il semble être un choix malheureux car dans [49] il est déjà dit qu'en cas de vraisemblance la convergence est attendue et dans [50] les cas de vraisemblance sont amplement discutés. Par conséquent, [53] proposent une robustification de la vraisemblance d'un modèle de mesure répétée à structure de variance connue et n'en tire pas profit. Il fait encore des demandes sur l'existence de la deuxième dérivée de la fonction de perte pour assurer l'expansion de Taylor comme il l'a fait [51, page 167] pour prouver la convergence de la régression simple. Cette exigence de deuxième dérivée et l'utilisation de l'expansion de Taylor est également explorée dans les esquisses de preuve proposées par [58, 105, 106]. Cependant, la preuve de la cohérence de l'estimateur robuste pour le modèle mixte le plus complet est [104], où l'auteur explore la fonction de perte de Huber et la pondération proposée par [15]. Il est étrange qu'un livre consacré à la partie asymptotique de modèles mixtes comme [56] avec un chapitre sur l'animal robuste ne cite ni [53] ni [104].

Outre la question de la robustesse, des valeurs aberrantes, il faut également prendre en compte les points d'influence et d'effet de levier. [15] a proposé d'approfondir l'étude de la matrice covariable ou de la matrice de dessin, car à partir de celle-ci, certaines observations pourraient avoir une influence excessive sur le modèle. De la façon dont la vraisemblance robuste est construite en utilisant la fonction Huber les valeurs aberrantes apparaissent dans les résidus, apparemment peu importe si elles proviennent d'une mesure discordante dans la variable réponse ou la matrice de dessin. Cependant, dans les études d'observation où l'on a peu de contrôle sur la matrice des covariables, cette subtilité peut

faire une différence. Car grâce à la matrice de projection, on peut déjà estimer quelles observations sont trop influentes dans le modèle et les pénaliser pour que les observations aient un poids plus équilibré dans le modèle final. Cette proposition de [15] n'est pas en concurrence avec les fonctions de [49], mais ils symharmonisent facilement dans une même vraisemblance comme elle a exploré [71] et avait déjà démontré et prouvé la consistance asymptotique [104].

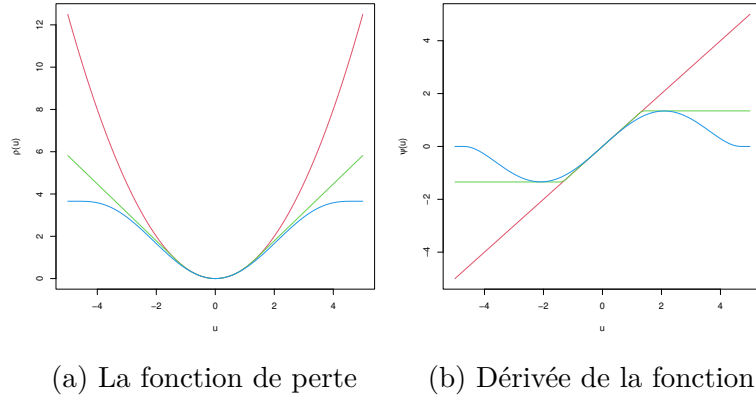


Figure 1.5: Fonction de perte quadratique (rouge), fonction de perte de Huber (vert) et fonction de perte de Tukey (bleu).

Dans la figure 1.5 sont les fonctions de perte quadratique, Huber et Tukey, la gauche et la droite, leurs dérivées. Clairement, les fonctions quadratiques et Huber sont convexes et la fonction Tukey ne l'est pas. Hors de l'intervalle entre $c/\sqrt{5}$ et $-c/\sqrt{5}$. La méfiance selon laquelle les fonctions non convexes ne génèrent pas d'estimateurs cohérents est ancienne [49] suggère que des contraintes supplémentaires doivent être imposées. [70, page 45] suggère, pour la fonction Tukey, que de bons points de départ sont suffisants pour la convergence. Déjà [78] est plus radical et exige que la plupart des déchets tombent dans la région convexe de la fonction non convexe et que celle-ci soit symétrique, ce qui est le cas de la fonction Tukey. Par conséquent, la valeur de c ne peut pas être trop petite. [51, page 103] nous avertit que l'utilisation de fonctions comme Andrews, Hampel et Tukey peut être très dangereuse et que même la suppression manuelle des observations impossibles peut être une alternative plus prudente. Cette étude des propriétés des fonctions était vitale pour ne pas proposer une combinaison de la fonction *check* pour estimer des quantités avec des fonctions non convexes, qu'apparemment c'est un chemin qui n'apporte pas de bons résultats.

Un chemin qu'on n'a pas suivi dans la thèse consiste à optimiser le paramètre tuning automatiquement [127, 8], mais il n'a pas été possible de trouver une quantité qui ne serait pas monotone ou qui ne causerait pas plus d'instabilité au modèle. Par conséquent, on préfère conserver des valeurs traditionnelles qui correspondent à une efficacité relative de 95 % par rapport à l'estimation moyenne par moindres carrés [82, page 357]. Établir une

région compacte avec un maximum est un bon point de départ pour ceux qui suivent cette problématique [31].

Toutefois, comme nous l'avons vu dans nos propres simulations, les fortes corrélations entre les covariables sont la principale faiblesse de toutes ces méthodes. Et les données environnementales de Victoria sont assez complexes et corrélées, donc on a préféré suivre la proposition des auteurs précédents de réduire la dimensionnalité des données par une technique robuste d'analyse de composant principal [17]. Il existe plusieurs techniques possibles pour ce faire qui explorent différentes façons d'estimer la variance des données [18, 80, 52].

Grâce à la combinaison de la méthodologie exposée, nous avons pu analyser la base de données [102] et considérer l'effet des polluants, malgré les observations divergentes. On a également pris en compte des facteurs propres à chaque enfant tels que : son niveau d'allergie et si l'enfant était ou pas fumeur passif. Une variable dichotomique indiquant ou non si l'asthme n'a pas été incluse, car les asthmatiques du groupe présentent un tableau contrôlé et utilisent des médicaments en cas de besoin. Par conséquent, les asthmatiques et les non-asthmatiques n'ont pas présenté de différence significative quant à la capacité respiratoire mesurée au cours des mois. De cette façon, on peut contribuer au débat sur la santé publique avec une telle étude qui combine un échantillon de facteurs d'enfants qui sont contrôlés par un dessin d'expérience, mais des informations climatologiques sont ajoutées avec des caractéristiques d'études d'observation. Il est également bien à noter qu'avec ces méthodologies robustes, il n'a pas été nécessaire de supprimer d'échantillon, ce qui serait un problème dans une taille d'échantillon relativement petite comme celle-ci.

Quel est l'effet des polluants de l'air sur la santé des enfants et des adolescents les plus vulnérables ? Par exemple, sur les 10 % d'enfants les plus vulnérables ? Ou le 1 % le plus vulnérable ? Ces questions structurent la deuxième partie de cette thèse, dans laquelle la question principale est la régression quantile appliquée aux données longitudinales.

Pour répondre à cette question, on a suivi un chemin tortueux, qui a commencé avec [107] qui proposait d'unifier la régression quantile et des copulations, en donnant une interprétation alternative aux percentiles, dans laquelle le percentile cesse d'être une valeur définie par l'analyste pour ajuster un hyperplan et devient un hyperparamètre du modèle. Cette méthode est très sensible car elle dépend d'une estimation et d'une transformation des déchets. En lisant des articles qui unissaient régression quantile et mesures répétées ou données longitudinales, il était possible de réaliser qu'il y avait un consensus : tous les articles citaient [65]. Ainsi, comme [66] est un article séminal pour la régression quantile, [65] est l'article fondamental pour la régression quantile pour les données longitudinales. Cependant, la lecture de cet article n'est pas facile car il est difficile de comprendre, puisque l'auteur traite les interceptions individuelles comme des effets fixes.

Certains travaux utilisent la distribution asymétrique de Laplace comme [61, 39,

40, 33]. L'expression fermée semblait solide, avait une vraisemblance, et le percentile est choisi par l'analyste résultant en des expressions dont l'asymétrie augmente à mesure que le percentile se rapproche des extrêmes zéro ou un. La méthode d'estimation par EM était déjà disponible dans R et pouvait être modifiée, y compris les prioris bayésiens pourraient être facilement incluses dans la méthode. Par la suite, le travail de [118] a été lu, on a remarqué qu'il proposait un M-quantile pour les modèles mixtes ou mieux, un M-quantile pour les effets aléatoires et promettait d'étendre la méthode pour les modèles mixtes dans les travaux futurs. [118] citait des œuvres comme [53, 97, 106]. Il proposait une modification de la robuste vraisemblance de Huggins en suivant la correction de type II proposée par [131] pour l'estimation robuste de la moyenne. Ce plugin quantil à la place d'une moyenne semble étrange et improbable. Une autre citation de valeur d'un tel travail est l'étude de [10]. Dans cet article inaugural de la méthode M-quantile pour la régression simple, il est non seulement plus cité parce que l'un des principaux appels de la M-régression pour les travaux appliqués est sa robustesse et la régression quantile est déjà robuste. Les autres avantages énumérés dans [68, 11, 62] sont moins voyants. Cependant, la même question se posait : est-il possible d'utiliser une fonction M-quantil comme plugin dans une vraisemblance construite pour l'effet moyen sans dégénérer cette vraisemblance ?

En retournant à [65], l'auteur explique que minimiser le BLUP d'un modèle avec des interceptions aléatoires équivaut à minimiser un modèle de régression simple avec de nombreux intercepteurs dans lesquels ils subissent une contrainte: la somme de ses modules ne peut dépasser une certaine quantité. Cette construction de la méthode LASSO permet de résoudre deux questions : estimer les coefficients des variables prédictives en tenant compte de l'effet individuel des unités d'échantillonnage dans un modèle longitudinal et ne pas laisser le nombre de paramètres augmenter de manière excessive. Cela répond au fait que LASSO et BLUP sont des techniques efficaces. L'auteur cite [98], qui attribue l'idée à un article qui semble ne rester que le résumé [46]. La proposition de [65] est d'effectuer une régression quantile pour un modèle avec plusieurs interceptions qui sont pénalisés par LASSO. Il devient donc clair pourquoi [65] n'utilise pas le terme LMM. Nonobstant, tant d'effets fixes étaient encore une question un peu confuse. Il manquait un chercheur comme [137] pour séparer les méthodes avec de la maîtrise. Les données longitudinales peuvent être divisées en deux types d'analyse : les effets aléatoires et les effets fixes. Les premiers sont généralement supposés être sans espoir et indépendants des autres variables prédictives. En raison de ces propriétés, ils peuvent être omis du processus d'estimation en termes de location et seulement inclus dans la structure de variance. Sur les secondes, aucune de ces deux hypothèses sont nécessaires, étant traitées comme des variables de dummies. Mais le nombre de paramètres augmente avec la taille de l'échantillon et l'estimation devient [2, 47, 38, 137]. L'idée de [46] est d'être flexible dans les hypothèses et d'assumer des effets fixes, mais d'être pratique dans l'estimation et

de les pénaliser pour LASSO ce qui équivaut à des effets aléatoires. Selon Koenker, cette démonstration d'équivalence et d'équilibre ouvre la porte à d'autres pénalités et propositions méthodologiques. C'est là que naquit la régression quantile avec des effets fixes et la pénalisation par LASSO [65]. L'extension naturelle serait de proposer une M-quantile pour des effets fixes au lieu d'effets aléatoires comme [118].

Jusqu'à présent, on a parlé de régression quantile sans entrer dans les détails, mais chaque méthode a son histoire. La régression quantile devient célèbre à partir de l'article de [66] qui généralise le concept non paramétrique de quantile invariable pour une relation conditionnelle entre au moins deux variables aléatoires ou une matrice de dessin et une variable de réponse. [66] souligne que contrairement au cas univarié dans lequel le quantil a une formule fermée, le cas bivarié était ouvert à la discussion, pourrait être estimé par la fonction *check*, mais pourrait également être estimé par ranks, entre autres possibilités qui peuvent encore être proposées. [86] propose le retour à la moyenne, plus précisément les expectiles qui sont une forme de quantil moyen. Utiliser une fonction quadratique asymétrique à la place de la fonction *check*. Sa méthode a de bonnes propriétés comme moins de risque de *quantile crossing* et une meilleure estimation des quantiles extrêmes. Cependant, cette méthode est sensible à la distribution de probabilité des erreurs et n'est pas robuste aux valeurs aberrantes. *Quantile crossing* est comme n ol'appelle lorsque l'estimation du coefficient quantile d'une régression n'est pas monotone croissant pour un $\mathbf{X}$ fixe. Par exemple, supposons que nous estimions les coefficients d'un retour pour deux centiles : 0,3 et 0,6. La valeur prédite par le coefficient 0.3 fois une valeur fixe de $\mathbf{X}$ doit être inférieure ou égale à la valeur prédite par un centile 0.6. Si cela ne se produit pas, l'hypothèse que la fonction de probabilité accumulée soit monotone croissante n'est pas satisfaite. Souvent, l'estimation empirique de la régression quantile peut entraîner un quantile crossing. La régression expectile n'est pas non plus exempte de ce problème, mais elle est moins fréquente. Puis, vient le travail de [10], qui propose un équilibre entre quantile et expectile : la M-quantile. Tout comme le travail de [49], il pèse entre la moyenne et la médiane avec différentes fonctions de perte, le travail de [10] pèse entre le quantil et l'expectile. Les auteurs mentionnent également que l'estimation n'est pas unique et peut être obtenue par la fonction de perte ou sa dérivée, ce qui peut entraîner des différences mineures en fonction du type de problème. Enfin, nous avons un article qui prouve que les trois propositions, bien qu'avec des noms différents, sont également quantis [59].

Dans la figure 1.6 figurent les fonctions *check*, M-quantile et quadratique asymétrique. La gauche est la fonction de perte et la droite est ses dérivées. Clairement, toutes les fonctions sont convexes, ce qui facilite grandement l'estimation et la preuve de cohérence. La dérivée de la fonction *check* a un point de discontinuité à l'origine, ce qui peut causer une instabilité numérique en fonction du type de problème à minimiser. Dans cet exemple, le percentile choisi est 0.3, mais des valeurs plus extrêmes entraînent une asymétrie plus

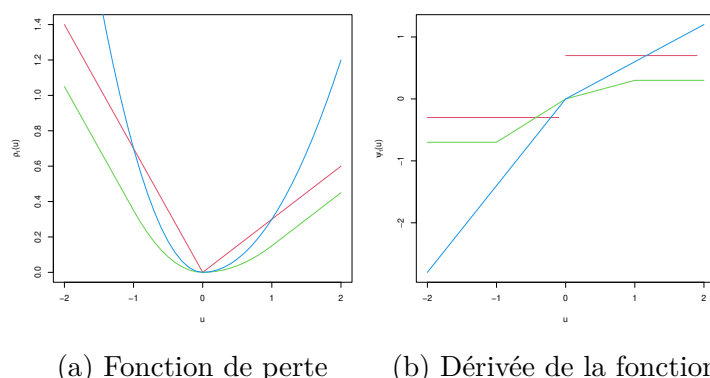


Figure 1.6: Fonction *check* (rouge), fonction M-quantile (vert) et fonction quadratique asymétrique (bleu) par $\tau = 0.3$.

forte. Les valeurs supérieures à 0,5 inversent le rapport d'asymétrie, par exemple 0,7 miroir horizon .

Table 1.5: Acronymes des modèles de régression en fonction de leurs caractéristiques et de leur forme d'estimation

Intercept(s)	Quantile	M-quantile	Expectile
Unique	QR	MQR	ER
Aléatoire	QRRE/LQMM	MQRRE	ERRE
Fixé	QRFE	MQRFE	ERFE
Pénalisé	PQRFE	MQRFE	PERFE

Jusqu'à présent, on discute sur trois types d'optimisation : quantile, M-quantile et expectile. En outre, on a quatre modèles possibles : régression simple, effets aléatoires, effets fixes et effets fixes avec pénalisation de type LASSO. Le croisement de ces formes d'utilisation par les types de modèles aboutit à 12 méthodes possibles, en plus des particularités et des restrictions que chaque auteur ajoute à sa méthode comme *splines*, EM, etc. talemment ce graphique. La comparaison de tant de méthodes aboutit à plusieurs acronymes et la lecture devient fatigante. Par conséquent, dans la table 1.5 a été rédigé un résumé. Sur les caractéristiques des intercepts : s'il y a un seul intercept, nous avons une régression simple, s'ils sont aléatoires, nous avons des effets aléatoires ou des modèles mixtes, s'ils sont fixes, nous avons une régression avec de nombreuses variables dummies, enfin les intercepteurs pénalisés sont fixes, mais avec la restriction est de type LASSO. Voici une brève mention des auteurs qui ont proposé leurs méthodes:

- QR est proposé par [66],
- LQMM est proposé par [39],
- QRFE est proposé par [65],

- PQRFE est proposé par [65],
- MQR est proposé par [10],
- MQRRE est proposé par [118],
- MQRFE est proposée par nous dans cette thèse,
- PMQRFE est proposé par nous dans cette thèse,
- ER est proposé par [86],
- ERRE est proposé par [7, 123],
- ERFE est proposé par nous dans cette thèse,
- PERFE est proposé par nous dans cette thèse.

Un chemin qui n'a pas été suivi dans la thèse consiste à optimiser le paramètre tuning de M-quantile automatiquement comme certains auteurs l'ont fait pour la M-estimation de la moyenne [127]. Même un choix analytique de la constante de tuning est un problème ouvert, puisque la plupart des travaux comparent la M-quantile avec le *check* et non avec la quadratique asymétrique, alors la relation classique d'efficacité doit être mise à jour [82, page 357]. Ainsi, recycler des valeurs comme $c = 1.345$ semble une "copie et colle" inappropriée, mais répétée dans la littérature [100, 9, 118]. L'un des rares travaux dans lesquels l'impact de plusieurs valeurs de tuning constant est analysé est celui de [62]. C'est pourquoi $c = 1$ a été choisi comme [69, 62].

Les résultats asymptotiques de cohérence et de normalité des estimateurs sont prouvés par [65] pour les modèles QRFE et PRQFE. Vérifier le paramètre de contrainte est laissé ouvert par l'auteur, mais [73] prouve la cohérence du paramètre tuning. Koenker utilise les convergences explicitées par [64] et une identité de [63]. Malheureusement, cette identité ne vaut que pour la fonction *check* et ne peut pas être appliquée dans M-quantile ou expectile. Néanmoins, les résultats asymptotiques des nouveaux modèles ont été partiellement résolus. Nous avons déjà prouvé la cohérence et la normalité des estimateurs de type expectile, c'est-à-dire pour ERFE et PERFE, mais cela n'a pas encore été fait pour M-quantile : MQRFE et PMQFE. La forme quadratique de l'expectile aide dans les épreuves. On pense que la cohérence de M-quantile peut être prouvée, car des cas particuliers tels que la fonction Huber ont déjà fait l'objet d'une enquête approfondie. Même pour les cas avec pénalité LASSO [108], et les modèles avec effets aléatoires [104, 106].

En l'absence d'une preuve mathématique définitive, nous avons de bons résultats en simulant Monte Carlo et en l'appliquant à des données réelles qui semblent prouver des hypothèses. Pour l'illustration de la méthode avec des données réelles, on a recherché des

données de panel dans lesquelles on pouvait être très sûr qu'il n'y avait pas d'erreurs et dont la relation entre les covariables était déjà un consensus dans la littérature spécifique. Bien que cette partie du texte ait commencé par poser une question sur l'effet des polluants de l'air sur les enfants les plus vulnérables en faisant allusion à des questions de santé publique, la base de données choisie pour cette deuxième partie est de nature économique. De cette façon, la méthode est utile pour une autre science appliquée comme l'économie. En outre, la source est la Banque Mondiale, et pour être dans un tel référentiel, la base de données a subi des révisions exhaustives, la fin concerne les indices des pays membres de l'Organisation de Coopération et de Développement Économiques (OCDE). La littérature économique a déjà expliqué la relation entre le pourcentage de la haute technologie dans les exportations et les variables prédictives comme l'investissement dans l'éducation et la science et le développement.

Une autre base de données, dont la structure est provocante, est le Millennium Cohort Study (MCS), analysé par [118, 41, 4]. Le MCS constitue une étape importante dans les recherches longitudinales dans le Royaume-Uni où près de 19.000 enfants ont fait partie des échantillons pour participer à l'étude qui a commencé en 2001 et suit actif jusqu'à nos jours. Plus de 3000 questions sont posées chaque année aux parents, aux enseignants, à la famille et à l'enfant lui-même. Une telle quantité d'informations peut être utilisée pour étudier des questions de santé publique, de développement physique et psychologique, de relations socio-économiques, entre autres. [118] n'a étudié que deux variables réponses et 11 variables prédictives covariables, ce qui ne représente que la pointe de l'iceberg de ce qu'il y a à analyser dans le MCS. Les grandes banques de données nous amènent à la troisième question structurante de la thèse .

Quel est l'effet des polluants de l'air sur la santé des enfants et des adolescents les plus vulnérables dans lesquels 5 % sont vulnérables aux polluants et aux variables prédictives ? 5% est une très grande quantité et nous n'avons aucune information a priori pour les sélectionner. Une telle question structure à la troisième partie de la thèse, dans laquelle la question principale est la régression quantile appliquée aux données longitudinales de grande dimension.

Pour répondre à cette question, un chemin plus direct a été parcouru, car il existe un livre récemment publié qui a fourni une carte mise à jour des options disponibles jusqu'ici pour ce type de problème [13]. Dans le livre sont présentes des techniques telles que LASSO [114], LASSO adaptatif [140], Elastic-net [141] pour la régression avec des erreurs du cadre, M-estimation [49], entre autres.

La solution proposée est une modification mineure du modèle original de [65]. Où les variables exploratoires sont pénalisées ainsi que les intercepts par LASSO. Cela semble naturel, mais cela entraîne des conséquences comme : comment estimer le paramètre de restriction. Nous n'avons plus la garantie montrée par [73] que l'écart type des déchets divisé par l'écart type des intercepts converge vers la vraie valeur λ . Car maintenant

le paramètre tuning s'étend aussi aux prédicteurs. Il semble que nous devions recourir à la stratégie que la plupart des analystes utilisent : *cross-validation* (CV). Pourtant, on préfère utiliser le critère *Bayesian Information Criteria* (BIC) pour deux raisons : parce que la régression quantile a un biais plus explicatif que prédictif. Il est certain que la prédiction au 10% entile comportera de grandes erreurs quadratiques ou même absolues ; et les données longitudinales ont un rapport de dépendance intrinsèque. Il est clair qu'il est possible de contourner cette limitation et de réafficher les données longitudinales, mais s'il est possible d'éviter une autre source d'erreurs, je préfère suivre cette voie.

Table 1.6: Acronymes des modèles de régression selon le type de sanction.

Méthode	Intercepts	LASSO dans α	LASSO dans β	Adaptatif
QR	Non	Non	Non	Non
QRFE	Oui	Non	Non	Non
PQRFE/LQRFE	Oui	Oui	Non	Non
DPQRFE	Oui	Oui	Oui	Non
ALQRFE	Oui	Oui	Oui	Oui

Maintenant, le nombre de méthodes n'est pas si important, mais on pense qu'un tableau peut aider à mieux comprendre la contribution de chaque auteur et la nomenclature. Dans le Tableau 1.6 nous avons les acronymes de la régression quantile simple (QR) proposée par [66], régression quantile avec effets fixes et pénalités par [65] (PQR) proposée par [65], régression quantile avec des effets fixes et pénalisation sur les interceptions et les variables exploratoires par LASSO (LQRFE) proposée par nous, régression quantile avec effets fixes et pénalisation sur les intercepts et les variables exploratoires par LASSO en deux étapes (DLQRFE) proposée par [77], qui n'en demeure pas moins un sous-cas de Group LASSO ver [138], et la régression quantile avec des effets fixes et la pénalisation sur les intercepts et les variables exploratoires par LASSO adaptatif (ALQRFE) proposée par nous. La régression quantile simple avec LASSO comme proposition par [134] est un cas particulier.

- QR est proposé par [66],
- QRFE est proposé par [65],
- PQRFE est proposé par [65],
- LQRFE est proposée par nous dans cette thèse,
- DPQRFE est proposé par [77],
- ALQRFE est proposée par nous dans cette thèse.

Tout comme [118, 4] ont limité ses études de MCS à l'Angleterre, l'analyse effectuée s'est limitée à l'Irlande du Nord, car l'Irlande est le pays avec le plus petit échantillon,

un peu plus de 600 enfants tous les deux ans étant donné que les données manquantes. Le modèle fixe le plus simple de tous nécessite déjà 600 interceptions à estimer, si nous limitons la base de données à un peu plus de 1200 covariables, nous aurons déjà une taille élevée, car ce sont plus de paramètres que d'observations. Quoique, nous n'avons pas besoin d'aller aussi loin pour avoir des problèmes de confusion et être nécessaire de sélectionner les covariables ont déjà été supprimées. Dans notre exemple, on a considéré environ 100 covariables de MCS. Dans ce cas, il a été observé que la corrélation entre les variables exploratoires était faible. Cependant, si elle était modérée ou élevée, comme c'est le cas le plus souvent, la restriction par LASSO serait déconseillée, et il serait plus approprié de procéder avec un LASSO adaptatif. Cette variante a été écrite plus généralement par LASSO pour aider les futurs chercheurs appliqués, car il existe plusieurs bases de données longitudinales similaires de grande valeur pour les études épidémiologiques et sociales pour ne citer que deux exemples, l'un brésilien et l'autre français : [le 1982 Birth Cohort](#) , développé par des chercheurs de l'Université Fédérale de Pelotas (UFPEL) à Pelotas, au sud du Brésil avec un échantillon initial de près de 6000 enfants, et [Étude Longitudinale Française depuis l'Enfance \(ELFE\)](#). France avec un échantillon initial de plus de 18000 enfants.

Les deux réponses variables choisies par [118] sont : *strengths and difficulties questionnaire* (SDQ), externe et interne qui indiquent le degré de désordre social et de souffrance interne des enfants et des adolescents. En ajoutant de nombreuses variables exploratoires, nous trouvons de nouveaux prédicteurs en plus de ceux que la littérature spécialisée établit habituellement. Un phénomène intéressant s'est produit : en incluant des questions intimes de chaque famille, les variables générales telles que les conditions du quartier, la conjoncture économique ou le statut ethnique ont cessé d'être importantes.

La normalité asymptotique des estimateurs pour le nouveau modèle a été facile à obtenir, car la partie la plus laborieuse avait déjà été faite par [63, 64, 65]. En outre, des études de simulation de Monte Carlo ont été réalisées, qui montrent des variables indépendantes LASSO maîtrise le LASSO adaptatif et le modèle fixe sans restriction.

Bien qu'il ait toujours ouvert de nouvelles questions en utilisant les polluants de l'air et la capacité respiratoire des enfants, en fait les vraies questions auxquelles on répond dans les chapitres de cette thèse changent en fonction des données réelles analysées. Ainsi, la vraie question pour la deuxième partie serait : quel est l'effet de l'investissement dans l'éducation et la science sur la qualification technologique des exportations des pays les plus avancés, des pays médians et des pays les moins avancés de l'OCDE ? De même, la troisième partie répond à la question : quelles sont les variables les plus pertinentes dans le score SDQ interne et externe des enfants les plus vulnérables aux traumatismes familiaux ou sociaux ?

Les méthodes de calcul ont été brièvement décrites dans chaque chapitre. Tous ont été développés dans R et RCPP footnoteLe paquet RCPP offre une intégration directe

et rapide entre R et C++ [95, 27]. L'optimisation a été réalisée par des variations de la méthode Newton-Raphson dans laquelle le cœur de la fonction a été écrit en RCPP et les fonctions auxiliaires en R pour accélérer le processus. Peu à peu, les codes montent sur ma page personnelle sur GitHub où ils seront à la disposition de ceux qui sont intéressés. La documentation des fonctions et une vignette ou un bref tutoriel sont sur les plans futurs.

Cette thèse est le fruit d'un accord de cotutela entre l'Université Fédérale du Minas Gerais (UFMG) et l'Université Paris-Saclay (UPS) et suit donc quelques particularités comme avoir une introduction en trois langues. En outre, elle n'a pas de grandes excentricités, elle suit un format de recueil d'articles dans l'ordre chronologique dans lequel ils ont été écrits, mais qui possède également une logique théorique méthodologique. Je ne pouvais pas mettre en œuvre des méthodes de calcul pour une régression quantile pour les données longitudinales si je n'avais pas d'abord écrit des codes pour un LMM robuste et testé cette performance par des simulations de Monte Carlo. Ce pas en arrière nécessaire a abouti au premier manuscrit qui a également été le matériel défendu dans l'évaluation à mi parcours.

Le reste de cette thèse suit la structure suivante : le chapitre 2 est principalement consacré à discuter de modèles LMM robustes. La base de données de Victoria sur l'effet des polluants de l'air sur les enfants avec et sans asthme était le problème motivant. Les principaux résultats étant rassemblés dans le manuscrit qui a été soumis le 13 juin 2022 à une revue scientifique des statistiques. À la suite de l'examen de qualification, une étude de simulation avec des pondérations a été ajoutée pour minimiser l'impact des observations les plus influentes, car, dans ce contexte, elles n'ont pas contribué à une amélioration intensive de la technique. Elles n'ont pas été ajoutées au manuscrit, seulement citées, mais sont présentes dans la thèse. Dans le chapitre 3 je discute des méthodes de régression quantile pour les données de tableau de bord. Je mets l'accent sur le traitement par effets fixes et les nouvelles méthodes présentées par la communauté scientifique qui utilisent des fonctions de perte alternatives à la fonction *check*. Ce manuscrit n'a encore été soumis à aucune revue scientifique. Le chapitre 5 est réservé à la conclusion générale de la thèse. Il y a un mélange d'informations dans l'annexe, dans la section A.1 sont les facteurs de correction pour le LMM robuste présenté dans le chapitre 3. Dans la section A.2 sont les simulations utilisant les poids de Cantoni. Dans la section A.3 se trouve une liste de covariables et comment les extraire du référentiel numérique MCS. Dans la section A.4 se trouve l'aide des paquets que j'ai écrits et un bref commentaire pour les installer. Dans la section A.5 se trouve la preuve de l'identité de Knight, qui est présentée dans [63] sans preuve. La section A.6 présente la performance de calcul des méthodes en termes de temps d'estimation.

Chapter 2

A longitudinal study of the influence of air pollutants on respiratory health. A robust multivariate approach

Abstract

This paper aims to evaluate the statistical association between exposure to air pollution and forced expiratory volume in the first second (FEV_1) in both asthmatic and non-asthmatic children and teenagers, in which the response variable FEV_1 was repeatedly measured on a monthly basis, characterizing a longitudinal experiment. Due to the nature of the data, a robust linear mixed model (RLMM), combined with a RPCA, is proposed to handle the multicollinearity among the covariates and the impact of extreme observations (high levels of air contaminants) on the estimates. The Huber and Tukey loss functions are considered to obtain robust estimators of the parameters in the LMM. A finite sample size investigation is conducted under the scenario where the covariates follow linear time series models with and without additive outliers (AO). The impact of the time-correlation and the outliers on the estimates of the fixed effect parameters in the LMM is investigated. In the real data analysis, the robust model strategy evidenced that RPCA exhibits three PC, mainly related to relative humidity (Hmd), particulate matter with a diameter smaller than $10 \mu\text{m}$ (PM_{10}) and particulate matter with a diameter smaller than $2.5 \mu\text{m}$ ($PM_{2.5}$).

Keywords: Linear mixed model, Principal component analysis, M -estimation, robustness, outliers, asthma.

2.1 Introduction

The development in cities not only benefits the local economy, such as by creating jobs and contributing to urban development, but it also generates various residues which lead to environmental and health problems, affecting the quality of life of the population among others. The World Health Organization (WHO) conceptualizes health as a “state of complete physical, mental and social well-being, and not simply the absence of disease or illness” [133]. As is widely addressed in the environmental epidemiological studies, children are more vulnerable to air pollution since their respiratory and immunologic systems are immature. Consequently, they suffer more severe mortality and morbidity [112, 1, 28]. In this context, the statistical association between air quality variables and health effects must be computed with caution, independently of the statistical regression and time series models used. In the statistical models in epidemiological studies, a standard strategy is to consider air pollutants and weather as covariate variables. In general, these variables are time-correlated and present multicollinearity. Ignoring these phenomena in the modelling steps can lead to a wrong model selection and have severe consequences on the analysis of the impact of the pollutants on health, like a false-positive conclusion of the population health risk. Additionally, high levels of pollutants frequently appear in the pollutant variables, but their impact is often ignored in the literature. However, these observations can be identified as AO which affect the estimation of some statistical characteristics of the data, like the mean, the variance, and the correlation. Hence, robust estimation methods are needed to get reliable statistical models see, for example, [96, 17].

Many studies have recently paid attention to these issues in quantifying the association between pollutants and adverse health effects. For example, PCA has been proposed to mitigate multicollinearity in the predictor or regressor variables. [126] studied the combined effects of pollutants on daily mortality using a generalized additive model (GAM) with a robust PCA. They concluded that the relative risk (RR) estimates were more pronounced when the multivariate robust PCA technique was used. The application of PCA generally requires the data to be obtained through independent replications. However, as addressed in [139], if the covariates are time-correlated the PC are also autocorrelated. In this context, to handle the multicollinearity among the covariates and the autocorrelation of the PC, [109] and [55] have combined the PCA technique and multivariate time series in the GAM model to quantify the impact of the pollutants on respiratory diseases. They showed that the estimation of the RR was more pronounced than indicated previously in the literature. This corroborates that statistical tools must be used with caution to quantify linear and non-linear statistical associations between response and predictor variables.

The same issues appear in longitudinal studies where repeated measurements are collected over time. General regression models with multiple sources of errors, denoted as

LMMs, have been suggested in the literature to analyse this type of data, see for example [120] for a review. In the early 90s, many researches on robust estimation methods for repeated measurements were conducted, from a theoretical and an applied point of view. [53] introduced the M-estimation approach and since then, significant progress has been made towards proposing robust methodologies for analysing longitudinal data in different areas of application, especially because, nowadays, there is a large amount of multivariate data, and also computational and software facilities. Recently, as a robust alternative technique for longitudinal data analysis, special attention has been paid to quantile regression methods see, e.g., [65, 36].

The goal of this paper is to quantify the predictor variables of air pollution, weather, the logarithm of total immunoglobulin E (IgE), among others, with FEV_1 , both in asthmatic and non-asthmatic children and teenagers up to 18 years old. FEV_1 values were repeatedly measured monthly in an experiment held at the Public Health Center of Praia do Suá, in Vitória, Capital of State of Espírito Santo (ES), Brazil. Biological, social, and economic level variables of each unit (child) were also considered as covariates in the modelling strategy. This data set is presented in [102, page 31]. Since the response variable is longitudinal, a particular case of the LMM discussed by [53] is proposed, but considering the covariates as a multiple time series data with abrupt observations, such as high peaks of the pollutants. As previously mentioned, these observations produce the same effect on the estimates as AO do. Since the estimation of the covariance and correlation matrices are strongly influenced by AO, the estimation of eigenvectors and eigenvalues is also affected, and thus classical PCA is sensitive to AO.

In this context, a RLMM, combined with a RPCA, is proposed to handle simultaneously, the multicollinearity among the covariates and the impact of high peaks of the air contaminants on the estimates. The classical LS estimator and the standard PCA are also considered for comparison. To the best of our knowledge, this modelling strategy has not been suggested yet for the case of repeated measurement data with multivariate time series covariates and outliers. In addition, to empirically verify the impact of the outliers on the estimates of the fixed effects and the variances of the source errors, a finite sample size investigation is conducted under the scenario where the covariates follow linear time series models, either with or without AO.

The remaining sections of this paper are organized as follows. Section 2.2 discusses the LMM and its robust estimation. Section 2.3 presents a Monte Carlo simulation study to evaluate the efficiency of the LS method and M -estimation with Huber and Tukey loss functions, under different scenarios. Section 2.4 addresses the real data problem, and conclusions are in Section 2.5. Appendix A.1 explicit some quantities of interest and Appendix A.2 develops the simulations to deeply understand the impact of weighted likelihood.

2.2 Linear mixed model

We denote by $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{im_i})^T$ the vector of measurements taken at times $t_{i1}, \dots, t_{im_i}$ of the i th subject for $i = 1, \dots, n$. Following [25] and [120, page 23], we assume that each measurement Y_{ij} of $\mathbf{Y}_i$, $j = 1, \dots, m_i$, follows the LMM

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\boldsymbol{\gamma}_i + U_{ij}, \quad (2.1)$$

where $\mathbf{X}_{ij}$ is a known $(1 \times d)$ design vector, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_{d-1})^T$ is a vector of d unknown but fixed parameters to be estimated, $\mathbf{Z}_{ij}$ is a known $(1 \times q)$ design vector for the random effect, $\boldsymbol{\gamma}_i$ is a vector of q unobservable random effects assumed to follow a normal distribution with mean 0 and $(q \times q)$ covariance matrix $\mathbf{D}$, i.e., $\boldsymbol{\gamma}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$, and $\mathbf{U}_i = (U_{i1}, \dots, U_{im_i})^T \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \boldsymbol{\Omega}_i)$ models an experimental error. When $q = 1$ and $Z_{ij} = 1$, γ_i represents a random intercept. When $\boldsymbol{\Omega}_i$ is the identity matrix, the components of $\mathbf{U}_i$ are independent. However, in real applications, it is usually assumed that these components are time-correlated and follow a first order autoregressive (AR(1)) model. In this context, we suppose here that the (j, k) th entry of $\boldsymbol{\Omega}_i$ is given by $\nu^{|t_{ij} - t_{ik}|} / (1 - \nu^2)$ where $|\nu| < 1$. [120, page 99] suggest alternative correlation structures for the LMM.

It follows from (2.1) that

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\boldsymbol{\gamma}_i + \mathbf{U}_i$$

where $\mathbf{X}_i = [\mathbf{X}_{i1}^T, \dots, \mathbf{X}_{im_i}^T]^T$ is the $(m_i \times d)$ design matrix of subject i for fixed parameters and $\mathbf{Z}_i = [\mathbf{Z}_{i1}^T, \dots, \mathbf{Z}_{im_i}^T]^T$ is the $(m_i \times q)$ design matrix for the random effects of subject i . We assume that $\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_n, \mathbf{U}_1, \dots, \mathbf{U}_n$ are mutually independent. Then, $\mathbf{Y}_i$ is a Gaussian vector with mean $\mathbf{X}_i\boldsymbol{\beta}$ and covariance matrix

$$\boldsymbol{\Sigma}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \sigma_e^2\boldsymbol{\Omega}_i. \quad (2.2)$$

The standardized residuals $\mathbf{R}_i$ is defined as

$$\mathbf{R}_i = \boldsymbol{\Sigma}_i^{-1/2}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}). \quad (2.3)$$

Let $\boldsymbol{\theta} = (\beta_0, \dots, \beta_{d-1}, D_{11}, \dots, D_{qq}, \sigma_e^2, \nu)^T$ be the vector parameter with dimension $p = d + q(q+1)/2 + 2$. To estimate $\boldsymbol{\theta}$, we consider the likelihood function proposed by [53] in which we introduce the weights ω_{ij} defined below see, for example, [15]. The log-likelihood is written as

$$l(\boldsymbol{\theta}) = - \sum_{i=1}^n \left(\frac{\lambda}{2} \log |\boldsymbol{\Sigma}_i| + \sum_{j=1}^{m_i} \omega_{ij} \rho(R_{ij}) \right), \quad (2.4)$$

where ρ is a loss function with derivative $\rho' = \psi$, $\lambda = \mathbb{E}(R\psi(R))$ with $R \sim \mathcal{N}(0, 1)$, $\omega_{ij} = \sqrt{1 - h_{i(j)}}$, $h_{i(j)}$ is $i^{(j)}$ th component of the diagonal of the projection matrix $\mathbf{H} =$

$\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ where $\mathbf{X} = [\mathbf{X}_1^T, \dots, \mathbf{X}_n^T]^T$ is the $(N \times d)$ design matrix, $N = \sum_{i=1}^n m_i$, and

$$i^{(j)} = \begin{cases} j & \text{if } i = 1, \\ j + \sum_{k=1}^{i-1} m_k & \text{if } i \geq 2. \end{cases}$$

The role of the weights ω_{ij} and function ρ is to accommodate the outliers not only in the response variable $\mathbf{Y}_i$, but also in the covariate $\mathbf{X}_i$. When $\omega_{ij} = 1$, (2.4) becomes the likelihood function given in [53]. Function ρ is assumed to be non-monotonic and to possess bounded derivatives of a sufficiently high order. In particular, ψ should be continuous and bounded, see [51, page 165]. The M-estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ maximizes (2.4) where $\boldsymbol{\Sigma}_i$ and R_{ij} are calculated from (2.2) and (2.3), respectively, and $\mathbf{D}$ is constrained to be a covariance matrix. An alternative robust and weighted method to estimate mixed models was proposed by [71].

Under some regularity conditions [19, 53], the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ may be estimated by

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\theta}}) = G(\hat{\boldsymbol{\theta}})^{-1} \sum_{i=1}^n \left(\Lambda_i(\hat{\boldsymbol{\theta}}) \Lambda_i(\hat{\boldsymbol{\theta}})^T \right) G(\hat{\boldsymbol{\theta}})^{-1}, \quad (2.5)$$

where $G(\boldsymbol{\theta}) = \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$ and $\Lambda_i(\boldsymbol{\theta})$ is the vector of derivatives with respect to $\boldsymbol{\theta}$ of the i th summand in (2.4) [53, 131]. An estimated asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ may be obtained as

$$\widehat{\text{Cov}}(\hat{\boldsymbol{\beta}}) \approx \left(\frac{\bar{\omega}}{\varrho} \sum_{i=1}^n \mathbf{X}_i^T \hat{\boldsymbol{\Sigma}}_i^{-1} \mathbf{X}_i \right)^{-1}, \quad (2.6)$$

where $\hat{\boldsymbol{\Sigma}}_i$ is the estimate of $\boldsymbol{\Sigma}_i$ obtained by replacing the unknown parameters by their estimates in (2.2), $\bar{\omega} = N^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} \omega_{ij}$ is a known constant, and

$$\varrho = \frac{\mathbb{E}[\psi(R)^2]}{(\mathbb{E}[\psi'(R)])^2}. \quad (2.7)$$

The approximation in (2.6) is based on the inverse Hessian matrix see, for example, [51, page 173] and [42] for the classical linear and LMM regression models, respectively.

Remark 1. There are several possible choices of functions ρ which are known in the literature. Here, we consider the classical LS function $\rho_1(r) = \frac{1}{2}r^2$, the Huber function

$$\rho_2(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } |r| \leq \kappa_2, \\ \kappa_2|r| - \frac{1}{2}\kappa_2^2 & \text{if } |r| > \kappa_2, \end{cases}$$

where $\kappa_2 > 0$, and the Tukey bisquared function

$$\rho_3(r) = \begin{cases} \frac{\kappa_3^2}{6} (1 - (1 - (r/\kappa_3)^2)^3) & \text{if } |r| \leq \kappa_3, \\ \frac{\kappa_3^2}{6} & \text{if } |r| > \kappa_3, \end{cases}$$

where $\kappa_3 > 0$. We denote by (λ_k, ϱ_k) the value of (λ, ϱ) when $\rho = \rho_k$ for $k = 1, 2, 3$. Obviously, since $\psi_1(r) = r$, $\lambda_1 = \varrho_1 = 1$. The explicit expressions of (λ_2, ϱ_2) and (λ_3, ϱ_3) in terms of κ_2 and κ_3 , respectively, are given in the Appendix. These expressions are useful to compute (2.4) and (2.6).

