New Regression Methods for Measures of Central Tendency

A Thesis presented for the degree of
Doctor of Philosophy

by
Katerina Aristodemou

Supervised by
Dr Keming Yu

Department of Mathematical Sciences
School of Information Systems, Computing and Mathematics
Brunel University
England

2014

In memory of my beloved grandfather,

Michael Ivanovich Diadenko

Abstract

Measures of central tendency have been widely used for summarising statistical data, with the mean being the most popular summary statistic. However, in real-life applications it is not always the most representative measure of central location, especially when dealing with data which is skewed or contains outliers. Alternative statistics with less bias are the median and the mode.

Median and quantile regression has been used in different fields to examine the effect of factors at different points of the distribution. Mode estimation, on the other hand, has found many applications in cases where the analysis focuses on obtaining information about the most typical value or pattern. This thesis demonstrates that mode also plays an important role in the analysis of big data, which is becoming increasingly important in many sectors of the global economy.

However, mode regression has not been widely applied, even though there is a clear conceptual benefit, due to the computational and theoretical limitations of the existing estimators. Similarly, despite the popularity of the binary quantile regression model, computational straight forward estimation techniques do not exist.

Driven by the demand for simple, well-found and easy to implement inference tools, this thesis develops a series of new regression methods for mode and binary quantile regression. Chapter 2 deals with mode regression methods from the Bayesian perspective and presents one parametric and two non-parametric methods of inference. Chapter 3 demonstrates a mode-based, fast pattern-identification method for big data and proposes the first fully parametric mode regression method, which effectively uncovers the dependency of typical patterns on a number of covariates. The proposed approach is demonstrated through the analysis of a decade-long dataset on the Body Mass Index and associated factors, taken from the Health

Survey for England. Finally, Chapter 4 presents an alternative binary quantile regression approach, based on the nonlinear least asymmetric weighted squares, which can be implemented using standard statistical packages and guarantees a unique solution.

Declaration

I declare that the work presented in this thesis is my original research and has not been presented for a higher degree at any other university or institute.

.....

Katerina Aristodemou

Acknowledgements

I would like to express my sincere gratitude to Dr Keming Yu who has given me the opportunity to pursue a PhD. His knowledge, assistance and supervision have been invaluable. Also, I would like to thank my research collaborators for their valuable contributions. Furthermore, I would like to thank all the staff and fellow students at Brunel University.

A special thank you goes to my sister Eleni and my beloved husband Marinos for all their love and understanding. Without them this degree would not have been possible. Words cannot express how grateful I am to my parents Dafnis and Galina for their endless love and the sacrifices they made on my behalf. Finally, I would like to thank my family and friends for their permanent encouragement and support.

Authors Publication

1. Yu, K., Aristodemou, K. and Zudi, L. (2014), Bayesian Mode Regression. *Submitted for publication.*

2. Yu, K., Aristodemou, K., Becker, F. and Lord, J. (2014), Fully Parametric Mode Regression in Big Data of Body Mass Index. *Working paper.*

3. Aristodemou, K. and Yu, K. (2014) Binary Quantile Regression and Variable Selection: An Alternative Estimation Approach. *Submitted for publication.*

4. Aristodemou, K. and Yu, K. (2010) Bayesian Estimation of Error in Variables Models through Bayesian Quantile Regression with Application to the Permanent Income Hypothesis. Quantile Regression Workshop: Theory and Applications. Berlin, Germany *Paper presentation.*

5. Aristodemou, K. and Yu, K. (2008). CAViaR via Bayesian Nonparametric Quantile Regression, Workshop on Inference and Estimation in Probabilistic Time-Series Models, Isaac Newton Institute, University of Cambridge, Cambridge. *In Proceedings.*

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Processing and understanding large quantities of random data has always been a challenging task. Descriptive statistics have been extensively used to summarise sets of observations, in order to communicate large amounts of information in a simplified, sensible and concise form. Such summary statistics include measures of central tendency, distribution, and dispersion. A measure of central tendency (also referred to as measure of central location) is a summary measure that attempts to describe a dataset with a single value that represents the middle or the centre of the distribution and aims at providing a representative description of the entire distribution of scores. Although many measures of central tendency have been recognised and used, three of these measures are of particular importance: (1) the mean, (2) the median, and (3) the mode.

- The **population mean** is the average value of all the measurements in the population. It is estimated by the sample mean which is equal to the sum of all the values in a sample divided by the number of observations in the sample. The sample mean is the most commonly used measure of central tendency.

- The **population median** is the point in the population above which and below 50% of the scores lie. It is estimated by the sample median, which is the middle value in an ordered sequence of observations in a sample.

- The **population mode** is the most likely value of the population. It is estimated by the sample mode, which is the value that occurs most often (has the

highest frequency) in a sample.

Depending on the problem at hand, different measures of central tendency may be appropriate. The choice of the most appropriate measure depends mainly on the following three factors:

1. **Level of measurement of the data**: In case of interval-ratio variables, all three measures are suitable for analysis. For ordinal variables, both the mode and the median are appropriate whereas for nominal variables, the mode is the only measure that can be used.

2. **Shape of the distribution**: In a symmetrical distribution the mean, the median and the mode coincide, thus the choice depends on the level of measurement of the data and on the objective of the analysis. In a skewed distribution, or in the presence of outliers, the mean is pulled in the direction of the tail, dragging it away from the typical value and making it a less representative measure of central tendency. However, both the median and the mode are robust to the presence of outliers. The mode, being the peak of the distribution retains its position, whereas the median is influenced much less by the skewed values. Usually, in skewed data, the median is located between the mean and the mode.

3. **Objective of the analysis/ research question**: When the objective of the analysis is to identify the average value in a dataset, then the mean is the most appropriate measure. The median provides information on the middle value and combined with other quantiles, can provide a complete picture of the distribution of the data. The mode is the most appropriate measure when the objective is to identify either the most typical value, i.e. to identify patterns, or the value that occurs most often.

Unlike descriptive statistics, which are used to describe the characteristics of a single variable, inferential statistics are used to make predictions or inferences about a population on the basis of a sample. Examples of inferential statistics include, among others, regression analysis, logistic regression, analysis of variance (ANOVA), correlation analysis, structural equation modelling and survival analysis.

Regression analysis is one of the most commonly used statistical techniques in social, behavioural and physical sciences. Its main objective is to quantify the relationship between a response variable y and a set of explanatory variables $\boldsymbol{x}$ through a mathematical model which can be used for inference, prediction and hypothesis testing.

1.0.1 Mean Regression

Conventional regression models aim at inferring the relationship between one or more explanatory variables, X, and the response variable y given $X = x$, by estimating a mean regression function, $m(x)$ which provides an estimate of the conditional expectation $E(y|x)$. The standard mean regression function is defined as

$$Y = m(x) + \epsilon,$$

under the assumption $E(\epsilon|x) = 0$. The aim of the analysis is to estimate the mean regression function that provides the best fit for the data. The method of least squares is a standard approach to determine the best fit by minimising the sum of squared residuals. Least squares methods can provide a solution for both linear and nonlinear functional forms of $m(x)$. In the case of a linear functional form, the estimation is performed through ordinary least squares (OLS) which has a closed form solution, whereas in the case of a nonlinear functional form, a closed form solution is not available and the problem is solved via iterative optimisation methods.

In linear regression models the mean regression function is modelled as a linear function of the explanatory variable, such that,

$$m(x) = \boldsymbol{x}'\boldsymbol{\beta},$$

where $\boldsymbol{\beta}$ is a set of unknown parameters to be estimated. The OLS procedure is the simplest and most common type of estimation procedure used for statistical analysis. OLS is the Best Linear Unbiased Estimator (BLUE) under the Gauss-Markov Assumptions:

1. The model must be linear in the parameters.

2. $E(\epsilon|\boldsymbol{x}) = 0$

3. $Var(\epsilon|\boldsymbol{x}) = \sigma^2$ (homoscedasticity assumption.)

4. $Cov(\epsilon_i, \epsilon_j|\boldsymbol{x}) = 0 \quad \forall \quad i \neq j.$

5. No perfect multicollinearity between independent variables.

The conditional mean $E\left[y_i|x_i\right] = \boldsymbol{x}_i'\boldsymbol{\beta}$ is estimated by solving the following minimisation problem:

$$\hat{\boldsymbol{\beta}} = \arg\min \sum_{i=1}^{n} \left(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}\right)^2. \tag{1.0.1}$$

A normality assumption is not required for the consistency of the OLS estimates. However, under the normality assumption $\epsilon|\boldsymbol{x} \sim N(0, \sigma^2\mathbf{I})$, the OLS estimator is equivalent to the maximum likelihood estimator (MLE). The method of maximum likelihood (ML) chooses the values of the parameters that are most consistent with the data.