Remark 2. We denote by $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ the LS, Huber and Tukey estimate of β , respectively. we deduce from (2.6) that the asymptotic relative efficiency of $\hat{\beta}_2$ and $\hat{\beta}_3$ with respect to $\hat{\beta}_1$ is approximately equal to $1/\varrho_2$ and $1/\varrho_3$, respectively. This can be a guideline to choose κ_2 and κ_3 . The standard choice is $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$, which gives a relative efficiency of 95% for $\hat{\beta}_2$ and $\hat{\beta}_3$, respectively, see, e.g., [44, page 383], [119, page 123] and [82, page 357].

Remark 3. Under some assumptions, the asymptotic properties of the above estimators are well-established in the literature. All estimators are consistent and asymptotically Gaussian [53, 97, 42, 15]. If the loss function is convex, then the estimator converges to the global maximum, otherwise, as in the case of the Tukey loss function, a good initial point is required to ensure that the estimator converges to the true solution [70, page 45].

Remark 4. For each $k = 1, 2, 3$, the maximization of $l(\theta)$ where $\rho = \rho_k$ in (2.4) is achieved numerically by successive iterations. The initial value $\hat{\beta}_{0k}$ of β is obtained by minimizing

$$l_0(\beta) = \sum_{i=1}^n \sum_{j=1}^{m_i} \rho_k(Y_{ij} - \mathbf{X}_{ij}\beta),$$

σ_{01}^2 is the sample variance of the residuals $(Y_{ij} - \mathbf{X}_{ij}\hat{\beta}_{01})$ and σ_{0k} is the median absolute deviation of the residuals $(Y_{ij} - \mathbf{X}_{ij}\hat{\beta}_{0k})$ for $k = 2, 3$. The initial values of D is D_{0k} where $D_{0k,ii} = \sigma_{0k}^2/(q+1)$ for $i = 1, \dots, q$ and $D_{0k,ij} = 0$ if $i \neq j$, the initial value of σ_e^2 is $\sigma_{0k,e}^2 = \sigma_{0k}^2/(q+1)$ and the initial value of ν is ν_0 where ν_0 is the same for the three loss functions and is sampled from the uniform distribution on the interval $(0, 1)$. From the computational aspect, $l(\theta)$ is maximized using the L-BFGS-B method which is a quasi-Newton procedure allowing box constraints, i.e., each variable is restricted inside lower and upper bounds. This ensures that the variances are positive. The gradient of $l(\theta)$ is approximated by finite differences. Our codes are developed in R language [95] however, we also use the package RcppArmadillo to accelerate the calculations [27]. This allows to work in a big data context [14]. Our codes are available in an R package [22].

2.3 Monte Carlo simulations

In this section, a simulation study with finite sample sizes is conducted to verify the performance of the estimators in LMM when the covariates are cross-correlated time series with AO. In the context of time series data, AO are particularly dangerous since they have

a strong impact on sample estimates such as sample mean and sample autocorrelations. Additionally, as previously mentioned, the pollution series may present observations with high levels of pollutant concentrations which may produce sample densities with heavy tails, and these observations provoke the same effect on the estimates as AO do. We consider the following particular case of model (2.1),

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \gamma_i + U_{ij}, \quad (2.8)$$

where the measurements are supposed to be taken at times $t_{i,j} = j$, $\mathbf{X}_{ij} = (X_{ij}^{(0)}, X_{ij}^{(1)}, X_{ij}^{(2)})$ with $X_{ij}^{(0)} = 1$ and $(X_{ij}^{(1)}, X_{ij}^{(2)})^T$ follows the first order vector autoregressive (VAR(1)) model

$$\begin{bmatrix} X_{i,j}^{(1)} \\ X_{i,j}^{(2)} \end{bmatrix} = \begin{bmatrix} \phi_{1,1} & 0 \\ 0 & \phi_{2,2} \end{bmatrix} \begin{bmatrix} X_{i,j-1}^{(1)} \\ X_{i,j-1}^{(2)} \end{bmatrix} + \begin{bmatrix} W_{i,j}^{(1)} \\ W_{i,j}^{(2)} \end{bmatrix}, \quad \begin{bmatrix} W_{i,j}^{(1)} \\ W_{i,j}^{(2)} \end{bmatrix} \sim \mathcal{N} \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \xi \\ \xi & 1 \end{pmatrix} \right], \quad (2.9)$$

where $\phi_{1,1} = 0, 0.7$, $\phi_{2,2} = 0, 0.6$ and $\xi = 0, 0.5, 0.9$ see, for example, [12, pages 417-421]. In (2.8), $\boldsymbol{\beta} = (2, -0.5, 1)^T$, $\gamma_i \sim \mathcal{N}(0, 1)$ is a scalar random intercept, and $U_{ij} \sim \mathcal{N}(0, 1)$. Therefore, the parameter vector is $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, D, \sigma_e^2)^T$ with true value $\boldsymbol{\theta}_0 = (2, -0.5, 1, 1, 1)^T$. We take $n = 20, 40, 80$, $m_i = 5, 50$, and the weights $\omega_{ij} = 1$ and $\omega_{ij} = (1 - h_{i(j)})^{\frac{1}{2}}$ in (2.4). As previously mentioned, $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$.

To evaluate the robustness of each estimation method, we simulate $\mathbf{X}_{ij}$ with (2.9) and Y_{ij} with (2.8) where $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. Then, when we estimate $\boldsymbol{\theta}$, we replace in (2.3) $X_{ij}^{(2)}$ by the corrupted version $X_{ij}^{(2)*}$ defined by

$$X_{ij}^{(2)*} = X_{ij}^{(2)} + \mu B_{ij} K_{ij},$$

where B_{ij} is a Bernoulli random variable (RV) with $\mathbb{P}(B_{ij} = 1) = \delta$ where $\delta = 0, 0.01, \dots, 0.05$ and K_{ij} follows a Student's t -distribution with 3 degrees of freedom. The RVs B_{ij}, K_{ij} for $i = 1, \dots, n$ and $j = 1, \dots, m_i$ are mutually independent, and $\mu = 1, 2$ is the magnitude of AO. To compare the estimators, we calculate by Monte Carlo simulations for each component $\hat{\theta}_k$, $k = 1, \dots, 5$ of the parameter estimate $\hat{\boldsymbol{\theta}}$, the SM $\hat{\mu}(\hat{\theta}_k)$, SMSE $\hat{\sigma}^2(\hat{\theta}_k)$, and SCP with 95% of confidence $\hat{v}(\hat{\theta}_k)$. Let $M = 1000$ be the number of replications and $\hat{\theta}_{k,m}$ be the estimate of θ_k obtained in the m th experiment for $m = 1, \dots, M$. We have

$$\hat{\mu}(\hat{\theta}_k) = \frac{1}{M} \sum_{m=1}^M \hat{\theta}_{k,m}, \quad \hat{\sigma}^2(\hat{\theta}_k) = \frac{1}{M} \sum_{m=1}^M (\theta_k - \hat{\theta}_{k,m})^2, \quad \hat{v}(\hat{\theta}_k) = \frac{1}{M} \sum_{m=1}^M w_{k,m},$$

where $w_{k,m} = 1$ if $\theta_k \in [\hat{\theta}_{k,m} - 1.96\hat{\sigma}(\hat{\theta}_k), \hat{\theta}_{k,m} + 1.96\hat{\sigma}(\hat{\theta}_k)]$ and 0 otherwise. Asymptotically, the SCP is equal to 0.95.

In our simulations, a slight improvement of the performance when using $\omega_{ij} = (1 - h_{i(j)})^{\frac{1}{2}}$ instead of $\omega_{ij} = 1$ in (2.4) was observed only for the smallest sample size $(n, m_i) = (20, 5)$ and $\xi = 0.9$, while in all other cases both log-likelihood functions led to similar estimates. This shows that in the scenarios considered in our empirical study, the

Huber and Tukey loss functions are sufficient to handle some potential outlier effects on the SM, SMSE and SCP. Based on this empirical evidence, the results presented here are related to $n = 40$, $m_i = 5, 50$ and $\omega_{ij} = 1$.

Figures 2.1 to 2.3 plot for $\xi = 0, 0.5, 0.9$, respectively, the SM and SCP of each parameter estimate as a function of the percentage of contamination δ using the three estimation methods when $m_i = 5$, $\mu = 1$ and $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$ in (2.9), which implies that $(X_{ij}^{(1)}, X_{ij}^{(2)})^T$ is not time-correlated. Table 2.1 display the SM, SMSE and SCP of each parameter estimate using the three estimation methods when $\xi = 0, 0.5$, $\delta = 0, 0.02$, $m_i = 5$, $\mu = 2$ and $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$ in (2.9). Tables 2.2 and 2.3 present the same results when $\xi = 0, 0.5$, $\delta = 0, 0.002$, $m_i = 50$, $\mu = 2$ and $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $(\phi_{1,1}, \phi_{2,2}) = (0.7, 0.6)$, respectively.

In Figure 2.1, $\xi = 0$ and therefore $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ are uncorrelated. $\hat{\mu}(\hat{\beta}_0)$ and $\hat{\mu}(\hat{\beta}_1)$ fluctuate randomly around the true values. $\hat{\mu}(\hat{\beta}_2)$ decreases and $\hat{\mu}(\hat{\sigma}_e^2)$ increases as δ increases, and $\hat{D}$ is slightly underestimated. About SCP, the three estimation methods provide a good estimate of β_0 , LS and Huber estimates of β_1 are always good and Tukey estimate of β_1 is acceptable. Concerning β_2 , LS estimate is acceptable only if there are no outliers, Huber estimate is acceptable when $\delta \leq 0.02$ and Tukey estimate is acceptable when $\delta \leq 0.03$. Regarding σ_e^2 , LS estimate is acceptable for $\delta \leq 0.04$, Huber estimate is always good and Tukey estimate is always acceptable. With respect to D , LS and Huber estimates are always acceptable and Tukey estimate is always good. Therefore, the three methods provide good estimates of β_0 and β_1 and Tukey estimate is the most resilient to outliers.

In Figure 2.2, $\xi = 0.5$ and therefore the correlation between $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ is moderate. Again $\hat{\mu}(\hat{\beta}_0)$ fluctuates randomly around the true value, $\hat{\mu}(\hat{\sigma}_e^2)$ slowly increases and D is slightly underestimated. However, β_1 is overestimated and β_2 is underestimated and the bias increases as δ increases. About SCP, the three estimation methods provide a good estimate of β_0 , except for LS when $\delta = 0$ and for Tukey when $\delta = 0.05$. About β_1 , Huber is always good, Tukey is mainly good except for $\delta = 0.01, 0.03$, and LS is acceptable if $\delta \leq 0.03$. Concerning β_2 , LS is acceptable only if $\delta \leq 0.01$, Huber if $\delta \leq 0.02$ and Tukey if $\delta \leq 0.03$. Regarding σ_e^2 , Huber is always good, Tukey is mainly acceptable, LS is acceptable if $\delta \leq 0.03$. About D , Tukey estimates are always good, Huber and LS are mainly acceptable.

In Figure 2.3, $\xi = 0.9$ and then the correlation between $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ is strong. Concerning the SM, β_0 , σ_e^2 and D display similar behaviors as previously. Again, β_1 is overestimated and β_2 is underestimated and the bias increases severely as δ increases. Regarding SCP, the three estimation methods provide a good estimate of β_0 , except for LS when $\delta = 0, 0.03$ and for Tukey when $\delta = 0.04$. Concerning β_1 , LS estimate is acceptable only if there are no outliers, Huber estimate is acceptable when $\delta \leq 0.02$ and Tukey estimate is acceptable when $\delta \leq 0.03$. For β_2 , LS estimate is acceptable only if there

are no outliers, Huber and Tukey estimates are acceptable when $\delta \leq 0.02$. Regarding σ_e^2 , Huber is always good, LS is mainly good, and Tukey is always acceptable. About D , Tukey is always good, Huber is mainly good, and LS is mainly acceptable.

These simulations show that the use of one of the robust methods is strongly recommended when the presence of outliers is suspected in the data. When the covariates present moderate, low or no correlation and the percentage of outliers is less than 2%, Huber estimates are the best choice. If the covariates are strongly correlated or the percentage of outliers is more than 2%, then Tukey estimates are the best choice. Note that if the contamination level of a covariate is 4% or higher, then this covariate should not be included in the model since even Tukey estimation method does not provide an acceptable SCP. Also moderate and strong correlations should be avoid when possible.

Table 2.1 reports the SM, SMSE and SCP related to the simulations shown in Figures 2.1 and 2.2 when $\delta = 0$ (uncontaminated data). This table also displays the performance of the estimators with outlier's magnitude $\mu = 2$ and $\delta = 0.02$. These sample quantities corroborate the previous analysis of the performance of the estimators based on the plots. Now, the increase of the magnitude μ from 1 to 2 affects the preciseness of the estimators. For example, when $\xi = 0$, $\delta = 0.02$ and $\mu = 1$, the estimates of the $\hat{\sigma}^2(\hat{\beta}_2)$ were 0.0127, 0.0094 and 0.0087 for LS, Huber and Tuckey methods, respectively. Thus, comparing these values to the ones in Table 2.1 for $\xi = 0$, $\delta = 0.02$ and $\mu = 2$, we see that $\hat{\sigma}^2(\hat{\beta}_2)$ is almost multiplied by 4 for the LS method. In the same way, the corresponding SCP reduces substantially from 89.6% to 59.1%. Similar conclusions are drawn for the other cases. In general, the Tukey loss-function displayed more resistance against the increase of the outlier's magnitude.

Table 2.2 reports the simulation results when the size of the repeated measurements is $m_i = 50$. For a comparison purpose with the previous scenario, the expected number of outliers was fixed to 4 by choosing $\delta = 0.002$. Similar conclusions to the previous case can be derived. The SM, SMSE and SCP are strongly affected by the outliers, especially when $\xi = 0.5$ and $\delta = 0.002$. The MSE estimates decrease as m_i increases, which corroborates the asymptotic result that the estimators are consistent.

Table 2.3 displays the empirical quantities when $(\phi_{1,1}, \phi_{2,2}) = (0.7, 0.6)$ in (2.9), which implies that $(X_{ij}^{(1)}, X_{ij}^{(2)})^T$ is also time-correlated. Although the SM and SMSE behave similarly to the case in Table 2.2, the SCP present for the regression coefficients β_1 and β_2 much smaller relative frequency than the previous cases and, again, the Tukey method provides the most resistant estimator. The degradation of the SCP is more important when $\xi = 0.5$, i.e. when $X_{ij}^{(1)}$ and $X_{ij}^{(2)}$ are correlated.

This empirical study shows that the multivariate time series structure of the covariates, as well as the presence of atypical observations in the data, are phenomena that can not be ignored in the model strategy, otherwise, the conclusions can be totally erroneous and lead to severe consequences in terms of statistical inference.

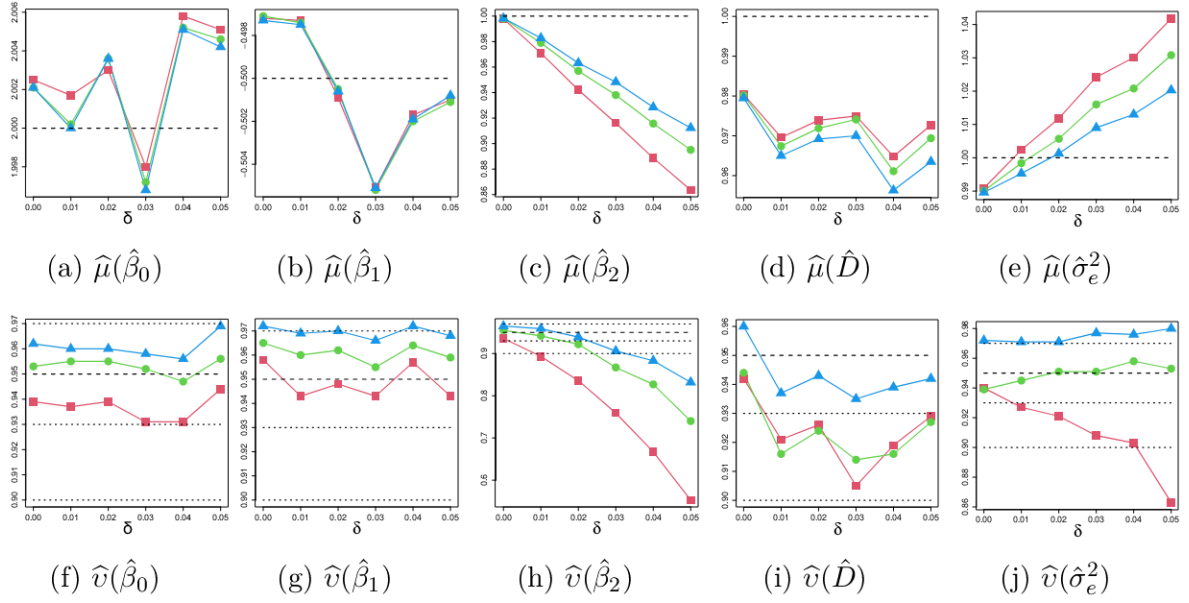


Figure 2.1: SM and SCP with 95% of confidence when $\xi = 0$, $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 1$. LS estimates (red square), Huber estimates with $\kappa_2 = 1.345$ (green circle) and Tukey estimates with $\kappa_3 = 4.685$ (blue triangle).

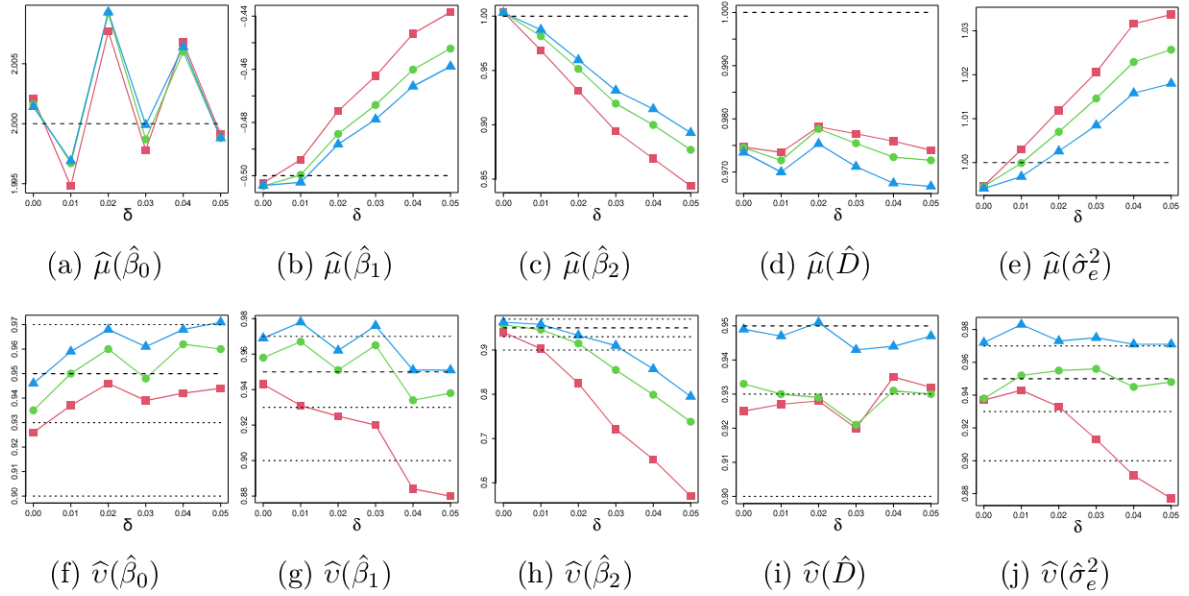


Figure 2.2: SM and SCP with 95% of confidence when $\xi = 0.5$, $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 1$. LS estimates (red square), Huber estimates with $\kappa_2 = 1.345$ (green circle) and Tukey estimates with $\kappa_3 = 4.685$ (blue triangle).

Table 2.1: SM, SMSE and SCP with 95% of confidence when $\xi = 0, 0.5$, $\delta = 0, 0.02$, $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 2$. LMM estimated by LS, Huber (H) and Tukey (T) with $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$, respectively.

Parameters	SM			SMSE			SCP		
$\xi = 0, \delta = 0$	LS	H	T	LS	H	T	LS	H	T
β_0	2.002	2.002	2.002	0.0289	0.0306	0.0307	0.939	0.953	0.962
β_1	-0.497	-0.497	-0.497	0.0057	0.0061	0.0061	0.958	0.965	0.972
β_2	0.998	0.998	0.998	0.0067	0.0070	0.0070	0.936	0.955	0.965
D	0.991	0.990	0.990	0.0031	0.0034	0.0035	0.940	0.939	0.972
σ_e^2	0.980	0.980	0.980	0.0188	0.0193	0.0193	0.942	0.944	0.960
$\xi = 0, \delta = 0.02$	LS	H	T	LS	H	T	LS	H	T
β_0	2.006	2.005	2.004	0.0302	0.0319	0.0324	0.947	0.963	0.972
β_1	-0.499	-0.499	-0.500	0.0070	0.0069	0.0069	0.943	0.965	0.971
β_2	0.860	0.920	0.952	0.0467	0.0186	0.0108	0.591	0.815	0.923
D	1.059	1.036	1.016	0.0113	0.0064	0.0043	0.788	0.932	0.983
σ_e^2	0.974	0.968	0.957	0.0192	0.0197	0.0210	0.927	0.927	0.933
$\xi = 0.5, \delta = 0$	LS	H	T	LS	H	T	LS	H	T
β_0	2.002	2.002	2.001	0.0328	0.0346	0.0347	0.926	0.935	0.946
β_1	-0.503	-0.504	-0.504	0.0084	0.0088	0.0088	0.943	0.958	0.969
β_2	1.004	1.003	1.003	0.0086	0.0092	0.0092	0.940	0.957	0.963
D	0.995	0.995	0.994	0.0032	0.0035	0.0036	0.937	0.938	0.972
σ_e^2	0.975	0.975	0.974	0.0195	0.0202	0.0205	0.925	0.933	0.949
$\xi = 0.5, \delta = 0.02$	LS	H	T	LS	H	T	LS	H	T
β_0	1.999	2.000	1.999	0.0291	0.0305	0.0313	0.942	0.957	0.960
β_1	-0.419	-0.449	-0.468	0.0229	0.0146	0.0113	0.802	0.902	0.951
β_2	0.830	0.891	0.930	0.0642	0.0320	0.0197	0.567	0.772	0.892
D	1.049	1.032	1.014	0.0090	0.0061	0.0045	0.819	0.944	0.981
σ_e^2	0.977	0.970	0.957	0.0192	0.0200	0.0214	0.933	0.932	0.948

2.4 Real data analysis

In this section, we apply the model and estimation methods discussed before in order to quantify the statistical association between FEV_1 , in both asthmatic and non-asthmatic children and teenagers, aged 7 to 18 years, with PM_{10} , $PM_{2.5}$ and sulfur dioxide (SO_2) pollutants, Hmd, temperature in degrees Celsius (Tnp), IgE and passive smoking (PS). The pollutants and the weather variables were measured at the air quality automatic monitoring network (AQAMN) of the great Vitória region (GVR), which is a densely populated region with approximately 1,900,000 inhabitants in an area of 2319 km² located on the east coast in the ES, Brazil (latitude 20°19S, longitude 40°20W) and has a humid tropical climate, with average temperatures ranging from 24°C to 30°C. The GVR is a port region and industrialized area. The AQAMN of the GVR consists of eight monitoring stations: two in Serra (Laranjeiras and Carapina), three in Vitória city (Jardim Camburi, Praia do Suá and Vix-Centro), two in Vila Velha (VV-Centro and Ibes), and one in

Table 2.2: SM, SMSE and SCP with 95% of confidence when $\xi = 0, 0.5$, $\delta = 0, 0.002$, $(n, m_i) = (40, 50)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 2$. LMM estimated by LS, Huber (H) and Tukey (T) with $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$, respectively.

Parameters	SM			SMSE			SCP		
$\xi = 0, \delta = 0$	LS	H	T	LS	H	T	LS	H	T
β_0	1.999	1.999	2.000	0.0256	0.0270	0.0269	0.934	0.957	0.963
β_1	-0.501	-0.502	-0.502	0.0005	0.0006	0.0006	0.948	0.955	0.966
β_2	1.000	0.999	0.999	0.0005	0.0005	0.0005	0.937	0.956	0.969
D	1.000	1.000	1.000	0.0003	0.0003	0.0003	0.950	0.963	0.983
σ_e^2	0.976	0.977	0.977	0.0130	0.0132	0.0132	0.927	0.928	0.950
$\xi = 0, \delta = 0.002$	LS	H	T	LS	H	T	LS	H	T
β_0	2.000	2.000	2.000	0.0255	0.0269	0.0270	0.948	0.970	0.974
β_1	-0.499	-0.500	-0.500	0.0005	0.0005	0.0005	0.957	0.971	0.977
β_2	0.980	0.994	0.997	0.0027	0.0006	0.0006	0.819	0.966	0.974
D	1.009	1.004	1.001	0.0007	0.0003	0.0003	0.870	0.971	0.981
σ_e^2	0.983	0.982	0.981	0.0123	0.0128	0.0128	0.932	0.931	0.949
$\xi = 0.5, \delta = 0$	LS	H	T	LS	H	T	LS	H	T
β_0	1.993	1.993	1.993	0.0250	0.0262	0.0262	0.947	0.956	0.966
β_1	-0.501	-0.500	-0.500	0.0008	0.0008	0.0008	0.936	0.955	0.966
β_2	1.001	1.001	1.001	0.0007	0.0008	0.0008	0.948	0.961	0.971
D	1.000	1.000	1.000	0.0003	0.0003	0.0003	0.951	0.958	0.982
σ_e^2	0.979	0.979	0.979	0.0128	0.0131	0.0131	0.929	0.927	0.943
$\xi = 0.5, \delta = 0.002$	LS	H	T	LS	H	T	LS	H	T
β_0	2.002	2.003	2.003	0.0246	0.0254	0.0254	0.935	0.957	0.966
β_1	-0.488	-0.497	-0.499	0.0014	0.0008	0.0007	0.899	0.962	0.972
β_2	0.975	0.993	0.997	0.0035	0.0009	0.0008	0.800	0.943	0.966
D	1.009	1.003	1.001	0.0006	0.0003	0.0003	0.880	0.959	0.987
σ_e^2	0.988	0.987	0.986	0.0138	0.0142	0.0143	0.922	0.925	0.945

Cariacica. The response variable FEV_1 was measured monthly, from July to December 2017, at the Public Health Center of Praia do Suá, Vitória, ES, Brazil. All environmental covariates refer to the daily averages of the day before the spirometry test. Other noise covariates, such as body mass index (BMI), age and gender were dropped from the final model since, in the first steps of the model strategy, these variables were not statistically significant.

The longitudinal study involved 82 children, and for each child, 6 measurements of FEV_1 were obtained. The age range of the study is justified by the fact that children of these ages are more susceptible to the effects of air pollution and have a higher prevalence of respiratory diseases. This data set is presented in [102].

Before analysing the statistical association between pollution and FEV_1 , we verify if the high peaks of concentrations present in the data have some impact on the correlation matrix of the environmental covariates. Let $\mathbf{M}_i$ be the (6×5) matrix of the 6 measurements of the 5 environmental covariates, Hmd, Tmp, SO_2 , PM_{10} and $PM_{2.5}$, cor-

Table 2.3: SM, SMSE and SCP with 95% of confidence when $\xi = 0, 0.5$, $\delta = 0, 0.002$, $(n, m_i) = (40, 50)$, $(\phi_{1,1}, \phi_{2,2}) = (0.7, 0.6)$, $\mu = 2$. LMM estimated by LS, Huber (H) and Tukey (T) with $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$, respectively.

Parameters	SM			SMSE			SCP		
$\xi = 0, \delta = 0$	LS	H	T	LS	H	T	LS	H	T
β_0	1.999	1.999	2.000	0.0252	0.0263	0.0261	0.939	0.962	0.969
β_1	-0.501	-0.501	-0.501	0.0004	0.0004	0.0004	0.907	0.928	0.938
β_2	0.999	0.999	0.999	0.0004	0.0004	0.0004	0.916	0.938	0.947
D	1.000	0.999	0.999	0.0002	0.0003	0.0003	0.965	0.962	0.987
σ_e^2	0.976	0.975	0.975	0.0133	0.0137	0.0137	0.924	0.923	0.935
$\xi = 0, \delta = 0.002$	LS	H	T	LS	H	T	LS	H	T
β_0	2.005	2.004	2.005	0.0262	0.0268	0.0269	0.938	0.961	0.965
β_1	-0.500	-0.500	-0.500	0.0004	0.0004	0.0004	0.919	0.939	0.945
β_2	0.989	0.997	0.999	0.0013	0.0005	0.0004	0.811	0.916	0.942
D	1.010	1.004	1.001	0.0008	0.0003	0.0003	0.876	0.968	0.985
σ_e^2	0.980	0.979	0.978	0.0127	0.0130	0.0131	0.938	0.934	0.952
$\xi = 0.5, \delta = 0$	LS	H	T	LS	H	T	LS	H	T
β_0	1.999	1.999	1.999	0.0252	0.0263	0.0261	0.939	0.962	0.969
β_1	-0.501	-0.501	-0.501	0.0006	0.0006	0.0006	0.899	0.920	0.927
β_2	1.000	1.000	1.000	0.0005	0.0006	0.0006	0.908	0.932	0.939
D	1.000	0.999	0.999	0.0002	0.0003	0.0003	0.965	0.962	0.986
σ_e^2	0.976	0.975	0.975	0.0133	0.0137	0.0137	0.924	0.923	0.935
$\xi = 0.5, \delta = 0.002$	LS	H	T	LS	H	T	LS	H	T
β_0	2.005	2.004	2.005	0.0262	0.0268	0.0269	0.938	0.961	0.965
β_1	-0.493	-0.498	-0.499	0.0009	0.0006	0.0006	0.871	0.935	0.940
β_2	0.985	0.996	0.998	0.0019	0.0007	0.0006	0.776	0.905	0.924
D	1.010	1.004	1.001	0.0007	0.0003	0.0003	0.877	0.968	0.985
σ_e^2	0.980	0.979	0.978	0.0127	0.0131	0.0131	0.939	0.934	0.952

responding to the i th child, and $\mathbf{M} = [\mathbf{M}_1^T, \mathbf{M}_2^T, \dots, \mathbf{M}_{g_2}^T]^T$ be the (492×5) matrix of all environmental covariates. Table 2.4 displays respectively below and above the diagonal, the classical and robust correlations of the columns of $\mathbf{M}$. The robust correlation is based on S_n , a robust variance estimator [99, 79, 103]. As can be seen, the correlations are very different, and in general, the robust ones are numerically larger. This may indicate that there are high levels of concentration that are causing the same effect as AO. To clarify this issue, the original data were modified by replacing the 10% smallest and greatest peaks of the pollutants by their respective means. The sample correlations of the modified data are shown in Table 2.5, and differently from the results in Table 2.4, both sample correlation functions display close values. Thus, this empirical result illustrates the effect of the large peaks of the concentrations on the sample correlations and the occurrence of the multicollinearity phenomenon.

To minimize the effects of outliers, we use a RPCA method before the regression analysis in the LMM in order to quantify the statistical association between FEV_1 and

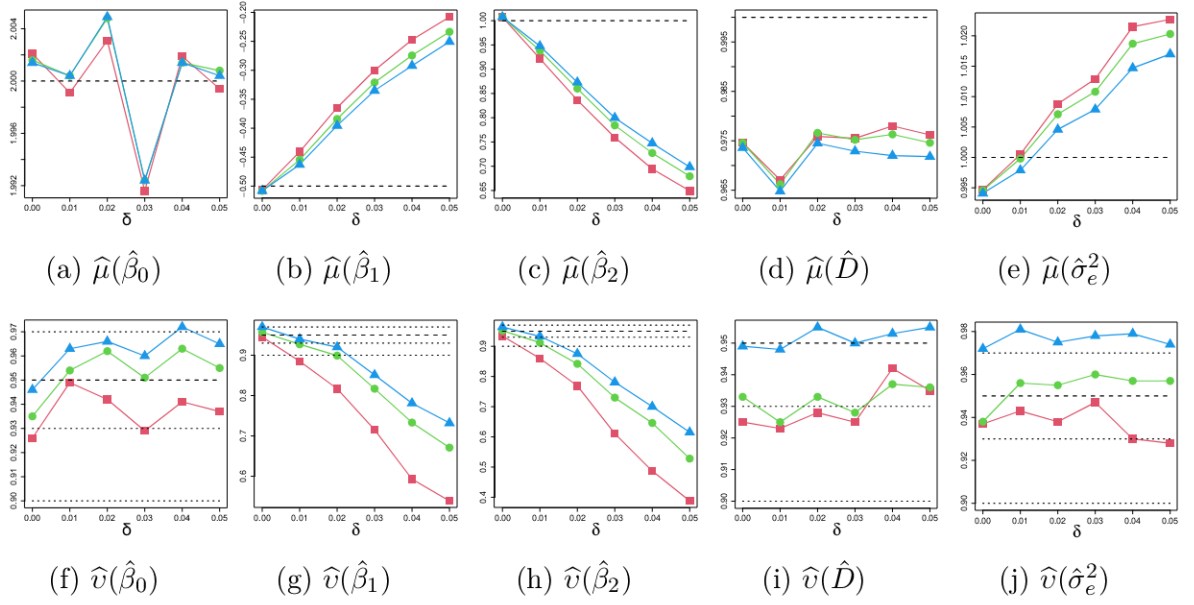


Figure 2.3: SM and SCP with 95% of confidence when $\xi = 0.9$, $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 1$. LS estimates (red square), Huber estimates with $\kappa_2 = 1.345$ (green circle) and Tukey estimates with $\kappa_3 = 4.685$ (blue triangle).

Table 2.4: Correlations between environmental covariates.

		Robust correlation				
		Hmd	Tmp	SO ₂	PM ₁₀	PM _{2.5}
Classical correlation	Hmd		-0.572	-0.324	-0.044	0.273
	Tmp	-0.469		0.378	0.231	0.594
	SO ₂	0.033	0.433		0.601	0.101
	PM ₁₀	-0.245	0.149	0.417		0.526
	PM _{2.5}	-0.074	0.389	0.264	0.648	

Table 2.5: Correlation between environmental covariates after replacing the 10% smallest and greatest values by the average of each covariate.

		Robust correlation				
		Hmd	Tmp	SO ₂	PM ₁₀	PM _{2.5}
Classical correlation	Hmd		-0.445	0.094	0.451	0.231
	Tmp	-0.286		0.045	-0.153	0.447
	SO ₂	0.131	0.053		0.487	-0.054
	PM ₁₀	0.398	0.090	0.448		0.218
	PM _{2.5}	0.237	0.385	-0.048	0.268	

the covariates. Table 2.6 displays the loadings $\mathbf{L}^{(k)}$ for $k = 1, 2, 3$ of the three first PC of the classical PCA [119, page 304], and RPCA [52] of $\mathbf{M}$. The CPV of the classical PCA is smaller than the RPCA, suggesting a number of regressors larger than three in the model to achieve a CPV larger than 90%. In the case of the classical PCA, the first PC is essentially an average of all variables except Hmd, whereas the most important

contributions to the second and third PC are from Hmd and SO₂, respectively. The situation is very different in the RPCA where the loadings values differ from classical PCA and each PC is mainly explained by only one variable: the main contributions to the first, second and third PC are from Hmd, PM₁₀ and PM_{2.5}, respectively. Note that the variables PM₁₀ and SO₂ are highly correlated (see Table 2.4).

Table 2.6: Loadings of the three first PC of the classical PCA and RPCA, and CPV.

	PCA			RPCA		
	$\mathbf{L}^{(1)}$	$\mathbf{L}^{(2)}$	$\mathbf{L}^{(3)}$	$\mathbf{L}^{(1)}$	$\mathbf{L}^{(2)}$	$\mathbf{L}^{(3)}$
Hmd	-0.436	-0.795	0.297	-0.972	0.011	-0.048
Tmp	0.704	0.481	0.373	0.184	-0.116	0.358
SO ₂	0.639	-0.279	0.635	0.134	0.194	-0.469
PM ₁₀	0.767	-0.344	-0.393	0.021	0.945	-0.059
PM _{2.5}	0.765	-0.317	-0.311	-0.060	0.235	0.803
CPV	0.454	0.686	0.862	0.826	0.913	0.967

We adjust the following LMM to our data

$$Y_{ij} = \beta_0 + \beta_1 P_{ij}^{(1)} + \beta_2 P_{ij}^{(2)} + \beta_3 P_{ij}^{(3)} + \beta_4 \text{IgE}_i + \beta_5 \text{PS}_i + \gamma_i + U_{ij}, \quad (2.10)$$

where, for $i = 1, \dots, 82$ and $j = 1, \dots, 6$, Y_{ij} is the FEV₁ measurement of i th child at j th trial, $P_{ij}^{(k)}$ is the j th element of $\mathbf{P}_i^{(k)} = \mathbf{M}_i \mathbf{L}^{(k)}$ for $k = 1, 2, 3$, IgE_i is the allergic condition of i th child, PS_i is a binary variable indicating if the i th child is a passive smoker or not, $\gamma_i \sim \mathcal{N}(0, D)$ is a scalar random intercept and U_{ij} is the measurement error of the model. We assume that $\mathbf{U}_i = (U_{i,1}, \dots, U_{i,6})^T \sim \mathcal{N}(\mathbf{0}, \sigma_e^2 \mathbf{\Omega})$ where the (j, k) th entry of $\mathbf{\Omega}$ is $\Omega_{jk} = \nu^{|j-k|} / (1 - \nu^2)$ with $|\nu| < 1$.

Let $\boldsymbol{\theta} = (\beta_0, \dots, \beta_5, D, \sigma_e^2, \nu)^T$. Since $n = 82$ is large enough, the empirical results in Section 2.3 suggest to take $\omega_{ij} = 1$ in (2.4). Six different models can be fitted according to the choice of ρ in (2.4) and whether the loadings $\mathbf{L}^{(k)}$ of the PC correspond to the classical PCA or the RPCA. For instance, LS-PCA means that we use $\rho = \rho_1$ and the loadings of the classical PCA, and Tukey-RPCA means that we use $\rho = \rho_3$ and the loadings of the RPCA. Table 2.7 displays the estimates of $\boldsymbol{\theta}$, their SE and p -values in the fitted models LS-PCA, LS-RPCA, Huber-RPCA and Tukey-RPCA. For the four models, the intercept β_0 and the parameter ν are significant and the parameters β_3 and β_4 are not significant with a p -value larger than 18%. The SE of the estimates of β_1 and β_2 obtained by the LS-PCA are much larger compared to those of the robust approaches, which display very close estimates and standard deviation values. In general, the Huber-RPCA and Tukey-RPCA fitted models are quite similar by presenting very close parameter estimates and p -values. The LS-RPCA also shows similar estimates to these two robust approaches, except the estimate of β_1 which is not significant with a p -value larger than 10%. All methods find that the PS (parameter β_5) give a significant contribution to the response

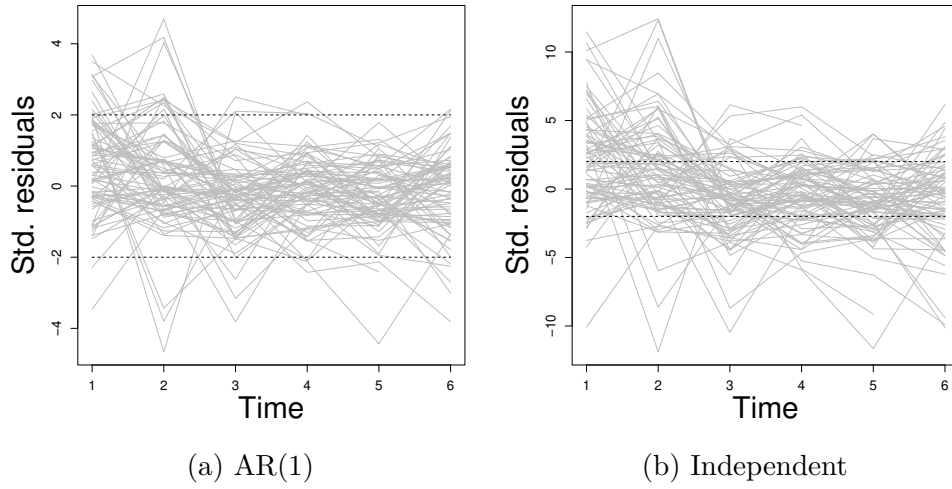


Figure 2.4: Estimated standardized residuals of Huber-RPCA model with AR(1) and independent errors.

variable with a p -value smaller than 7, 5%. For the four models, the estimates of variances σ_e^2 and D are significant [130].

Table 2.7: Estimated parameters, SE and p -values in the fitted LMM LS-PCA, LS-RPCA, Huber-RPCA and Tukey-RPCA.

	LS-PCA		LS-RPCA		Huber-RPCA		Tukey-RPCA	
	coeff. (s.e.)	p-value	coeff. (s.e.)	p-value	coeff. (s.e.)	p-value	coeff. (s.e.)	p-value
β_0	103.62 (5.00)	<0.001	93.74 (7.78)	<0.001	92.59 (8.21)	<0.001	92.84 (8.18)	<0.001
β_1	-1.34 (0.30)	<0.001	-0.11 (0.07)	0.115	-0.13 (0.07)	0.058	-0.14 (0.07)	0.052
β_2	-0.54 (0.41)	0.188	-0.22 (0.08)	0.003	-0.21 (0.08)	0.008	-0.21 (0.08)	0.008
β_3	-0.96 (0.48)	0.046	0.20 (0.15)	0.180	0.18 (0.15)	0.225	0.17 (0.14)	0.222
β_4	-0.85 (0.86)	0.324	-0.66 (0.86)	0.440	-0.92 (0.94)	0.328	-1.11 (0.96)	0.247
β_5	-6.89 (3.07)	0.025	-6.49 (3.06)	0.034	-6.61 (3.23)	0.041	-5.86 (3.29)	0.075
D	10.69 (1.21)	<0.001	9.86 (1.44)	<0.001	10.17 (1.37)	<0.001	9.77 (1.38)	<0.001
σ_e^2	9.89 (0.37)	<0.001	10.16 (0.39)	<0.001	9.15 (0.45)	<0.001	8.49 (0.51)	<0.001
ν	0.40 (0.08)	<0.001	0.51 (0.08)	<0.001	0.48 (0.09)	<0.001	0.49 (0.10)	<0.001

In Figure 2.4a, we plot for each $i = 1, \dots, 82$, the estimated standardized residuals $\hat{R}_{ij}$ for $j = 1, \dots, 6$ obtained by replacing Σ_i and β by their estimates in (2.3). For all estimation methods, these residuals behave similarly and we display only the residuals obtained with the Huber-RPCA model. To see the influence of the AR(1) parameter ν in model (2.10), we have also fitted a Huber-RPCA model with independent errors $U_{ij} \sim N(0, \sigma_e^2)$. The corresponding estimated standardized residuals are plotted in Figure 2.4b. There is a clear difference between Figures 2.4a and 2.4b; the variability on the right side is much smaller than on the left, which is a crucial point in favor of the AR(1) structure, see [120, pages 125-127].

2.5 Conclusion

This paper proposes to combine RPCA and robust estimation in a LMM in the context where the covariates present outliers and are time-correlated. Monte Carlo simulations show that Huber and Tukey estimation methods are very competitive with the classical LS when there are no outliers in the data. On the other hand, the performances of the robust methods are much better than the LS in the presence of outliers. The real data set analysis reveals that the respiratory capability of children is mainly impacted by pollutants, Hmd and PS. The high levels of the pollutants create a similar effect as AO on the estimates. To mitigate this phenomenon and the multicollinearity effects on the regression estimates, we consider RPCA and RLMM.

Chapter 3

Alternative methods to quantile regression for panel data

Abstract

Linear fixed effect models are a general way of fitting panel or longitudinal data with a distinct intercept for each unit. Based on expectile and M -quantile approaches, we propose alternative regression estimation methods to estimate the parameter of Linear fixed effect models. The estimation functions are penalized by the least absolute shrinkage and selection operator (LASSO) to reduce the dimensionality of the data. Some asymptotic properties of the estimators are established and finite sample size investigations are conducted to verify the empirical performances of the estimation methods. Furthermore, the computational implementations of the procedures are discussed and a real economic panel data from Organisation for Economic Cooperation and Development (OECD) is analysed to show the usefulness of the methods in a practical problem.

Keywords: Quantile regression, Expectile, M -estimation, repeated measures, LASSO.

3.1 Introduction

Panel data, longitudinal data and repeated measures are widespread in applied sciences such as econometrics, medicine, and engineering. They must be analyzed with models that consider the dual-source of variability between and within observational units. There are two main approaches to handling this kind of data: *random effects*, which require that explanatory variables and individual effects are independent, and *fixed effects*, which relaxes this assumption. The LMM might combine these two approaches, but usually the explanatory variables are treated as fixed effects, and individual intercepts are assumed

as random effects [72]. Random coefficients and fixed intercepts are allowed, but they are exceptions in applied research. The fixed effects specification is more robust than the random effects way. Still, analyses of the fixed effects panel data is more difficult due to the increasing number of parameters with the sample size [47, 137].

The least absolute deviation regression (LAD) estimates the median effect of regressors on the dependent variable rather than the mean effect as in the ordinary regression. The QR, which generalizes LAD to any percentile of interest is a favourable alternative method for at least three reasons. This returns a general relationship between predictors and response variable (it is not restricted to the median or mean effect), does not require Gaussian errors, and is less sensitive to heteroscedastic data [66] and [67, page 41].

The combination of the methods QR and LMM seems to be a natural and necessary strategy, since there are many longitudinal data whose variables are heteroscedastic and non-Gaussian. However, each branch has its prerequisites and assumptions; fortunately, QRFE and PQRFE bring together this link between QR and panel data [65]. A detailed overview is available in [81], but this field has made expressive progress in recent years [34, 77, 36, 113]. Under additional assumptions, it is possible to combine QR with random effects, but this is outside of the scope of this paper.

The estimation of quantiles can be done by the *check function*, which is synthetic, robust and easy to interpret [66]. However, each estimation process is not totally perfect and has weakness [4]. The check function shows numerical instability close to zero because it is a non-derivable point. It is subject to the *crossing quantile* phenomenon, in which the predicted response does not increase monotonously for a fixed value of the covariates. This fact violates the incremental principle of the cumulative distribution function (CDF) and, consequently, breaks the probability law [67, page 55]. The crossing quantile might get worse if there is a *curse of dimensionality*, which means a massive dimension of covariates [11]. Finally, extreme quantile estimates tend to display a progressive bias [124]. There are some alternative ways to correct these problems as expectile regression (ER), which is in fact a weighted least squared [86] and M -quantile regression (MQR), which generalizes the M -estimation proposed by [49] to quantile context [10]. Both alternative approaches are quantiles as it is proved by [59]. The ER used to be more efficient and more stable (their curves hardly cross each other), but more biased than QR [101]. MQR balances the trade-off between QR robustness and ER efficiency, see a deep discussion about these methods in [62]. This balance is tuned by a constant c as in the classical M -estimation methods proposed by [49], a deep discussion about c on quantile context is in [69, 62]. MQR is also more resilient to curse of dimensionality [69].

To the best of our knowledge, the papers that combine ER and random effects, i.e., approach to expectile regression with random effects (ERRE), are [123, 76]. [118, 4] merge MQR and random effects, i.e., these authors propose the M -quantile regression with random effects (MQRRE). Some studies follow the ideas of uniting QR and random

effects, that is a type of quantile regression with random effects (QRRE), which is called linear quantile mixed models (LQMM) by the authors [39, 40]. However, these works assume the random intercepts have no location contribution, are uncorrelated with the explanatory predictors, and occasionally follow a Gaussian distribution. The original proposition of QRFE is free of all these restrictions [65]. The QRFE allows individual intercepts to be different from zero because they are treated as fixed effects. Consequently, there is a weird side effect, which consists of an increasing number of parameters to be estimated. To avoid this elevated parametric set, QRFE takes use of LASSO to shrink their dimension [65]. The optimal value of LASSO's tuning parameter for this model is discussed by [73, 74].

In this paper, we propose new variants of quantile regressions for LMM with fixed effects by introducing alternative loss functions and penalizing the intercepts with LASSO. The expectile and M -quantile are the loss functions considered in the new proposed estimation methods, which are denoted as ERFE, MQRFE, PERFE and PMQRFE. The asymptotic properties of the estimators and their empirical evidences support the their use in real problems. The remainder of this paper is as follows. Section 3.2, the model and estimation methods are introduced and the asymptotic properties are discussed. Section 3.3 discusses the computational details and presents simulations. Section 3.4 is committed to a real data application of interest in the recent economics literature. Section 3.5 summarizes our conclusions.

3.2 Model and estimation methods

3.2.1 The model

Panel data analysis might be divided into two types of models: mixed models with random or fixed effects. The former is generally assumed to have zero expectation and it presents independence between intercepts and explanatory covariate parameters. Under these restrictions, random effects models may omit the location effect of intercepts and only include them in the variance structure. The later is more general and it does not impose those assumptions. Thus, it treats the individual effect as parameters to be estimate. However, the number of parameters grows along with the sample size, and the estimation becomes inefficient [47, 137].

We denote by $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{im_i})^T$ the vector of measurements taken at times $t_{i1}, \dots, t_{im_i}$ of the i th subject for $i = 1, \dots, n$. Following [72, 25], we assume that each measurement Y_{ij} of $\mathbf{Y}_i$, $j = 1, \dots, m_i$, follows the LMM

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \alpha_i + U_{ij}, \quad i = 1, \dots, n \quad j = 1, \dots, m_i, \quad (3.1)$$

where $\mathbf{X}_{ij}$ is a known $(1 \times d)$ design vector, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^T$ is a vector of d unknown but fixed parameters to be estimated, α_i is the individual intercept (random or fixed) of the model and follow an unknown probability distribution, and U_{ij} is a RV. Thus, model (3.1) could be define in a vector form as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{U},$$

where $\mathbf{Y} = (\mathbf{Y}_1^T, \dots, \mathbf{Y}_n^T)^T$ is a vector of length $N = \sum_{i=1}^n m_i$. $\mathbf{X}_i = [\mathbf{X}_{i1}^T, \dots, \mathbf{X}_{im_i}^T]^T$ is a $(m_i \times d)$ matrix and $\mathbf{X} = [\mathbf{X}_1^T, \dots, \mathbf{X}_n^T]^T$ is a $(N \times d)$ matrix. $\mathbf{Z} = [\mathbf{Z}_1, \dots, \mathbf{Z}_n]$ is a $(N \times n)$ incident matrix, i.e., $\mathbf{Z}_i$ is a vector of zeros and ones. If $i = 1$, then $\mathbf{Z}_1$ is one from 1 to m_1 and zero otherwise. If $i > 1$, then $\mathbf{Z}_i$ is one from $m_{i-1} + 1$ to m_i , and zero otherwise. When $m_i = m$ for all i , $\mathbf{Z} = \mathbf{I}_n \otimes \mathbf{1}_m$, where $\otimes$ denotes the Kronecker product, $\mathbf{I}_n$ is the $(n \times n)$ identity matrix and $\mathbf{1}_m$ is the m -vector of ones. $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^T$ is the n -vector of individual intercepts. $\mathbf{U} = (\mathbf{U}_1^T, \dots, \mathbf{U}_n^T)^T$, where $\mathbf{U}_i^T = (U_{i1}, \dots, U_{im_i})$ is the N -vector of errors. We assume that the RV's $\boldsymbol{\alpha}$'s are independent and identically distributed (IID), as well as the RV's $\mathbf{U}$'s. Here, we assume that (3.1) is a fixed effect model. In this context, the conditional quantile regression of $\mathbf{Y}$ is given by

$$Q_{Y_{ij}}(\tau | \mathbf{X}_{ij}) = \alpha_i(\tau) + \mathbf{X}_{ij}\boldsymbol{\beta}(\tau),$$

where $\tau \in (0, 1)$, $\boldsymbol{\beta}(\tau)$ is the quantile regressor, $\alpha_i(\tau)$ is quantile individual intercept, and $\alpha_i(\tau) + \mathbf{X}_{ij}\boldsymbol{\beta}(\tau)$ is the 100τ th percentile of Y_{ij} [65].

3.2.2 Estimation methods: quantile, expectile and M -quantile approaches

The join estimation of $\boldsymbol{\alpha}(\tau)$ and $\boldsymbol{\beta}(\tau)$ requires solve the following equation

$$\hat{\boldsymbol{\alpha}}(\tau), \hat{\boldsymbol{\beta}}(\tau) = \arg \min_{\boldsymbol{\alpha} \in \mathbb{R}^n, \boldsymbol{\beta} \in \mathbb{R}^d} \sum_{i=1}^n \sum_{j=1}^{m_i} \rho_\tau(Y_{ij} - \alpha_i - \mathbf{X}_{ij}\boldsymbol{\beta}), \quad (3.2)$$

where $\rho_\tau(u) = |\tau - 1_{u \leq 0}| \rho(u)$, $u \in \mathbb{R}$, and $\rho(\cdot)$ is a continuously differentiable convex loss function. Here, we discuss three options for $\rho(\cdot)$,

- i) The most common choice for $\rho(\cdot)$ is the absolute loss function $\rho_1(u) = |u|$, then $\rho_\tau(\cdot)$ is the check function [66],
- ii) An alternative for $\rho(\cdot)$ is the squared loss function $\rho_2(u) = u^2$, then $\rho_\tau(\cdot)$ is the asymmetric least squared function [86],
- iii) A worth choice for $\rho(\cdot)$ is the Huber loss function $\rho_3(u) = (c|u| - \frac{1}{2}c^2)1_{|u| > c} + \frac{1}{2}u^2 1_{|u| \leq c}$, where $c \in (0, \infty)$, then $\rho_\tau(\cdot)$ is the M -quantile function [10].

The $\rho_3(\cdot)$ is a weight between the two previous functions, and its $\rho_\tau(\cdot)$ has properties that balance them, i.e., it is more robust than the squared and more efficient than the absolute [62]. The solution of (3.2) is $\hat{\alpha}(\tau), \hat{\beta}(\tau)$ and correspond to QRFE if ρ_1 [65], ERFE if ρ_2 , and MQRFE if ρ_3 , the last two are novelties.

Remark 5. The QR [66], ER [86] and MQR [10] are special cases of QRFE, ERFE and MQRFE where $\alpha = \mathbf{0}$, respectively.

As discussed previously, a weakness of fixed models is the large number of parameters to be estimated and the accumulation of estimation errors [137]. To fix this, [65] and used the ideas of [114] to penalized the intercepts by LASSO. Then, the estimation procedure is given by

$$\hat{\alpha}(\tau, \lambda), \hat{\beta}(\tau, \lambda) = \arg \min_{\alpha \in \mathbb{R}^n, \beta \in \mathbb{R}^d} \sum_{i=1}^n \sum_{j=1}^{m_i} \rho_\tau(Y_{ij} - \alpha_i - \mathbf{X}_{ij}^T \beta) + \lambda \sum_{i=1}^n |\alpha_i|, \quad (3.3)$$

where λ is a tuning parameter, which used to be estimated by cross-validation (CV) or information criteria (IC). However, [65] recommends to choose λ as a standard deviation ratio $\sigma(\mathbf{U})/\sigma(\alpha)$, further [73, 74] reinforced this claim. The original proposal to (3.3) is with $\rho_1(\cdot)$ only, but we extend it to three different loss functions. Thus, we have the following estimation methods

- i) If $\rho(\cdot) = \rho_1(\cdot)$, then the solution of (3.3) is the PQRFE proposed by [65],
- ii) If $\rho(\cdot) = \rho_2(\cdot)$, then the solution of (3.3) is the PERFE proposed by us,
- iii) If $\rho(\cdot) = \rho_3(\cdot)$, then the solution of (3.3) is the PMQRFE proposed by us.

3.2.3 Asymptotic results for expectile regression with fixed effects

The asymptotic results for QRFE and PQRFE are established by [65]. Here we concentrate on the case where $\rho_\tau(u) = g(u)\rho_2(u)$ and $g(u) = |\tau - 1_{u \leq 0}|$, i.e., the ERFE and PERFE cases. We restrict our attention to balanced design $m_i = m$ for all $i = 1, \dots, n$. Let us impose some regularity conditions:

A.1 Let $\varepsilon_{ij}(\tau) = \alpha_i(\tau) + \mathbf{X}_{ij}\beta(\tau) - Y_{ij}$ is IID,

A.2 The limit form of the following matrices are positive definite

$$\mathbf{V} = \lim_{\substack{m \rightarrow \infty \\ n \rightarrow \infty}} \frac{1}{m} \begin{bmatrix} \mathbf{Z}^T \mathbf{H} \mathbf{Z} & \mathbf{Z}^T \mathbf{H} \mathbf{X} / \sqrt{n} \\ \mathbf{X}^T \mathbf{H} \mathbf{Z} / \sqrt{n} & \mathbf{X}^T \mathbf{H} \mathbf{X} / n \end{bmatrix}, \quad \mathbf{\Gamma} = \lim_{\substack{m \rightarrow \infty \\ n \rightarrow \infty}} \frac{1}{m} \begin{bmatrix} \mathbf{Z}^T \mathbf{G} \mathbf{Z} & \mathbf{Z}^T \mathbf{G} \mathbf{X} / \sqrt{n} \\ \mathbf{X}^T \mathbf{G} \mathbf{Z} / \sqrt{n} & \mathbf{X}^T \mathbf{G} \mathbf{X} / n \end{bmatrix},$$

where $\mathbf{G} = \text{diag}(g(\varepsilon_{11}(\tau) - \mathbf{W}_{11}\delta), \dots, g(\varepsilon_{nm}(\tau) - \mathbf{W}_{nm}\delta))$, $\mathbf{H} = \text{diag}(g_{11}^2 \varepsilon_{11}(\tau)^2, \dots, g_{nm}^2 \varepsilon_{nm}(\tau)^2)$, $g_{ij} = g(\varepsilon_{ij}(\tau) - \mathbf{W}_{ij}\delta)$, $\mathbf{W} = [\mathbf{Z}/\sqrt{m}, \mathbf{X}/\sqrt{n}]$, $\delta = (\delta_0^T, \delta_1^T)^T$, δ_0^T is a m -vector, δ_1^T is a d -vector,

A.3 $\max_{i \geq 1, j \geq 1} \|\mathbf{X}_{ij}\| < M_0 < \infty$.

A.4 There exists a constant $c > 0$ such that $\left(\frac{n^c}{m}\right) \rightarrow 0$ as $n \rightarrow \infty$,

A.5 Let us consider $\rho_\tau(u) = g(u)\rho_2(u)$, $g(u) = |\tau - 1_{u \leq 0}|$, $\rho_2(u) = u^2$ and $u \in \mathbb{R}$,

A.6 The tuning parameter $\lambda_m/\sqrt{m} \rightarrow \lambda_0 \geq 0$ as $m \rightarrow \infty$.

Theorem 1. (ERFE asymptotic normality) Under the conditions A.1-A.5, $\sqrt{N}(\hat{\beta}(\tau) - \beta(\tau))$ is Gaussian random vector with zero mean and variance equal to the lower $(d \times d)$ block of $\Gamma^{-1}\mathbf{V}\Gamma^{-1}$.

Theorem 2. (PERFE consistency) Under assumptions A.1-A.6, then the component $\hat{\delta}_1$ has the same limiting distribution as the component δ_1 of limiting distribution

$$V_0(\delta) = \delta^T \Gamma \delta - 2\delta^T \mathbf{B} + \lambda_0 \delta^T \mathbf{s},$$

where $\mathbf{B}$ is a Gaussian vector with zero mean and matrix of covariance equal to $\mathbf{V}$, $\mathbf{s} = (\mathbf{s}_0^T, \mathbf{s}_1^T)^T$, $\mathbf{s}_0 = (\text{sgn}(\alpha_1), \dots, \text{sgn}(\alpha_n))^T$, $\mathbf{s}_1$ is a d -vector of zeros, and $\text{sgn}(u) = 1_{u>0} - 1_{u<0}$.

The proof of Theorem 1 is in the appendix. The Theorem 2 is a consequence of Theorem 1 and of [65, Theorem 2].

Theorem 1 establish the conditions for consistency and asymptotic normality of the estimator $\hat{\beta}(\tau)$ define in (3.2), and Theorem 2 the convergence of the estimator $\hat{\beta}(\tau, \lambda)$ define in (3.3), i.e., the ERFE and the PERFE. The convergence of the QRFE and the PQRFE is known. Consequently, we believe that MQRFE and PMQRFE follow similar behavior, because $\rho_3(\cdot)$ is a combination of $\rho_1(\cdot)$ and $\rho_2(\cdot)$. There are classical works as [136] and recent papers as [9, 108], but the asymptotics for MQRFE and PMQRFE is still an open problem.

3.3 Monte Carlo study

3.3.1 Computational considerations

In this subsection, we explain how we estimate $\boldsymbol{\theta}(\tau) = (\boldsymbol{\alpha}(\tau)^T, \boldsymbol{\beta}(\tau)^T)^T \in \mathbb{R}^p$, $p = n + d$ by several different methods. We call as QR, ER and MQR estimates of $\boldsymbol{\beta}(\tau)$ the solution of (3.2) restricted by $\alpha_i = 0 \ \forall i = 1, \dots, n$ if ρ_1 , ρ_2 and ρ_3 , respectively. As well as QRFE, ERFE and MQRFE estimates of $\boldsymbol{\beta}(\tau)$ and $\boldsymbol{\alpha}(\tau)$ the solution of (3.2) without restriction. Further as PQRFE, PERFE and PMQRFE estimates of $\boldsymbol{\beta}(\tau)$ and $\boldsymbol{\alpha}(\tau)$ the solution of (3.3) restricted by the tuning parameter λ which also must be estimated. For simplicity, we put a subscripted index to identify the estimation method from one to nine in the following order: QR, ER, MQR, QRFE, ERFE, MQRFE, PQRFE, PERFE and

PMQRFE, e.g., $\hat{\alpha}(\tau)_4$ corresponds to QRFE procedure. We call the zero index as the initial value, e.g., $\alpha(\tau)_0$. We use the same scheme for the empirical residuals

$$\hat{\varepsilon}_{ij}(\tau)_k = Y_{ij} - \hat{\alpha}_i(\tau)_k - \mathbf{X}_{ij}\hat{\beta}(\tau)_k, \quad (3.4)$$

where $\hat{\beta}(\tau)_k$ is the estimated $\beta(\tau)$ by the k th method.

To obtain $\hat{\beta}(\tau)_k$, $k = 1, 2, 3$ or in other words QR, ER and MQR we need to minimize (3.2) under the restriction that $\alpha(\tau) = \mathbf{0}$ and choose ρ_1 , ρ_2 and ρ_3 , respectively. To achieve $\hat{\beta}(\tau)_k$, $k = 4, 5, 6$ or QRFE, ERFE and MQRFE we need to minimize (3.3) without restrictions on $\alpha(\tau)$ and choose ρ_1 , ρ_2 and ρ_3 , respectively. To get $\hat{\beta}(\tau)_k$, $k = 7, 8, 9$ or PQRFE, PERFE and PMQRFE we need to minimize (3.3) with a restriction tuning parameter λ previously estimated and choose ρ_1 , ρ_2 and ρ_3 , respectively. Usually, λ is estimated by CV or Bayesian information criterion (BIC), but there is a key result for median regression given by [73]. The standard deviation of the errors divide by the standard deviation of the intercepts converge in probability to the optimal tuning parameter, i.e., $\sigma(\mathbf{U})/\sigma(\alpha) \rightarrow \lambda$ at least for $\tau = 0.5$ [73]. Thus, an obvious strategy is to replace U_{ij} by the empirical residual $\hat{\varepsilon}_{ij}(\tau)_k$ from (3.4), and the correspondent $\hat{\alpha}_i(\tau)_k$. This has a strong computational advantage compared with CV or BIC, which is time. Because those methods require reestimating several times, and this optimal parameter needs just two steps. We write a pseudocode as Algorithm 1 to make the methodological nestle structure more intuitive; the example is about PQRFE, but is very similar to PERFE and PMQRFE.

Data: $\mathbf{Y}, \mathbf{X}$

Result: Obtain $\hat{\beta}(\tau)_7$ and $\hat{\alpha}(\tau)_7$ by PQRFE.

while QRFE for $\tau = 0.5$ **do**

 take initial values $\beta(\tau)_0 = \mathbf{0}$ and $\alpha(\tau)_0 = \mathbf{0}$

 minimize (3.2)

 obtain $\hat{\lambda}(\tau)_4 = \hat{\sigma}(\hat{\mathbf{R}}(\tau)_4)/\hat{\sigma}(\hat{\alpha}(\tau)_4)$

 return $\hat{\beta}(0.5)_4$, $\hat{\alpha}(0.5)_4$ and $\hat{\lambda}(0.5)_4$

end

while PQRFE **do**

 take initial values $\beta(\tau)_0 = \hat{\beta}(0.5)_4$, $\alpha(\tau)_0 = \hat{\alpha}(0.5)_4$ and $\hat{\lambda}(0.5)_4$

 minimize (3.3) under restriction $\lambda = \hat{\lambda}(0.5)_4$

 return $\hat{\beta}(\tau)_7$ and $\hat{\alpha}(\tau)_7$

end

Algorithm 1: Example of optimizing by PQRFE

The covariance presented in Theorem 1 is asymptotic, we approximate it by $d \times d$ lower block of $\hat{\Gamma}^{-1}\hat{\mathbf{V}}\hat{\Gamma}^{-1}$, where

$$\hat{\mathbf{V}} = \frac{1}{m} \begin{bmatrix} \mathbf{Z}^T \hat{\mathbf{H}} \mathbf{Z} & \mathbf{Z}^T \hat{\mathbf{H}} \mathbf{X} / \sqrt{n} \\ \mathbf{X}^T \hat{\mathbf{H}} \mathbf{Z} / \sqrt{n} & \mathbf{X}^T \hat{\mathbf{H}} \mathbf{X} / n \end{bmatrix}, \quad \hat{\Gamma} = \frac{1}{m} \begin{bmatrix} \mathbf{Z}^T \hat{\mathbf{G}} \mathbf{Z} & \mathbf{Z}^T \hat{\mathbf{G}} \mathbf{X} / \sqrt{n} \\ \mathbf{X}^T \hat{\mathbf{G}} \mathbf{Z} / \sqrt{n} & \mathbf{X}^T \hat{\mathbf{G}} \mathbf{X} / n \end{bmatrix},$$

where $\hat{\mathbf{G}} = \text{diag}(g(\hat{\varepsilon}_{11}(\tau)), \dots, g(\hat{\varepsilon}_{nm}(\tau)))$, $\hat{\mathbf{H}} = \text{diag}(\hat{g}_{11}^2 \hat{\varepsilon}_{11}(\tau)^2, \dots, \hat{g}_{nm}^2 \hat{\varepsilon}_{nm}(\tau)^2)$, $\hat{g}_{ij} = g(\hat{\varepsilon}_{ij}(\tau))$. This is a coherent approximation for ERFE and PERFE, but for QRFE and PQRFE [65] establishes $\mathbf{D}_1^{-1} \mathbf{D}_0 \mathbf{D}_1^{-1}$, which are

$$\mathbf{D}_0 = \lim_{\substack{m \rightarrow \infty \\ n \rightarrow \infty}} \frac{1}{m} \begin{bmatrix} \mathbf{Z}^T \mathbf{Z} & \mathbf{Z}^T \mathbf{X} / \sqrt{n} \\ \mathbf{X}^T \mathbf{Z} / \sqrt{n} & \mathbf{X}^T \mathbf{X} / n \end{bmatrix}, \quad \mathbf{D}_1 = \lim_{\substack{m \rightarrow \infty \\ n \rightarrow \infty}} \frac{1}{m} \begin{bmatrix} \mathbf{Z}^T \mathbf{F} \mathbf{Z} & \mathbf{Z}^T \mathbf{F} \mathbf{X} / \sqrt{n} \\ \mathbf{X}^T \mathbf{F} \mathbf{Z} / \sqrt{n} & \mathbf{X}^T \mathbf{F} \mathbf{X} / n \end{bmatrix},$$

where $\mathbf{F} = \text{diag}(f_{ij}(\varepsilon_{ij}(\tau)))$, f_{ij} is the density of $\varepsilon_{ij}(\tau)$, and could be approximated by $\hat{\mathbf{F}} = \text{diag}(\hat{f}_{ij}(\hat{\varepsilon}_{ij}(\tau)))$, using kernel function as Epanechnikov [60]. $\hat{\mathbf{D}}_1$ is similar to $\mathbf{D}_1$, but replacing $\mathbf{F}$ to $\hat{\mathbf{F}}$. To be coherent, we propose replace $\hat{\mathbf{V}}$ to $\hat{\mathbf{V}}_1$ for MQRFE and PMQRFE, which is

$$\hat{\mathbf{V}}_1 = \frac{1}{m} \begin{bmatrix} \mathbf{Z}^T \hat{\mathbf{H}}_1 \mathbf{Z} & \mathbf{Z}^T \hat{\mathbf{H}}_1 \mathbf{X} / \sqrt{n} \\ \mathbf{X}^T \hat{\mathbf{H}}_1 \mathbf{Z} / \sqrt{n} & \mathbf{X}^T \hat{\mathbf{H}}_1 \mathbf{X} / n \end{bmatrix},$$

where $\hat{\mathbf{H}}_1 = \text{diag}(\hat{g}_{11}^2 \rho_3(\hat{\varepsilon}_{11}), \dots, \hat{g}_{nm}^2 \rho_3(\hat{\varepsilon}_{nm}))$.

All the computational techniques discussed here are available in an R package [23]. We use RcppArmadillo to accelerate the calculations [27].

3.3.2 Monte Carlo design and measures of accuracy

In this section, a simulation study with finite sample sizes is conducted to verify the performance of the estimators QR, QRFE, PQRFE, MQRFE, PMQRFE, ERFE and PERFE under six different scenarios, e.g., a combination of homoscedasticity and heteroscedasticity with normality, heavy tail and skew examples which are quite frequent in real data applications. Heavy tail distributions or a mixture of distributions with AO may affect the estimation of some statistical characteristics of the data, like the mean, the variance, and the correlation. Hence, robust estimation methods are needed to get reliable statistical models [17].

The first model is a simple case where the subjects α_i and design matrix X_{ij} have a location effect, and we call this case as location-shift model (LSM) in (3.5). The second model presents a design matrix with a location and scale effect, and we name it location-scale-shift model (LSSM) in (3.6). In this way, the response variable Y_{ij} is generated by

$$Y_{ij} = \alpha_i + X_{ij}\beta + U_{ij}, \quad (3.5)$$

and in the second case by

$$Y_{ij} = \alpha_i + X_{ij}\beta + (1 + X_{ij}\gamma)U_{ij}, \quad (3.6)$$

where $X_{ij} = W_i + V_{ij}$, both W_i and V_{ij} follow standard Gaussian. There are three variants for each model:

- i) Normal case, both $\boldsymbol{\alpha}$ and $\mathbf{U}$ follow IID standard Gaussian,

- ii) Student's t case, both $\boldsymbol{\alpha}$ and $\mathbf{U}$ follow IID Student's t -distribution with three degrees of freedom,
- iii) Chi-squared case, both $\boldsymbol{\alpha}$ and $\mathbf{U}$ follow IID Chi-squared distribution with three degrees of freedom.

All the cases we assume $\beta = 0$ and $\gamma = \frac{1}{10}$. Consequently, $\beta(\tau) = 0$ for (3.5), but $\beta(\tau) = 0.1F_U^{-1}(\tau)$ for (3.6), this experimental design is originally proposed by [65].

The sample size follows a balanced design with $n = 50$, $m = 5$, it was already showed that increase n and/or m improves the estimated $\beta(\tau)$ and $\hat{\lambda}$ is more directly affected by m [73]. Investigating unbalanced scenarios is not very productive, as it is known that they do not substantially increase the challenge to estimate [32]. As the "curse of dimensionality" and crossing quantile are hotspots, and it is impossible to investigate them without studying several quantiles, we explore nine different values $\tau = 0.1, 0.2, \dots, 0.9$ [11, 93, 62]. Since the asymptotic extremes of MQR are already being investigated, we did not seek to optimize c , and we chose a shared value among the choices of previous authors as $c = 1$ [69, 62]. Each reported entry is based on $S = 1000$ Monte Carlo replications.

To compare the estimators, we calculate by Monte Carlo simulations for each component $\hat{\theta}_k(\tau)$, $k = 1, \dots, p$ of the parameter estimate $\hat{\boldsymbol{\theta}}(\tau)$, the empirical bias, root mean squared error (RMSE) and coverage probability (CP) with $100(1-\delta)\%$ of confidence by

- i) Bias($\hat{\theta}_k(\tau)$) = $\frac{1}{S} \sum_{s=1}^S (\hat{\theta}_{sk}(\tau) - \theta_k(\tau))$,
- ii) RMSE($\hat{\theta}_k(\tau)$) = $\left(\frac{1}{S} \sum_{s=1}^S (\hat{\theta}_{sk}(\tau) - \theta_k(\tau))^2 \right)^{1/2}$,
- iii) CP($\hat{\theta}_k(\tau), \delta$) = $\frac{1}{S} \sum_{s=1}^S (1_{\theta_k(\tau) \in [\hat{\theta}_{sk}(\tau) \pm \widehat{\text{SE}}(\hat{\theta}_{sk}(\tau))\phi^{-1}(1-\delta/2)])$,

where $\hat{\theta}_{sk}(\tau)$ is the s th simulated estimation of $\theta_k(\tau)$, ϕ is the standard normal distribution, $\widehat{\text{SE}} = (\widehat{\Sigma}_k/N)^{1/2}$ is the estimated SE, $\widehat{\Sigma}_k$ is the k th element of diagonal of $\widehat{\Sigma} = \widehat{\mathbf{\Gamma}}^{-1}\widehat{\mathbf{V}}\widehat{\mathbf{\Gamma}}^{-1}$ for ERF, $\widehat{\Sigma} = \widehat{\mathbf{D}}_1^{-1}\widehat{\mathbf{D}}_0\widehat{\mathbf{D}}_1^{-1}$ for QRF, and $1 - \delta \in (0, 1)$ is the confidence level, which we fixed on 0.05 [83].

3.3.3 Empirical results

The empirical investigation starts from the point left open by [65], i.e., estimating λ . Assuming the optimal value shown by [73] for λ is $\sigma(\mathbf{U})/\sigma(\boldsymbol{\alpha})$ and that the distributions of $\mathbf{U}$ and $\boldsymbol{\alpha}$ are always paired, then the optimal $\lambda = 1$. We investigated the three methods by which λ is estimated with $\tau = 0.5$. Figure 3.1 shows $\hat{\lambda}(0.5)_k$ box-plots in the six scenarios by the three methods $k = 4, 5, 6$. Normal LSM shows a spread between 0.6 and 1.3, Student's t LSM displays $0.1 < \hat{\lambda}(0.5)_k < 3.2$, Chi-squared LSM shows a spread between 0.4 and 1.5. The normal distribution is the one that allows values closer to one,

followed by Chi-squared. The three methods display a median estimation close to the real value of λ , ERFE has a slightly more pronounced underestimation than the others. The LSSM case is very similar.

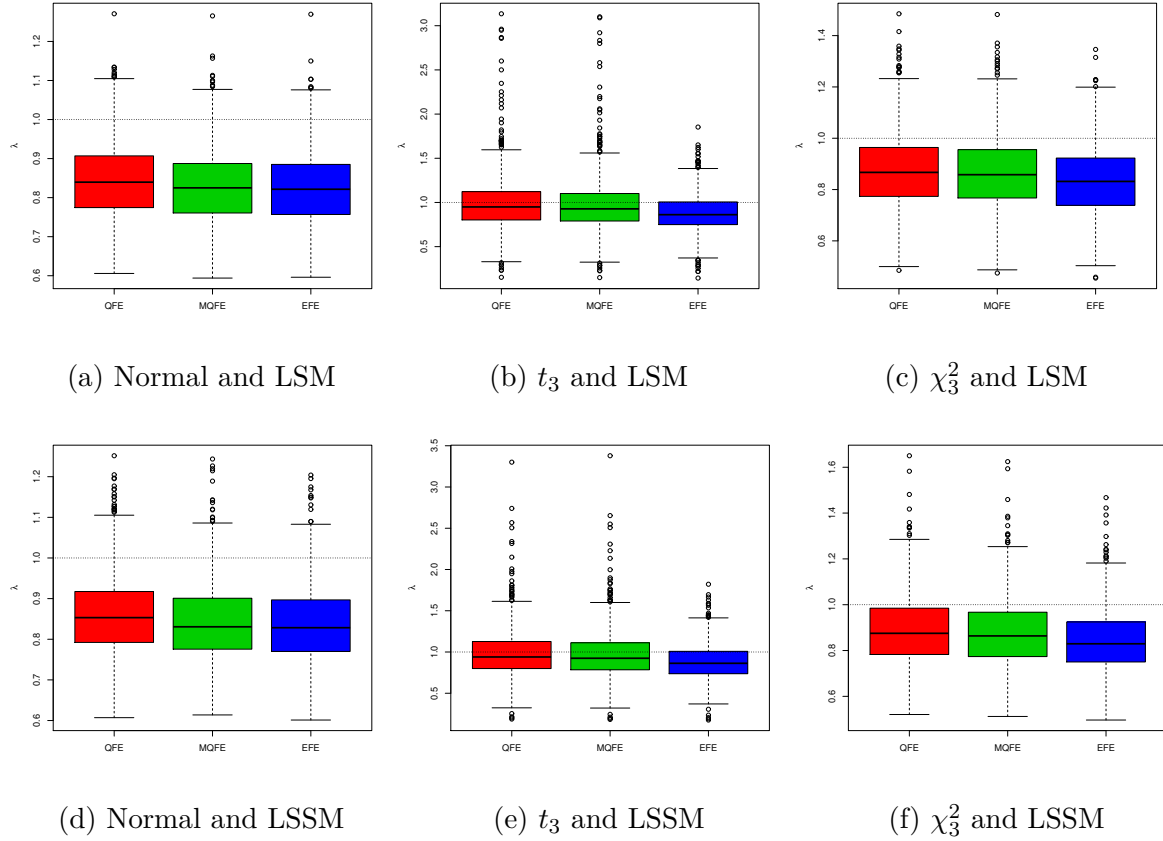


Figure 3.1: Within each boxplot λ is estimated by QRFE (red at the left), MQRFE (green at the middle), ERFE (blue at the right); and the plots follow LS (above) and LSS (below) models and Normal (on the left), t_3 (on the center) χ_3^2 (on the right) distributions.

The second part of our empirical investigation compares the estimation of $\beta(\tau)$ by the available methods $k = 3, \dots, 9$. Even though $k = 3$ is model misspecification, we want to emphasize this inadequacy by empirical results. Figure 3.2 shows the three measurements (Bias, RMSE, CP) for the three distributions of the LSM for nine τ . Even though the Figure seems synthetic, it is essential to say that to plot each row of the Figure 9,000 artificial datasets are generated and 63,000 point estimates of $\beta(\tau)$ are performed. Bias appears to be a random function for Normal; Student's t and Chi-squared, which is a good result, i.e., no method has a systematic bias. About RMSE the worst result is present by QR as expected, it is the price for omitting alpha from the model. For Normal case, QRFE and PQRFE have high RMSE, followed by MQRFE and PMQRFE. In other words, the lowest RMSE is performed by PERFE and followed by ERFE. About the CP, QR is disastrous, PQRFE is below 0.9 for extreme quantiles $\tau = 0.1, 0.9$. The other methods fluctuate between 1 and 0.9. Concerning Student's t QR has high RMSE, soon

followed by ERFE, PERFE, QRFE, PQRFE, MQRFE and PMQRFE, the last dominates the rest. About CP, QR is disastrous again, and all the others are quite similar. For the Chi-squared case, we have a curious point: the RMSE of PMQRFE and PQRFE are worst than QR for $\tau < 0.4$, also for the CP. Soon after, the QR presents bad indices. Then MQRFE and QRFE present the best statistics for $\tau < 0.6$ and ERFE and PERFE the best results for $\tau \leq 0.6$. Here, there seems to be an asymmetry influence of F on the estimates with LASSO penalty, at least for two of them, this result is weird, but it is predicted by [108] on the assumption B.3. which requires F to be even for LASSO procedure under M -estimation. Figure 3.3 presents results very similar to Figure 3.2. Except that the bias of the LSSM presents an increasing curve shape. This non-random fit indicates the presence of an unexplained source of variability in the analysis. Such a fact is in agreement with model 2, which has heteroscedasticity. Summarizing PERFE is the best method for Gaussian data, PMQRFE and PQRFE are the best in the case of heavy tail data like Student's t. MQRFE and QRFE are the best methods for skew data, such as Chi-squared.

3.4 Real data application

After verifying the effectiveness of the methods on simulated data, the next natural step is to validate this methodology on real data. The actual world never presents the same challenges as theory; on the one hand, no database meets all the assumptions of the model; on the other hand, some methodological details for which we dedicate a long time of adjustment may be irrelevant or trivial in practice. A classical economic application of quantile regression on panel data consists of studying the gross domestic product (GDP) growth of the OECD countries [37]. This work inspires us, but we choose to investigate the economic quality of the countries in terms of the high technology aggregate. This class of panel data commonly presents non-stationary variables and non-linear relationships between them, which configures a severe impediment to using the traditional linear models [6].

We work with a public database, which is available on the World Bank [To explore the data](#)¹. The response variable is HTE and we use EDU [111, 89, 85] and R&D [3, 117, 6] as predictors. We restrict our analysis to the 38 OECD members in 2022 [To know about OECD](#)², and the analyzed period corresponds to 12 years, starting from 2007. If there is no missing, the observational design would be balanced with $n = 38$, $m = 12$ and $N = 456$. However, there are 111 missings, and the observational design is unbalanced with $n = 35$, Japan, Turkey and United States of America (USA) are eliminated because they had no information about EDU, $m_i = 2, \dots, 12$ and final $N = 345$.

¹Source visited at May 26, 2022

²Source visited at May 26, 2022

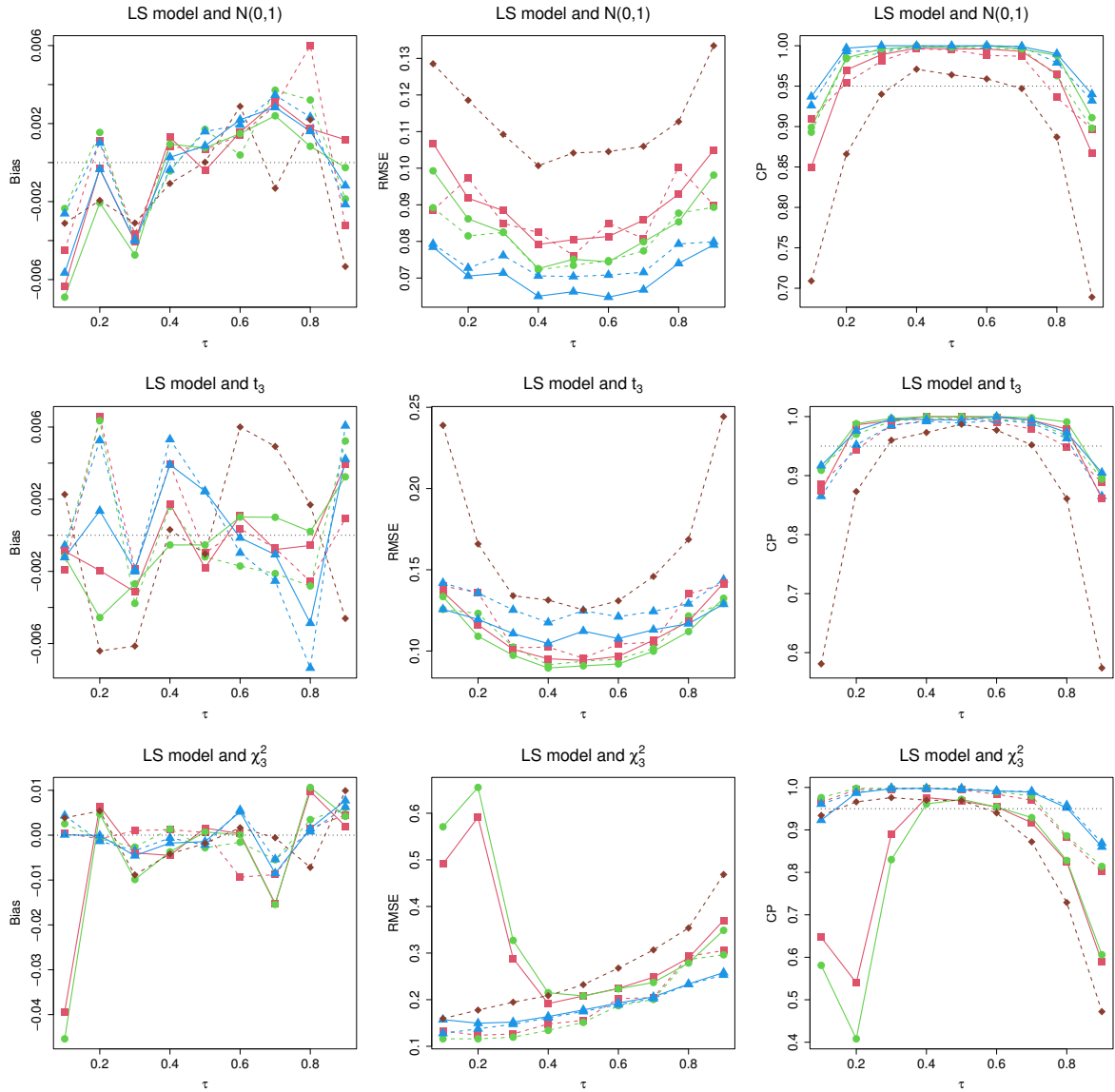


Figure 3.2: Bias, RMSE and CP with 95% of confidence, LS model estimate by QR (dotted line with cyan rhombus), QRFE (dashed line with red square), MQRFE (dashed line with green circle), ERFE (dashed line with blue triangle), PQRFE (solid line with red square), PMQRFE (solid line with green circle), PERFE (solid line with blue triangle).