Let $x_i, ..., x_n$ be a random sample of independent and identically distributed (iid) observations and $y_i \sim f(y_i; \boldsymbol{\beta}, x_i)$, where f is a known probability density function, then the MLE of $\boldsymbol{\beta}$ can be obtained by optimising the log-likelihood function, $L(\boldsymbol{\beta}|y)$, where

$$L(\boldsymbol{\beta}|y) = \frac{1}{n}\sum_{i=1}^{n} \log f(y_i; \boldsymbol{\beta}, x_i).$$

Under the distributional assumption, the ML method can be used to estimate the regression parameters in any given regression model. Furthermore, MLE estimates enjoy standard large sample properties (consistency and asymptotic normality).

1.0.2 Quantile Regression

In many real-world applications, the estimation of the conditional mean proves to be inadequate for describing the behaviour of the conditional distribution of the response variable y. This is particularly true for asymmetric response distributions and distributions which contain outliers. Data with such characteristics can be found in many fields, including econometric, survival analysis and ecology.

Quantile regression estimates either the conditional median or other quantiles of the response variable y. It provides an alternative approach to estimate models with skewed data, as it is able to provide a complete picture of the conditional distribution of the response variable when a set of quantiles is modelled. This is particularly useful when the effect of the covariates on the upper or lower quantiles of the response variable vary differently from the centre or in cases where modelling the extremes of the conditional distribution is of special interest, e.g. in the analysis of financial or environmental data. The main advantage of quantile regression over least-squares regression is its flexibility for modelling data with heterogeneous conditional distributions. It makes no distributional assumption about the error term in the model and it is less sensitive to the presence of outliers in the dependent variable.

An additional limitation of the least-squares regression is the assumption that the covariates affect only the location of the conditional distribution of the response, and not its scale or any other aspect of its distributional shape (homoscedasticity), an assumption that often fails in practice. A major advantage of quantile regression is its capability of capturing both a location and a scale shift in the response variable, by allowing the regression parameters to vary at various points of the conditional distribution; thus allowing the examination of the way covariates influence the location and scale of the entire response distribution.

Since the seminal paper of Koenker and Bassett (1978), quantile regression gradually became a complimentary approach for the traditional conditional mean estimation method and today it has become a dominant approach in empirical work in several fields of study.

As introduced by Koenker and Bassett (1978), the classical quantile regression model, corresponding to the linear model in (1.0.1) is defined as:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta}(\tau) + \epsilon_i$$

where $(0 < \tau < 1)$ represents the quantile level.

The τ^{th} conditional quantile function is defined as $Q_\tau\left[y_i|\boldsymbol{x}_i\right] = \boldsymbol{x}_i'\boldsymbol{\beta}(\tau)$ by assuming that the $Q_\tau[\epsilon_i|\boldsymbol{x}_i] = 0$ (in contrast to the assumption of $E[\epsilon_i|\boldsymbol{x}_i] = 0$ in mean

regression).

Estimates for $\boldsymbol{\beta}(\tau)$ are obtained by solving the following minimisation problem:

$$\hat{\boldsymbol{\beta}}(\tau) = \arg\min \sum_{i=1}^{n} \rho_\tau \left(y_i - \boldsymbol{x}_i' \boldsymbol{\beta}(\tau)\right)$$

where ρ_τ is the quantile check function defined as:

$$\rho_\tau(u) = \tau\, u I_{[0,\infty)}(u) - (1-\tau)\, u I_{(-\infty,0)}(u) \tag{1.0.2}$$

where $I(\cdot)$ is the indicator function.

1.0.3 Mode Regression

Like mean, quantile and variance, mode is also an important measure of central tendency. Many practical questions, particularly in the analysis of big data, often focus on "which element (gene or file or signal) is the most typical among all elements in a network?" In such cases, mode regression is able to provide a summary of how the regressors affect the conditional mode and is completely different from other models that are based on conditional mean, conditional quantile or conditional variance. The mean or median of two densities may be identical, while the shapes of the two densities can be quite different. The mode preserves some of the important features, such as wiggles, of the underlying distribution function, whereas the mean and the median tend to average out the data.

1.1 Motivation

The motivation behind the work in this thesis is twofold. First, it is the reinforcement of the importance of mode as an important measure of central tendency, especially in light of its suitability for the big data analysis. Second, it is the lack of simple, well-founded and easy to implement statistical methods for mode regression and binary quantile regression. The sub-sections below describe the identified gaps in literature in these two areas.

1.1.1 Mode Estimation and Mode Regression

A mode estimator is often defined as the maximum of the estimated distribution density. Conditional mode estimation is typically carried out by conditional density estimation via different nonparametric methods. Mode estimation and regression can play an important role in terms of identifying the typical value or pattern in a dataset but also in terms of inference and prediction.

Despite its advantages, mode regression has not been adequately studied in the literature. Lee (1989,1993) explored direct inference for mode regression and focused on the case where the dependent variable is truncated. This work introduced a method of estimating the conditional mode of y given $\boldsymbol{x}$, $mode(y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$, by minimising with respect to $\boldsymbol{\beta}$, as σ approaches 0, a loss function $K(y; \boldsymbol{x}'\boldsymbol{\beta}, \sigma)$. However, this method has not been well-applied due to a lack of proper inference tools. Recently, Kemp and Silva (2012) proposed a semi-parametric mode regression estimator for the case in which the variable of interest is unbounded, continuous and observable over its entire support and Yao and Li (2014) proposed an Expectation-Maximisation algorithm in order to estimate the regression coefficients of modal linear regression. However, this research involves either semiparametric or nonparametric estimation of regression parameters, it has a slow rate of convergence and it is subject to bandwidth selection; thus it has little, if any, practical use. Furthermore, a direct Bayesian method for mode regression is not available even though, there is clear practical motivation from this perspective.

An example of the applicability of mode estimation and regression in real life is the analysis of big data. Big data analysis is the process of examining large amounts of data of different types to uncover hidden patterns, unknown correlations and other useful information. Big data is becoming increasingly important in every sector and function of the global economy. Recently, there has been considerable amount of research effort devoted to managing and analysing such types of data. Often, in big data analysis, fast and accurate identification of patterns is required. Even though this can be achieved through available data-mining/pattern-finding algorithms, mode can serve as a quick, effective and statistically meaningful alternative technique for pattern recognition. Identifying the typical value or pat-

tern is an important step in big data analysis. However, this is not the sole aim of statistical analysis or the only scientific objective. Quantifying the relationship between the pattern and other covariates is often desirable. Mode regression, which models the relationship between the typical value and a set of explanatory variables, could achieve this goal.

1.1.2 Binary Quantile Regression

Quantile regression has been widely used in many areas such as economics, ecology and finance, as an alternative to mean regression in cases of skewed data and it has been applied to different types of models, such as time series models, survival analysis models, censored regression and count data models.

Binary quantile regression was first introduced by Manski (1975, 1985). In these papers the Maximum Score Estimator (MSE) was developed, which requires very weak assumptions governing the relation of errors to regressors and can accommodate for heteroscedasticity of unknown form. However, this work faces important technical drawbacks in both optimising the objective function and in making inference on the regression parameters. To overcome some of these shortcomings Horowitz (1992) developed a Smoothed Maximum Score Estimator (SMSE) for the linear median regression case, which can be computed using standard optimisation routines. Kordas (2006) extended this estimator to a family of conditional quantile functions giving the opportunity for a complete understanding of the conditional distribution of the latent response variable given covariates.

Even though both the maximum score and the smoothed maximum score estimators have desirable asymptotic properties, they are difficult to implement in practice. The maximum score estimator has a discontinuous objective function (step-function) and hence it cannot be solved with gradient-based optimisation methods. The objective function of the smoothed maximum score estimator can have several local maxima, thus requiring stochastic search algorithms to identify the global maximum (e.g. the simulated annealing algorithm suggested by Horowitz (1992)). Due to the complex structure of the objective functions of these estimators, standard optimisation heuristics cannot guarantee global convergence. In addition, even though

algorithms for solving both the MSE and the SMSE are readily available, these are not included in standard software packages.

1.2 Contributions

In an effort to address the identified gaps in the area of mode regression and to provide an alternative, easy to implement, quantile estimation approach for the popular binary model, several research contributions have been achieved. The overall aim of this work was the development of estimators that enjoy good finite and large sample properties but also computational simplicity.

The contributions of this thesis can be summarised as follows.

- **Parametric Bayesian mode regression**: A parametric Bayesian mode regression model has been developed, where the likelihood function is based on a uniform probability density. Bayesian inference has the advantage of being able to provide the entire posterior distribution of the parameters under investigation and to allow uncertainty to be taken into account when making predictions. Furthermore, the Markov Chain Monte Carlo (MCMC) sample can be used to estimate the covariance matrix, and other asymptotic quantities of classical estimates. The proposed estimator enjoys good finite sample performance and large sample properties.