Before starting to fit regression models, we performed a brief EDA. Figure 3.4a and 3.4b present a scattering pattern opening in a cone, typical of data with heteroscedasticity, so QRFE could be a good technique to deal with it. Figure 3.4c shows large dispersion between countries (they are sorted alphabetically by ISO code, e.g., Switzerland is CHE). Each country appears to present its own average; therefore, a longitudinal model with an α_i for each individual seems necessary. Figure 3.4d shows the asymmetry of the response variable and discourages using methods with a penalty, which are sensitive to skewness. A brief pool of statistical tests is performed to reinforce the EDA. One-sample normality Kolmogorov-Smirnov test was performed for HTE, residuals of QRFE

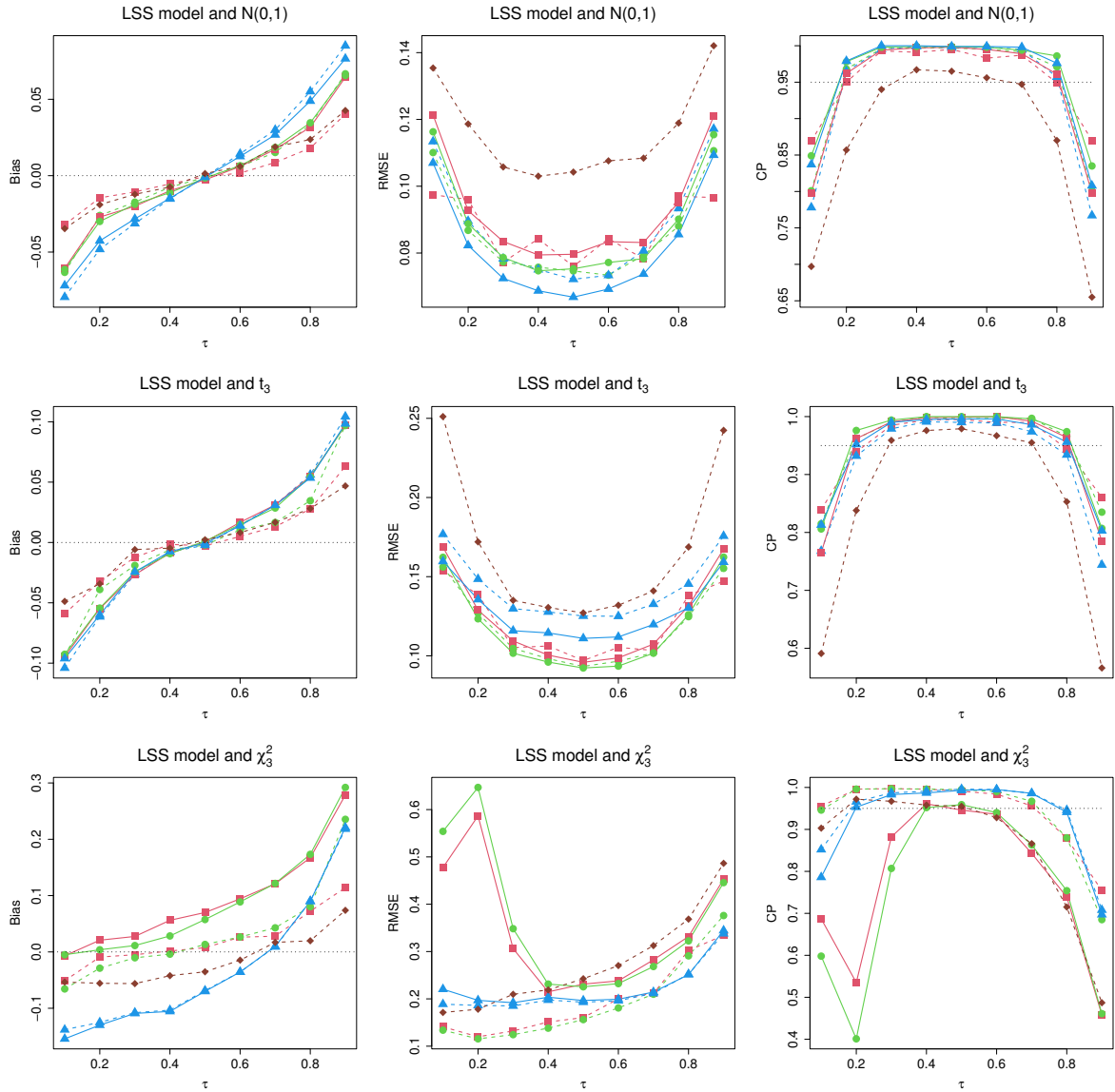


Figure 3.3: Bias, RMSE and CP with 95% of confidence, LSS model estimate by QR (dotted line with cyan rhombus), QRFE (dashed line with red square), MQRFE (dashed line with green circle), ERFE (dashed line with blue triangle), PQRFE (solid line with red square), PMQRFE (solid line with green circle), PERFE (solid line with blue triangle).

and ERFE with p-values always smaller than 0.001; and skewness of 0.896, 1.488 and 0.687, respectively. These results reinforce the hypothesis of non-normality and non-symmetry. Summarizing the data seems to require a quantile analysis incorporating individual intercepts and robustness against skewness, as QRFE, MQRFE and ERFE do. The simulations show that the best methods for chi-squared distribution are MQRFE, QRFE and ERFE, respectively, see Figure 3.2.

We adjust the following model to this data

$$Y_{ij} = \alpha_i + \beta_1 \text{EDU}_{ij} + \beta_2 \text{R\&D}_{ij} + U_{ij}, \quad (3.7)$$

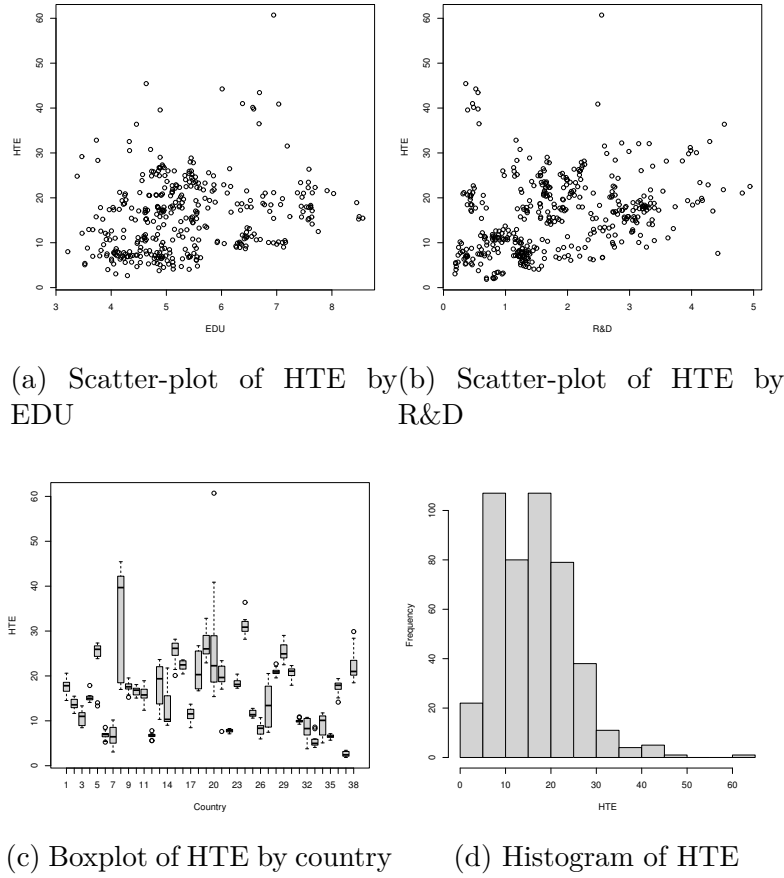


Figure 3.4: EDA of HTE, EDU and R&D

where, for $i = 1, \dots, 35$ and $j = 1, \dots, 12$, Y_{ij} is the HTE output of i th country at j th year, EDU_{ij} is the yearly national investment on educational, $R\&D_{ij}$ is the yearly national investment on research, α_i is the country base and U_{ij} is the measurement error of the model.

We fit the model 3.7 by several methods, but we show the two most suitable according to our previous EDA analysis. Figure 3.5 shows the central fit of the estimated $\beta(\tau)$ and its confidence intervals for the three methods. $\beta_1(\tau)$ corresponds to the influence of EDU on HTE, it is constant around 1.5 for QRFE and MQRFE, but it is not different of zero for ERFE. The QRFE is not significantly different of zero at $\tau = 0.4$, MQRFE is significantly different of zero for all τ between 0.1 and 0.9. The $\beta_2(\tau)$ is quantile coefficient of R&D, it floats around 1 and 6 for QRFE and MQRFE, and around 4 and 11 for PERFE. The three methods are significantly different of zero for all τ between 0.1 and 0.9. The balance between robustness and efficiency allowed the MQRFE to be the only method that always shows significant results in accordance with the economic literature. Thus R&D and EDU investments have a positive impact on HTE. We conclude that R&D and EDU constitute investments that influences the composition of a mature economy focused on innovation, whose export of products with high technology is not exclusively

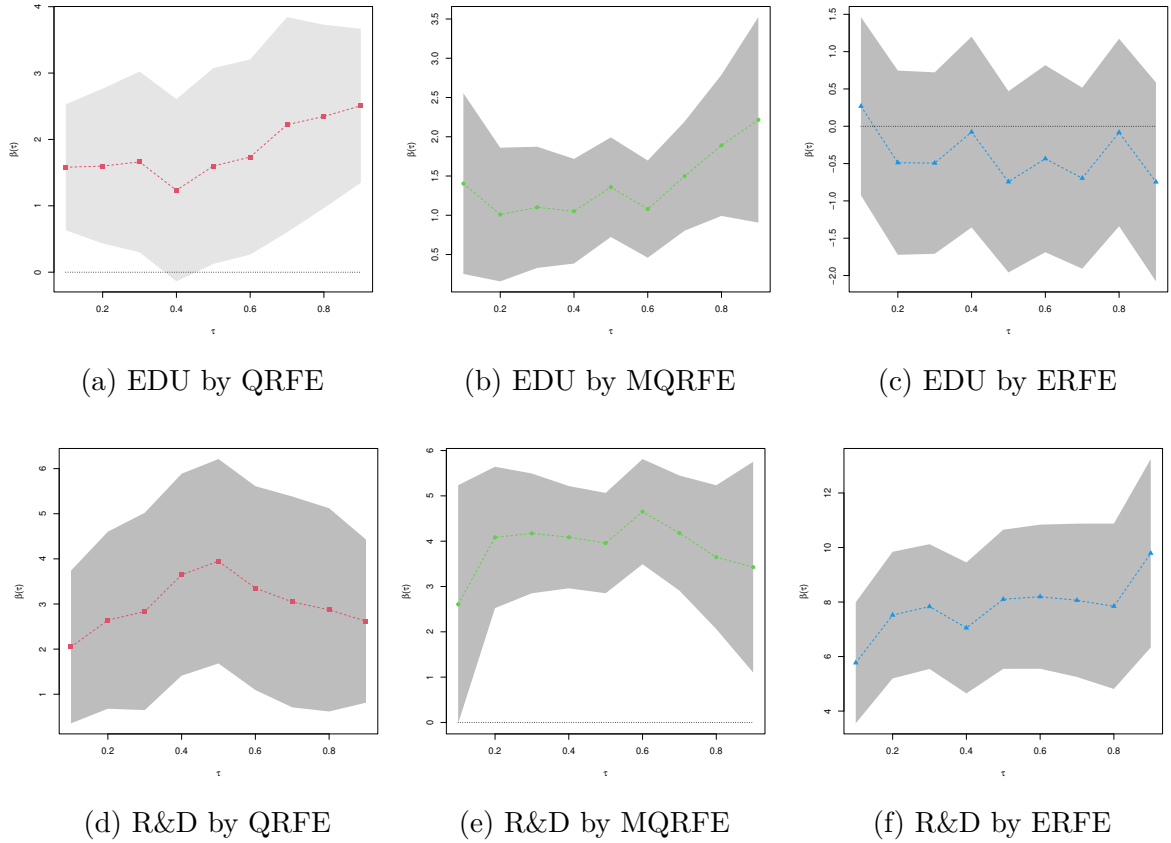


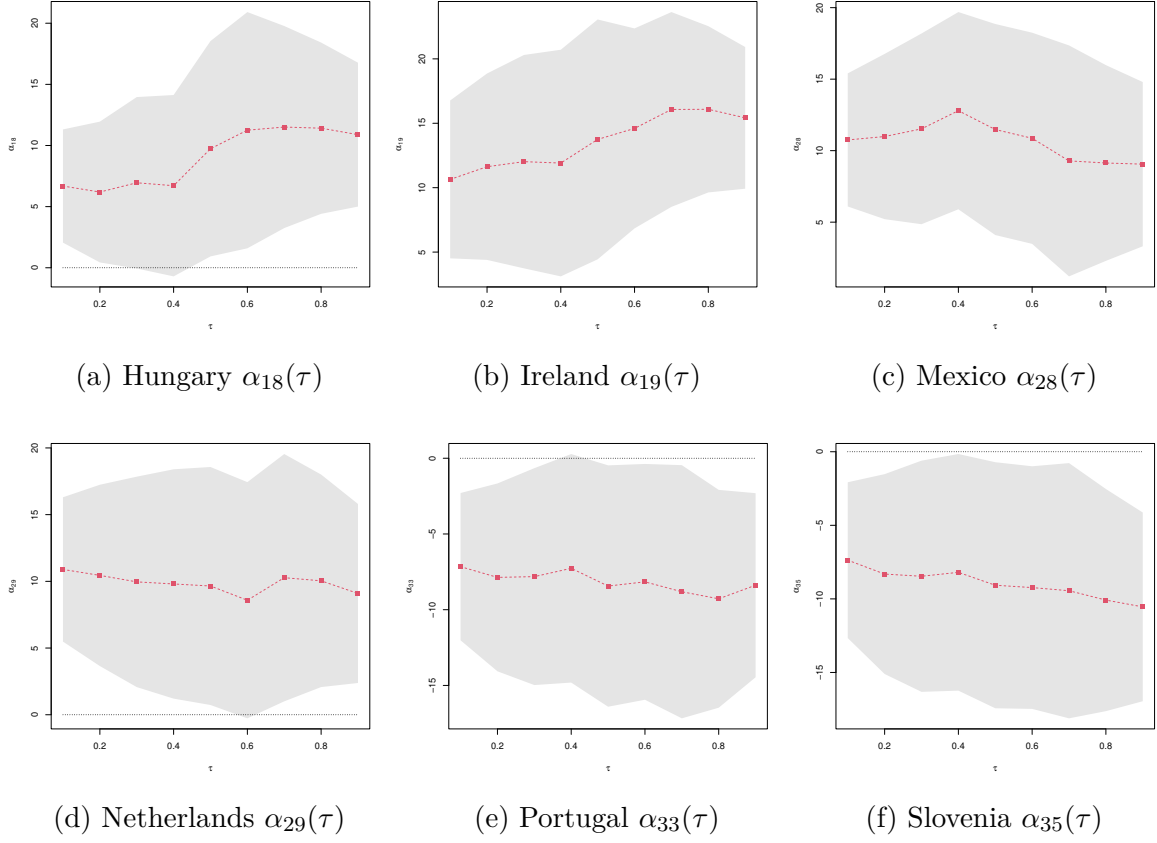
Figure 3.5: EDU and R&D coefficients and confidence intervals by QRFE, MQRFE and ERFE

dependent on commodities.

Additionally, Figure 3.6 display some $\alpha_i(\tau)$ different of zero: Hungary and Ireland have increasing $\alpha_{18}(\tau)$ and $\alpha_{19}(\tau)$, Mexico and Netherlands have positive constant $\alpha_{28}(\tau)$ and $\alpha_{29}(\tau)$, Portugal and Slovenia have negative fixed $\alpha_{33}(\tau)$ and $\alpha_{35}(\tau)$.

3.5 Conclusion

This paper explores alternatives to the check function to enlarge the scope of quantile regression for panel data. We present four new methods: ERFE, PERFE, MQRFE, and PMQRFE. We show the asymptotic properties of two procedures: ERFE and PERFE. We display how to implement the full family of methods computationally. The Monte Carlo empirical study slices the best scenario for each method. Under normality, we recommend PERFE, if the errors present heavy tails, we suggest PMQRFE, and under a skew distribution, we claim that MQRFE is the best choice. These recommendations are possible because we extend the investigated τ range from one or two points, as in the previous literature, to nine. Concern about heteroscedasticity does not impact the methods. There is no dominant method, which is the best option for any case, but each

Figure 3.6: $\alpha_i(\tau)$ coefficients and confidence intervals by MQRFE

one is the most accurate for a specific context. A real data analysis is performed about a traditional economic issue: the impact of human resources and research investment on qualified exportation by countries of OECD. This is an example of skew distribution where MQRFE shows that EDU and R&D have positive influence on HTE. This result agrees with the Economics literature.

3.6 Appendix

Here we present the proof of Theorem 1. Let us consider an objective function as

$$\begin{aligned}
V_{nm}(\boldsymbol{\delta}) &= \sum_{i=1}^n \sum_{j=1}^m \rho_{\tau}(Y_{ij} - \mathbf{Z}_{ij}\boldsymbol{\alpha}(\tau) - \mathbf{X}_{ij}\boldsymbol{\beta}(\tau) - \mathbf{Z}_{ij}\boldsymbol{\delta}_0/\sqrt{m} - \mathbf{X}_{ij}\boldsymbol{\delta}_1/\sqrt{N}) - \rho_{\tau}(Y_{ij} - \mathbf{Z}_{ij}\boldsymbol{\alpha}(\tau) - \mathbf{X}_{ij}\boldsymbol{\beta}(\tau)) \\
&= \sum_{i=1}^n \sum_{j=1}^m \rho_{\tau}(\varepsilon_{ij}(\tau) - \mathbf{W}_{ij}\boldsymbol{\delta}) - \rho_{\tau}(\varepsilon_{ij}(\tau)) \\
&= \sum_{i=1}^n \sum_{j=1}^m g(\varepsilon_{ij}(\tau) - \mathbf{W}_{ij}\boldsymbol{\delta})\rho_2(\varepsilon_{ij}(\tau) - \mathbf{W}_{ij}\boldsymbol{\delta}) - g(\varepsilon_{ij}(\tau))\rho_2(\varepsilon_{ij}(\tau)) \\
&= \sum_{i=1}^n \sum_{j=1}^m g_{ij}(\varepsilon_{ij}(\tau)^2 - 2\varepsilon_{ij}(\tau)\mathbf{W}_{ij}\boldsymbol{\delta} + (\mathbf{W}_{ij}\boldsymbol{\delta})^2) - g_{ij}^*(\varepsilon_{ij}(\tau)^2) \\
&= \mathbf{e}^T \mathbf{G} \mathbf{e} - 2\mathbf{e}^T \mathbf{G} \mathbf{W} \boldsymbol{\delta} + \boldsymbol{\delta}^T \mathbf{W}^T \mathbf{W} \boldsymbol{\delta} - \mathbf{e}^T \mathbf{G}^* \mathbf{e} \\
&= -2\mathbf{e}^T \mathbf{G} \mathbf{W} \boldsymbol{\delta} + \boldsymbol{\delta}^T \mathbf{W}^T \mathbf{W} \boldsymbol{\delta} - \mathbf{e}^T (\mathbf{G} - \mathbf{G}^*) \mathbf{e}
\end{aligned}$$

where $\mathbf{Z}_{ij} = (Z_{m(i-1)+j,1}, \dots, Z_{m(i-1)+j,n})$, $\mathbf{W}_{ij} = (\mathbf{Z}_{ij}/\sqrt{m}, \mathbf{X}_{ij}/\sqrt{N})$, $g_{ij} = g(\varepsilon_{ij}(\tau) - \mathbf{W}_{ij}\boldsymbol{\delta})$, $g_{ij}^* = g(\varepsilon_{ij}(\tau))$, $\mathbf{e} = (\varepsilon_{11}(\tau), \dots, \varepsilon_{nm}(\tau))^T$, $\mathbf{G} = \text{diag}(g(\varepsilon_{11}(\tau) - \mathbf{W}_{11}\boldsymbol{\delta}), \dots, g(\varepsilon_{nm}(\tau) - \mathbf{W}_{nm}\boldsymbol{\delta}))$, $\mathbf{G}^* = \text{diag}(g(\varepsilon_{11}(\tau)), \dots, g(\varepsilon_{nm}(\tau)))$. Note that $\hat{\boldsymbol{\delta}} = (\hat{\boldsymbol{\delta}}_0^T, \hat{\boldsymbol{\delta}}_1^T)^T = (\sqrt{m}(\hat{\boldsymbol{\alpha}}(\tau) - \boldsymbol{\alpha}(\tau))^T, \sqrt{N}(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau))^T)^T$ minimizes $V_{nm}(\boldsymbol{\delta})$. By condition A.2 and A.3 imply Lindeberg condition and

$$-2\mathbf{e}^T \mathbf{G} \mathbf{W} \boldsymbol{\delta} \xrightarrow{\mathcal{L}} \mathbf{B} \boldsymbol{\delta},$$

where $\mathbf{B} = \mathcal{N}(\mathbf{0}, \mathbf{V})$, $\mathbf{V} = \mathbf{W}^T \mathbf{H} \mathbf{W}$, $\mathbf{H} = \text{diag}(g_{11}^2 \varepsilon_{11}(\tau)^2, \dots, g_{nm}^2 \varepsilon_{nm}(\tau)^2)$.

$$\mathbb{E}(\boldsymbol{\delta}^T \mathbf{W}^T \mathbf{G} \mathbf{W} \boldsymbol{\delta}) = \boldsymbol{\delta}^T \boldsymbol{\Gamma} \boldsymbol{\delta}$$

where $\boldsymbol{\Gamma} = \mathbf{W}^T \mathbf{G} \mathbf{W}$, the condition A.3 implies

$$\text{Var}(\boldsymbol{\delta}^T \mathbf{W}^T \mathbf{G} \mathbf{W} \boldsymbol{\delta}) \rightarrow 0,$$

and also that

$$\mathbb{E}(\mathbf{e}^T (\mathbf{G} - \mathbf{G}^*) \mathbf{e}) \rightarrow 0.$$

Thus $\hat{\boldsymbol{\delta}}_1 = \sqrt{N}(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau))$ is normal with zero mean and variance equal to the lower $(d \times d)$ block of $\Gamma^{-1} \mathbf{V} \Gamma^{-1}$, this is an extension of [86, Theorem 3], it is similar to [65, Theorem 1], except that does not depend of the density distribution of $\varepsilon_{ij}(\tau)$.

Chapter 4

Adaptive LASSO Quantile Regression with Fixed Effects with an application to the Millennium Cohort Study

Abstract

The QRFE analysis of longitudinal data involves a different intercept for each unit. When the number of subjects is much larger than the number of repeated measurements per individual and the number of explanatory variables is large, we propose to reduce the dimension by a double LASSO penalization. We extend to longitudinal data the adaptive LASSO quantile regression proposed by [142]. The method enjoys the oracle properties of asymptotic normality and proper model selection. We analyse the finite sample size performance of the method with Monte Carlo simulations, and the Millennium Cohort Study (MCS) is used to illustrate the methodology with a real longitudinal data.

Keywords: Quantile regression, longitudinal data, oracle property, variable selection.

4.1 Introduction

The MCS is an observational and multidisciplinary cohort study designed to watch the lives of children born at the new millennium dawn. The survey began with an original sample size of 18,818 cohort members of the four countries of the United Kingdom (UK). Now the study is in its eighth sweep, and the members have an average age of 22 years. There is a range of information about health, physical development, cognitive ability, educational attainment, and socioeconomic characteristics of the sampling members and

their families. This great effort in data gathering makes MCS an unparalleled and valuable resource for investigating health outcomes and social inequalities [16]. As MCS provides such vast data collection, each scientific study investigates a small part of its content. Neighborhood deprivation, pregnancy planning, parental distress, and child adiposity trajectories are topics explored by [87, 30, 110, 116]. Missing data is a broadly relevant topic because, at each sweep, the sample size decreases [45, 91, 92].

A cohort study is a specific kind of longitudinal data where each individual is observed repeatedly. A common way to analyze this type of collection is by LMM where a random intercept fits the individual location, which can not be explained totally by the general predictors [72, 25]. Of course, this analysis can be more complex and include random coefficients, cross-sectional data, or heteroscedastic errors [24, pages 6-7]. The choice of the appropriate LMM allows to quantify the global average effect of explanatory variables and the individual effects. However, instead of explaining the mean effect of an outcome, one may be interested in extreme behaviour of the outcome given the regressors. Quantile regression is commonly applied to this end [66, 65]. The MCS has been used for testing new quantile methods such as the conditional quantile regression with bootstrap [41], the univariate and multivariate M -quantile regression with random effects [118, 4]. These works contributed also to understand the missing observations mechanism in the data base and the extreme behaviour of two outcomes, namely “SDQ-I” and “SDQ-E” in England.

There is an equivalence between best linear unbiased predictions (BLUP) and LMM with penalized random intercepts [65]. This result has a long history that dates back to [46] as mentioned by [98]. This equivalence opens the doors to other penalization in the quantile regression context. [65] proposes a quantile regression with repeated measures without penalization and with LASSO penalization of the intercepts. The former method is QRFE, and the latter is PQRFE. Both procedures are robust to AO as any quantile regression and can be used even though the errors display skewness and heteroscedasticity. [73] and [74] study how to optimize the LASSO tuning parameter and obtain precise results for the median regression. [37], [35], and [36] approximate the asymptotic variance of the regression coefficients by bootstrap. Shrinking the intercept parameters allow to reduce the estimation error in the regressor parameter vector. Now, when the number of predictors is large, a quantile regression method with a penalization on both intercept and predictor parameters is natural. This is proposed by [77], and is denoted as double penalized quantile regression with fixed effects (DPQRFE). However, LASSO is known to present an overestimation behavior in the sense that it selects variables which are not relevant [13, page 25]. An attractive approach to correct this bias is to use the adaptive LASSO [140]. This approach consists of two steps, first obtaining the weights by a consistent method and then, minimizing the penalized objective function. [142] provides an adaptive LASSO for quantile regression which has oracle properties: asymptotic

normality and proper model selection.

This paper proposes an adaptive LASSO method for quantile regression with longitudinal data where intercept and predictor parameters are penalized. This method is named ALQRFE. We investigate its finite sample behaviour in the case of skewness and heavy tail errors, and we apply our methodology to the MCS data studied by [118, 4]. These authors focus on two outcomes SDQ-I and SDQ-E of cohort members from 2004 to 2008 in England. Here, we consider Northern Ireland, which is less studied and smaller than England. This increases the challenge of fitting a longitudinal model with fewer observations per parameter. We add social, economic, and family predictors to check if they are relevant explanatory variables for the outcomes of interest.

The rest of this article is organized as follows. Section 4.2 describes the MCS data. In Section 4.3, the model and estimation methods are introduced, and the asymptotic theory is presented. Section 4.4 discusses the computational aspects and the precision of the proposed methods by Monte Carlo simulations in different scenarios. Section 4.5 is dedicated to the application of our method to the real data. Section 4.6 concludes the paper and the Appendix displays the list of considered covariates and how to extract them from MCS databases.

4.2 Data

We work with a public database, which is freely available to researchers under standard access conditions on the UK Data Service, [click to explore the data](#).¹ The MCS target population is any baby born between September 1st, 2000, and August 31st, 2001 (for England and Wales), and between November 24th, 2000, and January 11th, 2002 (for Northern Ireland and Scotland). The baby also must have been born alive, still living in UK when it is nine months old, and whose family is eligible to receive any child benefit from the government. The initial sweep contains 18,552 families from all UK, but only 15,590 (second sweep), 15,246 (third) and 13,857 (fourth) [16]. The non-responses increase over sweeps, are not equally distributed in the population, and vulnerable families tend to be more challenging to re-interview [45, 91]. Thus, a model considering the longitudinal effect is fundamental to avoiding bias [118]. We restrict our analysis to North Ireland during sweeps 2, 3, and 4. North Ireland is the smallest country of UK, it has the highest dimensionality among the four countries, i.e., the greatest ratio parameter per observation.

We study two outcomes, SDQ-I and SDQ-E, both scores range between zero and twenty points, where high values indicate a progressive difficulty and suffering, see [118, 4]. We also use similar predictors as the previous authors: age (in years), squared age, adverse life events (ALE), family poverty measured by socio-economic disadvantage (SED) [115], mother's depression measured by Kessler score (KS), maternal education separate in uni-

¹Source visited on June 13, 2022

versity degree, general certificate of secondary education (GCSE) and no-qualification, ethnicity (non-white (baseline) or white), gender (female (baseline) or male), neighbourhood deprivation measured by index of multiple deprivation (IMD) [87], Stratum (disadvantage (baseline) or advantage). North Ireland does not have an Ethnic stratum like England; this is slightly different from the study of previous authors. We also include 84 additional covariates. We mention them by blocks: family composition (from 12 to 22), family income (from 23 to 25, 28, and 29), other characteristics (26 and 27), playful and educational stimuli at home (from 30 to 43), values cultivated by the family (from 44 to 51), child's general behavior and mood (from 52 to 89 and 92), capacity of the grandparents to help (90), drug consumption (91 and 94), beliefs and life satisfaction (93 and 95). Table A.2 shows the variable codes to obtain the previous 11 covariates and Table A.3 displays the variable codes to achieve the new 84 covariates. Table A.1 shows the datasets where each covariate can be found. After removing rows with missing data, our dataset has 99 columns (95 covariates, two outcomes, sweep and child identification) and 2001 rows (673 for the 2nd sweep, 673 for the 3rd sweep, and 655 for the 4th sweep) and involves 702 children from North Ireland.

Both outcomes are non Gaussian, as shown in Figure 4.1. The mode, median, and mean are very distant from each other and confirm a strong skewness and a right strong heavy tail as in the England case [118]. Consequently, methods sensitive to AO and skewness must be avoided.

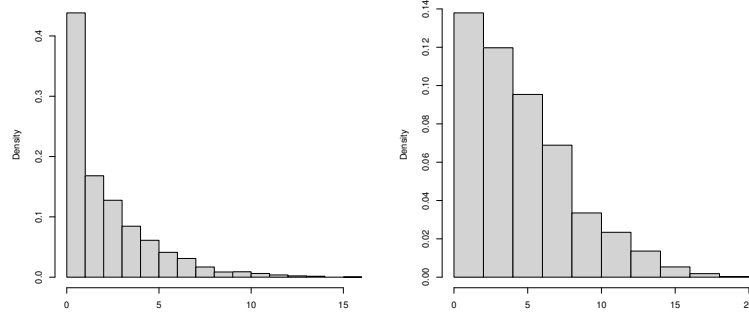


Figure 4.1: Histogram for SDQ-I (left) and SDQ-E (right)

4.3 Estimation methods

We denote by $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{im_i})^T$ the vector of measurements taken at sweeps $1, \dots, m_i$ of the i th subject for $i = 1, \dots, n$. Following [72] and [24, page 5], we assume that each measurement Y_{ij} of $\mathbf{Y}_i$, $j = 1, \dots, m_i$, follows the LMM

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \alpha_i + U_{ij}, \quad i = 1, \dots, n \quad j = 1, \dots, m_i, \quad (4.1)$$

where $\mathbf{X}_{ij}$ is a $(1 \times d)$ vector of exploratory covariates, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)^T$ is a vector of d unknown but fixed parameters to be estimated, α_i are the random intercepts or the fixed effect of the model and follow an unknown probability distribution, and U_{ij} is a RV usually IID.

It follows from (4.1) that

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\alpha} + \mathbf{U},$$

where $\mathbf{Y} = (\mathbf{Y}_1^T, \dots, \mathbf{Y}_n^T)^T$ is a vector of length $N = \sum_{i=1}^n m_i$. $\mathbf{X}_i = [\mathbf{X}_{i1}^T, \dots, \mathbf{X}_{im_i}^T]^T$ is a $(m_i \times d)$ matrix, and $\mathbf{X} = [\mathbf{X}_1^T, \dots, \mathbf{X}_n^T]^T$ is a $(N \times d)$ observational matrix. $\mathbf{Z} = [\mathbf{Z}_1, \dots, \mathbf{Z}_n]$ is a $(N \times n)$ incident matrix, i.e., $\mathbf{Z}_i$ is a vector of zeros and ones. If $i = 1$, then $\mathbf{Z}_1$ is one from 1 to m_1 and zero otherwise. If $i > 1$, then $\mathbf{Z}_i$ is one from $m_{i-1} + 1$ to m_i , and zero otherwise. For example, if $m_i = m$ for all i , then $\mathbf{Z} = \mathbf{I}_n \otimes \mathbf{1}_m$, where $\mathbf{I}_n$ is the identity matrix of dimension $(n \times n)$ and $\mathbf{1}_m$ is a m -vector of ones. Finally, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^T$ is the n -vector of intercepts, and $\mathbf{U} = (\mathbf{U}_1^T, \dots, \mathbf{U}_n^T)^T$ is the N -vector of errors where $\mathbf{U}_i^T = (U_{i1}, \dots, U_{im_i})$.

4.3.1 Quantile regression with fixed effects

When modeling longitudinal data, the intercept $\boldsymbol{\alpha}$ can be assumed to be random or non random. In the first case, explanatory variables and subjects are assumed to be independent, the subject effect is treated as random without location impact and is only related to the variance structure. In the second case, the individual effects are treated as dummy variables. However, the number of parameters increases with the sample size, which induces some difficulties [2, 38, 137]. In the context of dependent quantile regression of $\mathbf{Y}$ with fixed effects, for any $\tau \in (0, 1)$, the conditional quantile of Y_{ij} is

$$Q_{Y_{ij}}(\tau | \mathbf{X}_{ij}) = \alpha_i(\tau) + \mathbf{X}_{ij}\boldsymbol{\beta}(\tau),$$

where $\boldsymbol{\beta}(\tau)$ is the quantile regressor, $\alpha_i(\tau)$ is the quantile individual intercept, and $\alpha_i(\tau) + \mathbf{X}_{ij}\boldsymbol{\beta}(\tau)$ is the 100 τ th percentile of Y_{ij} . Here we adopted the same notation as [67, page 278] and [36], $\boldsymbol{\alpha}(\tau)$ and $\boldsymbol{\beta}(\tau)$ are estimated jointly as a $p = (n + d)$ vector parameter $\boldsymbol{\theta}(\tau) = (\boldsymbol{\alpha}(\tau)^T, \boldsymbol{\beta}(\tau)^T)^T$ by

$$\hat{\boldsymbol{\theta}}(\tau) = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \sum_{i=1}^n \sum_{j=1}^{m_i} \rho_\tau(Y_{ij} - \alpha_i - \mathbf{X}_{ij}\boldsymbol{\beta}), \quad (4.2)$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T)^T$ and $\rho_\tau(u) = u(\tau - 1_{\{u < 0\}})$, $u \in \mathbb{R}$, is the well-known check function [66].

The solution $\hat{\boldsymbol{\theta}}(\tau)$ of (4.2) corresponds to QRFE. The term fixed effects is commonly used to refer only to the predictors in a LMM. However, here it is employed to emphasize that the individual intercepts are treated as fixed effects in a quantile regression for longitudinal data. The QR considered by [66] is a special case of QRFE, where $\boldsymbol{\alpha}(\tau) = \mathbf{0}$.

4.3.2 Penalized quantile regression with fixed effects

As mentionned by [65], there is an equivalence between the BLUP of a LMM with random effects and regression estimate where intercepts are penalized. As a generalization, [65] proposes quantile regression with a LASSO penalization of the intercepts, denoted as PQRFE. This improves the estimator's performance taking advantage of sparsity. Now, in the case of quantile regression with a large number of regressors, [142] propose to penalize the predictor vector by LASSO. We propose to penalize intercepts and the predictor vector in the context of quantile regression with fixed effects. The estimation procedure of LQRFE is given by

$$\hat{\theta}(\tau, \lambda) = \arg \min_{\theta \in \mathbb{R}^p} \sum_{i=1}^n \sum_{j=1}^{m_i} \rho_{\tau}(Y_{ij} - \alpha_i - \mathbf{X}_{ij}\beta) + \lambda \sum_{i=1}^n |\alpha_i|, \quad (4.3)$$

where λ is a tuning parameter estimated by CV or by IC.

[142] propose also to replace LASSO quantile regression by an adaptive version which has oracle properties. Here, we generalize their method to longitudinal data. Then, the ALQRFE is given by

$$\hat{\theta}(\tau, \lambda, \omega) = \arg \min_{\theta \in \mathbb{R}^p} \sum_{i=1}^n \sum_{j=1}^{m_i} \rho_{\tau}(Y_{ij} - \alpha_i - \mathbf{X}_{ij}\beta) + \lambda \sum_{i=1}^n \hat{\omega}_{1i} |\alpha_i| + \lambda \sum_{k=1}^d \hat{\omega}_{2k} |\beta_k|, \quad (4.4)$$

where $\omega = (\omega_1^T, \omega_2^T)^T$, $\omega_1 = (\omega_{11}, \dots, \omega_{1n})^T$ and $\omega_2 = (\omega_{21}, \dots, \omega_{2d})^T$. If $\omega = \mathbf{0}$, then ALQRFE coincides with QRFE. If $\omega_1 = \mathbf{1}$ and $\omega_2 = \mathbf{0}$, then ALQRFE is equal to LQRFE proposed by [65]. If $\omega = \mathbf{1}$, then ALQRFE is equivalent to DPQRFE proposed by [77]. The weights in (4.4) are given by $\hat{\omega}_{1i} = |\hat{\alpha}_i(\tau)|^{-1}$ and $\hat{\omega}_{2k} = |\hat{\beta}_k(\tau)|^{-1}$ where $\hat{\alpha}_i(\tau)$ and $\hat{\beta}_k(\tau)$ are consistent estimates of $\alpha_i(\tau)$ and $\beta_k(\tau)$, respectively.

In the litterature, the terminology “low” or “classical” dimension refers usually to the case where the number of parameters p is much smaller than the number of observations N , and “high” dimension means the opposite [13, pages 1 and 78]. Some authors consider $p \geq N$ as high dimension [29] and [122, page 194]. However, penalized methods are necessary even in moderate dimension where $0.2N < p < N$ [43].

4.3.3 Asymptotic results

For notation simplicity, we restrict our attention to balanced design, i.e., $m_i = m$ for all $i = 1, \dots, n$, and thus $\mathbf{Z} = \mathbf{I}_n \otimes \mathbf{1}_m$ and $\mathbf{Z}^T \mathbf{Z} = m \mathbf{I}_n$. We use standard notations, the symbol “ $\xrightarrow{\mathbb{P}}$ ” means convergence in probability, “ $\xrightarrow{\mathcal{L}}$ ” denotes convergence in distribution, $\text{sgn}(u)$ is the sign of u . Let $\mathcal{B} = \{k : \beta_k(\tau) \neq 0\}$ with cardinality $\text{card}(\mathcal{B}) = d_0 \leq d$, and $\mathcal{B}_n = \{k : \hat{\beta}_k(\tau, \lambda, \omega) \neq 0\}$. We impose the same regularity conditions as [60]:

(A1) For $i = 1, \dots, n$, define the $(m \times (d+1))$ matrix $\mathbf{A}_i = [\mathbf{Y}_i, \mathbf{X}_i]$. We assume that for each fixed i , the rows of $\mathbf{A}_i$ are IID, and that the matrices $\mathbf{A}_1, \dots, \mathbf{A}_n$ are mutually independent.

(A2) For each $\delta > 0$,

$$\epsilon_\delta = \inf_{i \geq 1} \inf_{|\alpha_i(\tau)| + \|\beta(\tau)\|_1 = \delta} \mathbb{E} \left[\int_0^{\alpha_i(\tau) + \mathbf{X}_{i1}\beta(\tau)} \{F_i(s|\mathbf{X}_{i1}) - \tau\} ds \right] > 0,$$

where $\|\cdot\|_1$ stands for the l_1 norm, $F_i(u|\mathbf{x})$ denotes the conditional distribution function of U_{ij} given $\mathbf{X}_{ij} = \mathbf{x}$. We assume that $F_i(u|\mathbf{x})$ has density $f_i(u|\mathbf{x})$. Let $f_i(u)$ denote the marginal density of U_{ij} .

(A3) There exists a constant M such that $\sup_{i \geq 1} \|\mathbf{X}_{i1}\| < M$, almost surely,

(A4) (a) For each i , $f_i(u|\mathbf{x})$ is continuously differentiable with respect to u for each $\mathbf{x}$ and let $f'_i(u|\mathbf{x}) = \partial f_i(u|\mathbf{x})/\partial u$; (b) there exist constants C_f and L_f such that $f_i(u|\mathbf{x}) < C_f$ and $|f'_i(u|\mathbf{x})| < L_f$ uniformly over $(u, \mathbf{x})$ and $i \geq 1$; (c) $f_i(0)$ is bounded from below by some positive constant independent of i ,

(A5) Put $\gamma_i = \mathbb{E}[f_i(0|\mathbf{X}_{i1})\mathbf{X}_{i1}]/f_i(0)$ and $\Gamma_n = n^{-1} \sum_{i=1}^n \mathbb{E}[f_i(0|\mathbf{X}_{i1})\mathbf{X}_{i1}(\mathbf{X}_{i1}^T - \gamma_i^T)]$. (a) Γ_n is nonsingular for each n , and the limit $\Gamma = \lim_{n \rightarrow \infty} \Gamma_n$ exists and is nonsingular; (b) the limit $V = \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \mathbb{E}[(\mathbf{X}_{i1} - \gamma_i)(\mathbf{X}_{i1} - \gamma_i)^T]$ exists and is nonsingular.

The following theorem can be proved using [140, Theorem 2], [142, Theorem 3.1] and [60, Theorems 3.1 and 3.2].

Theorem 3. Under conditions (A1) – (A5), if $\lambda_n/\sqrt{n} \rightarrow 0$, $\lambda_n \rightarrow \infty$, $n^2(\log n)^3/m \rightarrow 0$ as $n \rightarrow \infty$, but m grows at most polynomially in n , then the ALQRFE satisfies the following oracle properties:

i) Consistency in variable selection: $\lim_{n \rightarrow \infty} \mathbb{P}(\mathcal{B}_n = \mathcal{B}) = 1$,

ii) Asymptotic normality: the $\sqrt{N}(\hat{\beta}_{\mathcal{B}}(\tau, \lambda, \boldsymbol{\omega}) - \beta_{\mathcal{B}}(\tau)) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \tau(1-\tau)\Gamma_{\mathcal{B}}^{-1}V_{\mathcal{B}}\Gamma_{\mathcal{B}}^{-1})$,

where $\Gamma_{\mathcal{B}}$ and $V_{\mathcal{B}}$ are the positive definite matrices defined in assumption (A5) where $\mathbf{X}$ is replaced by the $(N \times d_0)$ matrix $\mathbf{X}_{\mathcal{B}}$ obtained by deleting the columns c_k of $\mathbf{X}$ for $k \notin \mathcal{B}$.

4.4 Monte Carlo study

4.4.1 Computational considerations

QR and QRFE are simpler to implement than LQRFE and ALQRFE. The implementation of LQRFE and ALQRFE are described in Algorithms 2 and 3. We use R language

[95], but the core of the code is written in RCPP to improve computational time [26], and the minimization is made by a quasi-Newton method with box constraints [21].

The LQRFE and ALQRFE require a tuning parameter chosen by CV or IC, such as the BIC. Splitting the data into two groups, training and validation as required by the CV technique may not be trivial in cases such as longitudinal data where there is an intrinsic dependence affinity between observational units. In addition, QR is not performed for the prediction and presents low accuracy for τ different from the median. For these reasons, we advocate using alternative methods to CV, such as BIC. We estimate the tuning parameter by BIC as [75, 74] do for quantile regression. For LQRFE, we define

$$\text{BIC}_1(\mathbf{Y}, \lambda) = \sum_{i=1}^n \sum_{j=1}^{m_i} \rho_\tau(Y_{ij} - \hat{\alpha}_i(\tau, \lambda) - \mathbf{X}_{ij} \hat{\boldsymbol{\beta}}(\tau, \lambda)) + \log(N)(\text{card}(\mathcal{A}_n^{(1)}) + d),$$

where $\mathcal{A}_n^{(1)} = \{i : \hat{\alpha}_i(\tau, \lambda) \neq 0\}$ is the set of nonzero intercepts by LQRFE method. We fix $0 < \lambda_{\min} < \lambda_{1,\max}$ and restrict $\lambda_1 \in [\lambda_{\min}, \lambda_{1,\max}]$. [13, page 38] suggest to take $\lambda_{\min} = \sqrt{\log(p)/N}$, and a natural choice is $\lambda_{1,\max} = m \max(\tau, 1 - \tau)$, because any λ_1 larger than this value implies that $\hat{\alpha}_i(\tau, \lambda) = 0$ for all $i = 1, \dots, n$, see [74]. For ALQRFE, we define

$$\text{BIC}_2(\mathbf{Y}, \lambda, \boldsymbol{\omega}) = \sum_{i=1}^n \sum_{j=1}^{m_i} \rho_\tau(Y_{ij} - \hat{\alpha}_i(\tau, \lambda, \boldsymbol{\omega}) - \mathbf{X}_{ij} \hat{\boldsymbol{\beta}}(\tau, \lambda, \boldsymbol{\omega})) + \log(N)[\text{card}(\mathcal{A}_n^{(2)}) + \text{card}(\mathcal{B}_n)],$$

where $\mathcal{A}_n^{(2)} = \{i : \hat{\alpha}_i(\tau, \lambda, \boldsymbol{\omega}) \neq 0\}$ is the set of nonzero intercepts by ALQRFE method. We restrict $\lambda_2 \in [\lambda_{\min}, \lambda_{2,\max}]$ where $\lambda_{2,\max} = \min\{m \max(\tau, 1 - \tau), 2 \max_{1 \leq k \leq d} |\mathbf{X}_k^T \mathbf{Y}|/N\}$, see [74].

BIC_1 and BIC_2 are calculated for K ordered values $\lambda_1, \lambda_2, \dots, \lambda_K$ of λ , where $\lambda_1 = \lambda_{\min}$, $\lambda_K = \lambda_{1,\max}$ for BIC_1 , and $\lambda_K = \lambda_{2,\max}$ for BIC_2 . In both cases, like [84], we take $\lambda_{k+1} = \eta^k \lambda_1$ for $k = 1, \dots, K - 1$, where

$$\eta = \frac{\log \lambda_K - \log \lambda_1}{K - 1}.$$

We now turn to estimating the asymptotic variance-covariance matrix. The estimation of Γ and V depends on the conditional densities, and their sample analogs do not directly estimate them because the conditional densities are unknown. We consider the kernel estimation of the matrices Γ and V . Let $K : \mathbb{R} \rightarrow \mathbb{R}$ denote a kernel function (probability density function). Let $\{h_n\}$ denote a sequence of positive numbers (bandwidths) such that $h_n \rightarrow 0$ as $n \rightarrow \infty$. We use the notation $K_{h_n}(u) = h_n^{-1} K(u/h_n)$. Let $\hat{u}_{ij} = Y_{ij} - \hat{\alpha}_{ij} - \mathbf{X}_{ij} \hat{\boldsymbol{\beta}}$, which can be viewed as an estimator of u_{ij} , where τ is omitted to simplify the notation. Γ and V can be estimated by

$$\hat{\Gamma} = N^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} K_{h_n}(\hat{u}_{ij}) \mathbf{X}_{ij} (\mathbf{X}_{ij} - \hat{\boldsymbol{\gamma}}_i)^T, \quad \hat{V} = N^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} (\mathbf{X}_{ij} - \hat{\boldsymbol{\gamma}}_i) (\mathbf{X}_{ij} - \hat{\boldsymbol{\gamma}}_i)^T, \quad (4.5)$$

Data: $\mathbf{Y}, \mathbf{X}, \mathbf{Z}$

Result: Obtain $\hat{\beta}(\tau, \lambda)$ and $\hat{\alpha}(\tau, \lambda)$ by LQRFE

take initial values $\beta(\tau)_0 = \mathbf{0}$ and $\alpha(\tau)_0 = \mathbf{0}$

declare $k = K$

while $1 \leq k$ **do**

 minimize (4.3) with $\lambda = \lambda_{(k)}$

 update $\hat{\alpha}(\tau, \lambda_{(k)})$ and $\hat{\beta}(\tau, \lambda_{(k)})$

 calculate $\text{BIC}_1(\mathbf{Y}, \lambda_{(k)})$

$k = k - 1$

end

return $\hat{\alpha}(\tau, \lambda)$ and $\hat{\beta}(\tau, \lambda)$ which minimizes BIC_1 .

Algorithm 2: Implementation of LQRFE

Data: $\mathbf{Y}, \mathbf{X}, \mathbf{Z}$

Result: Obtain $\hat{\beta}(\tau, \lambda, \omega)$ and $\hat{\alpha}(\tau, \lambda, \omega)$ by ALQRFE

run a QRFE model and obtain $\hat{\alpha}(\tau)$ and $\hat{\beta}(\tau)$

take weights $\hat{\omega}_1 = (|\hat{\alpha}_1(\tau)|^{-1}, \dots, |\hat{\alpha}_n(\tau)|^{-1})^T$ and $\hat{\omega}_2 = (|\hat{\beta}_1(\tau)|^{-1}, \dots, |\hat{\beta}_d(\tau)|^{-1})^T$

take initial values $\beta(\tau)_0 = \mathbf{0}$ and $\alpha(\tau)_0 = \mathbf{0}$

declare $k = K$

while $1 \leq k$ **do**

 minimize (4.4) with $\lambda = \lambda_{(k)}$

 update $\hat{\alpha}(\tau, \lambda_{(k)}, \hat{\omega})$ and $\hat{\beta}(\tau, \lambda_{(k)}, \hat{\omega})$

 calculate $\text{BIC}_2(\mathbf{Y}, \lambda_{(k)}, \hat{\omega})$

$k = k - 1$

end

return $\hat{\alpha}(\tau, \lambda, \omega)$ and $\hat{\beta}(\tau, \lambda, \omega)$ which minimizes BIC_2 .

Algorithm 3: Implementation of ALQRFE

where

$$\hat{\gamma}_i = \frac{1}{m_i \hat{f}_i} \sum_{j=1}^{m_i} K_{h_n}(\hat{u}_{ij}) \mathbf{X}_{ij}, \quad \hat{f}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} K_{h_n}(\hat{u}_{ij}),$$

good choices of kernel estimator as Epanechnikov guarantees the consistency of $\hat{\Gamma}$ and $\hat{V}$. We estimate the ALQRFE covariance in a similar way as (4.5), but after a dimension reduction of $\mathbf{X}$, denoted by $\mathbf{X}^{(\mathcal{B})}$. Let us consider

$$\hat{\Gamma}_{\mathcal{B}} = N^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} K_{h_n}(\hat{u}_{ij}) \mathbf{X}_{ij}^{(\mathcal{B})} (\mathbf{X}_{ij}^{(\mathcal{B})} - \hat{\gamma}_i)^T, \quad \hat{V}_{\mathcal{B}} = N^{-1} \sum_{i=1}^n \sum_{j=1}^{m_i} (\mathbf{X}_{ij}^{(\mathcal{B})} - \hat{\gamma}_i) (\mathbf{X}_{ij}^{(\mathcal{B})} - \hat{\gamma}_i)^T, \quad (4.6)$$

where

$$\hat{\gamma}_{i,\mathcal{B}} = \frac{1}{m_i \hat{f}_i} \sum_{j=1}^{m_i} K_{h_n}(\hat{u}_{ij}) \mathbf{X}_{ij}^{(\mathcal{B})},$$

where $\mathbf{X}_{ij}$ is replaced by the $(1 \times d_0)$ row vector $\mathbf{X}_{ij}^{(\mathcal{B})}$ obtained by deleting the columns c_k of $\mathbf{X}_{ij}$ for $k \notin \mathcal{B}_n$.

4.4.2 Design and measures

In this section, a simulation study with finite sample sizes is conducted to verify the performance of the estimators QRFE, LQRFE and ALQRFE under six different scenarios, e.g., a combination of three errors CDF with two dimension sizes. We explore different kinds of errors, such as Gaussian, heavy tail, and skewness, which are quite frequent in real data applications. Heavy tail distributions and/or a mixture of distributions with AO may affect the estimation of some statistical characteristics of the data, like the mean, the variance, and the correlation. Hence, robust estimation methods are needed to get reliable statistical models [96, 17].

All the models follow (4.1), $\mathbf{U}$ are IID following a F_U which are normal, Student's t or Chi-squared, the model is balanced with constant repetitions $m_i = m = 5$ and $n = 50$, $\mathbf{Z} = \mathbf{I}_n \otimes \mathbf{1}_m$, $\mathbf{X}$ is a $(N \times d)$ matrix following a multivariate Gaussian distribution with zero mean and correlation of 0.9 between each column, $d = 4, 40$, α_0 are following a $B(\pi_1)F_\alpha$, where F_α is normal, t_3 or χ_3^2 and $B(\pi_1)$ is a Bernoulli with probability of success equal to $\pi_1 = 0.2$, β_0 follows a $B(\pi_2)F_\beta$, where F_β is a standard normal and the Bernoulli has probability equal to $\pi_2 = 0.75, 0.1$, $\mathbf{Y}$ is the final result. A similar experimental design with $d = 1$ and without sparsity is originally proposed by [65]. Previous work investigates deterministic sparsity [77]. To summarize, the three variants of the model are:

- i) Normal case, both α and $\mathbf{U}$ follow IID standard Gaussian, α with a sparsity of 80%,
- ii) Student's t case, both α and $\mathbf{U}$ follow IID t_3 , α with a sparsity of 80%,
- iii) Chi-squared case, both α and $\mathbf{U}$ follow IID χ_3^2 , α with a sparsity of 80%.

The two dimension designs are:

- i) Dense ($\pi_2 = 0.75$), and low ($d = 4$, which correspond to a ratio of 0.216 parameters per observation),
- ii) Sparse ($\pi_2 = 0.1$), and moderate ($d = 40$, which correspond to a ratio of 0.36 parameters per observation).

Examining unbalanced scenarios (where $\exists i : m_i \neq m$) is not very fruitful because they are not substantially more difficult to fit than balanced designs [32]. However, extreme quantiles are progressively more complicated to estimate as far they are from 0.5 [125]. For this reason, we investigate nine distinct values as $\tau = 0.1, 0.2, \dots, 0.9$. The reported entries are based on $S = 600$ Monte Carlo replications in all cases.

To compare the estimators, we calculate the MAE and MSE by Monte Carlo simulations as

$$\text{i) MAE}(\hat{\beta}(\tau)) = \frac{1}{Sd} \sum_{s=1}^S \sum_{k=1}^d |\hat{\beta}_{sk}(\tau) - \beta_k(\tau)|,$$

$$\text{ii) } \text{MSE}(\hat{\beta}(\tau)) = \frac{1}{Sd} \sum_{s=1}^S \sum_{k=1}^d (\hat{\beta}_{sk}(\tau) - \beta_k(\tau))^2,$$

where $\hat{\beta}_{sk}(\tau)$ is the s th simulated estimation of $\beta_k(\tau)$. As all the models are on the same scale, MAE is the best way to compare them [132, 54].

4.4.3 Empirical results

The Figure 4.2 shows the MSE results for three distributions times two dimensions. The dense and low situation does not show a large difference between the methods, except for the Chi-squared distribution, where LQRFE performs very poorly as τ increases. These results are expected because, in the low and dense situation, we have four covariates, and three of them might be different from zero on average. So the difference between a method with and without variable selection should not be large. The skew distribution reveals the fragility of LQRFE method. It is very biased for large τ . Remind the skew behavior for χ_3^2 implies its quantiles are monotonic growing, for example the first quartile $\chi_3^2(0.25)^{-1} \approx 1.212$ and the the third quartile $\chi_3^2(0.75)^{-1} \approx 4.108$. This strong bias of LQRFE is awkward, but it is predicted by [108] on the assumption B.3. which requires F to be even for LASSO procedure under M -estimation. Consequently, we reinforce the necessity of consistent weights for the ALQRFE method. The sparse and moderate situation displays a vast difference between the methods. For Gaussian and Student's t distributions ALQRFE dominates the scenario, and for Chi-squared again LQRFE displays a low performance. This first result is expected because, under a sparse and moderate set, we have 40 predictors. Still, only four differ from zero on average, so the oracle property should make a decisive advantage for ALQRFE.

The Figure 4.3 displays the MAE results, which are more robust than MSE. Although MAE and MSE agree in some situations, for example, in the Gaussian and Student's t cases with $d = 40$, the MAE metric is more clear in the majority of the cases. The dense and low situation shows ALQRFE performance dominates the other methods for all distributions. This is impressive in a scenario where the difference between the true model is just one parameter less in a cluster of four. Additionally, LQRFE displays better results than QRFE for the Student's t distribution, but QRFE displays more satisfactory effects than LQRFE for Chi-squared. The sparse and moderate situation displays a colossal contrast between the methods. ALQRFE performance dominates the other methods for all distributions. Again, LQRFE and QRFE display a switch realization for the Student's t and Chi-squared distributions.

In summarizing, ALQRFE is the best method under multicollinearity, dense or sparse predictors, low or moderate dimension, and Gaussianity, heavy tails or skew distributions. Furthermore, ALQRFE is the best method for any τ .

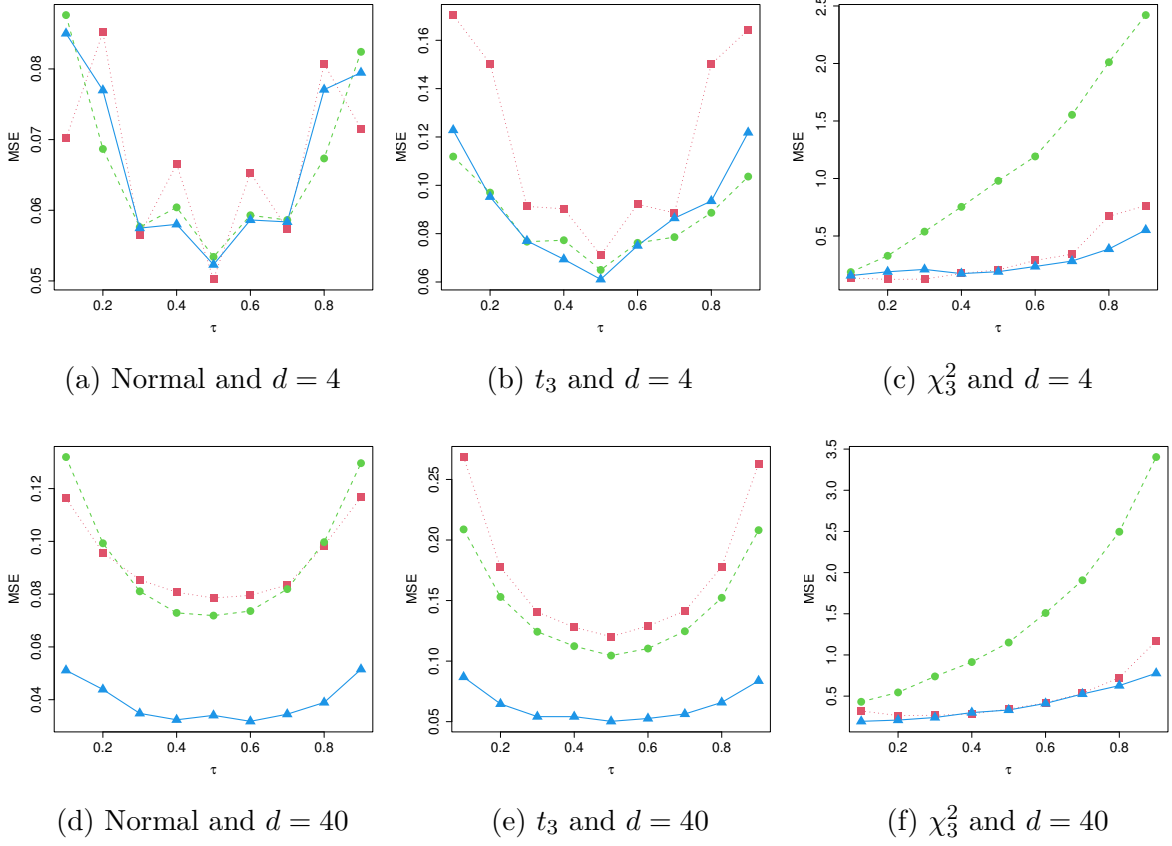


Figure 4.2: MSE of estimates by QRFE (red square), LQRFE (green circle) and ALQRFE (blue triangle), to Normal, t_3 and χ_3^2 models with $d = 4, 40$.

4.5 Real data application

In this section, we analyze the database present in the previous Section 4.2. A fixed effects method requires estimating an intercept for every 702 children, plus 95 explanatory covariates. There are 2001 observations, which result in a parameter per observation ratio of 0.398, a clear case of moderate dimensionality [43]. For this reason, penalized methods such as LQRFE and ALQRFE are preferable than non penalized as QRFE. Between LQRFE and ALQRFE, we recommend the later by its oracle properties and the fragility of LQRFE under skewness, see Figure 4.3f, as we expected after view the outcomes' histograms in Figure 4.1.

In Tables 4.1 and 4.3 only 11 of the QRFE model are shown for SDQ-E and SDQ-I in Northern Ireland, the results being compared with the 11 covariates estimated by MQRRE for England [118, 4]. In Tables 4.2 and 4.4 only significant covariates in at least one $\tau = 0.1, 0.5, 0.9$ by ALQRFE are shown, SDQ-E and SDQ-I in Northern Ireland. Since the ALQRFE method selects the true model with a limiting probability equal to one, then the significant covariates are new findings. They can be precious indicators for

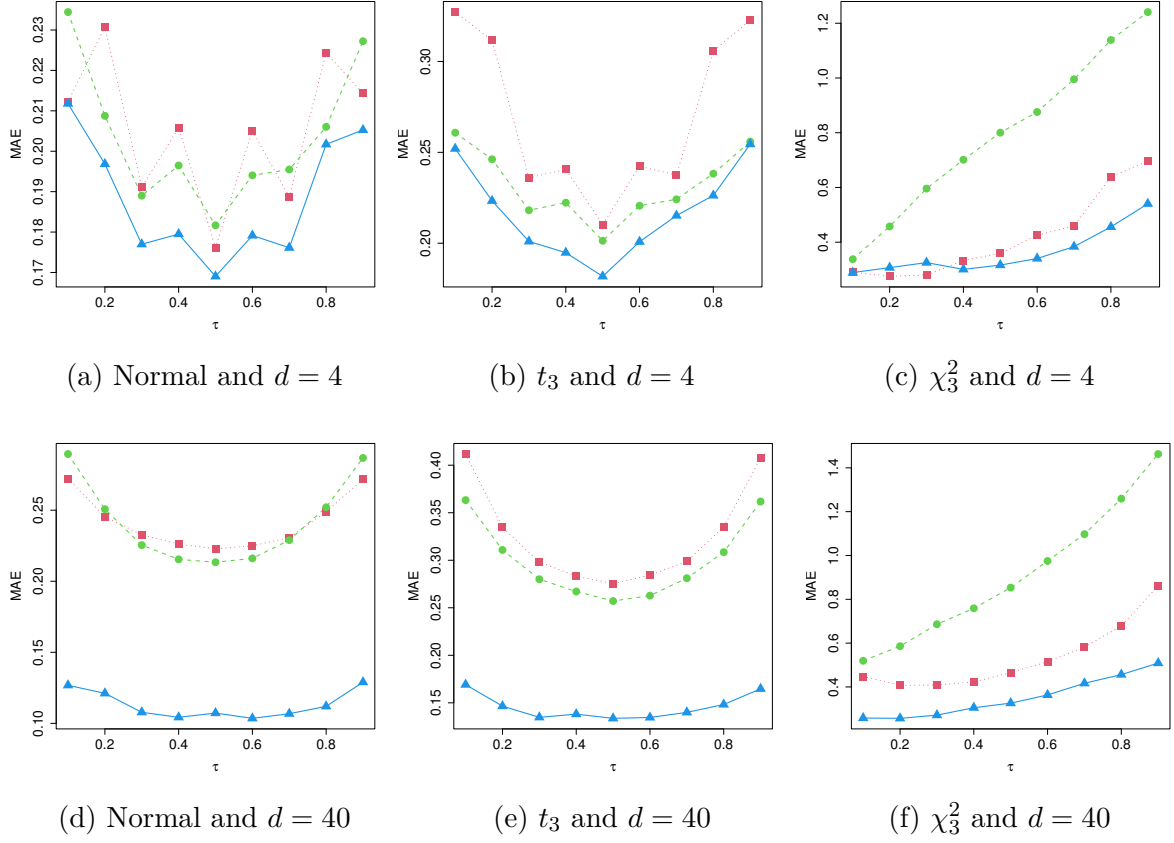


Figure 4.3: MAE of estimates by QRFE (red square), LQRFE (green circle) and ALQRFE (blue triangle), to Normal, t_3 and χ_3^2 models with $d = 4, 40$.

teachers and psychologists who care for these children closely.

4.5.1 First outcome: externalizing difficulties

Even though the QRFE runs for all covariates, we just display the 11 comparable one in Table 4.1. From these, six, five and seven covariates present a significant p-value at 95% of confidence to $\tau = 0.1, 0.5, 0.9$, respectively. Let us enumerate them with their impact: negative $\beta_1(\tau)$ (age), positive $\beta_2(\tau)$ (squared age), positive $\beta_3(0.9)$ (adverse life events), positive $\beta_5(0.9)$ (mother's depression), negative $\beta_6(\tau)$ (mother's high education), negative $\beta_7(\tau)$ (mother's median education), positive $\beta_8(\tau)$ (white), positive $\beta_9(\tau)$ (male). Remind that a negative impact on SDQ-E means protection against externalizing difficulties, i.e., reducing the external metric. The mother's high and median education levels are protection factors as the whiteness and maleness of the children are risk factors for SDQ-E. This result has 100% of agreement with the MQRRE for SDQ-E in England [118].

The Table 4.2 displays only significant covariates in at least one $\tau = 0.1, 0.5, 0.9$ for SDQ-E by ALQRFE. There are 36, 16 and 52 significant coefficients at 95% of confidence to $\tau = 0.1, 0.5, 0.9$, respectively. For comparison, we have 56, 47 and 59 sig-

nificant coefficients for SDQ-E by QRFE. Clearly, ALQRFE is more parsimonious than QRFE. It is screaming that the most vulnerable children to high SDQ-E scores present many significant predictors. However, ALQRFE is still more conservative than QRFE, and only $\beta_1(\tau)$ (age), $\beta_5(0.9)$ (mother's depression), $\beta_8(\tau)$ with $\tau = 0.1, 0.9$ (white), and $\beta_9(0.1)$ (male) are significant with negative (the first) or positive (the remain) impact on the outcome. The other covariates are cite only if they are significant for all τ s, let us enumerate them with their impact: positive $\beta_{26}(\tau)$ (mother's height), interchangeable $\beta_{28}(\tau)$ (language), positive $\beta_{46}(\tau)$ (negotiation), positive $\beta_{52}(\tau)$ (be understood), positive $\beta_{53}(\tau)$ (temper), negative $\beta_{55}(\tau)$ (obedient), positive $\beta_{58}(\tau)$ (fidgeting), positive $\beta_{63}(\tau)$ (distracted), negative $\beta_{69}(\tau)$ (patience), negative $\beta_{73}(\tau)$ (view), positive $\beta_{75}(\tau)$ (affection), negative $\beta_{78}(\tau)$ (physical affection), positive $\beta_{84}(\tau)$ (angry), positive $\beta_{88}(\tau)$ (manipulative), negative $\beta_{89}(\tau)$ (sharer of feelings). Interestingly, if we can look at each child with psycho-behavioral indicators more deeply, then socioeconomic context and neighborhood vulnerabilities become less suitable to predict SDQ-E. Of course, these variables are still relevant for policymakers, but more particular variables may be vital for psychologists, teachers, and parents.

4.5.2 Second outcome: internalizing difficulties

The Table 4.3 follows a similar engine as Table 4.1, i.e., estimations for SDQ-I by QRFE. There are four, three and four significant covariates for $\tau = 0.1, 0.5, 0.9$, respectively. The estimates of $\beta_8(\tau)$, $\beta_9(\tau)$ and $\beta_{11}(\tau)$ are all significant to all quantiles. The $\beta_8(\tau)$ (white) and $\beta_{11}(\tau)$ (disadvantage stratum) have a positive impact on SDQ-I, while $\beta_9(\tau)$ (male) has a negative impact. The gender and stratum effect agree between North Ireland and England as well as QRFE and MQRRE, but ethnicity does not.

The Table 4.4 is similar to 4.2, i.e., estimations for SDQ-I by ALQRFE. It presents 19, 27 and 30 significant covariates for $\tau = 0.1, 0.5, 0.9$, which are much fewer than in Table 4.2. Perhaps this is because it is more difficult to measure SDQ-I than SDQ-E, or this is a rarer event, as Figure 4.1 suggests. For comparison, we have 56, 50 and 56 significant coefficients for SDQ-I by QRFE. We have a second example, that ALQRFE is more parsimonious than QRFE. From the previous 11 covariates, there are zero, one and three significant covariates at 95% of confidence to $\tau = 0.1, 0.5, 0.9$, respectively, by ALQRFE model. Let us enumerate them and their impact: $\beta_5(0.9)$ (mother's depression) has a significant positive effect, $\beta_9(\tau)$ (white) has a significant negative effect, and $\beta_{11}(0.9)$ (disadvantageous stratum) has a significant positive impact on SDQ-I. The other covariates are cite only if they are significant for all τ s, let us enumerate them with their impact: negative $\beta_{35}(\tau)$ (reading help), positive $\beta_{43}(\tau)$ (visits), positive $\beta_{54}(\tau)$ (playing alone), positive $\beta_{56}(\tau)$ (worried), negative $\beta_{59}(\tau)$ (friends 1), negative $\beta_{62}(\tau)$ (friends 2), positive $\beta_{64}(\tau)$ (nervous), positive $\beta_{71}(\tau)$ (prefer adults), positive $\beta_{72}(\tau)$ (fears), positive $\beta_{78}(\tau)$

(physical affection), negative $\beta_{85}(\tau)$ (energy), positive $\beta_{86}(\tau)$ (bad mood), positive $\beta_{92}(\tau)$ (suffering). Only one covariate (physical affection) is significant for all τ in both outcomes with estimation by ALQRFE. An interesting remark is that it has a negative impact in SDQ-E and a positive influence in SDQ-I. All the other covariates are significant either in SDQ-E or SDQ-I, but not on both.

The QRFE finds almost the same risk and protective factors as MQRRE even within a large set of covariates which could confound the estimation. Therefore North Ireland policymakers face the same challenges as their colleagues in England. Using the ALQRFE method, we display more significant predictors of concern for Irish children's health. The ALQRFE brings exciting results to psychologists, teachers, and parents who wish to prevent children from suffering from SDQ-E and SDQ-I.

Table 4.1: Estimated coefficient, SE and p-value of selected covariates for SDQ-E model by QRFE to $\tau = 0.1, 0.5, 0.9$.

	$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
	Coef.	SE	p-value	Coef.	SE	p-value	Coef.	SE	p-value
β_1	-0.808	0.246	0.001	-0.803	0.377	0.033	-0.785	0.233	0.001
β_2	0.049	0.023	0.030	0.067	0.035	0.052	0.057	0.021	0.008
β_3	0.084	0.055	0.124	0.087	0.086	0.310	0.111	0.053	0.036
β_4	0.270	0.230	0.239	0.274	0.358	0.444	0.297	0.220	0.177
β_5	0.042	0.025	0.092	0.048	0.039	0.218	0.083	0.025	0.001
β_6	-0.267	0.064	*	-0.267	0.089	0.003	-0.268	0.054	*
β_7	-0.205	0.068	0.002	-0.203	0.096	0.033	-0.196	0.059	0.001
β_8	0.237	0.022	*	0.238	0.034	*	0.247	0.020	*
β_9	0.126	0.030	*	0.127	0.043	0.003	0.131	0.025	*
β_{10}	-0.036	0.059	0.546	-0.033	0.091	0.721	-0.012	0.056	0.830
β_{11}	-0.103	0.053	0.053	-0.102	0.080	0.203	-0.095	0.049	0.052

* p-value smaller than 0.001.

Table 4.2: Estimated coefficient, SE and p-value of selected covariates for SDQ-E model by ALQRFE to $\tau = 0.1, 0.5, 0.9$.

	$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
	Coef.	SE	p-value	Coef.	SE	p-value	Coef.	SE	p-value
β_1	-0.806	0.349	0.021	-0.798	1.271	0.530	-0.785	0.322	0.015
β_5	0.023	0.034	0.486	0.003	0.125	0.984	0.076	0.035	0.028
β_8	0.082	0.031	0.008	0.083	0.132	0.526	0.245	0.025	*
β_9	0.118	0.048	0.014	0.036	0.282	0.899	-0.004	0.032	0.899
β_{12}	-0.378	0.152	0.013	-0.374	0.573	0.514	-0.371	0.156	0.018
β_{13}	0.254	0.030	*	-0.001	0.128	0.991	0.276	0.025	*
β_{17}	0.423	0.065	*	0.456	0.268	0.090	0.536	0.053	*

β_{18}	0.468	0.061	*	0.342	0.255	0.179	0.587	0.049	*
β_{19}	0.462	0.064	*	0.039	0.263	0.882	0.564	0.053	*
β_{22}	0.254	0.180	0.158	0.293	0.619	0.637	0.384	0.175	0.028
β_{24}	-0.005	0.002	0.015	0.014	0.011	0.231	-0.014	0.003	*
β_{26}	0.474	0.051	*	0.461	0.224	0.040	0.505	0.041	*
β_{27}	0.241	0.562	0.668	0.253	2.914	0.931	0.307	0.06	*
β_{28}	-0.002	0.001	0.027	0.008	0.003	0.012	0.010	0.002	*
β_{31}	0.254	0.030	*	-0.001	0.128	0.991	0.276	0.025	*
β_{33}	-0.103	0.061	0.092	-0.048	0.351	0.891	-0.092	0.043	0.032
β_{35}	0.058	0.039	0.137	0.010	0.427	0.982	-0.101	0.026	*
β_{37}	-0.008	0.045	0.857	0.000	0.242	0.999	0.067	0.029	0.022
β_{39}	0.013	0.060	0.831	0.013	0.229	0.954	0.081	0.034	0.017
β_{42}	-0.019	0.005	*	-0.010	0.021	0.628	-0.075	0.005	*
β_{43}	-0.064	0.014	*	0.091	0.066	0.167	0.176	0.014	*
β_{44}	0.024	0.006	*	0.000	—	—	0.043	0.002	*
β_{45}	-0.009	0.003	0.003	-0.003	0.017	0.875	0.030	0.003	*
β_{46}	0.284	0.013	*	0.318	0.045	*	0.340	0.011	*
β_{48}	0.018	0.002	*	-0.002	0.012	0.842	0.0250	0.002	*
β_{50}	-0.002	0.030	0.936	0.001	0.122	0.993	-0.062	0.024	0.010
β_{52}	0.430	0.032	*	0.407	0.131	0.002	0.445	0.027	*
β_{53}	2.146	0.284	*	2.146	0.966	0.026	2.149	0.268	*
β_{55}	-1.672	0.178	*	-1.672	0.649	0.010	-1.672	0.176	*
β_{58}	2.469	0.352	*	2.469	1.166	0.034	2.470	0.372	*
β_{60}	2.319	0.269	*	2.319	2.159	0.283	2.319	0.093	*
β_{61}	0.436	0.152	0.004	0.349	1.233	0.777	0.442	0.149	0.003
β_{63}	2.589	0.283	*	2.589	0.968	0.007	2.591	0.293	*
β_{66}	2.371	0.428	*	2.371	1.627	0.145	2.372	0.442	*
β_{69}	-1.342	0.183	*	-1.342	0.681	0.049	-1.342	0.182	*
β_{70}	1.175	0.127	*	1.175	1.152	0.308	1.176	0.082	*
β_{73}	-1.616	0.179	*	-1.616	0.656	0.014	-1.616	0.17	*
β_{75}	0.456	0.031	*	0.507	0.124	*	0.586	0.024	*
β_{76}	-0.096	0.062	0.125	-0.301	0.041	*	-0.352	0.011	*
β_{77}	0.025	0.031	0.416	0.003	0.197	0.988	0.056	0.026	0.030
β_{78}	-0.199	0.062	0.001	-0.182	0.033	*	-0.217	0.012	*
β_{79}	0.084	0.039	0.030	0.087	0.134	0.519	0.105	0.032	0.001
β_{80}	0.049	0.032	0.124	0.060	0.194	0.759	0.090	0.029	0.002
β_{81}	0.030	0.041	0.461	0.005	0.169	0.978	0.092	0.031	0.003
β_{83}	0.004	0.041	0.914	-0.063	0.211	0.767	-0.100	0.027	*
β_{84}	0.608	0.018	*	0.528	0.067	*	0.617	0.015	*

β_{85}	-0.017	0.064	0.786	0.000	—	—	0.035	0.013	0.009
β_{86}	0.007	0.015	0.635	-0.003	0.045	0.938	0.031	0.014	0.026
β_{87}	0.014	0.016	0.354	0.000	0.059	0.994	-0.038	0.01	*
β_{88}	0.840	0.063	*	0.839	0.033	*	0.841	0.015	*
β_{89}	-0.264	0.043	*	-0.304	0.152	0.045	-0.309	0.031	*
β_{92}	0.107	0.035	0.002	0.096	0.274	0.725	0.099	0.032	0.002
β_{93}	-0.106	0.060	0.078	-0.016	0.309	0.958	-0.128	0.026	*

* p-value smaller than 0.001.

Table 4.3: Estimated coefficient, SE and p-value of selected covariates for SDQ-I model by QRFE to $\tau = 0.1, 0.5, 0.9$.

	$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
	Coef.	SE	p-value	Coef.	SE	p-value	Coef.	SE	p-value
β_1	0.153	0.222	0.489	0.157	0.332	0.637	0.164	0.227	0.472
β_2	-0.034	0.020	0.096	-0.015	0.031	0.634	0.023	0.021	0.283
β_3	0.017	0.049	0.720	0.019	0.074	0.795	0.022	0.051	0.668
β_4	0.258	0.224	0.250	0.260	0.337	0.441	0.264	0.217	0.224
β_5	0.057	0.026	0.026	0.062	0.040	0.117	0.068	0.025	0.007
β_6	0.010	0.068	0.886	0.010	0.103	0.924	0.010	0.068	0.887
β_7	-0.076	0.078	0.330	-0.075	0.120	0.530	-0.074	0.079	0.352
β_8	0.196	0.018	*	0.196	0.027	*	0.198	0.019	*
β_9	-0.195	0.022	*	-0.194	0.035	*	-0.194	0.024	*
β_{10}	-0.017	0.057	0.764	-0.015	0.085	0.857	-0.012	0.056	0.827
β_{11}	0.143	0.046	0.002	0.144	0.070	0.039	0.145	0.049	0.003

* p-value smaller than 0.001.

Table 4.4: Estimated coefficient, SE and p-value of selected covariates for SDQ-I model by ALQRFE to $\tau = 0.1, 0.5, 0.9$.

	$\tau = 0.1$			$\tau = 0.5$			$\tau = 0.9$		
	Coef.	SE	p-value	Coef.	SE	p-value	Coef.	SE	p-value
β_5	0.018	0.022	0.409	0.056	0.043	0.196	0.060	0.023	0.009
β_9	-0.001	0.061	0.993	-0.113	0.042	0.007	-0.178	0.027	*
β_{11}	0.000	0.068	0.998	0.134	0.085	0.116	0.104	0.049	0.034
β_{18}	0.001	0.258	0.997	0.000	0.078	0.998	0.159	0.073	0.030
β_{28}	0.000	—	—	-0.005	0.002	0.004	0.000	—	—
β_{34}	0.100	0.033	0.003	0.013	0.036	0.719	0.061	0.017	*
β_{35}	-0.378	0.031	*	-0.336	0.035	*	-0.371	0.019	*
β_{43}	0.211	0.011	*	0.199	0.017	*	0.230	0.009	*
β_{45}	0.002	0.002	0.437	-0.007	0.004	0.071	0.005	0.002	0.011

β_{48}	-0.001	0.002	0.643	0.012	0.003	*	0.000	0.002	0.755
β_{50}	0.100	0.121	0.407	0.135	0.039	0.001	0.049	0.034	0.157
β_{53}	0.397	0.168	0.018	0.375	0.333	0.259	0.364	0.177	0.041
β_{54}	1.660	0.325	*	1.658	0.601	0.006	1.660	0.318	*
β_{55}	0.262	0.108	0.015	0.000	—	—	0.270	0.114	0.018
β_{56}	3.140	0.747	*	3.140	0.069	*	3.140	0.654	*
β_{59}	-1.393	0.141	*	-1.377	0.278	*	-1.391	0.146	*
β_{60}	0.251	0.569	0.659	0.172	0.066	0.009	0.283	0.586	0.629
β_{62}	-1.133	0.156	*	-1.105	0.295	*	-1.131	0.158	*
β_{63}	0.415	0.187	0.026	0.413	0.378	0.275	0.125	0.192	0.513
β_{64}	2.196	0.203	*	2.196	0.396	*	2.196	0.206	*
β_{67}	0.953	0.880	0.279	0.930	0.038	*	0.953	0.617	0.123
β_{71}	2.152	0.293	*	2.151	0.586	*	2.152	0.312	*
β_{72}	2.279	0.294	*	2.279	0.574	*	2.279	0.288	*
β_{74}	1.053	0.556	0.058	1.023	0.094	*	1.053	0.551	0.056
β_{75}	0.072	0.121	0.549	0.145	0.041	*	0.129	0.035	*
β_{77}	0.082	0.107	0.445	-0.075	0.037	0.045	0.043	0.031	0.165
β_{78}	0.446	0.020	*	0.444	0.021	*	0.412	0.008	*
β_{79}	0.000	0.111	0.998	-0.327	0.041	*	-0.339	0.033	*
β_{80}	-0.001	0.104	0.991	0.121	0.039	0.002	0.093	0.031	0.002
β_{81}	-0.091	0.071	0.199	0.002	0.040	0.960	-0.105	0.026	*
β_{82}	0.179	0.013	*	-0.002	0.017	0.886	0.032	0.010	0.001
β_{83}	-0.051	0.072	0.473	-0.080	0.038	0.034	0.009	0.027	0.737
β_{84}	0.003	0.020	0.899	-0.119	0.021	*	-0.152	0.009	*
β_{85}	-0.469	0.023	*	-0.469	0.024	*	-0.455	0.017	*
β_{86}	0.613	0.023	*	0.608	0.027	*	0.611	0.015	*
β_{88}	0.000	0.009	0.990	-0.124	0.011	*	-0.123	0.006	*
β_{89}	0.000	—	—	-0.079	0.044	0.072	-0.084	0.030	0.005
β_{90}	0.086	0.015	*	-0.001	0.022	0.981	0.039	0.014	0.005
β_{92}	0.173	0.030	*	0.174	0.033	*	0.125	0.022	*

* p-value smaller than 0.001.

4.6 Conclusion

We propose a new method to analyze quantile regression for longitudinal data. The ALQRFE generalizes QRFE and enjoys oracle properties. The computational implementation of the new technique is presented here. The Monte Carlo empirical study shows that under the assumption of sparsity and multicollinearity, the ALQRFE method is the

best option. If the data dimension is low and dense, then the procedures are similar. However, if the data dimension is moderate and sparse, then ALQRFE estimates are much superior to others. The BIC seems a valid criterion to evaluate the tuning parameter. We analyze a crucial topic as the two outcomes SDQ-E and SDQ-I present in the massive survey about the new millennium children, the MCS study. Several new predictors are confirmed and display a more complex relation between family constitution, values, education, and income with the two outcomes. However, the socioeconomic issues present in previous works are still relevant for the general discussion.

Chapter 5

Conclusion and perspectives

What is the average effect of air pollutants on the health of children and adolescents from Vitória? In Chapter 2 we combine robust PCA and robust estimation in an LMM in the context where the covariates present outliers and are time-correlated. The respiratory ability of children is impacted by pollutants, mainly the PM_{10} , humidity, and passive smoking.

What is the effect of investment in education and science on the technological qualification of exports from the most advanced, medium and least developed OECD countries? In Chapter 3 we compare several alternative methods to quantile regression for panel data. Under a skew distribution like the yearly exportation, we affirm that MQRFE is the best choice. The estimated education and research and development coefficients are positive to all quantile, i.e., they positively influence the exportation's quality.

What are the most relevant variables to impact the internal and external SDQ scores of the most vulnerable children from Northern Ireland? In Chapter 4 we compare fixed effects with LASSO and adaptive LASSO for moderate dimension data. Under the sparsity assumption, the best method is the ALQRFE. Several new predictors are confirmed and display a more complex relation between family constitution, values, education, and income with the multivariate outcome. However, the socio-economic issues present in previous works are still relevant for the general discussion.

Closing a thesis does not mean the closing of research, and several open problems guide future perspectives. An interesting issue is exploring the boundaries of small sample sizes in both: RLMM and QRFE. It is a crucial step to change the infinities of asymptotic theory to practical values and stress to the minimum limit in both the number of subjects and the number of repetitions. We do not merge LASSO penalization and RLMM on our application because we do not wish to select pollutants, but observe their combined effect. However, to include the LASSO penalization on RLMM to model selection is a motivating approach. As well as to incorporate Cantoni's weights in quantile regression to mitigate the leverage points problem. Because QR is robust against univariate outliers, but not against multivariate extreme observations. The PQRFE and PMQRFE present poor performances under skewness data, maybe adaptive LASSO could improve it. The optimization might be improved. I used Newton-Raphson methods with numerical gra-

dient approximation, which can be replaced by analytical derivatives as they exist for RLMM and quantile regression methods. The MQR requires a tuning parameter c , which is not directly equivalent to the Huber function, thus it deserves a more extensive investigation. Unifying quantile regression theory to random effects, fixed effects, and mixed models is an ambitious goal. Furthermore, if we extend to M-quantile and expectile, and gather these methods in a single R package. Finally, we have ambitious challenges such as developing non-parametric strategies like splines and non-linear models for QRFE.

References

- [1] Emilia Zainal Abidin, Sean Semple, Irniza Rasdi, Sharifah Norkhadijah Syed Ismail, and Jon G. Ayres. The relationship between air pollution and asthma in Malaysian schoolchildren. *Air Qual. Atmos. Health*, 7(4):421–432, 2014.
- [2] Jason Abrevaya. Leapfrog estimation of a fixed-effects model with unknown transformation of the dependent variable. *J. Econom.*, 93(2):203–228, 1999.
- [3] Aye Mengistu Alemu. The Effect of R&D on High-Tech Product Export Competitiveness: Empirical Evidence from Panel Data of East Asian Economies. *STI Policy Review*, 3(1):46–62, 2012.
- [4] Marco Alfò, Maria Francesca Marino, Maria Giovanna Ranalli, Nicola Salvati, and Nikos Tzavidis. M-quantile regression for multivariate longitudinal data with an application to the Millennium Cohort Study. *J. R. Stat. Soc., C*, 70(1):122–146, January 2021.
- [5] Aleksandr Y. Aravkin, Anju Kambadur, Aurélie C. Lozano, and Ronny Luss. Sparse Quantile Huber Regression for Efficient and Robust Estimation. *preprint in ArXiv*, 2014.
- [6] Viorica Baesu, Claudiu Tiberiu Albulescu, Zoltan-Bela Farkas, and Anca Draghici. Determinants of the High-tech Sector Innovation Performance in the European Union: A Review. *Proc. Technol.*, 19:371–378, 2015. 8th International Conference Interdisciplinarity in Engineering, INTER-ENG 2014, 9-10 October 2014, Tirgu Mures, Romania.
- [7] Amadou Diogo Barry, Arthur Charpentier, and Karim Oualkacha. Quantile and Expectile Regression for random effects model. working paper or preprint, 2016.
- [8] Shahab Basiri, Esa Ollila, Gordana Draskovic, and Frédéric Pascal. Fusing Eigenvalues. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4968–4972, Brighton, United Kingdom, 2019. IEEE.
- [9] Annamaria Bianchi and Nicola Salvati. Asymptotic Properties and Variance Estimators of the M-quantile Regression Coefficients Estimators. *Commun. Stat. Theory Methods*, 44(11):2416–2429, 2015.