- **Nonparametric Bayesian mode regression**: The parametric Bayesian mode regression depends on the assumption of a uniform mode distribution, which may lead to inconsistent estimation due to model misspecification. With the aim of relaxing the distributional assumption and enhancing the flexibility of the model, two nonparametric Bayesian mode regression models have been developed. In the first model the estimator is based on the characterisation of the mode uniform distribution as a scale mixture of symmetric uniform distributions using a Dirichlet process prior for the model parameter σ. The method is nonparametric in the sense that it is not assumed that the prior belongs to any fixed class of distributions. The second method is an alternative estimation approach based on an empirical likelihood ratio. Empirical likelihood

is a nonparametric approach which combines the reliability of nonparametric methods with the flexibility and effectiveness of the likelihood approach and also demonstrates good large sample properties.

- **Fully parametric mode regression**: A simple fully parametric mode regression model has been developed, based on the Gamma density, which has both good theoretical properties and finite sample results and is easy to implement. The Gamma distribution is a very flexible density which can take several different shapes, thus making it suitable for data-driven statistical modelling. The estimation method involves first re-parameterising the Gamma density in terms of the mode of y and then introducing a regression-based functional form.

- **Mode estimation and big data analysis**: Big data usually includes datasets of a substantial size, which are difficult to capture, manage and analyse using existing software tools in a sensible amount of time. An alternative method for the analysis of big data has been proposed, which combines mode estimation and mode regression. Standard mode estimation techniques can serve as alternative tools for quick and meaningful pattern recognition in big data. Inference is made possible by isolating the data corresponding to these recognised patterns in a separate dataset to be used for analysis. Mode regression can then be applied to examine the relationships between the variables in this new dataset. The proposed methodology is demonstrated via the analysis of the results of the Health Survey for England for the years 1997-2010, which aims at exploring the effect of socio-economic characteristics and behavioural habits of adults in England on the typical Body Mass Index (BMI). Given the increasing focus on big data analytics, the timeliness of the proposed methodology can play a significant role in finding its way towards current and future application domains.

- **Binary quantile regression**: An alternative methodology for binary quantile regression, based on iteratively reweighted least squares has been developed. The method is computationally simple, is guaranteed to converge to a unique

solution and can be implemented with standard software packages, as the proposed estimator is based on standard gradient-based optimisation methods which generally converge much faster than stochastic search algorithms.

- **Variable selection for binary quantile regression**: Multicollinearity and overfitting are areas of concern in models with a large number of explanatory variables. Variable selection plays an important role in the model-building process. The proposed variable selection method is based on the modern adaptive lasso approach, which allows different shrinkage weights for different regression coefficients of independent variables. The method provides consistent variable selection and optimal prediction and also enjoys the oracle property.

- **Development of algorithms**: All the methods developed in this thesis have been implemented and tested in the free statistical software R. The algorithms for the implementation of the methods presented in this work can be made available to the statistical community upon request. In addition the possibility of developing R packages will be considered.

1.3 Thesis Structure

This Chapter introduced the basic principles of descriptive and inferential statistics, described the motivation for investigating new regression methods for measures of central tendency and presented the research contributions. The remainder of this thesis is organised as follows:

Chapter 2 - Bayesian Mode Regression: This Chapter introduces Bayesian mode regression by developing three different approaches: a parametric Bayesian method, a nonparametric Bayesian method and an empirical likelihood-based Bayesian method. It also provides their theoretic properties and application.

Chapter 3 - Fully Parametric Mode Regression for Big data Analysis: This Chapter initially demonstrates a fast mode-based pattern recognition method for Big data, and then introduces the first fully parametric method for mode regression, based on the Gamma density.

Chapter 4 - Binary Quantile Regression and Variable Selection: This Chapter demonstrates an alternative estimation approach for binary quantile regression and variable selection, which is efficient and can be implemented with standard software packages.

Chapter 5 - Conclusions and Future Work: The last Chapter concludes this thesis by summarising the work and discussing the contributions. Future directions of this research are also suggested.

Chapter 2

Bayesian Mode Regression

2.1 Introduction

Mode, the most likely value of a distribution, has wide applications in biology, astronomy, economics and finance. In these fields, it is not uncommon to encounter data distributions that are skewed or contain outliers. In those cases, the arithmetic mean may not be an appropriate statistic to represent the center of location of the data. Alternative statistics with less bias are the median and the mode. The mean or the median of two densities may be identical, while the shapes of the two densities can be quite different. The mode preserves some of the important features, such as wiggles, of the underlying distribution function, whereas the mean and the median tend to average out the data.

The mode has been used in modern science to identify the most frequent or the most typical element in certain network systems (Hedges and Shah (2003), Heckman et al. (2001), Kumar and Hedges (1998), Markov et al. (1997)). Mode estimation has attracted significant attention in the statistics literature for decades by various authors [Yasukawa (1926), Parzen (1962), Grenander (1965), Eddy (1980), Bickel and Fan (1996), Birgé (1997), Berlinet et al. (1998) and Meyer (2001) among others]. Moreover, identifying the typical value or pattern could be one of the most efficient statistical approaches for the analysis of big data.

However, mode estimation is more difficult than estimating the mean or the median. The mode estimator is often defined as the maximum of the estimated

distribution density, typically under nonparametric kernel estimation. Conditional mode estimation is typically carried out by conditional density estimation via different nonparametric methods [see for example Gasser et al. (1998), Hall and Huang (2001) and Hall et al. (2001), Brunner (1992), Ho (2006), Dunson et al. (2007)].

However, these nonparametric conditional density-based mode regression models do not provide a direct estimate of the conditional mode. The problem with these methods is twofold: the estimation of the conditional density may suffer from the well-known "curse of dimensionality" and, it is hard to describe and interpret the estimated conditional mode in terms of predictors or covariates.

Direct inference for mode regression was explored by Lee first in 1989, Lee (1989), and then in 1993, Lee (1993). However, it has not been well-applied due to lack of proper inference tools. Recently, Kemp and Silva (2012) relaxed Lee's restriction on truncated dependent variables and employed alternative kernel estimation. However, their regression coefficient estimator has slow convergence rate, involves bandwidth selection and provides only approximate Normal confidence intervals. Furthermore, Yao and Li (2014) proposed an Expectation-Maximisation algorithm in order to estimate the regression coefficients of the modal linear regression. These methods involve either semiparametric or nonparametric estimation methods. A direct Bayesian method for mode regression is not available even though there is a clear practical motivation from this perspective, given the practical and theoretical advantages of the Bayesian approach (e.g. incorporation of prior information to the analysis, estimation of the complete density distribution for the parameters of interest rather than a single point estimate as in the classical approach, delivery of exact inferences which do not rely on large sample approximations, etc.).

In conventional regression models, the method of least squares is usually applied to investigate the effect of the predictor variables on the conditional mean of the response variable. However, in the presence of outliers, the mean is pulled in the direction of the tail, making mean regression a less representative method of analysis. Mode regression, on the other hand, is robust to the presence of outliers. Quantile regression is an alternative approach to estimate models with skewed data, as it can provide a complete picture of the conditional distribution of the response variable

given the covariates. However, it cannot reveal any information about the typical value (mode).

Take the analysis of the adult Body Mass Index (BMI) used in this chapter as an example. BMI, defined by $BMI = \frac{weight(kg)}{height^2(cm)}$, is a measure of the relative weight and is used in a wide variety of contexts as a simple method to assess how much an individual's body weight deviates from what is normal or desirable for a person of his or her height. Such analysis is important as it is well-known that obesity has overtaken smoking as the biggest threat to people's health, in particular for middle-aged and old adults.

The dataset used in this chapter to demonstrate mode regression is taken from the Health Survey for England (HSE) 2011 teaching dataset. The Health Survey for England is a series of annual surveys about the health of people living in England, commissioned by the Department of Health. The sample contains observations for 4,138 individuals (1,814 males and 2,324 females) with two thirds being older than 40 years old. A BMI of $27kg/m^2$ for middle-aged and old adults can be classified as the cut-off point of unhealthy weight. An interesting question is how some covariates, such as units of alcohol and portions of fruit/vegetables consumed keep one's BMI in the healthy range. It would be safe to assume that the BMI for the majority of people in the data example falls in the desirable BMI range. Indeed, the typical BMI for the whole sample as well as separately for men or women are below $27kg/m^2$ (see Table 2.1), but the corresponding mean BMI and median BMI were near or greater than $27kg/m^2$. The plots in Figure 2.1 suggest that the location of the peak can be considered as the most representative measure of central tendency. Therefore, employing mode regression is preferable than mean and quantile regression for answering this scientific question.

This chapter introduces a fully Bayesian framework for direct mode regression by using three approaches: a parametric Bayesian method, a nonparametric Bayesian method and a nonparametric empirical likelihood based Bayesian method. The remainder of the chapter is organised as follows. Sections 2.2 and 2.3 introduce the three approaches, describe the theoretical and computational framework of these methods and give their mathematical justification. Section 2.4 illustrates the pro-

Table 2.1: Summary Statistics for the BMI dataset

Variable	Obs	Mean	SD	Median	Mode	Min	Max
			Total				
BMI	4138	27.7	5.13	26.9	26.1	15.9	56.0
age	4138	50.8	17.8	50	64	16	96
alcohol	4138	11.0	18.1	4.62	0	0	378.0
fruit&veg	4138	3.79	2.69	3.33	1	0	30
smoking	4138	1.22	0.58	1	1	1	3
male	4138	0.44	0.50	0	0	0	1
			Male				
BMI	1814	27.8	4.55	27.3	26.7	16.3	56.0
age	1814	51.5	17.7	51	64	16	94
alcohol	1814	15.0	21.4	8.57	0	0	378.0
fruit&veg	1814	3.67	2.65	3.33	1	0	29.3
smoking	1814	1.22	0.60	1	1	1	3
			Female				
BMI	2324	27.3	5.51	26.4	24.5	15.9	52.4
age	2324	50.17	17.8	49	64	16	96
alcohol	2324	7.84	14.4	2.48	0	0	378.0
fruit&veg	2324	3.88	2.71	3.5	2	0	30
smoking	2324	1.22	0.58	1	1	1	3

Note: *age* = person's age, *alcohol* = the total units of alcohol consumed per week, $fruit\&veg$ = the portion of fruit and vegetables consumed the previous day, *smoking* = the person's cigarette smoking status (0= Non-smoker, 1= Light smokers, under 10 a day, 2= Moderate smokers, 10 to under 20 a day, 3=Heavy smokers, 20 or more a day).

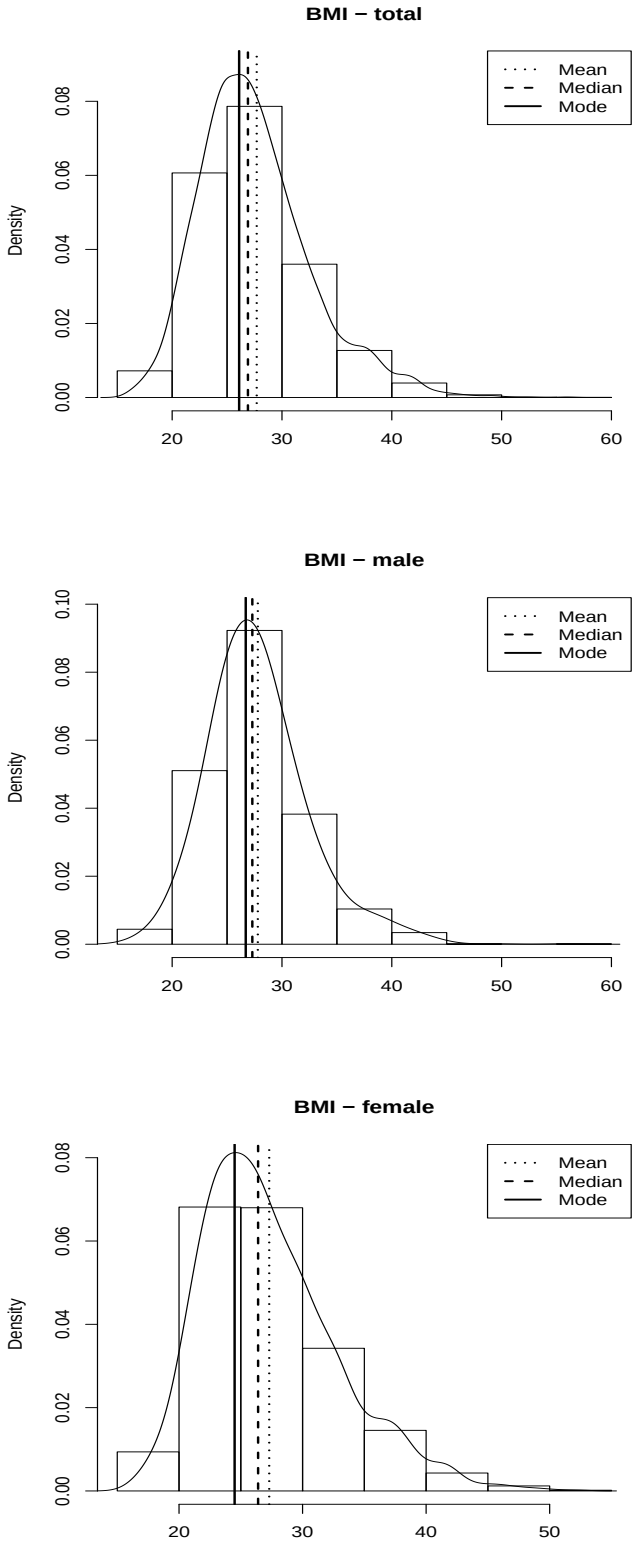


Figure 2.1: BMI dataset: BMI histograms for total, men and women

posed methods through three simulated case-studies and a real example. Concluding remarks are provided in Section 2.5.

2.2 Bayesian Mode Regression

2.2.1 Mode Estimation and Classical Mode Regression

Consider an arbitrary random variable Z, with distribution function $F_Z(z)$ and density function $f_Z(z)$. Let $K(Z;\cdot)$ be the *step-loss function* (Manski (1991)) defined by,

$$K(Z;\mu) = I\left[\frac{|Z-\mu|}{\sigma} > 1\right], \tag{2.2.1}$$

with $\sigma > 0$ and $I[A]$ being the indicator function of event A. If $f_Z(z)$ is symmetric around μ or if μ is the middle value of the interval of length 2σ that captures the most probability under $F_Z(z)$, then

$$\hat{\mu} = \text{argmin}_\mu E\{K(Z;\mu)\}$$

is the mode of Z.

Therefore, given a sample $\{Z_1, \ldots Z_n\}$ from Z, let $\hat{\mu}$ be the estimator of the mode of Z, then,

$$\hat{\mu} = \text{argmin}_\mu \sum_{i=1}^{n} I\left[|Z_i - \mu| > \sigma\right] \quad \Leftrightarrow$$

$$\hat{\mu} = \text{argmax}_\mu \sum_{i=1}^{n} I\left[|Z_i - \mu| \leq \sigma\right] \quad \Leftrightarrow$$

$$\hat{\mu} = \text{argmax}_\mu \exp\left(\sum_{i=1}^{n} I\left[|Z_i - \mu| \leq \sigma\right]\right) \Leftrightarrow$$

$$\hat{\mu} = \text{argmax}_\mu \prod_{i=1}^{n} \exp\left(I\left[|Z_i - \mu| \leq \sigma\right]\right).$$

Consider the uniform probability density function, $f(u)$, such that

$$f_\sigma(u) = \frac{e}{2\sigma}\exp(-I[|u-\mu| \leq \sigma])I[|u-\mu| \leq \sigma], \tag{2.2.2}$$

for a window parameter $\sigma > 0$. Then maximising $I[|u-\mu| \leq \sigma]$ is equivalent to minimising $f_\sigma(u)$.

Lee (1989) introduced mode regression by defining the conditional mode of y given $\boldsymbol{x}$, as $mode(y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$ based on the loss function $K(y; \boldsymbol{x}'\boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is the regression parameter. That is, given a sample $\{(\boldsymbol{x}_1, y_1), ...(\boldsymbol{x}_n, y_n)\}$ from $(\boldsymbol{x}, y)$, as σ approaches 0, the parameter $\boldsymbol{\beta}$ in the conditional model of $y|\boldsymbol{x}$ is estimated by

$$\hat{\boldsymbol{\beta}} = \text{argmin}_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} K(y_i;\, \boldsymbol{x}_i'\boldsymbol{\beta}) \tag{2.2.3}$$

2.2.2 Bayesian Inference and the Markov Chain Monte Carlo (MCMC) Method

Bayesian inference is a powerful statistical method that can be used to estimate unknown parameters in regression models by constructing posterior densities conditional on observed data. Let $\boldsymbol{\omega}$ be the vector of unknown model parameters to be estimated. According to the Bayes' Theorem, the joint posterior distribution of the unknown parameters, $f(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{x})$, is given by:

$$f(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{x}) = \frac{f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\omega}) f(\boldsymbol{\omega})}{\int f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\omega}) f(\boldsymbol{\omega}) d\boldsymbol{\omega}} \Rightarrow \tag{2.2.4}$$

$$f(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{x}) \propto f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\omega}) f(\boldsymbol{\omega})$$

where, $f(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\omega})$ is the likelihood of the data given the unknown parameters and $f(\boldsymbol{\omega})$ is the joint prior distribution of the unknown parameters.

Evaluating the joint or marginal posterior densities by analytic or numerical methods can be extremely difficult. MCMC techniques can be easily applied to obtain samples from the posterior distribution of the unknown parameters.

A MCMC scheme constructs a Markov chain whose equilibrium distribution is the posterior distribution $f(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{x})$. After running the Markov chain for a burn-in period, one obtains samples from the limiting distribution, provided that the Markov chain has converged. The Metropolis-Hastings algorithm (Hastings (1970)) is one of the more prominent MCMC methods for simulating realisations from the posterior distribution of the unknown parameters. The steps of the Metropolis-Hastings algorithm are described in Algorithm 1.