-
- [10] Jens Breckling and Ray Chambers. M-quantiles. *Biometrika*, 75(4):761–771, 1988.
 - [11] Jens Breckling, Philip Kokic, and Oliver Lubke. A note on multivariate M-quantiles. *Stat. Probab. Lett.*, 55:39–44, 2001.
 - [12] P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer-Verlag, New York, USA, 2 edition, 1991.
 - [13] Peter Bühlmann and Sara van de Geer. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Springer, Berlin, Germany, 2011.
 - [14] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, 16(5):1190–1208, 1995.
 - [15] Eva Cantoni and Elvezio Ronchetti. Robust Inference for Generalized Linear Models. *J. Am. Stat. Assoc.*, 96(455):1022–1030, 2001.
 - [16] Roxanne Connelly and Lucinda Platt. Cohort Profile: UK Millennium Cohort Study (MCS). *Int. J. Epidemiol.*, 43(6):1719–1725, 02 2014.
 - [17] H. H. C Cotta, V. A. Reisen, P Bondon, and F. P. Prezotti. Identification of redundant air quality monitoring stations using robust principal component analysis. *Environ. Model. Assess.*, 25:521–530, 2020.
 - [18] C Croux and G Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87(3):603–618, 09 2000.
 - [19] Martin Crowder. On Consistency and Inconsistency of Estimating Equations. *Econom. Theory*, 2(3):305–330, 1986.
 - [20] Martin Crowder. On Use of a Working Correlation Matrix in Using Generalized Linear Models for Repeated Measures. *Biometrika*, 82:407–410, 1995.
 - [21] Ian Meneghel Danilevich, Pascal Bondon, and Valderio A. Reisen. *alqrfe: Adaptive Lasso Quantile Regression with Fixed Effects*, 2022. R package version 1.1.
 - [22] Ian Meneghel Danilevich, Pascal Bondon, and Valderio A. Reisen. *rlmm: Robust Linear Mixed Models*, 2022. R package version 1.1.
 - [23] Ian Meneghel Danilevich, Valderio A. Reisen, and Pascal Bondon. *pqrfe: Penalized Quantile Regression with Fixed Effects*, 2022. R package version 1.1.
 - [24] Eugene Demidenko. *Mixed models: theory and applications with R*. Wiley, New Jersey, USA, 2 edition, 2013.

-
- [25] P. Diggle. An Approach to the Analysis of Repeated Measurements. *Biometrics*, 44(4):959–971, 1988.
- [26] Dirk Eddelbuettel and James Joseph Balamuta. Extending R with C++: A Brief Introduction to Rcpp. *Am. Stat.*, 72(1):28–36, 2018.
- [27] Dirk Eddelbuettel and Conrad Sanderson. RcppArmadillo: Accelerating R with high-performance C++ linear algebra. *Comput. Stat. Data Anal.*, 71:1054–1063, March 2014.
- [28] Graziella Favarato, H. Ross Anderson, Richard Atkinson, Gary Fuller, Inga Mills, and Heather Walton. Traffic-related pollution and asthma prevalence in children. Quantification of associations with nitrogen dioxide. *Air Qual. Atmos. Health*, 7(4):459–466, 2014.
- [29] Peter Filzmoser and Klaus Nordhausen. Robust linear regression for high-dimensional data: An overview. *WIREs Comput. Stat.*, 13(4):e1524, 2021.
- [30] Anika Flower, Jill Shawe, Judith Stephenson, and Pat Doyle. Pregnancy planning, smoking behaviour during pregnancy, and neonatal outcome: UK millennium cohort study. *BMC Pregnancy Childbirth*, 13:1471–2393, 2013.
- [31] J. S. Friedman, J. Jorgenson, and J. Kramer. An effective bound for the Huber constant for cofinite Fuchsian groups. *Math. Comput.*, 80(274):1163–1163, may 2011.
- [32] Liya Fu and You-Gan Wang. Quantile regression for longitudinal data with a working correlation model. *Comput. Stat. Data Anal.*, 56(8):2526–2538, 2012.
- [33] C. E. Galarza, V. H. Lachos, and D. Bandyopadhyay. Quantile regression in linear mixed models: a stochastic approximation EM approach. *Stat. Interface*, 10(3):471–482, 2017.
- [34] C.E. Galarza, L.M. Castro, F. Louzada, and V. Lachos. Quantile regression for nonlinear mixed effects models: a likelihood based perspective. *Stat. Pap.*, 61:1281–1307, 2020.
- [35] A. Galvão and G. Montes-Rojas. On Bootstrap Inference for Quantile Regression Panel Data: A Monte Carlo Study. *Econometrics*, 3(3):654–666, 2015.
- [36] A. F. Galvão, J. Gu, and S. Volgushev. On the unbiased asymptotic normality of quantile regression with fixed effects. *J. Econom.*, 218(1):178–215, 2020.
- [37] Antonio F. Galvão. Quantile regression for dynamic panel data with fixed effects. *J. Econom.*, 164(1):142–157, 2011. Annals Issue on Forecasting.

-
- [38] Joseph C. Gardiner, Zhehui Luo, and Lee Anne Roman. Fixed effects, random effects and GEE: What are the differences? *Stat. Med.*, 28(2):221–239, 2009.
 - [39] M. Geraci and M. Bottai. Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, 8(1):140–154, 2007.
 - [40] M. Geraci and M. Bottai. Linear quantile mixed models. *Stat. Comput.*, 24:461–479, 2014.
 - [41] Marco Geraci. Estimation of regression quantiles in complex surveys with data missing at random: An application to birthweight determinants. *Stat. Methods Med. Res.*, 25(4):1393–1421, 2016. PMID: 23613490.
 - [42] Paramjit S. Gill. A robust mixed linear model analysis for longitudinal data. *Stat. Med.*, 19(7):975–987, 2000.
 - [43] Xiao Guo and Guang Cheng. Moderate-Dimensional Inferences on Quadratic Functionals in Ordinary Least Squares. *J. Am. Stat. Assoc.*, 0(0):1–20, 2021.
 - [44] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley and Sons, New York, USA, 1986.
 - [45] Denise Hawkes and Ian Plewis. Modelling non-response in the National Child Development Study. *J. R. Stat. Soc., A*, 169(3):479–491, 2006.
 - [46] C.R. Henderson. Estimation of Genetic Parameters. *Ann. Math. Stat.*, 21:309–310, 1950.
 - [47] Joel L. Horowitz and Sokbae (Simon) Lee. Semiparametric estimation of a panel data proportional hazards model with fixed effects. *J. Econom.*, 119(1):155–198, 2004.
 - [48] P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233. Berkeley University of California Press, 1967.
 - [49] Peter J. Huber. Robust Estimation of a Location Parameter. *Ann. Stat.*, 1(53):73–101, 1964.
 - [50] Peter J. Huber. Robust Regression: Asymptotics, Conjectures and Monte Carlo. *Ann. Stat.*, 1(5):799–821, 1973.
 - [51] Peter J. Huber. *Robust Statistics*. Wiley and Sons, New York, US, 1981.

- [52] Mia Hubert, Peter J Rousseeuw, and Karlien Vanden Branden. ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, 47(1):64–79, 2005.
- [53] R. M. Huggins. A Robust Approach to the Analysis of Repeated Measures. *Biometrics*, 49(3):715–720, 1993.
- [54] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *Int. J. Forecast.*, 22(4):679–688, 2006.
- [55] Márton Ispány, Valderio A. Reisen, Glauro C. Franco, Pascal Bondon, Higor H. A. Cotta, Paulo R. P. Filho, and Faradiba S. Serpa. On generalized additive models with dependent time series covariates. In Rojas, I., Pomares, H., Valenzuela, and O., editors, *Time Series and Forecasting: Contributions to Statistics*, Contributions to statistics, pages 289–308. Springer series, 2018.
- [56] Jiming Jiang. *Asymptotic Analysis of Mixed Effects Models: Theory, Applications, and Open Problems*. CRC Press, Taylor & Francis Group, New York, USA, 2017.
- [57] Jiming Jiang. *Robust Mixed Model Analysis*. World Scientific Publishing Company, 2019.
- [58] Jiming Jiang and Weihong Zhang. Robust estimation in generalised linear mixed models. *Biometrika*, 88(3):753–765, 10 2001.
- [59] M. C. Jones. Expectiles and M-quantiles are quantiles. *Stat. Probab. Lett.*, 20(2):149–153, 1994.
- [60] Kengo Kato, Antonio F. Galvao, and Gabriel V. Montes-Rojas. Asymptotics for panel quantile regression models with individual effects. *J. Econom.*, 170(1):76–91, 2012.
- [61] Yu Keming and Zhang Jin. A Three-Parameter Asymmetric Laplace Distribution and Its Extension. *Commun. Stat. Theory Methods*, 34(9-10):1867–1879, 2005.
- [62] Joonpyo Kim and Hee-Seok Oh. Pseudo-quantile functional data clustering. *J Multivar Anal*, 178:104626, 2020.
- [63] Keith Knight. Limiting Distributions for L_1 Regression Estimators under General Conditions. *Ann. Stat.*, 26:755–70, 1998.
- [64] Keith Knight and Wenjiang Fu. Asymptotics for Lasso-Type Estimators. *Ann. Stat.*, 28:1356–1378, 2000.
- [65] R. Koenker. Quantile regression for longitudinal data. *J. Multivar. Anal.*, 91(1):74–89, 2004.

-
- [66] R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica*, 46:33–50, 1978.
 - [67] Roger Koenker. *Quantile Regression*. Cambridge University Press, 2005.
 - [68] P. Kokic, R. Chambers, J. Breckling, and S. Beare. A Measure of Production Performance. *J. Bus. Econ. Stat.*, 15(4):445–451, 1997.
 - [69] Philip Kokic, Jens Breckling, and Oliver Lubke. *A New Definition of Multivariate M-quantiles*, pages 15–22. Birkhäuser, Basel, Switzerland, 2002.
 - [70] Manuel Koller. *Robust Estimation of Linear Mixed Models*. Ph.d., ETH, Zurich, Switzerland, 2013.
 - [71] Manuel Koller and Werner A. Stahel. Sharpening Wald-type inference in robust regression for small samples. *Comput. Stat. Data Anal.*, 55(8):2504–2515, 2011.
 - [72] Nan M. Laird and James H. Ware. Random-Effects Models for Longitudinal Data. *Biometrics*, 38(4):963–974, 1982.
 - [73] Carlos Lamarche. Robust penalized quantile regression estimation for panel data. *J. Econom.*, 157(2):396–408, 2010.
 - [74] Carlos Lamarche and Thomas Parker. Wild Bootstrap Inference for Penalized Quantile Regression for Longitudinal Data, 2021. working paper or preprint.
 - [75] Eun Ryung Lee, Hohsuk Noh, and Byeong U. Park. Model Selection via Bayesian Information Criterion for Quantile Regression Models. *J. Am. Stat. Assoc.*, 109(505):216–229, 2014.
 - [76] Dongyu Li, Lei Wang, and Weihua Zhao. Estimation and inference for multikink expectile regression with longitudinal data. *Stat. Med.*, 41(7):1296–1313, 2022.
 - [77] H. Li, Y. Liu, and Y. Luo. Double Penalized Quantile Regression for the Linear Mixed Effects Model. *J. Syst. Sci. Complex.*, 33:2080–2102, 2020.
 - [78] Po-Ling Loh and Martin J. Wainwright. Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima. *J. Mach. Learn. Res.*, 16(19):559–616, 2015.
 - [79] Yanyuan Ma and Marc G. Genton. Highly robust estimation of the autocovariance function. *J. Time Ser. Anal.*, 21(6):663–684, 2000.
 - [80] Yanyuan Ma and Marc G. Genton. Highly Robust Estimation of Dispersion Matrices. *J. Multivar. Anal.*, 78(1):11–36, 2001.
 - [81] Maria Marino and Alessio Farcomeni. Linear quantile regression models for longitudinal experiments: An overview. *Metron*, 73:229–247, 08 2015.

-
- [82] Ricardo A. Maronna, R. Douglas Martin, and Víctor J. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, New Jersey, USA, 2006.
- [83] Josmar Mazucheli, Victor Leiva, Bruna Alves, and Andre Menezes. A New Quantile Regression for Modeling Bounded Data under a Unit Birnbaum–Saunders Distribution with Applications in Medicine and Politics. *Symmetry*, 13(4):1–21, 09 2020.
- [84] Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.*, 34(3):1436 – 1462, 2006.
- [85] Arta Mulliqi. The role of education in explaining technology-intensive exports: a comparative analysis of transition and non-transition economies. *East. J. Eur. Stud.*, 12:141–172, 2021.
- [86] W. Newey and J. Powell. Asymmetric Least Squares Estimation and Testing. *Econometrica*, 55(4):819–847, 1987.
- [87] Michael Noble, Gemma Wright, George Smith, and Chris Dibben. Measuring Multiple Deprivation at the Small-Area Level. *Environ. Plan A*, 38(1):169–185, 2006.
- [88] Mohamed Ouhourane, Yi Yang, Andrea L. Benedet, and Karim Oualkacha. Group penalized quantile regression. *Stat. Methods Appt.*, 2021.
- [89] Jesus Peña-Vinces and David Audretsch. Tertiary education and science as drivers of high-technology exporting firms growth in developing countries. *J. Technol. Transf.*, 46:1734–1757, 2021.
- [90] José C. Pinheiro and Douglas M. Bates. *Mixed-Effect Models in S and S-plus*. Springer-Verlag New York, 2000.
- [91] Ian Plewis. Non-Response in a Birth Cohort Study: The Case of the Millennium Cohort Study. *Int. J. Soc. Res. Methodol.*, 10(5):325–334, 2007.
- [92] Ian Plewis, Sosthenes Ketende, Heather Joshi, and Gareth Hughes. The Contribution of Residential Mobility to Sample Loss in a Birth Cohort Study: Evidence from the First Two Waves of the UK Millennium Cohort Study. *J. Off. Statist.*, 24(3):365–385, 09 2008.
- [93] Monica Pratesi, M. Giovanna Ranalli, and Nicola Salvati. Nonparametric M-quantile regression using penalised splines. *J. Nonparametr. Stat.*, 21(3):287–304, 2009.
- [94] A. Qu, B. Lindsay, and B. Li. Improving Generalised Estimating Equations Using Quadratic Inference Functions. *Biometrika*, 87(4):823–836, 2000.

-
- [95] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.
- [96] V. A. Reisen, C. Lévy-Leduc, H. A. C. Cotta, T. T. A. Albuquerque, and W. Stummer. Long-memory model under outliers: An application to air pollution levels. In Prashant Kumar, Bhola R Gurjar, and J.N. Govil, editors, *Air and Noise Pollution*, volume 3 of *Environmental Science and Engineering*, pages 211–243. Springer, Studium Press LLC, USA, 2017.
- [97] Alice M. Richardson and Alan H. Welsh. Robust Restricted Maximum Likelihood in Mixed Linear Models. *Biometrics*, 51(4):1429–1439, 1995.
- [98] G. K. Robinson. That BLUP is a Good Thing: The Estimation of Random Effects. *Stat. Sci.*, 6(1):15 – 32, 1991.
- [99] Peter Rousseeuw and B. C. van Zomeren. Unmasking Multivariate Outliers and Leverage Points. *J. Am. Stat. Assoc.*, 85(411):633–639, 1990.
- [100] N. Salvati, M. G. Ranalli, and M. Pratesi. Small area estimation of the mean using non-parametric M-quantile regression: a comparison when a linear mixed model does not hold. *J. Stat. Comput. Simul.*, 81(8):945–964, 2011.
- [101] Sabine Schnabel and Paul Eilers. Optimal expectile smoothing. *Comput. Stat. Data Anal.*, 53:4168–4177, 10 2009.
- [102] Faradiba Sarquis Serpa. *Modelo Linear Misto com Interações e Componentes Principais para Avaliar o Efeito Múltiplo de Poluentes e Variáveis Climáticas na Saúde Respiratória*. Ph.d., Universidade Federal do Espírito Santo, Vitória, Brazil, 2019.
- [103] G. Shevlyakov and P. Smirnov. Robust Estimation of the Correlation Coefficient: An Attempt of Survey. *Austrian J. Stat.*, 40:147–156, 2011.
- [104] Sanjoy K. Sinha. Robust Analysis of Generalized Linear Mixed Models. *J. Am. Stat. Assoc.*, 99(466):451–460, 2004.
- [105] Sanjoy K. Sinha. Robust inference in generalized linear models for longitudinal data. *Can. J. Stat.*, 34(2):261–278, 2006.
- [106] Sanjoy K. Sinha and J. N. K. Rao. Robust small area estimation. *Can. J. Stat.*, 37(3):381–399, 2009.
- [107] Luke B. Smith, Montserrat Fuentes, Penny Gordon-Larsen, and Brian J. Reich. Quantile regression for mixed models with an application to examine blood pressure trends in China. *Ann. Appl. Stat.*, 9(3):1226 – 1246, 2015.

-
- [108] Ezequiel Smucler and Victor Yohai. Robust and sparse estimators for linear regression models. *Comput. Stat. Data Anal.*, 111:116–130, 08 2017.
- [109] Juliana Souza, Valderio Reisen, Glauro Franco, Márton Ispány, Pascal Bondon, and Jane Meri Santos. Generalized additive model with principal component analysis: An application to time series of respiratory disease and air pollution data. *J. R. Stat. Soc. Ser. C Appl. Stat.*, 67(2):453–480, 2018.
- [110] Dana Šumilo, Jennifer J Kurinczuk, Maggie E Redshaw, and Ron Gray. Association between limiting longstanding illness in mothers and their children: findings from the UK Millennium Cohort Study. *BMJ Open*, 3(12), 2013.
- [111] Edinaldo Tebaldi. The determinants of high-technology exports: A panel data analysis. *Atl. Econ. J.*, 39, 12 2011.
- [112] George D. Thurston, Howard Kipen, Isabella Annesi-Maesano, John Balmes, Robert D. Brook, Kevin Cromar, Sara De Matteis, Francesco Forastiere, Bertil Forsberg, Mark W. Frampton, Jonathan Grigg, Dick Heederik, Frank J. Kelly, Nino Kuenzli, Robert Laumbach, Annette Peters, Sanjay T. Rajagopalan, David Rich, Beate Ritz, Jonathan M. Samet, Thomas Sandstrom, Torben Sigsgaard, Jordi Sunyer, and Bert Brunekreef. A joint ERS/ATS policy statement: what constitutes an adverse health effect of air pollution? An analytical framework. *Eur. Respir. J.*, 49:1600419, 2017.
- [113] Yuzhu Tian, Liyong Wang, Manlai Tang, and Maozai Tian. Weighted composite quantile regression for longitudinal mixed effects models with application to AIDS studies. *Commun. Stat. Simul. Comput.*, 50(6):1837–1853, 2021.
- [114] Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc., B*, 58(1):267–288, 1996.
- [115] Q. Q. Tiet, H. R. Bird, M. Davies, C. Hoven, P. Cohen, P. S. Jensen, and S. Goodman. Adverse life events and resilience. *J. Am. Acad. Child Adolesc. Psychiatry*, 37(11):1191–1200, 1998.
- [116] Kristiane Tommerup and Rebecca E. Lacey. Maternal and Paternal Distress in Early Childhood and Child Adiposity Trajectories: Evidence from the Millennium Cohort Study. *Obesity*, 29(5):888–899, 2021.
- [117] Ufuk Turen and Yunus Gökmen. The determinants of high technology exports volume: A panel data analysis of eu-15 countries. *Int. J. Econ. Manag. Soc. Sci.*, 2(3):217–232, 01 2013.

- [118] N. Tzavidis, N. Salvati, T. Schmid, E. Flouri, and E. Midouhas. Longitudinal analysis of the strengths and difficulties questionnaire scores of the Millennium Cohort Study children in England using M-quantile random-effects regression. *J. R. Stat. Soc., A*, 179(2):427–452, 2016.
- [119] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, USA, 4 edition, 2002.
- [120] Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, New York, USA, 2000.
- [121] J. Wagener, S. Volgushev, and H. Dette. The quantile process under random censoring. *Math. Methods Stat.*, 21:1934–8045, 2012.
- [122] Martin J Wainwright. *High-dimensional statistics: a non-asymptotic viewpoint*. Cambridge series on statistical and probabilistic mathematics 48. Cambridge University Press, Cambridge, UK, 2019.
- [123] L. Schulze Waltrup and G. Kauermann. Smooth expectiles for panel data using penalized splines. *Stat. Comput.*, 27:271–282, 2017.
- [124] Linda Schulze Waltrup, Fabian Sobotka, Thomas Kneib, and Goran Kauermann. Expectile and quantile regression - David and Goliath? *Stat. Modelling*, 15(5):433–456, 2015.
- [125] Huixia Judy Wang, Deyuan Li, and Xuming He. Estimation of High Conditional Quantiles for Heavy-Tailed Distributions. *J. Am. Stat. Assoc.*, 107(500):1453–1464, 2012.
- [126] Yaping Wang and H. Pham. Analyzing the effects of air pollution and mortality by generalized additive models with robust principal components. *Int. J. Syst. Assur. Eng. Manag.*, 2:253–259, 2011.
- [127] You-Gan Wang, Xu Lin, Min Zhu, and Zhidong Bai. Robust Estimation Using the Huber Function with a Data-Dependent Tuning Constant. *J. Comput. Graph. Stat.*, 16(2):468–481, 2007.
- [128] R. W. M. Wedderburn. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, 61(3):439–447, 12 1974.
- [129] William S. Wei. *Time Series Analysis: Univariate and Multivariate Methods*. Pearson, London, UK, 2 edition, 01 2006.
- [130] Stefan Wellek. A critical evaluation of the current “p-value controversy”. *Biom. J.*, 59(5):854–872, 2017.

-
- [131] A. H. Welsh and A. M. Richardson. Approaches to the robust estimation of mixed models. In *Handbook of Statistics*, volume 15, chapter 13, pages 343–384. Elsevier, Amsterdam, 1997.
- [132] Cort J. Willmott and Kenji Matsuura. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.*, 30(1):79–82, 2005.
- [133] World Health Organization. *WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: global update 2005: summary of risk assessment*. Genève, Switzerland, 2006.
- [134] Yichao Wu and Yufeng Liu. Variable selection in quantile regression. *Stat. Sin.*, 19(2):801–817, 04 2009.
- [135] Kelvin K. W. Yau and Anthony Y. C. Kuk. Robust estimation in generalized linear mixed models. *J. R. Stat. Soc., B*, 64(1):101–117, 2002.
- [136] Victor J. Yohai and Ricardo A. Maronna. Asymptotic Behavior of M-Estimators for the Linear Model. *Ann. Statist.*, 7(2):258–268, 1979.
- [137] Jinhong You and Xian Zhou. Asymptotic theory in fixed effects panel data seemingly unrelated partially linear regression models. *Econ. Theory*, 30(2):407–435, 04 2014.
- [138] Ming Yuan and Yi Lin. Model Selection and Estimation in Regression With Grouped Variables. *J. R. Stat. Soc., B*, 68:49–67, 02 2006.
- [139] Bartolomeu Zamprogno, Valdério A. Reisen, Pascal Bondon, Higor H. Aranda Cotta, and Neyval C. Reis Jr. Principal component analysis with autocorrelated data. *J. Stat. Comput. Simul.*, 90(12):2117–2135, 2020.
- [140] Hui Zou. The Adaptive Lasso and Its Oracle Properties. *J. Am. Stat. Assoc.*, 101(476):1418–1429, 2006.
- [141] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc., B*, 67(2):301–320, 2005.
- [142] Hui Zou and Ming Yuan. Composite quantile regression and the oracle model selection theory. *Ann. Statist.*, 36(3):1108–1126, 2008.

Appendix A

A.1 Correction factors

We establish explicit expressions of λ and ϱ in terms of κ when $\psi = \rho_2$ and $\psi = \rho_3$. Let $R \sim \mathcal{N}(0, 1)$ and f denote the probability density of R . Let $\kappa > 0$, $n \in \mathbb{N}$ and

$$I_n = \int_{-\kappa}^{\kappa} r^{2n} f(r) dr.$$

Then, $I_0 = \mathbb{P}(|R| \leq \kappa)$ and since $f'(r) = -rf(r)$, we have for $n \geq 1$,

$$I_n = - \int_{-\kappa}^{\kappa} r^{n-1} f'(r) dr = u_n + (2n-1)I_{n-1},$$

where $u_n = -2\kappa^{2n-1}f(\kappa)$. By iteration, we get

$$I_n = u_n + (2n-1)u_{n-1} + (2n-1)(2n-3)u_{n-2} + \cdots + 3 \cdot 5 \cdots (2n-1)(u_1 + I_0).$$

It was shown by [42] that

$$\lambda_2 = \mathbb{E}(R\psi_2(R)) = I_0.$$

Let

$$\begin{aligned} \lambda_3 &= \mathbb{E}[R\psi_3(R)] = \int_{-\kappa}^{\kappa} r^2 \left(1 - \left(\frac{r}{\kappa}\right)^2\right)^2 f(r) dr = \int_{-\kappa}^{\kappa} \left(r^2 - \frac{2r^4}{\kappa^2} + \frac{r^6}{\kappa^4}\right) f(r) dr \\ &= I_1 - \frac{2I_2}{\kappa^2} + \frac{I_3}{\kappa^4} = -\frac{2f(\kappa)}{\kappa} \left(\frac{15}{\kappa^2} - 1\right) + \left(1 - \frac{6}{\kappa^2} + \frac{15}{\kappa^4}\right) I_0. \end{aligned}$$

We have

$$\psi_2'(r) = \frac{d^2 \rho_2(r)}{dr^2} = \begin{cases} 1 & \text{if } |r| \leq \kappa, \\ 0 & \text{if } |r| > \kappa, \end{cases}$$

and

$$\psi_3'(r) = \frac{d^2 \rho_3(r)}{dr^2} = \begin{cases} \frac{5r^4}{\kappa^4} - \frac{6r^2}{\kappa^2} + 1 & \text{if } |r| \leq \kappa, \\ 0 & \text{if } |r| > \kappa. \end{cases}$$

Then

$$\begin{aligned} \mathbb{E}(\psi_2'(R)) &= I_0, \\ \mathbb{E}(\psi_2(R)^2) &= \int_{-\kappa}^{\kappa} r^2 f(r) dr + \kappa^2 \left[\int_{-\infty}^{-\kappa} f(r) dr + \int_{\kappa}^{\infty} f(r) dr \right] \\ &= I_1 + \kappa^2(1 - I_0) = \kappa^2 - 2\kappa f(\kappa) + (1 - \kappa^2)I_0, \end{aligned}$$

$$\begin{aligned}\mathbb{E}(\psi'_3(R)) &= \int_{-\kappa}^{\kappa} \left(\frac{5r^4}{\kappa^4} - \frac{6r^2}{\kappa^2} + 1 \right) f(r) dr = \frac{5I_2}{\kappa^4} - \frac{6I_1}{\kappa^2} + I_0 \\ &= \frac{2f(\kappa)}{\kappa} \left(1 - \frac{15}{\kappa^2} \right) + \left(1 - \frac{6}{\kappa^2} + \frac{15}{\kappa^4} \right) I_0,\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}(\psi_3(R)^2) &= \int_{-\kappa}^{\kappa} \left(\frac{r^{10}}{\kappa^8} - \frac{4r^8}{\kappa^6} + \frac{6r^6}{\kappa^4} - \frac{4r^4}{\kappa^2} + r^2 \right) f(r) dr = \frac{I_5}{\kappa^8} - \frac{4I_4}{\kappa^6} + \frac{6I_3}{\kappa^4} - \frac{4I_2}{\kappa^2} + I_1 \\ &= \frac{2f(\kappa)}{\kappa} \left(1 - \frac{13}{\kappa^2} + \frac{105}{\kappa^4} - \frac{945}{\kappa^6} \right) + \left(1 - \frac{12}{\kappa^2} + \frac{90}{\kappa^4} - \frac{420}{\kappa^6} + \frac{945}{\kappa^8} \right) I_0.\end{aligned}$$

The explicit expressions of ϱ when $\psi = \psi_2$ and $\psi = \psi_3$ follow easily from (2.7) and the above calculations.

A.2 Weighted robust linear mixed models: additional simulation

A.2.1 Proposed simulation

In this section, a simulation study with finite sample sizes of weak-stationary time series is conducted to verify the performance of the estimators in LMM under the scenario in which the covariates are time and cross-correlated with and without AO. The choice of the use of this type of outliers is because, in terms of times series data, these are more dangerous in terms of destroying the sample properties of the mean and variance among others sample statistical quantities [96]. As previously addressed, the pollution series may present observations with high levels of pollutant concentrations which may produce sample densities with heavy tails, and these observations can be identified as AO. For this purpose, we consider the following Model (A.1),

$$Y_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \gamma_i + U_{ij}, \quad (\text{A.1})$$

where, for all $i = 1, \dots, n$ and $j = 1, \dots, m_i$, $\mathbf{X}_{ij} = (X_{ij}^{(0)}, X_{ij}^{(1)}, X_{ij}^{(2)})^T$ with $X_{ij}^{(0)} = 1$ and $\mathbf{X}_{ij}^{(1,2)} = (X_{ij}^{(1)}, X_{ij}^{(2)})^T$ forms a VAR(1) Gaussian model with 2×2 diagonal auto-regressive coefficient matrix Ψ_1 such that $\phi_{1,1} = -0.3$, $\phi_{1,2} = 0.5$, $\phi_{2,1} = 0.2$ and $\phi_{2,2} = -0.2$ or $\Psi_0 = \mathbf{0}$, and the innovations have a covariance matrix with $\sigma_{1,1} = 1$, $\sigma_{1,2} = 0, 0.5, 0.9$, $\sigma_{2,2} = 1$ see [129, pages 387-394]. $\boldsymbol{\beta} = (2, -0.5, 1)^T$, $\gamma_i \sim \mathcal{N}(0, D)$ is a scalar random intercept, and $U_{ij} \sim \text{AR}(1)$ with parameters $\sigma_e^2 = 1$ and $\nu = 0, 0.15$. We consider $n = 20, 40, 80$ and $m_i = 5, 50$. As previously mentioned, $\kappa_2 = 1.345$ and $\kappa_3 = 4.685$.

To verify the performance of the estimators under the contaminated process, the variable $X_{ij}^{(2)}$ with AO is defined as

$$X_{ij}^{(2)*} = X_{ij}^{(2)} + \mu B_{ij} K_{ij},$$

where B_{ij} is a Bernoulli RV with probability of outlier's occurrences $\delta \geq 0$ and K_{ij} follows a t -Student distribution with 3 degrees of freedom. B_{ij} and K_{ij} are mutually independent for all $i = 1, \dots, n$ and $j = 1, \dots, m_i$, and $\mu = 0.06, 0.6$ is the magnitude of AO.

1. Small sample size:

Here we explore small sample sizes and autoregressive errors, so $n = 20, 40$, $m_i = 5$, correlation between X_1 and X_2 is $\sigma_{1,2} = 0, 0.5, 0.9$, $\Psi = \Psi_0$, probability of contamination equal to $\delta = 0, 0.02, 0.04, 0.06, 0.08, 0.1$ and $\mu = 0.6$. The parameter vector is $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, D, \sigma_e^2, \nu)^T$ with true value $\boldsymbol{\theta}_0 = (2, -0.5, 1, 1, 1, 0.15)^T$. So the goal is to analysis very small sample sizes like Maria Helena has interest and the cross effect of outliers and correlation.

- a) Scenario 1.1: $n = 20$, $m_i = 5$, $\sigma_{1,2} = 0$;
- b) Scenario 1.2: $n = 20$, $m_i = 5$, $\sigma_{1,2} = 0.5$;
- c) Scenario 1.3: $n = 20$, $m_i = 5$, $\sigma_{1,2} = 0.9$;
- d) Scenario 1.4: $n = 40$, $m_i = 5$, $\sigma_{1,2} = 0$;
- e) Scenario 1.5: $n = 40$, $m_i = 5$, $\sigma_{1,2} = 0.5$;
- f) Scenario 1.6: $n = 40$, $m_i = 5$, $\sigma_{1,2} = 0.9$;

2. Large sample size with Time Series:

Here we explore large sample sizes, time series and independent errors, so $n = 20, 80$, $m_i = 50$, $\sigma_{1,2} = 0$, $\Psi = \Psi_0, \Psi_1$, probability of contamination equal to $\delta = 0, 0.025, 0.05, 0.075, 0.1$ and $\mu = 0.06$. The parameter vector is $\boldsymbol{\theta} = (\beta_0, \beta_1, \beta_2, D, \sigma_e^2)^T$ with true value $\boldsymbol{\theta}_0 = (2, -0.5, 1, 1, 1)^T$. So the goal is to analysis large sample sizes like Liudas has interest and the cross effect of outliers and time series.

- a) Scenario 2.1: $n = 20$, $m_i = 50$, Ψ_0 ;
- b) Scenario 2.2: $n = 20$, $m_i = 50$, Ψ_1 ;
- c) Scenario 2.3: $n = 80$, $m_i = 50$, Ψ_0 ;
- d) Scenario 2.4: $n = 80$, $m_i = 50$, Ψ_1 ;

A.2.2 Results

A.2.3 Conclusion of simulation

1. Small sample size:

The SMSE and SCP of $\hat{\beta}_0$, $\hat{D}$, $\hat{\sigma}_e^2$ and $\hat{\nu}$ are flat and more or less similar between all estimator methods for all scenarios. The $\hat{\mu}(\hat{\sigma}_e^2) \approx 0.92$ and $\hat{\mu}(\hat{\nu}) \approx 0.07$, both below the true value, but their SCP are satisfactory. The SMSE for $n = 40$ is always smaller than for $n = 20$ for all parameters, thus we observe a favorable result

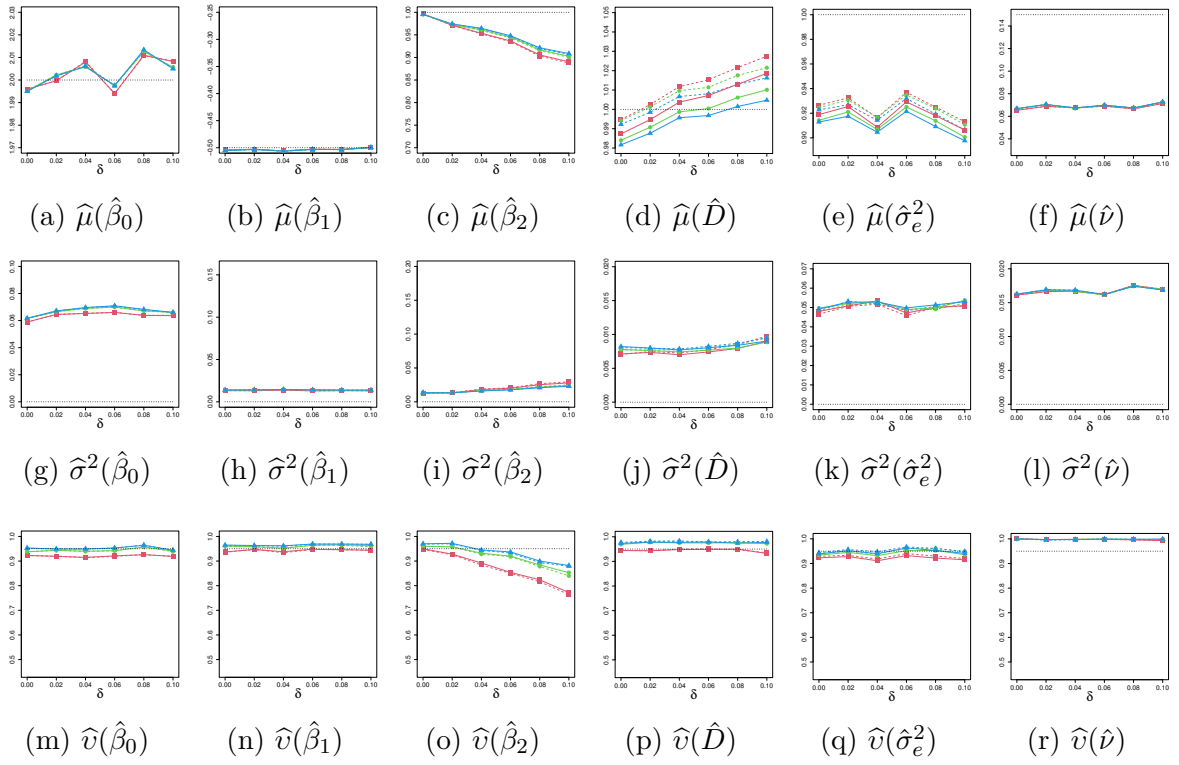


Figure A.1: The SM, SMSE and SCP with 95% of confidence of estimates on scenario 1.1. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.

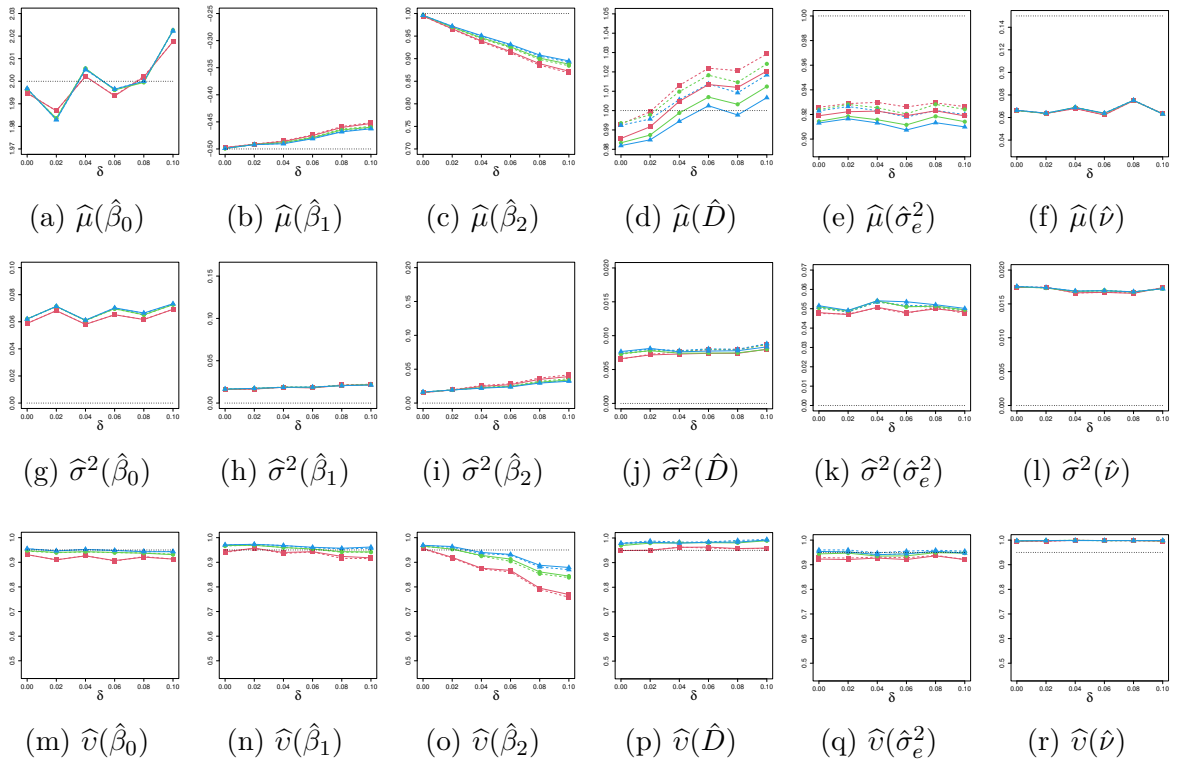


Figure A.2: The SM, SMSE and SCP with 95% of confidence of estimates on scenario 1.2. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.

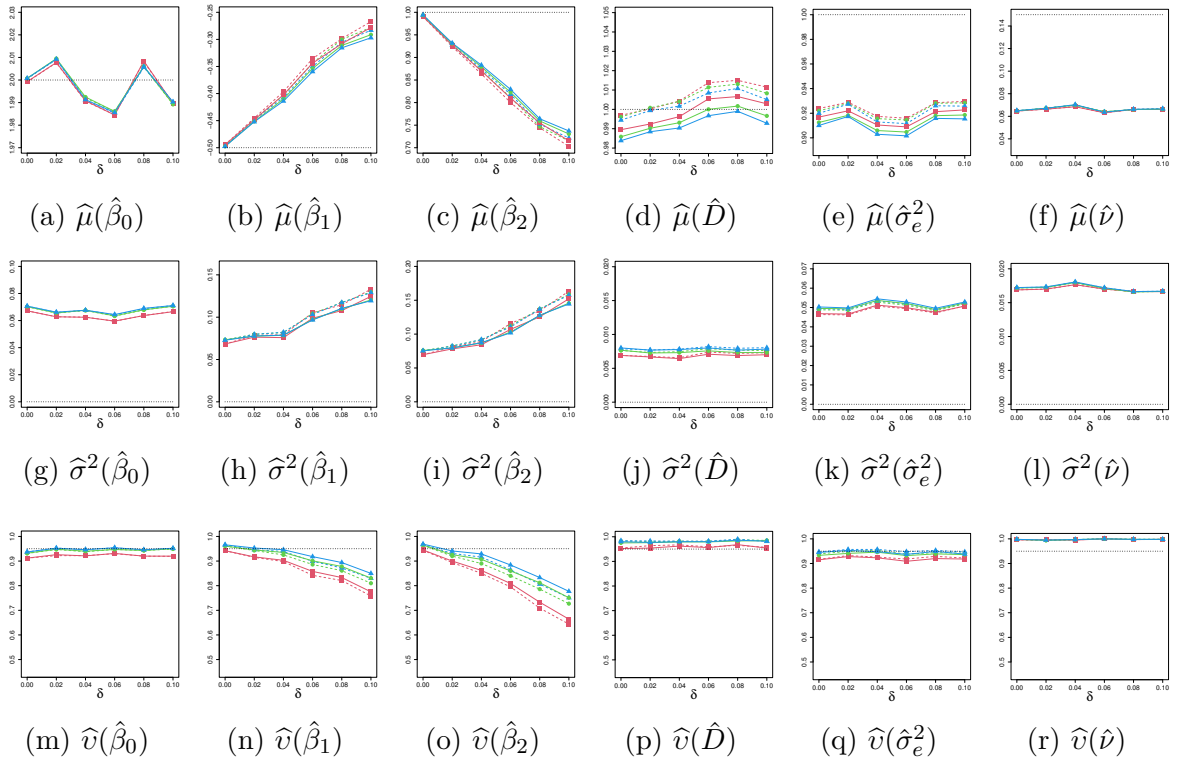


Figure A.3: The SM, SMSE and SCP with 95% of confidence of estimates on scenario 1.3. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.

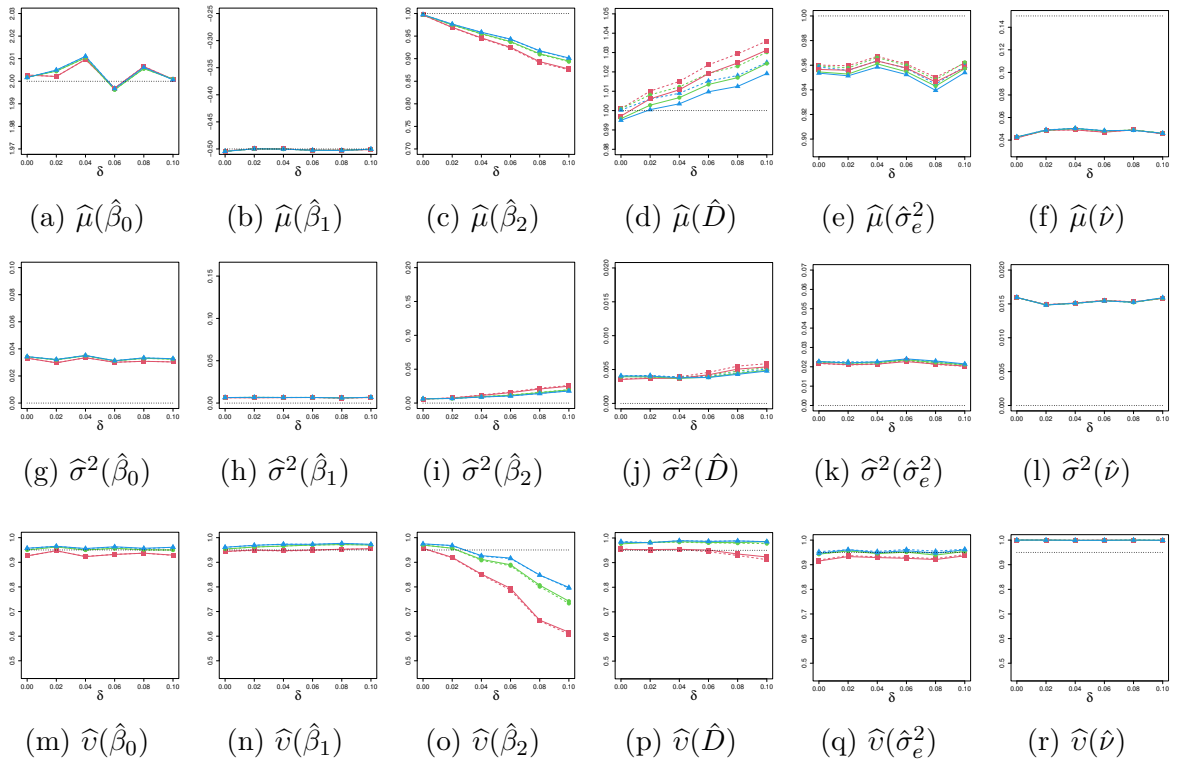


Figure A.4: The SM, SMSE and SCP with 95% of confidence of estimates on scenario 1.4. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.

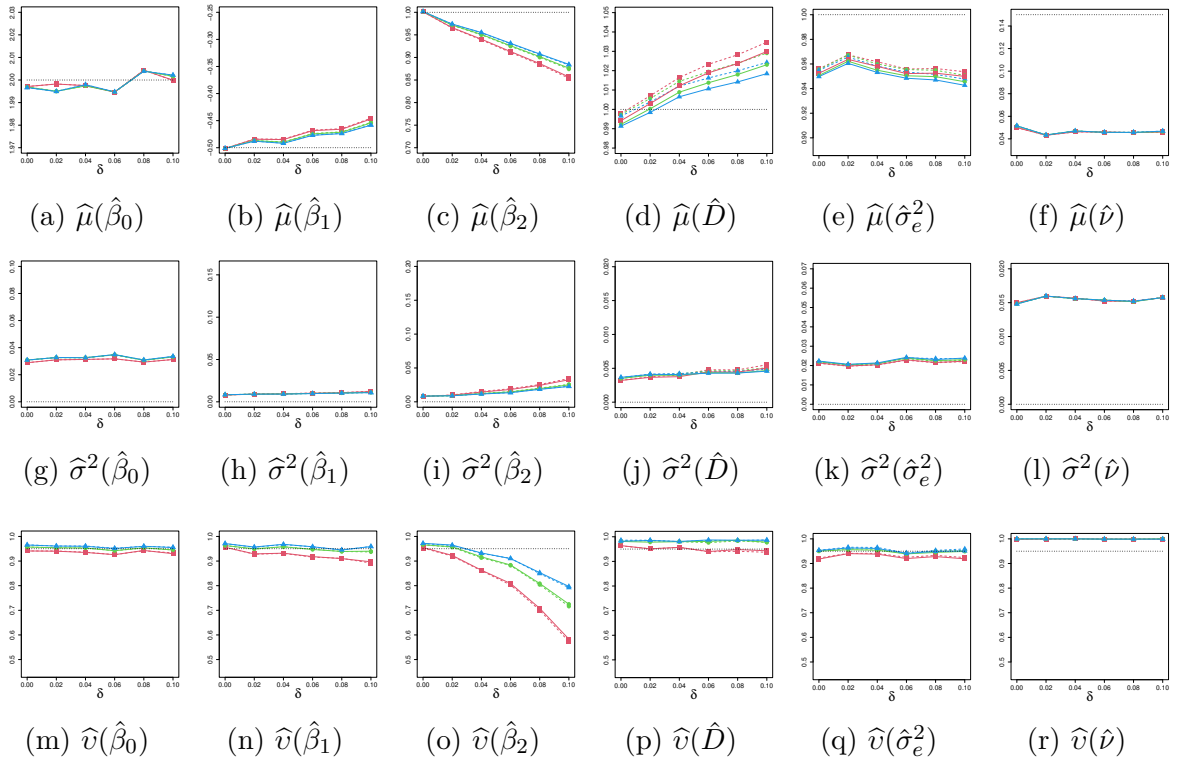


Figure A.5: The SM, SMSE and SCP with 95% of confidence of estimates on scenario 1.5. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.

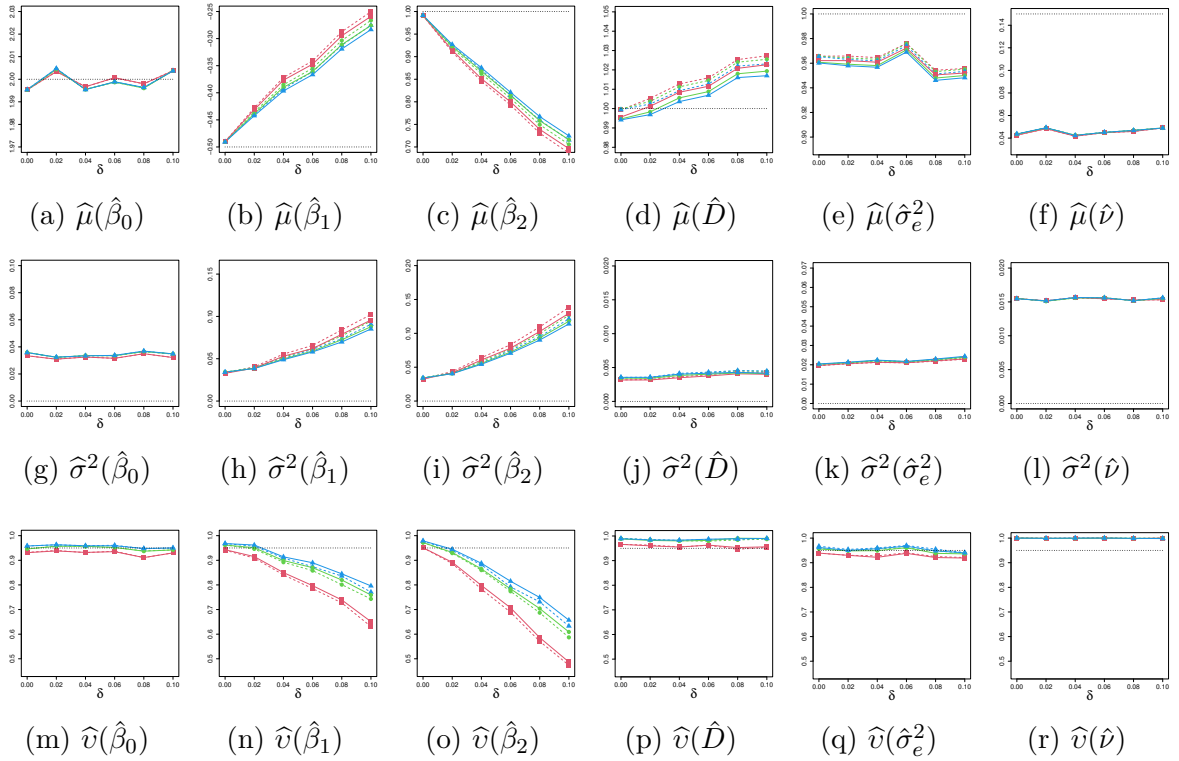


Figure A.6: The SM, SMSE and SCP with 95% of confidence of estimates on scenario 1.6. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.

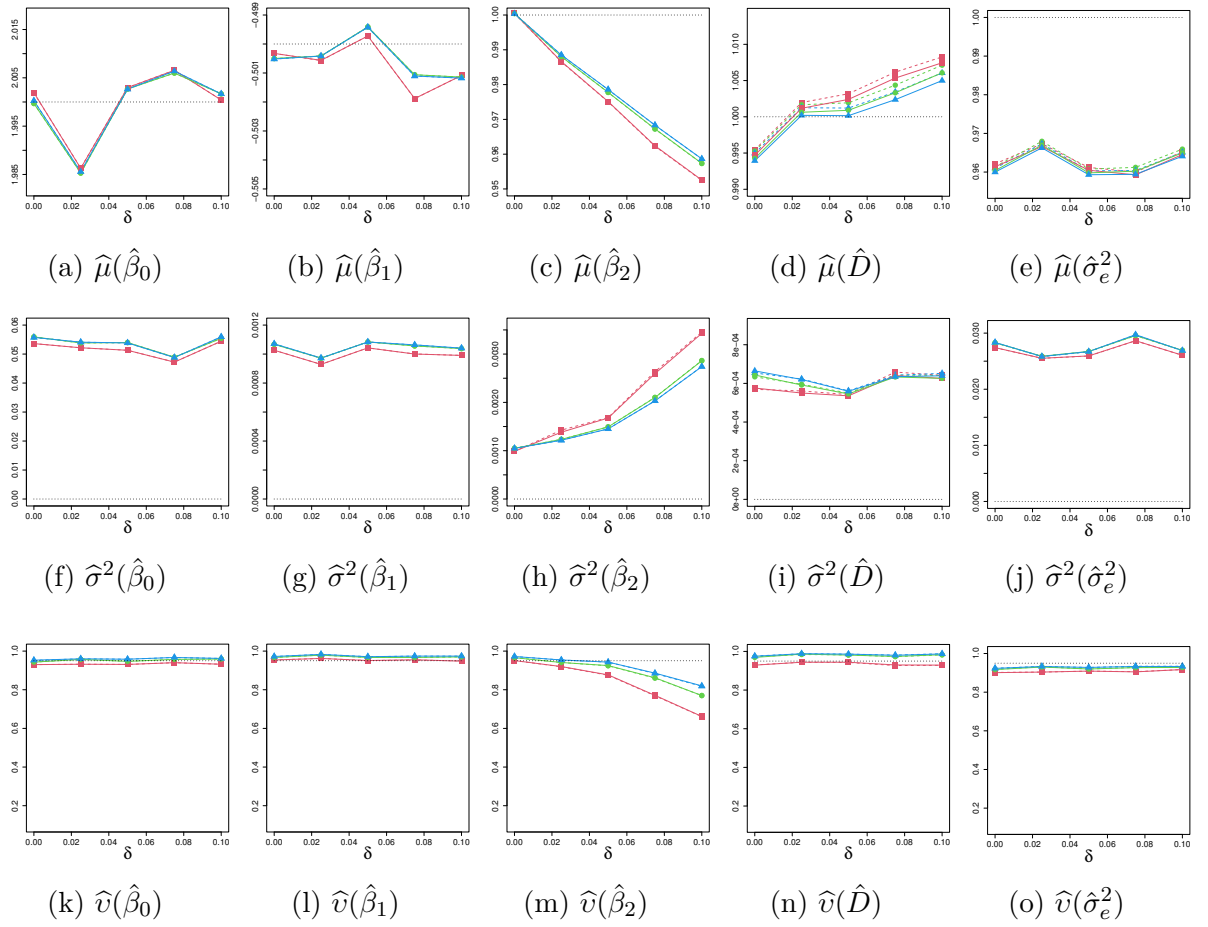


Figure A.7: The SM, SMSE and SCP with 95% of confidence of estimates on scenario 2.1. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed. Time series with $m_i = 50$

about asymptotic convergence. When we analyze scenario 3 (with high correlation and $n = 20$) weighted Tukey provides the best mean, SMSE and SCP for $\hat{\beta}_1$ and $\hat{\beta}_2$. When we analyze scenario 6 (with high correlation and $n = 40$) weighted and non-weighted Tukey provide the best mean, SMSE and SCP for $\hat{\beta}_1$ and $\hat{\beta}_2$. Though, robust loss function improve the solutions and weighted method helps in very small sample sizes.

2. Large sample size with Time Series:

The mean, SMSE and SCP of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}_e^2$ are flat and more or less similar between all estimator methods for all scenarios. The SMSE for $n = 80$ is always smaller than for $n = 20$ for all parameters, thus we observe a favorable result about asymptotic convergence. If we compare scenario 2.1 (Ψ_0) with 2.2 (Ψ_1) and 2.3 (Ψ_0) with 2.4 (Ψ_1), there is almost no difference. Also weighted and non-weighted methods provide very similar results as $N = 1,000$ and $N = 4,000$ are both large sample sizes. For all scenarios Tukey provides the best results in terms of mean, SMSE and SCP. Though, robust loss function improve the solutions and weighted method is

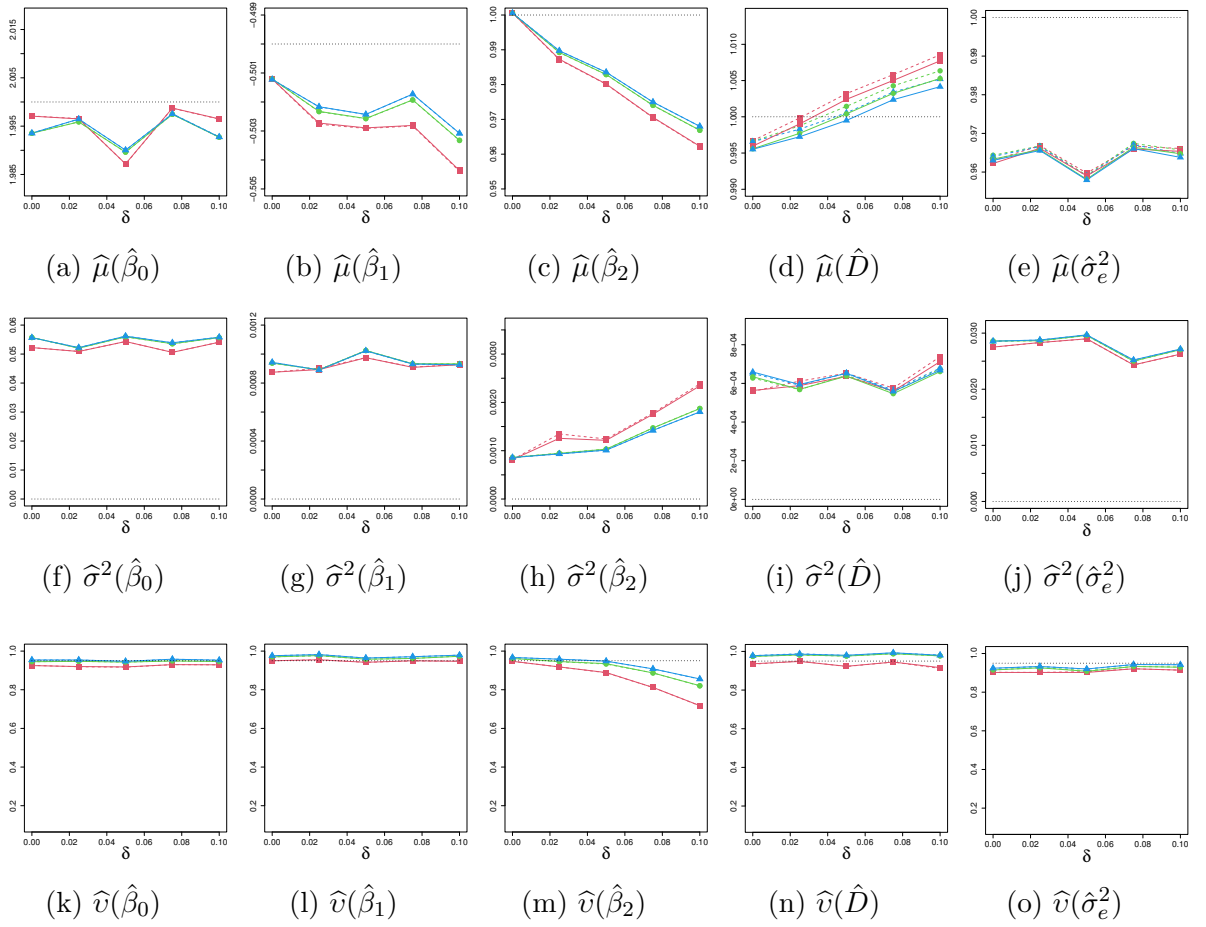


Figure A.8: The SM, SMSE and SCP with 95% of confidence of estimates on scenario 2.2. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.

irrelevant for large sample sizes.

A.3 Millennium Cohort Study covariates

The covariates used in this study come from 10 possible databases (numbers and labels in Table A.1) from four research sweeps: 2001, 2004, 2006 and 2008. The covariate codes in Tables A.2, A.3 and A.4 usually start with “B” to indicate that they belong to the second sweep (2004), the information about the first sweep used to start with “A”, the third sweep comes with the letter “C” and the fourth sweep with “D” or “DD”. Obviously, not all issues are repeated every year. With this detailed map of covariates, we believe that any scientist can reproduce our analysis easily.

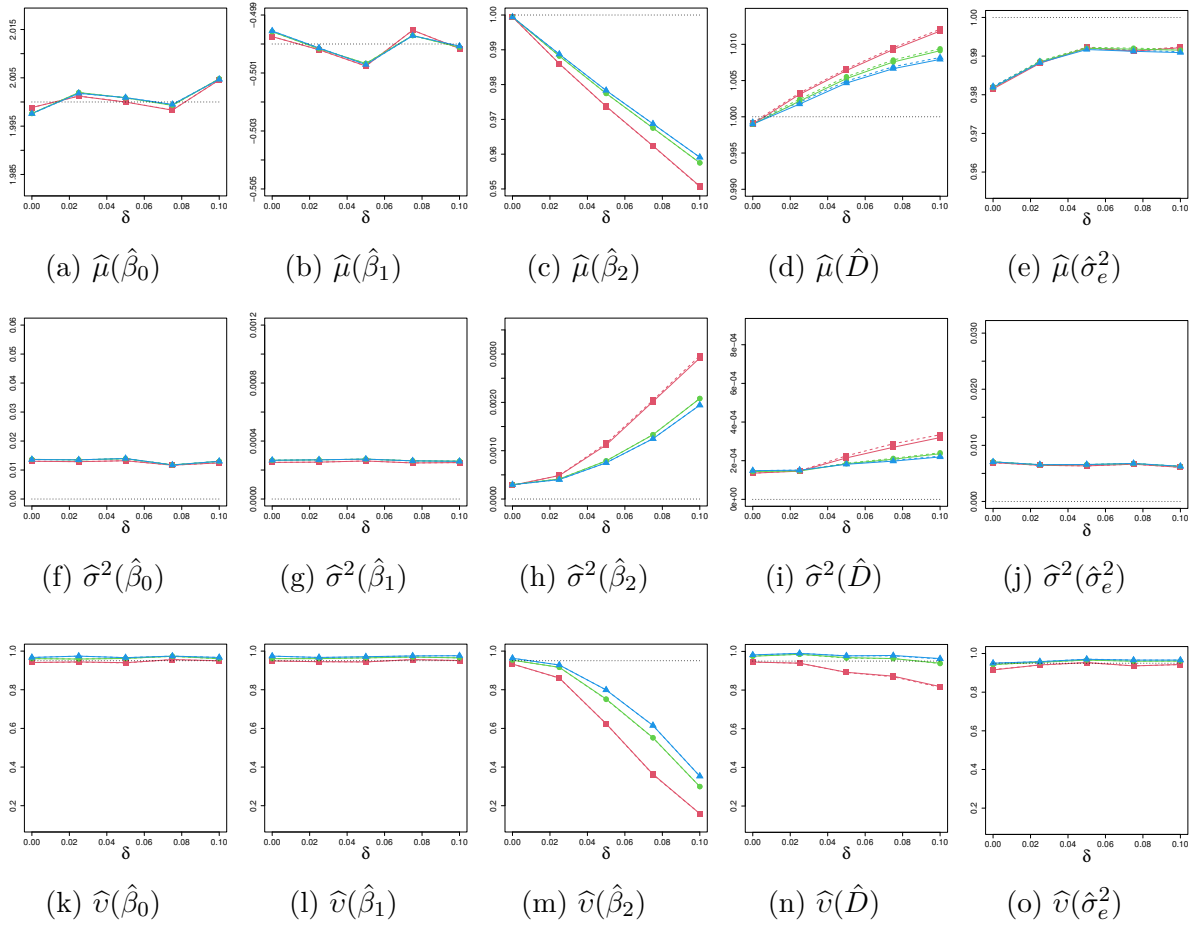


Figure A.9: The SM, SMSE and SCP with 95% of confidence of estimates on scenario 2.3. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.

Table A.1: Number and label of the databases

Number	Label
1	CM Cognitive Assessment
2	CM Derived
3	CM Interview
4	CM Teacher Survey
5	Family Derived
6	Geographically Linked Data
7	Hhgrid
8	Parent CM Interview
9	Parent Derived
10	Parent Interview

A.4 R packages

I developed three R packages for all main functions presented in this thesis. They are available on GitHub, and they can easily download and install in Ubuntu and Windows

Operator Systems if an R compiler is available. I do not include the Monte Carlo procedures, some graphics, and other auxiliary functions in the packages. The manuals of the R packages that I developed are in this appendix.

A.4.1 Robust linear mixed models

The Robust linear mixed models package is available on my personal page, [click here to open the package environment on GitHub](#).¹ and download the file `rlmm1.0.tar.gz`. Thus, install it on your R compiler and load it, e.g.,

```
install.packages("~/Downloads/rlmm1.0.tar.gz", repos = NULL, type = "source")  
library(rlmm)
```

¹Last update at August 24, 2022

Package ‘rlmm’

December 1, 2022

Type Package

Title Robust Linear Mixed Models

Version 1.1

Date 2022-12-01

Description This package solves robust linear mixed models using Huber, Tukey and Least Squared loss function.

License GPL (>= 2)

Encoding UTF-8

Imports Rcpp (>= 1.0.5), MASS (>= 7.3-49)

LinkingTo Rcpp, RcppArmadillo

Suggests tinytest (>= 1.3.1)

RoxygenNote 7.2.1

NeedsCompilation yes

Author Ian Meneghel Danilevich [aut, cre]
(<<https://orcid.org/0000-0003-4541-0524>>),
Pascal Bondon [aut],
Valderio A. Reisen [aut]

Maintainer Ian Meneghel Danilevich <iandanilevich@gmail.com>

R topics documented:

rlmm-package	2
rmm	2
Index	5

r1mm-package	<i>Robust Linear Mixed Models</i>
--------------	-----------------------------------

Description

This package solves robust linear mixed models using Huber, Tukey and Least Squared loss function.

Package Content

Index of help topics:

r1mm-package	Robust Linear Mixed Models
rmm	Robust linear model

Maintainer

NA

Author(s)

NA

rmm	<i>Robust linear model</i>
-----	----------------------------

Description

Estimate parameters of a robust linear mixed models.

Usage

```
rmm(
  y,
  x,
  z,
  subj,
  cor = "Ind",
  loss = "Huber",
  c = 1.345,
  weight = TRUE,
  tol = 1e-04
)
```

Arguments

y	Numeric vector, outcome.
x	Numeric matrix, covariates
z	Numeric vector or matrix, a vector of 1's for random intercepts, a vector of one covariate for random slopes.
subj	Numeric vector, identifies the unit to which the observation belongs.
cor	Factor, "Ind" for independent residuals, "AR1" for autoregressive first order residuals.
loss	Factor, "LS" least squared loss function, "Huber" loss function or "Tukey" loss function.
c	Numeric, positive real number, common choices are 1.345 for Huber and 4.685 for Tukey.
weight	Factor, if TRUE uses the Cantoni's weights, if FALSE does not.
tol	Numeric scalar, internal value, small value.

Value

beta Numeric vector, exploratory variables' coefficients.
 sigma Numeric, standard deviation.
 Psi Numeric matrix, covariance structure.
 phi Numeric, scalar of covariance structure.
 SE Numeric vector, exploratory variables' standard errors.
 loss Factor, "LS" least squared loss function, "Huber" loss function or "Tukey" loss function.
 cor Factor, "Ind" for independent residuals, "AR1" for autoregressive first order residuals.

References

Danilevich, I.M., Bondon, P., Reisen, V.A., Serpa, F.S. (2022), "A longitudinal study of the influence of air pollutants on respiratory health. A robust multivariate approach". Journal, vol number pages.
 Gill, P.S. (2000), "A robust mixed linear model analysis for longitudinal data", Stat. Med., 19: 975-987.
 Cantoni, E., Ronchetti, E. (2001), "Robust Inference for Generalized Linear Models", J. Am. Stat. Assoc., 96 (455): 1022-1030.

Examples

```
n = 10
m = 5
d = 4
N = n*m
x = matrix(rnorm(d*N), ncol=d, nrow=N)
subj = rep(1:n, each=m)
gamma = rnorm(n)
z = rep(1,N)
```

```
beta = rnorm(d)
eps = rnorm(N)
y = as.vector(x %*% beta + rep(gamma, each=m) + eps)
m1 = rmm(y, x, z, subj, cor="Ind", loss="Huber", c=1.345, weight = FALSE)
m1
```

Index

* **package**

rlmm-package, [2](#)

rlmm (rlmm-package), [2](#)

rlmm-package, [2](#)

rmm, [2](#)

A.4.2 Penalized quantile regression with fixed effects

The Penalized quantile regression with fixed effects package is available on CRAN mirror, [click here to open the package environment](#).² You can install it directly from CRAN to your R compiler, e.g.,

```
install.packages("pqrfe", dependencies=TRUE)
library(pqrfe)
```

²Last update at October 20, 2022

Package ‘pqrfe’

October 13, 2022

Type Package

Title Penalized Quantile Regression with Fixed Effects

Version 1.0

Date 2022-08-23

Description

Quantile regression with fixed effects is a general model for longitudinal data. Here we proposed to solve it by several methods. The estimation methods include three loss functions as: check, asymmetric least square and asymmetric Huber functions; and three structures as: simple regression, fixed effects and fixed effects with penalized intercepts by LASSO.

License GPL (>= 2)

Imports Rcpp (>= 1.0.5), MASS (>= 7.3-49)

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 7.2.1

NeedsCompilation yes

Author Ian Meneghel Danilevich [aut, cre] (0000-0003-4541-0524)

Maintainer Ian Meneghel Danilevich <iandanilevich@gmail.com>

R topics documented:

pqrfe-package	2
check_lambda	4
choice_p	5
clean_data	5
d_psi_als	6
d_psi_mq	6
f_den	7
f_tab	7
loss_er	8
loss_erfe	8
loss_erlasso	9
loss_mqr	10
loss_mqrfe	10

loss_mqrlasso	11
loss_qr	11
loss_qrfe	12
loss_qrlasso	12
mpqr	13
optim_er	14
optim_erfe	14
optim_erlasso	15
optim_mqr	15
optim_mqrfe	16
optim_mqrlasso	17
optim_qr	17
optim_qrfe	18
optim_qrlasso	19
plot_taus	19
pqr	20
print.PQR	22
psi_als	22
psi_mq	23
q_cov	23
rho_koenker	24
rho_mq	24
sgf	25

Index 26

pqrfe-package	<i>Penalized Quantile Regression with Fixed Effects</i>
---------------	---

Description

Quantile regression with fixed effects is a general model for longitudinal data. Here we proposed to solve it by several methods. The estimation methods include three loss functions as: check, asymmetric least square and asymmetric Huber functions; and three structures as: simple regression, fixed effects and fixed effects with penalized intercepts by LASSO.

Details

The DESCRIPTION file:

```

Package:    pqrfe
Type:       Package
Title:      Penalized Quantile Regression with Fixed Effects
Version:    1.0
Date:       2022-08-23
Authors@R: person(family = "Danilevicz", given = "Ian Meneghel", email = "iandanilevicz@gmail.com", role = c("aut", "
Description: Quantile regression with fixed effects is a general model for longitudinal data. Here we proposed to solve it t
License:    GPL (>= 2)

```

Imports: Rcpp (>= 1.0.5), MASS (>= 7.3-49)
 LinkingTo: Rcpp, RcppArmadillo
 RoxygenNote: 7.2.1
 Author: Ian Meneghel Danilevich [aut, cre] (0000-0003-4541-0524)
 Maintainer: Ian Meneghel Danilevich <iandanilevich@gmail.com>

Index of help topics:

check_lambda	check lambda
choice_p	choice model
clean_data	Clean missings
d_psi_als	D Psi ALS
d_psi_mq	D Psi M-quantile
f_den	Kernel density
f_tab	Tabular function
loss_er	Loss expectile regression
loss_erfe	Loss expectile regression with fixed effects
loss_erlasso	Loss lasso expectile regression with fixed effects
loss_mqr	Loss M-quantile regression
loss_mqrfe	Loss M-quantile regression with fixed effects
loss_mqrlasso	Loss lasso M-quantile regression with fixed effects
loss_qr	Loss quantile regression
loss_qrfe	Loss quantile regression with fixed effects
loss_qrlasso	Loss lasso quantile regression with fixed effects
mpqr	multiple penalized quantile regression
optim_er	optim expectile regression
optim_erfe	optim expectile regression with fixed effects
optim_erlasso	optim expectile regression with fixed effects and LASSO
optim_mqr	optim M-quantile regression
optim_mqrfe	optim quantile regression with fixed effects
optim_mqrlasso	optim M-quantile regression with fixed effects and LASSO
optim_qr	optim quantile regression
optim_qrfe	optim quantile regression with fixed effects
optim_qrlasso	optim quantile regression with fixed effects and LASSO
plot_taus	plot multiple penalized quantile regression
pqr	Penalized quantile regression with fixed effects
pqrfe-package	Penalized Quantile Regression with Fixed Effects
print.PQR	Print an PQR
psi_als	Psi ALS

psi_mq	Psi M-quantile
q_cov	Covariance
rho_koenker	Rho Koenker
rho_mq	Rho M-quantile
sgf	Identify significance

This section should provide a more detailed overview of how to use the package, including the most important functions.

Author(s)

NA

Maintainer: NA

References

This optional section can contain literature or other references for background information.

See Also

Optional links to other man pages

Examples

```
## Optional simple examples of the most important functions
## Use \dontrun{} around code to be shown but not executed
```

check_lambda	<i>check lambda</i>
--------------	---------------------

Description

check lambda

Usage

```
check_lambda(lambda, infb, supb)
```

Arguments

lambda	Numeric, value of lambda.
infb	Numeric, lower bound of lambda.
supb	Numeric, upper bound of lambda.

Value

lambda Numeric, valid value of lambda.

choice_p	<i>choice model</i>
----------	---------------------

Description

choice model

Usage

```
choice_p(effect)
```

Arguments

effect Factor, simple, fixed or lasso.

Value

penalty Numeric, 1, 2 and 3.

clean_data	<i>Clean missings</i>
------------	-----------------------

Description

Clean missings

Usage

```
clean_data(y, x, id)
```

Arguments

y Numeric vector, outcome.
x Numeric matrix, covariates
id Numeric vector, identifies the unit to which the observation belongs.

Value

list with the same objects y, x, id, but without missings.

Examples

```

n = 10
m = 4
d = 3
N = n*m
L = N*d
x = matrix(rnorm(L), ncol=d, nrow=N)
subj = rep(1:n, each=m)
alpha = rnorm(n)
beta = rnorm(d)
eps = rnorm(N)
y = x %*% beta + matrix(rep(alpha, each=m) + eps)
y = as.vector(y)
x[1,3] = NA
clean_data(y=y, x=x, id=subj)

```

d_psi_als	<i>D Psi ALS</i>
-----------	------------------

Description

Derivative of Psi asymmetric least square

Usage

```
d_psi_als(x, tau)
```

Arguments

x	generic vector
tau	percentile

d_psi_mq	<i>D Psi M-quantile</i>
----------	-------------------------

Description

Derivative of psi M-quantile

Usage

```
d_psi_mq(x, tau, c)
```

f_den

7

Arguments

x	generic vector
tau	percentile
c	tuning

f_den	<i>Kernel density</i>
-------	-----------------------

Description

Kernel density

Usage

```
f_den(x)
```

Arguments

x	Numeric vector.
---	-----------------

Value

y vector, kernel density estimation.

Examples

```
x = rnorm(10)
f_den(x)
```

f_tab	<i>Tabular function</i>
-------	-------------------------

Description

Tabular function

Usage

```
f_tab(N, n, d, theta, sig2, kind)
```

Arguments

N	sample size.
n	length of alpha.
d	length of beta.
theta	Numeric vector.
sig2	Numeric vector.
kind	Numeric, 1 means alpha, 2 means beta

loss_er	<i>Loss expectile regression</i>
---------	----------------------------------

Description

Loss expectile regression

Usage

```
loss_er(beta, x, y, tau, N, d)
```

Arguments

beta	initial values
x	design matrix
y	vector output
tau	percentile
N	sample size
d	columns of x

loss_erfe	<i>Loss expectile regression with fixed effects</i>
-----------	---

Description

Loss expectile regression with fixed effects

Usage

```
loss_erfe(theta, x, y, z, tau, n, d, mm)
```

Arguments

theta	initial values
x	design matrix
y	vector output
z	incident matrix
tau	percentile
n	N sample size
d	columns of x
mm	n columns of z

loss_erlasso	<i>Loss lasso expectile regression with fixed effects</i>
--------------	---

Description

Loss lasso expectile regression with fixed effects

Usage

```
loss_erlasso(theta, x, y, z, tau, n, d, mm, lambda)
```

Arguments

theta	initial values
x	design matrix
y	vector output
z	incident matrix
tau	percentile
n	N sample size
d	columns of x
mm	n columns of z
lambda	constriction parameter

loss_mqr	<i>Loss M-quantile regression</i>
----------	-----------------------------------

Description

Loss M-quantile regression

Usage

```
loss_mqr(beta, x, y, tau, N, d, c)
```

Arguments

beta	initial values
x	design matrix
y	vector output
tau	percentile
N	sample size
d	columns of x
c	tuning

loss_mqrfe	<i>Loss M-quantile regression with fixed effects</i>
------------	--

Description

Loss M-quantile regression with fixed effects

Usage

```
loss_mqrfe(theta, x, y, z, tau, n, d, mm, c)
```

Arguments

theta	initial values
x	design matrix
y	vector output
z	incident matrix
tau	percentile
n	N sample size
d	columns of x
mm	n columns of z
c	tuning

loss_mqrlasso	<i>Loss lasso M-quantile regression with fixed effects</i>
---------------	--

Description

Loss lasso M-quantile regression with fixed effects

Usage

```
loss_mqrlasso(theta, x, y, z, tau, n, d, mm, c, lambda)
```

Arguments

theta	initial values
x	design matrix
y	vector output
z	incident matrix
tau	percentile
n	N sample size
d	columns of x
mm	n columns of z
c	tuning
lambda	constriction parameter

loss_qr	<i>Loss quantile regression</i>
---------	---------------------------------

Description

Loss quantile regression

Usage

```
loss_qr(beta, x, y, tau, N, d)
```

Arguments

beta	initial values
x	design matrix
y	vector output
tau	percentile
N	sample size
d	columns of x

loss_qrfe	<i>Loss quantile regression with fixed effects</i>
-----------	--

Description

Loss quantile regression with fixed effects

Usage

```
loss_qrfe(theta, x, y, z, tau, n, d, mm)
```

Arguments

theta	initial values
x	design matrix
y	vector output
z	incident matrix
tau	percentile
n	N sample size
d	columns of x
mm	n columns of z

loss_qrlasso	<i>Loss lasso quantile regression with fixed effects</i>
--------------	--

Description

Loss lasso quantile regression with fixed effects

Usage

```
loss_qrlasso(theta, x, y, z, tau, n, d, mm, lambda)
```

Arguments

theta	initial values
x	design matrix
y	vector output
z	incident matrix
tau	percentile
n	N sample size
d	columns of x
mm	n columns of z
lambda	constriction parameter

mpqr	<i>multiple penalized quantile regression</i>
------	---

Description

Estimate QR for several taus

Usage

```
mpqr(x, y, subj, tau = 1:9/10, effect = "simple", c = 0)
```

Arguments

x	Numeric matrix, covariates
y	Numeric vector, outcome.
subj	Numeric vector, identifies the unit to which the observation belongs.
tau	Numeric vector, identifies the percentiles.
effect	Factor, "simple" simple regression, "fixed" regression with fixed effects, "lasso" penalized regression with fixed effects.
c	Numeric, 0 is quantile, Inf is expectile, any number between zero and infinite is M-quantile.

Value

Beta Numeric array, with three dimensions: 1) tau, 2) coef., lower bound, upper bound, 3) exploratory variables.

Examples

```
n = 10
m = 5
d = 4
N = n*m
L = N*d
x = matrix(rnorm(L), ncol=d, nrow=N)
subj = rep(1:n, each=m)
alpha = rnorm(n)
beta = rnorm(d)
eps = rnorm(N)
y = as.vector(x %*% beta + rep(alpha, each=m) + eps)

Beta = mpqr(x,y,subj,tau=1:9/10, effect="fixed", c = 1.2)
Beta
```

optim_er	<i>optim expectile regression</i>
----------	-----------------------------------

Description

This function solves a expectile regression

Usage

```
optim_er(beta, x, y, tau, N, d)
```

Arguments

beta	Numeric vector, initials values beta.
x	Numeric matrix, covariates.
y	Numeric vector, output.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.

Value

parametric vector and residuals.

optim_erfe	<i>optim expectile regression with fixed effects</i>
------------	--

Description

This function solves a expectile regression with fixed effects

Usage

```
optim_erfe(beta, alpha, x, y, z, tau, N, d, n)
```

Arguments

beta	Numeric vector, initials values beta.
alpha	Numeric vector, initials values alpha.
x	Numeric matrix, covariates.
y	Numeric vector, output.
z	Numeric matrix, incidence matrix.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.
n	Numeric integer, length of alpha.

Value

parametric vector and residuals.

optim_erlasso	<i>optim expectile regression with fixed effects and LASSO</i>
---------------	--

Description

This function solves a expectile regression with fixed effects and LASSO

Usage

```
optim_erlasso(beta, alpha, x, y, z, tau, N, d, n)
```

Arguments

beta	Numeric vector, initials values beta.
alpha	Numeric vector, initials values alpha.
x	Numeric matrix, covariates.
y	Numeric vector, output.
z	Numeric matrix, incidence matrix.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.
n	Numeric integer, length of alpha.

Value

parametric vector and residuals.

optim_mqr	<i>optim M-quantile regression</i>
-----------	------------------------------------

Description

This function solves a M-quantile regression

Usage

```
optim_mqr(beta, x, y, tau, N, d, c)
```

Arguments

beta	Numeric vector, initials values beta.
x	Numeric matrix, covariates.
y	Numeric vector, output.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.
c	Numeric, positive real value.

Value

parametric vector and residuals.

optim_mqrfe	<i>optim quantile regression with fixed effects</i>
-------------	---

Description

This function solves a quantile regression with fixed effects

Usage

```
optim_mqrfe(beta, alpha, x, y, z, tau, N, d, n, c)
```

Arguments

beta	Numeric vector, initials values beta.
alpha	Numeric vector, initials values alpha.
x	Numeric matrix, covariates.
y	Numeric vector, output.
z	Numeric matrix, incidence matrix.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.
n	Numeric integer, length of alpha.
c	Numeric, positive real value.

Value

parametric vector and residuals.

optim_mqrlasso	<i>optim M-quantile regression with fixed effects and LASSO</i>
----------------	---

Description

This function solves a M-quantile regression with fixed effects and LASSO

Usage

```
optim_mqrlasso(beta, alpha, x, y, z, tau, N, d, n, c)
```

Arguments

beta	Numeric vector, initials values beta.
alpha	Numeric vector, initials values alpha.
x	Numeric matrix, covariates.
y	Numeric vector, output.
z	Numeric matrix, incidence matrix.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.
n	Numeric integer, length of alpha.
c	Numeric, positive real value.

Value

parametric vector and residuals.

optim_qr	<i>optim quantile regression</i>
----------	----------------------------------

Description

This function solves a quantile regression

Usage

```
optim_qr(beta, x, y, tau, N, d)
```

Arguments

beta	Numeric vector, initials values.
x	Numeric matrix, covariates.
y	Numeric vector, output.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.

Value

parametric vector and residuals.

optim_qrfe	<i>optim quantile regression with fixed effects</i>
------------	---

Description

This function solves a quantile regression with fixed effects

Usage

```
optim_qrfe(beta, alpha, x, y, z, tau, N, d, n)
```

Arguments

beta	Numeric vector, initials values beta.
alpha	Numeric vector, initials values alpha.
x	Numeric matrix, covariates.
y	Numeric vector, output.
z	Numeric matrix, incidence matrix.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.
n	Numeric integer, length of alpha.

Value

parametric vector and residuals.

optim_qrlasso	<i>optim quantile regression with fixed effects and LASSO</i>
---------------	---

Description

This function solves a quantile regression with fixed effects and LASSO

Usage

```
optim_qrlasso(beta, alpha, x, y, z, tau, N, d, n)
```

Arguments

beta	Numeric vector, initials values beta.
alpha	Numeric vector, initials values alpha.
x	Numeric matrix, covariates.
y	Numeric vector, output.
z	Numeric matrix, incidence matrix.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.
n	Numeric integer, length of alpha.

Value

parametric vector and residuals.

plot_taus	<i>plot multiple penalized quantile regression</i>
-----------	--

Description

plot QR for several taus

Usage

```
plot_taus(
  Beta,
  tau = 1:9/10,
  D,
  col = 2,
  lwd = 1,
  lty = 2,
```

```

    pch = 1,
    cex.axis = 1,
    cex.lab = 1,
    main = ""
  )

```

Arguments

Beta	Numeric array, with three dimensions: 1) tau, 2) coef., lower bound, upper bound, 3) exploratory variables.
tau	Numeric vector, identifies the percentiles.
D	covariate's number.
col	color.
lwd	line width.
lty	line type.
pch	point character.
cex.axis	cex axis length.
cex.lab	cex axis length.
main	title.

Examples

```

n = 10
m = 5
d = 4
N = n*m
L = N*d
x = matrix(rnorm(L), ncol=d, nrow=N)
subj = rep(1:n, each=m)
alpha = rnorm(n)
beta = rnorm(d)
eps = rnorm(N)
y = as.vector(x %*% beta + rep(alpha, each=m) + eps)

Beta = mpqr(x,y,subj,tau=1:9/10, effect="lasso", c = Inf)
plot_taus(Beta,tau=1:9/10,D=1)

```

pqr

Penalized quantile regression with fixed effects

Description

Estimate parameters and tuning parameter.