Algorithm 1 The Metropolis Hastings Algorithm

1: Let T be the number of MCMC iterations.
2: Set $t = 0$ and initialise the parameter to be estimated by setting $\boldsymbol{\omega}(t) = \boldsymbol{\omega}(0)$.
3: Generate a proposal value, $\boldsymbol{\omega}_{new}$ from a chosen proposal density $q(\cdot|\boldsymbol{\omega}(t))$
4: Calculate the acceptance probability
5: $r = min\left\{1, \frac{q(\boldsymbol{\omega}(t)|\boldsymbol{\omega}_{new})}{q(\boldsymbol{\omega}_{new}|\boldsymbol{\omega}(t))} \frac{f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\omega}_{new})f(\boldsymbol{\omega}_{new})}{f(\boldsymbol{y},\boldsymbol{x}|\boldsymbol{\omega}(t))f(\boldsymbol{\omega}(t))}\right\}$.
6: Sample u from the uniform distribution $U(0, 1)$.
7: Set $\boldsymbol{\omega}(t + 1) = \boldsymbol{\omega}_{new}$ if $u < r$, otherwise $\boldsymbol{\omega}(t + 1) = \boldsymbol{\omega}(t)$
8: Set $t = t + 1$, if $t < T$ then return to step 2.

Given that the chain has converged, the posterior distribution of $\boldsymbol{\omega}$ is given by the frequency of appearance of the parameters in the Markov chain. This provides the complete density distribution of the estimated model parameters, rather than a single point estimate as in the classical approach. This is one of the major advantages of Bayesian inference.

Let S be the Markov chain drawn from the posterior distribution, $f(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{x})$, such as $S = (\boldsymbol{\omega}^{(1)}, \boldsymbol{\omega}^{(2)}, ..., \boldsymbol{\omega}^{(N)})$ where N is the number of draws after burn-in. Then, it is possible to compute the parameter estimates $\widehat{\boldsymbol{\omega}}$, by calculating a descriptive statistic of the Markov chain, S, e.g. the posterior mean is computed as: $\widehat{\boldsymbol{\beta}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\omega}^{(i)}$.

Chernozhukov and Hong (2003) showed that under general regularity conditions the posterior distribution concentrates at a rate $\frac{1}{\sqrt{n}}$ around the true parameter ω_0, that the estimators are consistent and asymptotically Normal and that the posterior quantiles or other relevant quantities provide asymptotically valid confidence intervals.

2.2.3 Parametric Bayesian Method

The conditional linear mode regression, denoted as $mode(y|\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{\beta}$, can be formulated as a standard regression model:

$$y = \boldsymbol{x}'\boldsymbol{\beta} + \epsilon$$

with $mode(\epsilon|\boldsymbol{x}) = 0$.

Lee (1989,1993) showed that, given a sample $\{(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_n, y_n)\}$ from $(\boldsymbol{x}, y)$, the classical mode regression estimator, $\hat{\boldsymbol{\beta}}$ is given by:

$$\hat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} I[|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}| \leq \sigma]. \tag{2.2.5}$$

Therefore, using equation (2.2.2), $\hat{\boldsymbol{\beta}}$ can be regarded as the maximum likelihood estimator of the "working" likelihood function

$$L(y|\boldsymbol{\beta}, \sigma) = \frac{e^n}{(2\sigma)^n} \prod_{i=1}^{n} \exp\left(-I[|y_i - \boldsymbol{x}_i'\boldsymbol{\beta}| \leq \sigma]\right) I\left[|u - \mu| \leq \sigma\right]. \tag{2.2.6}$$

Therefore, under a Bayesian framework, the joint posterior distribution of the unknown model parameters, $\boldsymbol{\beta}$ and σ, is given by

$$\pi(\boldsymbol{\beta}, \sigma|y) \propto L(y|\boldsymbol{\beta}, \sigma)\, \pi(\boldsymbol{\beta}, \sigma), \tag{2.2.7}$$

where $\pi(\boldsymbol{\beta}, \sigma)$ is the joint prior distribution of $\boldsymbol{\beta}$ and σ.

The Bayesian mode regression estimates, denoted as $\hat{\boldsymbol{\beta}}_B$ can be obtained using the marginal posterior distribution of $\boldsymbol{\beta}$, given by

$$\pi(\boldsymbol{\beta}|y) = \int \pi(\boldsymbol{\beta}, \sigma|y) d\sigma, \tag{2.2.8}$$

In a similar manner, an estimate of σ, denoted as $\hat{\sigma}$, can be obtained using the marginal posterior distribution of σ,

$$\pi(\sigma|y) = \int \pi(\boldsymbol{\beta}, \sigma|y) d\boldsymbol{\beta}, \tag{2.2.9}$$

Although a standard conjugate prior distribution is not available for the mode regression formulation, Markov Chain Monte Carlo (MCMC) methods may be used for extracting the posterior distributions of both $\boldsymbol{\beta}$ and σ.

2.2.4 Estimation of Covariance Matrix of Classical Estimates

Under the classical approaches of Lee (1989, 1993) and Kemp and Silva (2012), the covariance matrix, $\Sigma(\hat{\boldsymbol{\beta}})$ of the classical estimator $\hat{\boldsymbol{\beta}}$ and its inverse are often required but difficult to estimate or compute numerically, especially under a small or moderate sample size.

As mentioned in Section 2.2.2, an additional advantage of the proposed Bayesian approach is the ability of obtaining a natural and efficient estimator of $\Sigma(\boldsymbol{\beta})$ and other asymptotic quantities of $\boldsymbol{\beta}$, using the MCMC posterior sample. A consistent estimate of the inverse of the covariance matrix can be obtained by multiplying by n the variance-covariance matrix of this MCMC sequence (Chernozhukov and Hong (2003)).

A 95% Confidence interval (CI) for $\widehat{\boldsymbol{\beta}}$ can then be easily derived from this posterior distribution by taking the 0.05^{th} and 0.95^{th} quantiles of the Markov chain S.

2.2.5 Prior Selection and Proper Posteriors

In this sub-section, first it is demonstrated that almost all priors for $(\boldsymbol{\beta}, \sigma)$ could be used and yield a proper joint posterior. Then one practical selection for a prior for σ is provided.

Consider the following theorem.

Theorem 2.2.1. *Given the mode regression (2.2.3) and the 'working' likelihood (2.2.6), if the joint prior distribution $\pi(\boldsymbol{\beta}, \sigma)$ follows one of the following three choices:*

(1) $\pi(\boldsymbol{\beta}, \sigma) \propto 1$ (totally non-informative prior)

(2) $\pi(\boldsymbol{\beta}, \sigma) = \pi(\boldsymbol{\beta})\,\pi(\sigma|\boldsymbol{\beta})$ and one of $\pi(\boldsymbol{\beta})$ and $\pi(\sigma|\boldsymbol{\beta}) \propto 1$ and the other is a proper prior,

(3) $\pi(\boldsymbol{\beta}, \sigma) = \pi(\boldsymbol{\beta})\,\pi(\sigma|\boldsymbol{\beta})$ and both $\pi(\boldsymbol{\beta})$ and $\pi(\sigma|\boldsymbol{\beta})$ are proper priors,

then the posterior distribution of $\boldsymbol{\beta}$ and σ, $\pi(\boldsymbol{\beta}, \sigma|\boldsymbol{y})$, will be a proper distribution. In other words

$$0 < \int \pi(\boldsymbol{\beta}, \sigma|\boldsymbol{y})\, d\boldsymbol{\beta}\, d\sigma < \infty,$$

or, equivalently,

$$0 < \int L(\boldsymbol{y}|\boldsymbol{\beta}, \sigma)\, \pi(\boldsymbol{\beta}, \sigma)\, d\boldsymbol{\beta}\, d\sigma < \infty.$$

The proof can be found in Appendix A.1.

In practice one usually assumes that the components of $\boldsymbol{\beta}$ have independent improper uniform prior distributions which is a special case of the above theorem.

One Practical Selection of Prior on σ

If the conditional distribution is strictly unimodal and symmetric or if the regressors affect only the location of the distribution, then a consistent estimate of the mode can be obtained with a fixed σ (Lee (1989)). In practice, however, data with such characteristics is relatively rare. In addition, in such cases the added value of mode regression is rather limited as the mode coincides with the mean and the median. To extend mode regression to more interesting applications σ must be allowed to approach zero as the sample size goes to infinity.