Usage

```
pqr(x, y, subj, tau = 0.5, effect = "simple", c = 1)
```

Arguments

x	Numeric matrix, covariates
y	Numeric vector, outcome.
subj	Numeric vector, identifies the unit to which the observation belongs.
tau	Numeric scalar between zero and one, identifies the percentile.
effect	Factor, "simple" simple regression, "fixed" regression with fixed effects, "lasso" penalized regression with fixed effects.
c	Numeric, 0 is quantile, Inf is expectile, any number between zero and infinite is M-quantile.

Value

alpha Numeric vector, intercepts' coefficients.
 beta Numeric vector, exploratory variables' coefficients.
 lambda Numeric, estimated lambda.
 res Numeric vector, percentile residuals.
 tau Numeric scalar, the percentile.
 penalty Numeric scalar, indicate the chosen effect.
 c Numeric scalar, indicate the chosen c.
 sig2_alpha Numeric vector, intercepts' standard errors.
 sig2_beta Numeric vector, exploratory variables' standard errors.
 Tab_alpha Data.frame, intercepts' summary.
 Tab_beta Data.frame, exploratory variables' summary.
 Mat_alpha Numeric matrix, intercepts' summary.
 Mat_beta Numeric matrix, exploratory variables' summary.

References

Danilevicz, I.M., Bondon, P., Reisen, V.A. (2022), "Alternative methods to quantile regression for panel data". Journal, vol number pages.
 Koenker, R. (2004), "Quantile regression for longitudinal data", J. Multivar. Anal., 91(1): 74-89.

Examples

```
n = 10
m = 5
d = 4
N = n*m
x = matrix(rnorm(d*N), ncol=d, nrow=N)
subj = rep(1:n, each=m)
```

```

alpha = rnorm(n)
beta = rnorm(d)
eps = rnorm(N)
y = as.vector(x %*% beta + rep(alpha, each=m) + eps)
m1 = pqr(x=x, y=y, subj=subj, tau=0.75, effect="lasso", c = 0)
m1$Tab_beta

```

print.PQR	<i>Print an PQR</i>
-----------	---------------------

Description

Define the visible part of the object class PQR

Usage

```

## S3 method for class 'PQR'
print(x, ...)

```

Arguments

x	An object of class "PQR"
...	further arguments passed to or from other methods.

psi_als	<i>Psi ALS</i>
---------	----------------

Description

Psi asymmetric least square

Usage

```
psi_als(x, tau)
```

Arguments

x	generic vector
tau	percentile

psi_mq	<i>Psi M-quantile</i>
--------	-----------------------

Description

Psi M-quantile

Usage

```
psi_mq(x, tau, c)
```

Arguments

x	generic vector
tau	percentile
c	tuning

q_cov	<i>Covariance</i>
-------	-------------------

Description

Estimate Covariance matrix

Usage

```
q_cov(n, N, d, Z, X, tau, res, penalty, c)
```

Arguments

n	length of alpha.
N	sample size.
d	length of beta.
Z	Numeric matrix, incident matrix.
X	Numeric matrix, covariates.
tau	Numeric, identifies the percentile.
res	Numeric vector, residuals.
penalty	Numeric, 1 quantile regression, 2 quantile regression with fixed effects, 3 Lasso quantile regression with fixed effects
c	Numeric, tuning

rho_koenker	<i>Rho Koenker</i>
-------------	--------------------

Description

Rho Koenker

Usage

```
rho_koenker(x, tau)
```

Arguments

x	generic vector
tau	percentile

rho_mq	<i>Rho M-quantile</i>
--------	-----------------------

Description

Rho M-quantile

Usage

```
rho_mq(x, tau, c)
```

Arguments

x	generic vector
tau	percentile
c	tuning

`sgf`*Identify significance*

Description

Identify significance

Usage

```
sgf(x)
```

Arguments

x Numeric vector.

Value

y vector Factor, symbol flag of significant p-values.

Examples

```
n = 10
pvalue = rgamma(10,1,10)
sgf(pvalue)
```

Index

*Topic **package**
pqrfe-package, [2](#)

check_lambda, [4](#)
choice_p, [5](#)
clean_data, [5](#)

d_psi_als, [6](#)
d_psi_mq, [6](#)

f_den, [7](#)
f_tab, [7](#)

loss_er, [8](#)
loss_erfe, [8](#)
loss_erlasso, [9](#)
loss_mqr, [10](#)
loss_mqrfe, [10](#)
loss_mqrlasso, [11](#)
loss_qr, [11](#)
loss_qrfe, [12](#)
loss_qrlasso, [12](#)

mpqr, [13](#)

optim_er, [14](#)
optim_erfe, [14](#)
optim_erlasso, [15](#)
optim_mqr, [15](#)
optim_mqrfe, [16](#)
optim_mqrlasso, [17](#)
optim_qr, [17](#)
optim_qrfe, [18](#)
optim_qrlasso, [19](#)

plot_taus, [19](#)
pqr, [20](#)
pqrfe (pqrfe-package), [2](#)
pqrfe-package, [2](#)
print.PQR, [22](#)
psi_als, [22](#)
psi_mq, [23](#)
q_cov, [23](#)
rho_koenker, [24](#)
rho_mq, [24](#)
sgf, [25](#)

A.4.3 Adaptive Lasso Quantile Regression with Fixed Effects

The Adaptive Lasso Quantile Regression with Fixed Effects package is available on CRAN mirror, [click here to open the package environment](#).³ You can install it directly from CRAN to your R compiler, e.g.,

```
install.packages("alqrfe", dependencies=TRUE)
library(alqrfe)
```

³Last update at September 20, 2022

Package ‘alqrfe’

September 20, 2022

Type Package

Title Adaptive Lasso Quantile Regression with Fixed Effects

Version 1.0

Date 2022-09-19

Description

Quantile regression with fixed effects solves longitudinal data, considering the individual intercepts as fixed effects. The parametric set of this type of problem used to be huge. Thus penalized methods such as Lasso are currently applied. Adaptive Lasso presents oracle proprieties, which include Gaussianity and correct model selection. Bayesian information criteria (BIC) estimates the optimal tuning parameter lambda. Plot tools are also available.

License GPL (≥ 2)

Imports Rcpp ($\geq 1.0.5$), MASS ($\geq 7.3-49$)

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 7.2.1

NeedsCompilation yes

Author Ian Meneghel Danilevich [aut, cre]
(<https://orcid.org/0000-0003-4541-0524>)

Maintainer Ian Meneghel Danilevich <iandanilevich@gmail.com>

Repository CRAN

Date/Publication 2022-09-19 23:36:13 UTC

R topics documented:

alqrfe-package	2
bic_hat	4
clean_data	4
df_hat	5
f_den	6
f_tab	6
loss_alqr	7
loss_lqr	7

loss_qr	8
loss_qrfe	9
make_z	9
mqr	10
mqr_alpha	11
optim_alqr	12
optim_lqr	13
optim_qr	13
optim_qrfe	14
plot_alpha	15
plot_taus	16
print.ALQRFE	17
qr	17
q_cov	19
rho_koenker	20
sgf	20
Index	21

alqrfe-package	<i>Adaptive Lasso Quantile Regression with Fixed Effects</i>
----------------	--

Description

Quantile regression with fixed effects solves longitudinal data, considering the individual intercepts as fixed effects. The parametric set of this type of problem used to be huge. Thus penalized methods such as Lasso are currently applied. Adaptive Lasso presents oracle proprieties, which include Gaussianity and correct model selection. Bayesian information criteria (BIC) estimates the optimal tuning parameter lambda. Plot tools are also available.

Details

The DESCRIPTION file:

```

Package:      alqrfe
Type:         Package
Title:        Adaptive Lasso Quantile Regression with Fixed Effects
Version:      1.0
Date:         2022-09-19
Authors@R:    person(family = "Danilevicz", given = "Ian Meneghel", email = "iandanilevicz@gmail.com", role = c("aut", "cre"))
Description:   Quantile regression with fixed effects solves longitudinal data, considering the individual intercepts as fixed effects
License:      GPL (>= 2)
Imports:      Rcpp (>= 1.0.5), MASS (>= 7.3-49)
LinkingTo:    Rcpp, RcppArmadillo
RoxygenNote:  7.2.1
Author:       Ian Meneghel Danilevicz [aut, cre] (0000-0003-4541-0524)
Maintainer:   Ian Meneghel Danilevicz <iandanilevicz@gmail.com>

```

Index of help topics:

alqrfe-package	Adaptive Lasso Quantile Regression with Fixed Effects
bic_hat	Bayesian Information Criteria
clean_data	Clean missings
df_hat	degrees of freedom
f_den	Kernel density
f_tab	Tabular function
loss_alqr	Loss adaptive lasso quantile regression with fixed effects
loss_lqr	Loss lasso quantile regression with fixed effects
loss_qr	Loss quantile regression
loss_qrfe	Loss quantile regression with fixed effects
make_z	Incident matrix Z
mqr	multiple penalized quantile regression
mqr_alpha	multiple penalized quantile regression - alpha
optim_alqr	optim adaptive lasso quantile regression with fixed effects
optim_lqr	optim lasso quantile regression with fixed effects
optim_qr	optim quantile regression
optim_qrfe	optim quantile regression with fixed effects
plot_alpha	plot multiple penalized quantile regression - alpha
plot_taus	plot multiple penalized quantile regression
print.ALQRFE	Print an ALQRFE
q_cov	Covariance
qr	quantile regression
rho_koenker	Rho Koenker
sgf	Identify significance

This section should provide a more detailed overview of how to use the package, including the most important functions.

Author(s)

NA

Maintainer: NA

References

This optional section can contain literature or other references for background information.

See Also

Optional links to other man pages

Examples

```
## Optional simple examples of the most important functions
## Use \dontrun{} around code to be shown but not executed
```

bic_hat	<i>Bayesian Information Criteria</i>
---------	--------------------------------------

Description

Bayesian Information Criteria

Usage

```
bic_hat(res, theta, tau, N, p, inf)
```

Arguments

res	Numeric vector, residuals.
theta	Numeric vector, parameters.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
p	Numeric integer, parameter length.
inf	Numeric, internal small quantity.

Value

BIC value

clean_data	<i>Clean missings</i>
------------	-----------------------

Description

Clean missings

Usage

```
clean_data(y, x, id)
```

Arguments

y	Numeric vector, outcome.
x	Numeric matrix, covariates
id	Numeric vector, identifies the unit to which the observation belongs.

Value

list with the same objects y, x, id, but without missings.

Examples

```
n = 10
m = 4
d = 3
N = n*m
L = N*d
x = matrix(rnorm(L), ncol=d, nrow=N)
subj = rep(1:n, each=m)
alpha = rnorm(n)
beta = rnorm(d)
eps = rnorm(N)
y = x %*% beta + matrix(rep(alpha, each=m) + eps)
y = as.vector(y)
x[1,3] = NA
clean_data(y=y, x=x, id=subj)
```

df_hat	<i>degrees of freedom</i>
--------	---------------------------

Description

This function estimates the degrees of freedom

Usage

```
df_hat(theta, N, p, inf)
```

Arguments

theta	Numeric vector, parameters to be test
N	Numeric integer, sample size.
p	Numeric integer, length of theta.
inf	Numeric, internal small quantity.

Value

degrees of freedom

f_den	<i>Kernel density</i>
-------	-----------------------

Description

Kernel density

Usage

```
f_den(x, inf)
```

Arguments

x	Numeric vector.
inf	Numeric, internal small quantity.

Value

y vector, kernel density estimation.

Examples

```
x = rnorm(10)
f_den(x, 0.0001)
```

f_tab	<i>Tabular function</i>
-------	-------------------------

Description

Tabular function

Usage

```
f_tab(N, n, d, theta, sig2, kind, inf, digt)
```

Arguments

N	sample size.
n	length of alpha.
d	length of beta.
theta	Numeric vector.
sig2	Numeric vector.
kind	Numeric, 1 means alpha, 2 means beta
inf	Numeric scalar, internal value, small value.
digt	Numeric integer, round.

loss_alqr	<i>Loss adaptive lasso quantile regression with fixed effects</i>
-----------	---

Description

Loss adaptive lasso quantile regression with fixed effects

Usage

```
loss_alqr(theta, x, y, z, tau, n, d, mm, lambda, w)
```

Arguments

theta	initial values
x	design matrix
y	vector output
z	incident matrix
tau	percentile
n	N sample size
d	columns of x
mm	n columns of z
lambda	constriction parameter
w	weights

loss_lqr	<i>Loss lasso quantile regression with fixed effects</i>
----------	--

Description

Loss lasso quantile regression with fixed effects

Usage

```
loss_lqr(theta, x, y, z, tau, n, d, mm, lambda)
```

Arguments

theta	initial values
x	design matrix
y	vector output
z	incident matrix
tau	percentile
n	N sample size
d	columns of x
mm	n columns of z
lambda	constriction parameter

loss_qr	<i>Loss quantile regression</i>
---------	---------------------------------

Description

Loss quantile regression

Usage

```
loss_qr(beta, x, y, tau, N, d)
```

Arguments

beta	initial values
x	design matrix
y	vector output
tau	percentile
N	sample size
d	columns of x

loss_qrfe	<i>Loss quantile regression with fixed effects</i>
-----------	--

Description

Loss quantile regression with fixed effects

Usage

```
loss_qrfe(theta, x, y, z, tau, n, d, mm)
```

Arguments

theta	initial values
x	design matrix
y	vector output
z	incident matrix
tau	percentile
n	N sample size
d	columns of x
mm	n columns of z

make_z	<i>Incident matrix Z</i>
--------	--------------------------

Description

Create an Incident matrix Z

Usage

```
make_z(n, N, id)
```

Arguments

n	Numeric integer, number of incidents (subjects, units or individuals).
N	Numeric integer, sample size.
id	Numeric vector of integer, incident identification.

Value

Z matrix.

mqr	<i>multiple penalized quantile regression</i>
-----	---

Description

Estimate QR for several taus

Usage

```
mqr(x, y, subj, tau = 1:9/10, method = "qr", ngrid = 20, inf = 1e-08, digt = 4)
```

Arguments

x	Numeric matrix, covariates
y	Numeric vector, outcome.
subj	Numeric vector, identifies the unit to which the observation belongs.
tau	Numeric vector, identifies the percentiles.
method	Factor, "qr" quantile regression, "qrfe" quantile regression with fixed effects, "lqrfe" Lasso quantile regression with fixed effects, "alqr" adaptive Lasso quantile regression with fixed effects.
ngrid	Numeric scalar greater than one, number of BIC to test.
inf	Numeric scalar, internal value, small value.
digt	Numeric scalar, internal value greater than one, define "zero" coefficient.

Value

Beta Numeric array, with three dimmensions: 1) tau, 2) coef., lower bound, upper bound, 3) exploratory variables.

Examples

```
n = 10
m = 5
d = 4
N = n*m
L = N*d
x = matrix(rnorm(L), ncol=d, nrow=N)
subj = rep(1:n, each=m)
alpha = rnorm(n)
beta = rnorm(d)
eps = rnorm(N)
y = x %*% beta + matrix(rep(alpha, each=m) + eps)
y = as.vector(y)

Beta = mqr(x,y,subj,tau=1:9/10, method="qr", ngrid = 10)
Beta
```

mqr_alpha	<i>multiple penalized quantile regression - alpha</i>
-----------	---

Description

Estimate QR intercepts for several taus

Usage

```
mqr_alpha(
  x,
  y,
  subj,
  tau = 1:9/10,
  method = "qr",
  ngrid = 20,
  inf = 1e-08,
  digt = 4
)
```

Arguments

x	Numeric matrix, covariates
y	Numeric vector, outcome.
subj	Numeric vector, identifies the unit to which the observation belongs.
tau	Numeric vector, identifies the percentiles.
method	Factor, "qr" quantile regression, "qrfe" quantile regression with fixed effects, "lqrfe" Lasso quantile regression with fixed effects, "alqr" adaptive Lasso quantile regression with fixed effects.
ngrid	Numeric scalar greater than one, number of BIC to test.
inf	Numeric scalar, internal value, small value.
digt	Numeric scalar, internal value greater than one, define "zero" coefficient.

Value

Alpha Numeric array, with three dimmensions: 1) tau, 2) coef., lower bound, upper bound, 3) exploratory variables.

Examples

```
n = 10
m = 5
d = 4
N = n*m
L = N*d
x = matrix(rnorm(L), ncol=d, nrow=N)
```

```

subj = rep(1:n, each=m)
alpha = rnorm(n)
beta = rnorm(d)
eps = rnorm(N)
y = x %%% beta + matrix(rep(alpha, each=m) + eps)
y = as.vector(y)

Alpha = mqr(x,y,subj,tau=1:9/10, method="qr", ngrid = 10)
Alpha

```

optim_alqr

optim adaptive lasso quantile regression with fixed effects

Description

This function solves an adaptive lasso quantile regression with fixed effects

Usage

```
optim_alqr(beta, alpha, wbeta, walpha, x, y, z, tau, N, d, n, ngrid, inf)
```

Arguments

beta	Numeric vector, initials values
alpha	Numeric vector, initials values
wbeta	Numeric vector, beta weights
walpha	Numeric vector, alpha weights
x	Numeric matrix, covariates.
y	Numeric vector, output.
z	Numeric matrix, incidents.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.
n	Numeric integer, length of alpha.
ngrid	Numeric integer, number of iterations of BIC.
inf	Numeric, internal small quantity.

Value

parametric vector and residuals

optim_lqr	<i>optim lasso quantile regression with fixed effects</i>
-----------	---

Description

This function solves a lasso quantile regression with fixed effects

Usage

```
optim_lqr(beta, alpha, x, y, z, tau, N, d, n, ngrid, inf)
```

Arguments

beta	Numeric vector, initials values
alpha	Numeric vector, initials values
x	Numeric matrix, covariates.
y	Numeric vector, output.
z	Numeric matrix, incidents.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.
n	Numeric integer, length of alpha.
ngrid	Numeric integer, number of iterations of BIC.
inf	Numeric, internal small quantity.

Value

parametric vector and residuals

optim_qr	<i>optim quantile regression</i>
----------	----------------------------------

Description

This function solves a quantile regression

Usage

```
optim_qr(beta, x, y, tau, N, d)
```

Arguments

beta	Numeric vector, initials values.
x	Numeric matrix, covariates.
y	Numeric vector, output.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.

Value

parametric vector and residuals.

optim_qrfe	<i>optim quantile regression with fixed effects</i>
------------	---

Description

This function solves a quantile regression with fixed effects

Usage

```
optim_qrfe(beta, alpha, x, y, z, tau, N, d, n)
```

Arguments

beta	Numeric vector, initials values
alpha	Numeric vector, initials values
x	Numeric matrix, covariates.
y	Numeric vector, output.
z	Numeric matrix, incidents.
tau	Numeric scalar, the percentile.
N	Numeric integer, sample size.
d	Numeric integer, X number of columns.
n	Numeric integer, length of alpha.

Value

parametric vector and residuals

plot_alpha

*plot multiple penalized quantile regression - alpha***Description**

plot QR intercepts for several taus

Usage

```
plot_alpha(
  Beta,
  tau = 1:9/10,
  D,
  ylab = expression(alpha[1]),
  col = 2,
  lwd = 1,
  lty = 2,
  pch = 1,
  cex.axis = 1,
  cex.lab = 1,
  main = ""
)
```

Arguments

Beta	Numeric array, with three dimensions: 1) tau, 2) coef., lower bound, upper bound, 3) exploratory variables.
tau	Numeric vector, identifies the percentiles.
D	intercept's number.
ylab	y legend
col	color.
lwd	line width.
lty	line type.
pch	point character.
cex.axis	cex axis length.
cex.lab	cex axis length.
main	title.

Examples

```
n = 10
m = 5
d = 4
N = n*m
```

```

L = N*d
x = matrix(rnorm(L), ncol=d, nrow=N)
subj = rep(1:n, each=m)
alpha = rnorm(n)
beta = rnorm(d)
eps = rnorm(N)
y = x %*% beta + matrix(rep(alpha, each=m) + eps)
y = as.vector(y)

Beta = mqr_alpha(x,y,subj,tau=1:9/10, method="qr", ngrid = 10)
plot_alpha(Beta,tau=1:9/10,D=1)

```

plot_taus

*plot multiple penalized quantile regression***Description**

plot QR for several taus

Usage

```

plot_taus(
  Beta,
  tau = 1:9/10,
  D,
  col = 2,
  lwd = 1,
  lty = 2,
  pch = 1,
  cex.axis = 1,
  cex.lab = 1,
  main = ""
)

```

Arguments

Beta	Numeric array, with three dimensions: 1) tau, 2) coef., lower bound, upper bound, 3) exploratory variables.
tau	Numeric vector, identifies the percentiles.
D	covariate's number.
col	color.
lwd	line width.
lty	line type.
pch	point character.
cex.axis	cex axis length.
cex.lab	cex axis length.
main	title.

Examples

```

n = 10
m = 5
d = 4
N = n*m
L = N*d
x = matrix(rnorm(L), ncol=d, nrow=N)
subj = rep(1:n, each=m)
alpha = rnorm(n)
beta = rnorm(d)
eps = rnorm(N)
y = x %*% beta + matrix(rep(alpha, each=m) + eps)
y = as.vector(y)

Beta = mqr(x,y,subj,tau=1:9/10, method="qr", ngrid = 10)
plot_taus(Beta,tau=1:9/10,D=1)

```

print.ALQRFE

Print an ALQRFE

Description

Define the visible part of the object class ALQRFE

Usage

```

## S3 method for class 'ALQRFE'
print(x, ...)

```

Arguments

x	An object of class "ALQRFE"
...	further arguments passed to or from other methods.

qr

quantile regression

Description

Estimate quantile regression with fixed effects for one tau

Usage

```
qr(x, y, subj, tau = 0.5, method = "qr", ngrid = 20, inf = 1e-08, digt = 4)
```

Arguments

x	Numeric matrix, covariates
y	Numeric vector, outcome.
subj	Numeric vector, identifies the unit to which the observation belongs.
tau	Numeric, identifies the percentile.
method	Factor, "qr" quantile regression, "qrfe" quantile regression with fixed effects, "lqrfe" Lasso quantile regression with fixed effects, "alqr" adaptive Lasso quantile regression with fixed effects.
ngrid	Numeric scalar greater than one, number of BIC to test.
inf	Numeric scalar, internal value, small value.
digit	Numeric scalar, internal value greater than one, define "zero" coefficient.

Value

alpha Numeric vector, intercepts' coefficients.
 beta Numeric vector, exploratory variables' coefficients.
 lambda Numeric, estimated lambda.
 res Numeric vector, percentile residuals.
 tau Numeric scalar, the percentile.
 penalty Numeric scalar, indicate the chosen effect.
 sig2_alpha Numeric vector, intercepts' standard errors.
 sig2_beta Numeric vector, exploratory variables' standard errors.
 Tab_alpha Data.frame, intercepts' summary.
 Tab_beta Data.frame, exploratory variables' summary.
 Mat_alpha Numeric matrix, intercepts' summary.
 Mat_beta Numeric matrix, exploratory variables' summary.
 method Factor, method applied.

References

Koenker, R. (2004) "Quantile regression for longitudinal data", J. Multivar. Anal., 91(1): 74-89, <doi:10.1016/j.jmva.2004.05.006>

Examples

```
# Example 1
n = 10
m = 5
d = 4
N = n*m
L = N*d
x = matrix(rnorm(L), ncol=d, nrow=N)
subj = rep(1:n, each=m)
```

```

alpha = rnorm(n)
beta = rnorm(d)
eps = rnorm(N)
y = x %*% beta + matrix(rep(alpha, each=m) + eps)
y = as.vector(y)
m1 = qr(x,y,subj,tau=0.75, method="qrfe")
m1
m2 = qr(x,y,subj,tau=0.3, method="lqrfe", ngrid = 10)
m2

# Example 2, from MASS package
Rabbit = MASS::Rabbit
Rabbit$Treatment = ifelse(Rabbit$Treatment=="Control",0,1)
Rabbit$Animal = ifelse(Rabbit$Animal == "R1",1,ifelse(Rabbit$Animal == "R2",2,
ifelse(Rabbit$Animal == "R3",3,ifelse(Rabbit$Animal == "R4",4,5))))
X = matrix(cbind(Rabbit$Dose,Rabbit$Treatment), ncol=2)
m3 = qr(x=X, y=Rabbit$BPchange, subj=Rabbit$Animal,tau=0.5, method="alqrfe", ngrid = 10)
m3

```

q_cov

*Covariance***Description**

Estimate Covariance matrix

Usage

```
q_cov(alpha, beta, d, inf, n, N, res, method, tau, X, Z)
```

Arguments

alpha	Numeric vector.
beta	Numeric vector.
d	length of beta.
inf	Numeric scalar, internal value, small value.
n	length of alpha.
N	sample size.
res	Numeric vector, residuals.
method	Factor, "qr" quantile regression, "qrfe" quantile regression with fixed effects, "lqrfe" Lasso quantile regression with fixed effects, "alqr" adaptive Lasso quantile regression with fixed effects.
tau	Numeric, identifies the percentile.
X	Numeric matrix, covariates.
Z	Numeric matrix, incident matrix.

rho_koenker	<i>Rho Koenker</i>
-------------	--------------------

Description

Rho Koenker

Usage

```
rho_koenker(x, tau)
```

Arguments

x	generic vector
tau	percentile

sgf	<i>Identify significance</i>
-----	------------------------------

Description

Identify significance

Usage

```
sgf(x)
```

Arguments

x	Numeric vector.
---	-----------------

Value

y vector Factor, symbol flag of significant p-values.

Examples

```
n = 10
pvalue = rgamma(10,1,10)
sgf(pvalue)
```

Index

* **package**

alqrfe-package, [2](#)

alqrfe (alqrfe-package), [2](#)

alqrfe-package, [2](#)

bic_hat, [4](#)

clean_data, [4](#)

df_hat, [5](#)

f_den, [6](#)

f_tab, [6](#)

loss_alqr, [7](#)

loss_lqr, [7](#)

loss_qr, [8](#)

loss_qrfe, [9](#)

make_z, [9](#)

mqr, [10](#)

mqr_alpha, [11](#)

optim_alqr, [12](#)

optim_lqr, [13](#)

optim_qr, [13](#)

optim_qrfe, [14](#)

plot_alpha, [15](#)

plot_taus, [16](#)

print.ALQRFE, [17](#)

q_cov, [19](#)

qr, [17](#)

rho_koenker, [20](#)

sgf, [20](#)

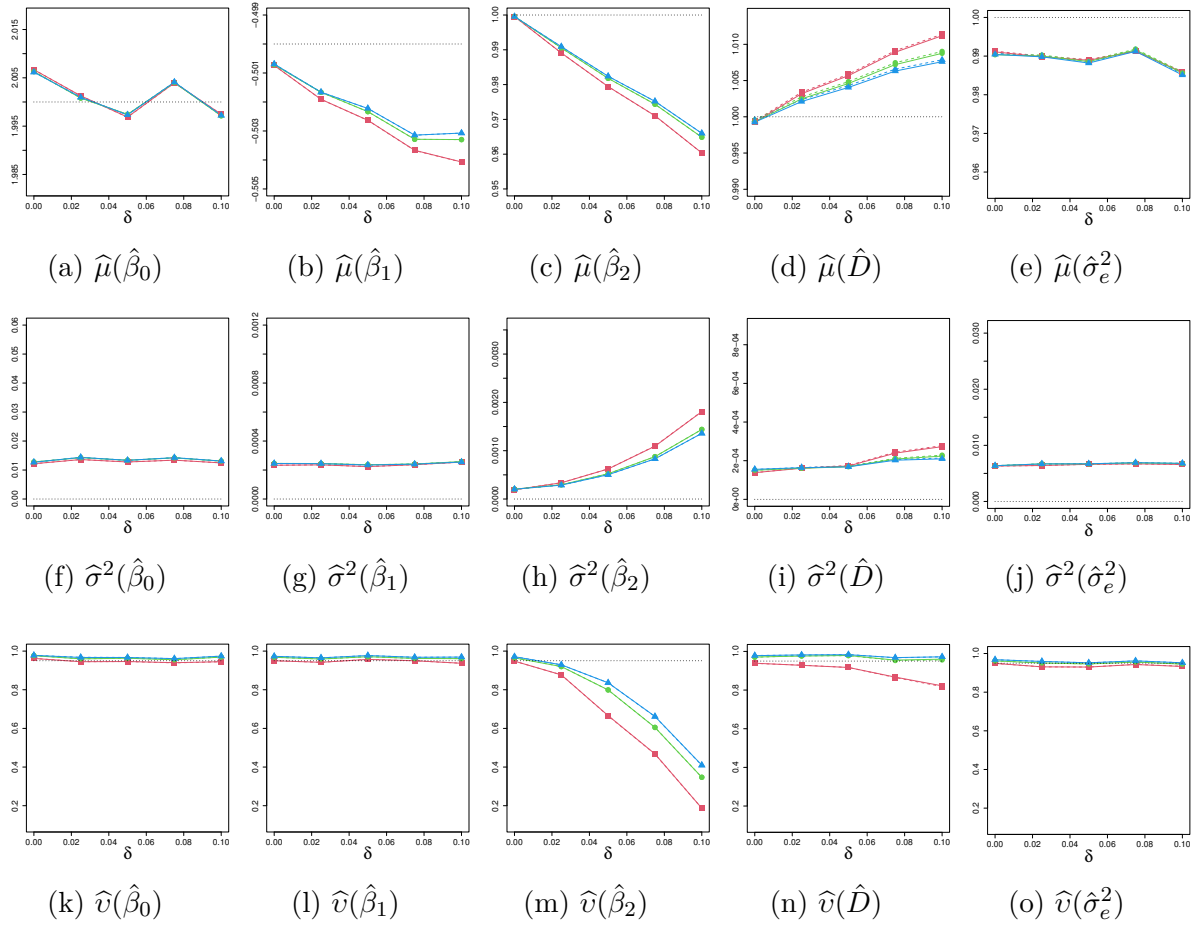


Figure A.10: The SM, SMSE and SCP with 95% of confidence of estimates on scenario 2.4. $\hat{\theta}_{LS}$ (red square), $\hat{\theta}_H$ (green circle) and $\hat{\theta}_T$ (blue triangle), weighted versions in solid line, non-weighted in dashed.

A.5 Knight's identity

The first Knight's identity is established by [63], the second is adapted by [65], several papers cite one or other identity as for example [121]. As far as we know, these identities where not proofed, to do this we slice each one in four cases and compare $\mathcal{LS}$ left side with $\mathcal{RS}$ right side.

The first Knight's identity is

$$|u - v| - |u| = -v(1_{u>0} - 1_{u<0}) + 2 \int_0^v (1_{u \leq s} - 1_{u \leq 0}) ds,$$

where $u \in \mathbb{R}$ and $v \in \mathbb{R}$.

Case 1: $u > 0$ and $u - v > 0$, implies

$$\mathcal{LS} = |u - v| - |u| = u - v - u = -v,$$

and

$$\mathcal{RS} = -v(1_{u>0} - 1_{u<0}) + 2 \int_0^v (1_{u \leq s} - 1_{u \leq 0}) ds = -v + 2 \cdot 0 = -v.$$

Table A.2: Old random variables

Number	Name	Code	Label	Dataset
*	SDQ-I	BPEER	DV SDQ Peer Problems	2
		BEMOTION	DV SDQ Emotional Symptoms	2
**	SDQ-E	BCONDUCT	DV SDQ Conduct Problems	2
		BHYPER	DV SDQ Hyperactivity/Inattention	2
1	Age	BCRKNAGE	DV Age in years:months for Bracken scoring	1
2	Age ²	BCRKNAGE	The square of the previous	1
3	ALE	BPFADI00	Whether non-res parent died	10
		BDOEDE00	DV OECD equivalised income	4
		BDHTYP00	DV Parents/Carers in Household	4
		BDOTHS00	DV Number of siblings of CM in household	4
		BPACCA00	Number of accidents or injuries	8
		BDCHNF00	DV Change in Natural father status	4
		BPMOAD00	When CM moved to current address (year)	10
		BDOTHS00	DV Number of siblings of CM in household	4
		BPDEAN00	Whether ever diagnosed with depression	10
4	SED	BDTOTS00	Number of siblings in house plus number of CM's	4
		BPROMA00	Number of rooms - grouped	10
		BDROOW00	DV Housing Tenure	4
		BPSTBE00	Whether receiving state benefits/payments	10
		BDOEDP00	DV OECD below 60% median poverty indicator	4
5	KS	BDKESS00	Kessler K6 Scale	10
6	Degree	BPACQU00	Highest Academic qualification	10
7	GCSE	BPACQU00	Highest Academic qualification	10
8	White	BDC06E00	DV Cohort Member Ethnic Group	2
9	Male	CHCSEX00	Cohort Member Sex	3
10	IMD	BIMDEMPN	IMD Overall Decile	6
11	Dis	BDSTRA00	Stratum	4

Case 2: $u > 0$ and $u - v \leq 0$, implies

$$\mathcal{LS} = v - u - u = v - 2u,$$

and

$$\mathcal{RS} = -v(1_{u>0}) + 2 \int_0^v (1_{u \leq s}) ds = -v + 2(v - u) = v - 2u.$$

Case 3: $u \leq 0$ and $u - v > 0$, implies

$$\mathcal{LS} = u - v + u = 2u - v,$$

and

$$\mathcal{RS} = -v(-1_{u<0}) + 2 \int_0^v (1_{u \leq s} - 1_{u \leq 0}) ds = v + 2 \int_0^v (1_{u \leq s} - 1) ds = v + 2(u - v) = 2u - v.$$

Case 4: $u \leq 0$ and $u - v \leq 0$, implies

$$\mathcal{LS} = v - u + u = v,$$

and

$$\mathcal{RS} = -v(-1_{u < 0}) + 2 \int_0^v (1_{u \leq s} - 1_{u \leq 0}) ds = v + 2 \int_0^v (1_{u \leq s} - 1) ds = v + 2(u - u) = v.$$

Which complete the argument.

The second Knight's identity is

$$\rho_\tau(u - v) - \rho_\tau(u) = -v\psi_\tau(u) \int_0^v (1_{u \leq s} - 1_{u \leq 0}) ds,$$

where $u \in \mathbb{R}$ and $v \in \mathbb{R}$, $\rho_\tau(\cdot)$ is the check function and $\psi_\tau(\cdot)$ is its derivative.

Case 1: $u > 0$ and $u - v > 0$, implies

$$\mathcal{LS} = (u - v)\tau - u\tau = -v\tau,$$

and

$$\mathcal{RS} = -v\tau + \int_0^v (0) ds = -v\tau.$$

Case 2: $u > 0$ and $u - v \leq 0$, implies

$$\mathcal{LS} = (u - v)(\tau - 1) - u\tau = -u + v - v\tau,$$

and

$$\mathcal{RS} = -v\tau + \int_0^v (1_{u \leq s}) ds = -v\tau + (v - u).$$

Case 3: $u \leq 0$ and $u - v > 0$, implies

$$\mathcal{LS} = (u - v)\tau - u(\tau - 1) = u\tau - v\tau - u\tau + u = -v\tau + u,$$

and

$$\mathcal{RS} = -v(\tau - 1) + \int_0^v (1_{u \leq s} - 1) ds = -v\tau + v + (u - v) = -v\tau + u.$$

Case 4: $u \leq 0$ and $u - v \leq 0$, implies

$$\mathcal{LS} = (u - v)(\tau - 1) - u(\tau - 1) = u\tau - u - v\tau + v - u\tau + u = v - v\tau,$$

and

$$\mathcal{RS} = -v(\tau - 1) + \int_0^v (1_{u \leq s} - 1) ds = -v(\tau - 1) + (v - v) = -v(\tau - 1).$$

Which finished the proof.

A.6 Computational performance

All the results are obtained using a desktop processor Intel Core i7-6700 CPU 3.40GHz $\times$ 8. This is a modest computer and is subject to inaccuracies. There are no use o parallel techniques in any of the methods. The LQRFE and ALQRFE methods which require a grid to calculate the tuning parameter, use a number of points equal to 10. The Table A.5 displays the time of the methods explained in Chapter 2, the Table A.6 shows the time of the methods exposed in Chapter 3, the Table A.7 exhibits the time of the procedures discussed in Chapter 4.

Table A.3: New random variables

Number	Code	Label	Dataset
12	BDOTHS00	DV Number of siblings of CM in household	5
13	BDNOCM00	DV Number of CM in household	5
14	BDTOTS00	DV Number of siblings in house plus number of CM's	5
15	BDNSIB00	DV Natural siblings of CM in household	5
16	BDHSIB00	DV Half siblings of CM in household	5
17	BDSSIB00	DV Step siblings of CM in household	5
18	BDASIB00	DV Adoptive siblings of CM in household	5
19	BDFSIB00	DV Foster siblings of CM in household	5
20	BDGPAR00	DV Grandparent of CM in household	5
21	BDOTHA00	DV Other adult in household	5
22	BDTOTP00	DV Number of people in household (including CM's)	5
23	BDOEDS00	DV Equivalised family income OECD scores	5
24	BDOEDE00	DV OECD equivalised income	5
25	BDOEDP00	DV OECD below 60% median poverty indicator	5
26	BDMHGT00	DV Natural mother Height in Metres	5
27	BDHLAN00	DV Language Spoken in household	5
28	BOEDEX00	DV PREDICTED weekly net family income	5
29	BOECDSC0	DV OECD Income Weighted Quintiles	5
30	BPOFRE00	How often do you read to the child	8
31	BPREEL00	Anyone else read to the child	8
32	BPREOF00	How often does anyone else read to the child	8
33	BPTOLI00	Anyone at home take child to the library	8
34	BPSDPA00	Anyone at home help child to learn sport etc	8
35	BPALPH00	Does anyone help CM with alphabet and reading?	8
36	BPNUMB00	Anyone at home teach child counting	8
37	BPOFCO00	How often at home try to teach child counting	8
38	BPSONG00	Anyone at home try teach child songs etc	8
39	BPOFSO00	How often teach child songs/poems/rhymes	8
40	BPDRAW00	Does child paint/draw at home	8
41	BPPAMA00	How often does CM paint/draw at home	8
42	BPEATW00	Child eaten with family past week	8
43	BPYOCH00	Been visited by friends with young children	8
44	BPINDE00	Values to instil: independence	8
45	BPOBRE00	Values to instil obedience and respect	8
46	BPNEGO00	Values to instil: art of negotiation	8
47	BPRSPE00	Values to instil: respect for elders	8
48	BPWESC00	Values to instil: doing well at school	8
49	BPREVA00	Values to instil: religious values	8
50	BPRULE00	Family has lots/not many rules	8
51	BPENFO00	Rules strictly/not strictly enforced	8
52	BPLANU00	Be understood speaking in own language by you	8
53	BPSDTT00	Often has temper tantrums	8
54	BPSDSP00	Tends to play alone	8
55	BPSDOR00	Generally obedient	8
56	BPSDMW00	Often seems worried	8
57	BPSDHU00	Helpful if someone is hurt, upset or ill	8
58	BPSDFS00	Constantly fidgeting	8
59	BPSDGF00	Has at least one good friend	8

Table A.4: New random variables (continue).

Number	Code	Label	Dataset
60	BPSDFB00	Fights with or bullies other children	8
61	BPSDUD00	Often unhappy	8
62	BPSDLC00	Generally liked by other children	8
63	BPSDDC00	Easily distracted	8
64	BPSDNC00	Nervous or clingy in new situations	8
65	BPSDKY00	Kind to younger children	8
66	BPSDOA00	Often argumentative with adults	8
67	BPSDPB00	CM picked on or bullied by other children	8
68	BPSDVH00	CM often volunteers to help others	8
69	BPSDST00	Can stop and think before acting	8
70	BPSDCS00	Can be spiteful to others	8
71	BPSDGB00	CM gets on better with adults than children	8
72	BPSDFE00	CM has many fears, is easily scared	8
73	BPSDTE00	CM sees tasks through to the end	8
74	BPSDEM00	If child has difficulties	8
75	BPPIAW00	Warm, affectionate relationship with child	8
76	BPPIST00	Struggle with child	8
77	BPPICO00	Child will seek comfort from me	8
78	BPPIPT00	Child uncomfortable with physical affection	8
79	BPPIVA00	Child values relationship with me	8
80	BPPIBP00	When praise child he/she beams with pride	8
81	BPPISI00	Child spontaneously shares information with me	8
82	BPPIAN00	Child easily becomes angry with me	8
83	BPPIET00	Easy to be in tune with child's feelings	8
84	BPPIAR00	Child angry or resistant after disciplined	8
85	BPPIDE00	Dealing with child drains my energy	8
86	BPPIBD00	Child bad mood, in for a long day	8
87	BPPIUC00	Child's feeling can be unpredictable towards me	8
88	BPPISM00	Child is sneaky/manipulative with me	8
89	BPPISE00	Child openly shares feelings/experiences	8
90	BPPPYK0G	Help from parents (Does NOT help in any of these ways)	10
91	BPSMUS0A	Current smoking cigarettes/cigars/pipe (Does not smoke)	10
92	BPCHSU00	Child suffers if mother works before school	10
93	BPNOSE00	Couples who have children should not separate	10
94	BPDRUG00	Used recreational drugs	10
95	BPWALI00	Life satisfaction	10

Table A.5: Average time (in seconds) of each method for model XX in each Monte Carlo scenario with $(n, m_i) = (40, 5)$, $(\phi_{1,1}, \phi_{2,2}) = (0, 0)$, $\mu = 1$, $d = 3$

Parameters	Methods		
	LS	Huber	Tukey
$\xi = 0$	1.068	1.068	1.075
$\xi = 0.5$	1.070	1.074	1.081
$\xi = 0.9$	1.066	1.074	1.078

Table A.6: Average time (in seconds) of each method for model XX in each Monte Carlo scenario with $(n, m_i) = (50, 5)$, $d = 1$

Parameters	Methods					
	QRFE	MQRFE	ERFE	PQRFE	PMQRFE	PERFE
$\gamma = 0, U_{ij}, T_i \sim N(0, 1)$	0.027	0.012	0.012	0.072	0.066	0.062
$\gamma = 0, U_{ij}, T_i \sim t_3$	0.027	0.013	0.011	0.073	0.066	0.061
$\gamma = 0, U_{ij}, T_i \sim \chi_3^2$	0.024	0.012	0.012	0.071	0.065	0.060
$\gamma = 0.1, U_{ij}, T_i \sim N(0, 1)$	0.028	0.011	0.013	0.072	0.066	0.060
$\gamma = 0.1, U_{ij}, T_i \sim t_3$	0.025	0.013	0.013	0.078	0.065	0.066
$\gamma = 0.1, U_{ij}, T_i \sim \chi_3^2$	0.025	0.012	0.012	0.069	0.065	0.060

Table A.7: Average time (in seconds) of each method for model XX in each Monte Carlo scenario with $(n, m_i) = (50, 5)$, $d = 4, 40$

Parameters	Methods		
	QRFE	LQRFE	ALQRFE
$d = 4, \pi_2 = 0.75, U_{ij}, T_i \sim N(0, 1)$	0.405	3.666	4.658
$d = 4, \pi_2 = 0.75, U_{ij}, T_i \sim t_3$	0.408	3.782	4.048
$d = 4, \pi_2 = 0.75, U_{ij}, T_i \sim \chi_3^2$	0.400	3.653	4.171
$d = 40, \pi_2 = 0.1, U_{ij}, T_i \sim N(0, 1)$	1.130	8.880	10.928
$d = 40, \pi_2 = 0.1, U_{ij}, T_i \sim t_3$	1.120	8.946	11.492
$d = 40, \pi_2 = 0.1, U_{ij}, T_i \sim \chi_3^2$	1.146	8.940	10.679