A suitable prior distribution for σ would be one with a positive support. To this end it is proposed to use either a Uniform(w_1, w_2) or a Gamma distribution with mean w, where, in both cases w_s can be determined using one of the following options, commonly used in bandwidth selection methods for kernel density estimation:

- The empirical rule, which states that, given a symmetric distribution, approximately 99.7% of the data values fall within three standard deviations (sd) of the mean, therefore, $w = 3sd$;
- Variations of Silverman's plug-in estimate for the bandwidth (Silverman (1986)), in which $w = 1.3643\delta n^{-0.2}[min(sd, IQR/1.349)]$, where, IQR is the sample inter quantile range and $\delta = 1.3510$ for a uniform kernel. To cover data with large number of outliers $IQR/1.349$ can be replaced by $1.4826MAD$, where MAD is the median absolute deviation.

Alternatively, as the next section demonstrates, a more flexible model can be developed by relaxing the distributional assumption on the prior for σ using a Dirichlet process prior. This leads to a flexible nonparametric mixture model. The method is nonparametric in the sense that it is not assumed that the prior belongs to any fixed class of distributions.

2.3 Nonparametric Bayesian Methods

In this section, two nonparametric Bayesian mode regression models are presented to avoid critical dependence on the assumption of a uniform distribution.

The methods allow the application of a likelihood approach without assuming that the data comes from a known family of distributions, thus reducing the possibility of inconsistent estimation due to misspecification, which may arise under the parametric Bayesian method.

2.3.1 Nonparametric Uniform Mixture Model

A nonparametric extension of the mode regression model can be constructed in the framework of finite mixture models. Under appropriate mixing and a sufficient large number of mixing components, any continuous density function on the real line can be approximated by a weighted sum of mixture distributions, such that,

$$f(y) = \sum_{j=1}^{k} \pi_j f_j(\cdot) \tag{2.3.10}$$

where $f_j(\cdot)$ are densities on $\mathbb{R}$ and π_j are mixing weights with $\sum_{j=1}^{k} \pi_j = 1$.

A strong unimodal density, $f(\cdot)$(with mode θ) is one that is non-decreasing on $(-\infty, \theta)$ and non-increasing on (θ, ∞) (Brunner (1992)). A density $f(\cdot)$ on $\mathbb{R}^+$ is non-increasing if and only if there exists a distribution function G such that $f(x|G) = \int \sigma^{-1} I_{[0<x<\sigma]} dG(\sigma)$ (Feller (1971)). Therefore, any unknown density $f(\cdot)$ (with mode θ), symmetric or not, can be represented as a scale mixture of symmetric uniform distributions, that is

$$f(x|\theta, G) = \int \frac{1}{2\sigma} I_{[-\sigma<x-\theta<\sigma]} dG(\sigma), \tag{2.3.11}$$

where G is the mixing distribution supported on $\mathbb{R}^+$.

This one-to-one mapping between f and G enables a nonparametric model for f through a nonparametric prior on G. A scale uniform Dirichlet process mixture for $f(\cdot, G)$ can be constructed by placing a Dirichlet prior on G (Kottas and Fellingham (2012)).

The Dirichlet Process (DP) was introduced by Ferguson (1973) and since then, it has been widely used in Bayesian nonparametric modelling. A $DP(M, G_0)$ is defined in terms of two parameters: G_0, which is the mean of the process, and the

concentration parameter M. The most commonly used representation of the DP is the "stick-breaking" representation (Sethuraman (1994)),

$$G(\cdot) \sim \sum_{i=1}^{\infty} w_i \delta_{\mu_i}(\cdot),$$

where $\mu_i \overset{iid}{\sim} G_0$ and $w_i = v_i \prod_{j<i}(1 - v_j)$, where $v_j \overset{iid}{\sim} Beta(1, M)$.

This representation states that each realisation of the DP can be represented as an infinite weighted sum of point masses. These points are a random sample from G_0 and the weights are constructed using the "stick-breaking" algorithm.

A nonparametric Bayesian mode regression model can be expressed in the hierarchical form:

$$\begin{aligned} &y_i|\boldsymbol{\beta}, \sigma_i \overset{iid}{\sim} f(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}; \sigma_i), i = 1 \cdots n \\ &\sigma_i|G \overset{iid}{\sim} G, i = 1 \cdots n \\ &G|M, d \sim DP(M, G_0(\cdot, d)) \\ &\boldsymbol{\beta}, M, d \sim p(\boldsymbol{\beta}), p(M), p(d), \end{aligned} \tag{2.3.12}$$

where, G is the mixing distribution, with base distribution G_0 and concentration parameter M and

$$f(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}; \sigma_i) = \frac{1}{2\sigma} I_{[-\sigma < y_i - x_i'\boldsymbol{\beta} < \sigma]}$$

is the density of a uniform distribution on $(-\sigma, \sigma)$.

2.3.2 Empirical Likelihood-based Bayesian Method

In addition to parametric and nonparametric likelihood, an empirical likelihood based method could be an alternative for Bayesian mode regression. The Empirical Likelihood (EL) method, introduced by Owen (1988, 1990), is a semi-parametric method of inference based on a data-driven likelihood ratio function. The method can be employed as an alternative to the bootstrap for constructing nonparametric confidence regions or hypothesis tests. Instead of re-sampling with equal probability weights like the bootstrap, the EL profiles a multinomial likelihood under a set of constraints which reflect the characteristics of the quantity of interests. EL

methods are known to enjoy good asymptotic properties, especially if the associated moment restrictions are of a sufficient smoothness. Like many estimation and inference procedures, e.g. Ordinary Least Square (OLS), Instrumental Variables (IV), and Generalised Method of Moments (GMM), EL is also based on the moment conditions:

$$E[g(\cdot)] = 0 \tag{2.3.13}$$

which can be estimated by $\widehat{g}(\cdot) = \sum p_i g(\cdot) = 0$, where p_i is called the implied probability associated with the observation x_i and $g(\cdot)$ is a vector of estimating functions. It can be shown that a solution to $\widehat{g}(\cdot) = 0$ exists for some choice of probabilities p_i such that $\sum_i p_i = 1$.

To derive an empirical likelihood for mode regression it is necessary to define some notations and a moment restriction. Lee (1993) generalised the mode regression estimator of Lee (1989), $\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}} E\{L(Y - \boldsymbol{x}'\boldsymbol{\beta})\}$, by using the rectangular kernel

$$L(Y;\mu) = \{(\sigma^2 - (Y-\mu)^2) I[|Y-\mu| < \sigma]\}.$$

Therefore, the moment restriction for the empirical likelihood can be obtained by the derivative

$$\frac{\partial}{\partial \mu} L(Y;\mu) = 2(Y-\mu) I[|Y-\mu| < \sigma].$$

Let $l(Y;\mu)$ be the derivative of $L(.;\mu)$ with respect to μ, then the mode, μ, of Y satisfies the moment restriction $E(l(Y;\mu)) = 0$.

Thus, under an empirical likelihood for mode regression $\mu = \boldsymbol{x}'\boldsymbol{\beta}$, for any proposed $\boldsymbol{\beta}$, to estimate the true p dimensional $\boldsymbol{\beta}_0$ the vector estimating functions $g(X, Y, \boldsymbol{\beta})$ with component $g_j(X, Y, \boldsymbol{\beta}) = l(Y; \boldsymbol{\beta}'X)\, X_j$ for $j = 1, .., p$ is used. Then, the profile empirical likelihood ratio is given by:

$$\mathfrak{R}(\boldsymbol{\beta}) = \max \left\{ \prod_{i=1}^{n} (n\, p_i) | \sum_{i=1}^{n} p_i\, g(X_i, Y_i, \boldsymbol{\beta}) = 0,\, p_i \geq 0,\, \sum_{i=1}^{n} p_i = 1 \right\}.$$

By a standard Lagrange multiplier argument,

$$\mathfrak{R}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \{n\, p_i(\boldsymbol{\beta})\}, \tag{2.3.14}$$

with the weights $p_i(\boldsymbol{\beta}) = \frac{1}{n(1+\hat{\lambda}(\boldsymbol{\beta})' g(X_i,Y_i,\boldsymbol{\beta}))}$, where the Lagrange multiplier $\hat{\lambda}(\boldsymbol{\beta})$ is the solution of λ to the following equation:

$$\sum_{i=1}^{n} \frac{g(X_i, Y_i, \boldsymbol{\beta})}{1 + \lambda^T g(X_i, Y_i, \boldsymbol{\beta})} = 0. \tag{2.3.15}$$

According to Qin and Lawless (1994), among others, the existence and uniqueness of $\hat{\lambda}(\boldsymbol{\beta})$ are guaranteed when the following two conditions are satisfied: (1) zero belongs to the convex hull of $\{g(X_i, Y_i, \boldsymbol{\beta}),\ i = 1, ..., n\}$ and (2) the matrix $\sum_{i=1}^{n}\{g(X_i, Y_i, \boldsymbol{\beta})g(X_i, Y_i, \boldsymbol{\beta})'\}$ is positive definite.

Under Bayesian inference, the empirical likelihood function $\mathfrak{R}(\boldsymbol{\beta})/n^n = \prod_{i=1}^{n}\{p_i(\boldsymbol{\beta})\}$ can be combined with a prior specification $\pi(\boldsymbol{\beta})$ on the parameter $\boldsymbol{\beta}$ to obtain the posterior distribution:

$$\pi(\boldsymbol{\beta}|data) \propto \pi(\boldsymbol{\beta})\, \mathfrak{R}(\boldsymbol{\beta}).$$

2.3.3 Asymptotic Properties of Bayesian Empirical Likelihood

Before establishing the asymptotic normality of the empirical likelihood-based Bayesian mode regression parameter estimates, the consistency of the empirical likelihood estimator must be established, which is a necessary condition for the asymptotic normality of the posterior. Since the criterion function $g(X, Y, \boldsymbol{\beta})$ results in a non-smooth estimating equation, a similar method to the one used by Molanes Lopez et al. (2009), among others, is employed to derive the asymptotic results.

Let $\hat{\boldsymbol{\beta}} = argmax_{\boldsymbol{\beta}} \mathfrak{R}(\boldsymbol{\beta})$ be the maximum empirical likelihood estimator (MELE) in a compact set of parameter space which contains the true parameter $\boldsymbol{\beta}_0$. Then note that the criterion function $g(X, Y, \boldsymbol{\beta})$ can be regarded as a special case of M-estimators as discussed in Chapter 5 of Van der Vaart (1998) and satisfies the conditions of theorem 5.7 in the book. Under some regularity conditions imposed on the marginal distribution of X and on the conditional distribution of Y given X, such as uniformly continuous and bounded, and since both $E\{g(X, Y, \boldsymbol{\beta})\}$ and $E\{g(X, Y, \boldsymbol{\beta})\, g(X, Y, \boldsymbol{\beta})'\} > 0$ are sufficiently smooth in a compact set of parameter space, which contains $\boldsymbol{\beta}_0$, the consistency condition C_3 of Molanes Lopez et al.

(2009) holds. Then the consistency of empirical likelihood estimates is established. Specifically, a rigorous statement of the conditions and theorem is as follows:

Assumption 1. There exists a neighbourhood $\mathcal{N}$ of $\boldsymbol{\beta}_0$ such that $P(\Re(\boldsymbol{\beta}) > 0) \to 1$ for any $\boldsymbol{\beta} \in \mathcal{N}$, as $n \to \infty$.

Assumption 2. The distribution function G_X has bounded support $\mathcal{X}$.

Assumption 3. The conditional distribution $F_X(t)$ of Y given X is twice continuously differentiable in t for all $X \in \mathcal{X}$.

Assumption 4. At any $X \in \mathcal{X}$, the conditional density function $F'_X(t) = f_X(t) > 0$ for t in a neighbourhood of $\boldsymbol{\beta}'_0 X$.

Assumption 5. $E\{g(X, Y, \boldsymbol{\beta}_0)\, g(X, Y, \boldsymbol{\beta}_0)'\} > 0$ is positive definite.

Assumption 6. $\log\{\pi(\boldsymbol{\beta})\}$ has bounded first derivative in a neighbourhood of $\boldsymbol{\beta}_0$.

Theorem 2.3.1. *Under Assumptions 1–5, the MELE $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}_0$.*

Assumptions 1-5 are standard conditions in this kind of asymptotic problems. For example, these conditions are similar to Assumptions 3.1-3.5 of Yang and He (2012, pp. 1110) for Bayesian empirical likelihood quantile regression. Assumption 1 guarantees that the interior of the convex hull of $\{g(X_i, Y_i, \boldsymbol{\beta}) : i = 1, \cdots, n\}$ for $\boldsymbol{\beta} \in \mathcal{N}$ contains the vector of zeros with probability tending to one. Assumption 4 ensures that $\boldsymbol{\beta}_0$ is indeed the unique solution for $Eg(X, Y, \boldsymbol{\beta}) = 0$. The proof of Theorem 2.3.1 is sketched in Appendix A.1.

The asymptotic normality of the posterior distribution $\pi(\boldsymbol{\beta}|data)$ could be established using the fact that the empirical log-likelihood ratio for $\boldsymbol{\beta}$ is well approximated by certain quadratics in the sense of Lemma 6 of Molanes Lopez et al. (2009) so that,

$$\Gamma_n(\boldsymbol{\beta}) \equiv -n^{-1} \sum_{i=1}^{n} \log(1 + \hat{\lambda}(\boldsymbol{\beta})' g(X_i, Y_i, \boldsymbol{\beta})) \tag{2.3.16}$$

$$= -\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)' V'_{12} V_{11}^{-1} V_{12} (\boldsymbol{\beta} - \boldsymbol{\beta}_0) + n^{-1/2} (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' V'_{12} V_{11}^{-1} W_n$$
$$- \frac{1}{2} n^{-1} W'_n V_{11}^{-1} W_n + o_P(n^{-1}), \tag{2.3.17}$$

with matrices

$$V_{11} = (E\{g_j(X, Y, \boldsymbol{\beta}_0)\, g_k(X, Y, \boldsymbol{\beta}_0)'\})_{j,k=1}^{p}$$

,

$$V_{12} = (-\frac{\partial}{\partial \beta_k} E\{g_j(X, Y, \boldsymbol{\beta})\}|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0})_{j,k=1}^{p},$$

and vector $W_n = n^{-1/2} \sum_{i=1}^{n} g(X_i, Y_i, \boldsymbol{\beta}_0))$.

Then,

Theorem 2.3.2. *Under Assumptions 1-6 and from* $\log \Re(\boldsymbol{\beta}) = n\Gamma_n(\boldsymbol{\beta})$*, the posterior density of* $\boldsymbol{\beta}$ *has the following expansion on any sequence of sets* $\{\boldsymbol{\beta} : \boldsymbol{\beta} - \boldsymbol{\beta}_0 = O(n^{-1/2})\}$*,*

$$\pi(\boldsymbol{\beta}|data) = \pi(\boldsymbol{\beta})\, \Re(\boldsymbol{\beta}) \propto \exp\{-\frac{1}{2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})' I_n (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + Q_n\} \tag{2.3.18}$$

with $I_n = nV_{12}' V_{11}^{-1} V_{12}$ *and empirical likelihood estimate* $\hat{\boldsymbol{\beta}}$ *and* $Q_n = o_p(1)$*. When* I_n *is positive definite,* $I_n^{1/2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$ *is converging in distribution to* $N(0, I)$*.*

The proof of Theorem 2.3.2 is sketched in Appendix A.1.

Finally, similarly to quantile regression, by Remark 3.2 of Yang and He (2012, pp. 1110), the posterior will be improper for flat priors on $\boldsymbol{\beta}$ in the Bayesian empirical likelihood approach for the proposed mode regression, and therefore flat priors on $\boldsymbol{\beta}$ should be avoided.

In the case of a prior distribution shrinking with n, it is possible to use a $\pi_n(\boldsymbol{\beta})$ which satisfies conditions similar to Assumption 3.7 of Yang and He (2012) as priors for the proposed mode regression (see Theorem 3.3 of Yang and He (2012) for details).

2.4 Numerical Experiments

In this section the proposed approaches to Bayesian mode regression are demonstrated through three simulated and one real examples. The first simulation example demonstrates the applicability of the proposed approach to mode estimation and the other two simulation examples are dedicated to mode regression. The real example investigates how factors such as gender, age, consumption of alcohol, consumption of fruit and vegetables and smoking can affect the body mass index (BMI).

2.4.1 Simulation Example 1

Mode estimation

In this sub-section the applicability of the proposed parametric Bayesian approach to mode estimation is demonstrated. The proposed methodology was applied to estimate the mode for samples generated from a series of distributions featuring different characteristics.

Specifically, the following five distributions were considered:

1. A symmetric distribution: Normal distribution, with mean 2 and standard deviation 0.5, with true mode at 2.

2. A symmetric distribution with heavy tails: Cauchy distribution, with location parameter 0 and scale parameter 2, with true mode at 0.

3. A symmetric distribution: Beta distribution with $\alpha = \beta = 2$, with true mode at 0.5.

4. An asymmetric distribution with heavy tails: χ^2 distribution with 4 degrees of freedom, with true mode at 2.

5. A discrete distribution: Poisson distribution with rate 2, with true mode at 1.

For each of these distributions n random observations for $n = 50$ and $n = 100$ were generated. Each simulation experiment was replicated 100 times. Realisations were simulated from the posterior distributions by means of a single-component Metropolis-Hastings algorithm. The parameter estimates were obtained using a random-walk Metropolis algorithm with a Gaussian proposal density centred at the current state of the chain. Convergence was assessed using time series plots and convergence diagnostics measures contained in the R package boa (Smith (2007)). Table 2.2 compares the MCMC posterior means (PM) with the value of the true mode (TM) of each of the distributions under investigation. Standard deviations (SD) and 95% Bayesian credible intervals (BCI) are also provided. All the three quantities computed were averaged over the 100 data sets.

Table 2.2: Simulation Example 1: True Mode (TM), Posterior Means (PM), Standard Deviations (SD) and 95% Bayesian Credible Intervals (BCI)

Sample size		Normal (2,0.5)	Cauchy (0,2)	Beta (2,2)	$\chi^2(4)$	Poisson (2)
50	T.M	2	0	0.5	2	1
	P.M	2.05	0.11	0.52	1.88	1.04
	SD	0.37	0.49	0.17	0.23	0.34
	95%BCI	(0.31,2.62)	(-1.00,0.84)	(0.23,0.82)	(0.51,1.49)	(1.00,2.81)
100	T.M	2	0	0.5	2	2
	P.M	1.91	0.21	0.46	1.91	1.52
	SD	0.22	0.46	0.16	0.13	0.32
	95%BCI	(1.51,2.27)	(-0.57,1.16)	(0.20,0.73)	(1.01,1.96)	(0.70,1.27)

As it can be seen from Table 2.2, the estimated results are very close to the true mode in all the cases considered in the simulation experiments.

2.4.2 Simulation Example 2

Data for the second simulation example was generated from the following regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{2.4.19}$$

where $x_i \sim N(0,1)$, $i = 1, ..., n$ for $n = 50, 100, 200$ and $\boldsymbol{\beta} = (1, 2)$. The following three specifications were considered for the model error ϵ (for relevant plots see Figure 2.2):

- Case 1: the standard Normal distribution, $\epsilon_i \sim N(0,1)$ - a symmetric error distribution.
- Case 2: a Fisher's Z distribution, $\epsilon_i \sim 1/2logZ$ with $Z \sim F_{2,2}$ - a skewed error distribution.

- Case 3: a Normal distribution with normally distributed outliers (contaminants) centred at twice the distance between the true mode and the 99^{th} percentile of the original Normal distribution and accounting for 20% of the total data points, $\epsilon_i \sim 0.80N(0, \frac{1}{4}) + 0.20N(2.5, \frac{1}{4})$ (Hedges and Shah (2003)) - an asymmetric error distribution.

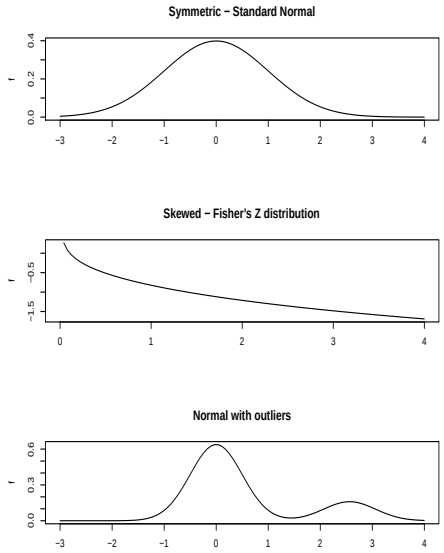


Figure 2.2: Density Plots of the three Distributions used in the Simulation Study

First, the parametric Bayesian mode regression (labelled PBMR) was fitted for all the three cases. Then, for demonstration and comparison purposes, the empirical likelihood based Bayesian mode regression (labelled ELBMR) and the nonparametric Bayesian mode regression (labelled NPBMR) were also fitted for the model under the asymmetric error specification (Case 3).

For the PBMR and ELBMR models, independent Normal distributions were used as priors of each component of $\boldsymbol{\beta}$, where the mean and standard derivation of the Normal prior are given by the classical estimator of Lee (1989, 1993) and its estimated standard error respectively. Realisations were simulated from the posterior distributions by means of a single-component Metropolis-Hastings algorithm. Each of the parameters was updated using a random-walk Metropolis algorithm with a Gaussian proposal density centred at the current state of the chain. The estimates

for the NPBMR model were obtained by fitting a truncated Dirichlet Process (DP) mixture model, which leads to a computationally straightforward approximation and can be easily implemented in the freely available WinBUGS software. Two parallel chains of equal length with different initial values were run for the model. The results were based on 10,000 iterations which followed a burn-in period of 40,000 for each chain.

The variance of the proposal density was chosen to provide an acceptance rate close to the optimal acceptance rate as defined in Roberts and Rosenthal (2001). Convergence was assessed using time series plots and convergence diagnostics measures contained in the R package boa (Smith (2007)). The estimates are posterior means using 10,000 iterations of the MCMC sampler (after 10,000 burn-in iterations). Figure 2.3 demonstrates the posterior trace plots for the model parameters estimated under the three proposed methods (for sample size n=50) and indicates good convergence of the chains.

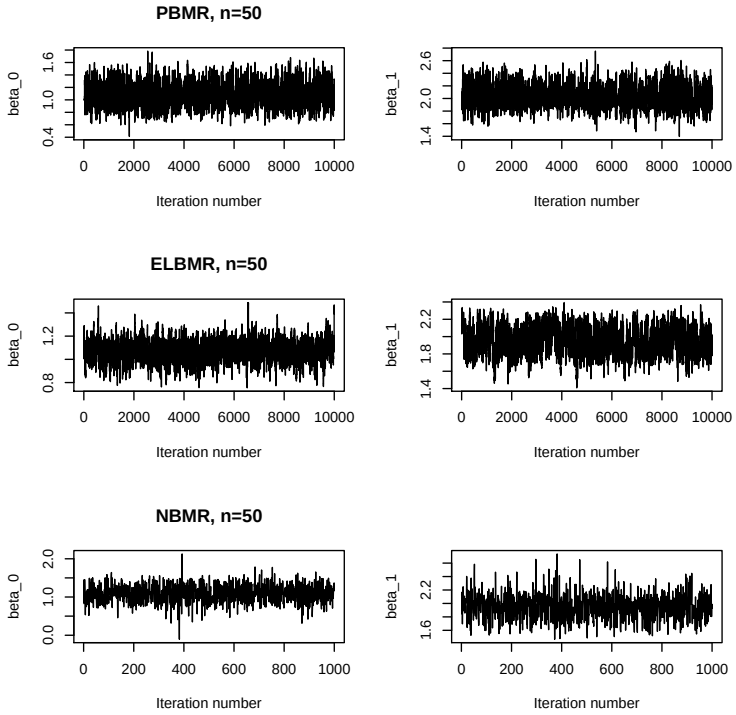


Figure 2.3: Asymmetric Distribution: Posterior Trace Plots for Model Parameters

Table 2.3: Simulation Example 2 - True Parameter Values (T.V.) and their Posterior Means, Standard Deviations (S.D.) and 95% Bayesian Credible Intervals (BCI)

		PBMR						ELBMR		NPBMR	
		Normal		Skewed		Asymmetric		Asymmetric		Asymmetric	
n		β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1	β_0	β_1
	T.V	1	2	1	2	1	2	1	2	1	2
	Mean	1.04	2.00	1.02	1.99	1.01	1.98	1.15	2.07	1.09	1.94
50	S.D.	0.22	0.24	0.17	0.19	0.19	0.19	0.07	0.08	0.24	0.19
	95%BCI	(0.60,1.46)	(1.54,2.46)	(0.68,1.35)	(1.62, 2.36)	(0.65,1.37)	(1.62,2.35)	(1.03, 1.30)	(1.94, 2.27)	(0.71,1.52)	(1.5,2.3)
	T.V	1	2	1	2	1	2	1	2	1	2
	Mean	1.00	2.02	0.98	2.00	1.00	2.00	0.89	2.04	1.06	2.00
100	S.D.	0.14	0.15	0.12	0.13	0.12	0.13	0.07	0.07	0.14	0.12
	95%BCI	(0.73,1.27)	(1.73,2.31)	(0.76,1.22)	(1.75,2.27)	(0.78,1.24)	(1.75,2.24)	(0.76, 1.01)	(1.91,2.17)	(0.82,1.31)	(1.84,2.21)
	T.V	1	2	1	2	1	2	1	2	1	2
	Mean	1.00	1.99	1.00	2.00	1.02	2.00	0.95	1.97	1.04	1.91
200	S.D.	0.09	0.10	0.09	0.09	0.09	0.09	0.05	0.04	0.07	0.06
	95%CI	(0.83,1.18)	(1.81,2.18)	(0.82,1.17)	(1.83,2.18)	(0.85,1.19)	(1.83,2.17)	(0.83, 1.06)	(1.89, 2.05)	(0.92,1.19)	(1.78,2.03)

Each simulation experiment was replicated 100 times. Table 2.3 compares the posterior means with the true values of β_0 and β_1 and gives standard deviations (SD) and 95% Bayesian credible intervals (BCI) for each of the cases considered in this example. In all the examples the three quantities computed were averaged over the 100 data sets. Figures 2.4, 2.5 and 2.6 show the posterior histograms of $\hat{\beta}_0$ and $\hat{\beta}_1$ respectively for the three simulation cases for different sample sizes, under the PBMR.

The results of the analysis indicate that the PBRM works well for the three cases considered, as all the absolute biases for the estimated parameters turn out to be in the range [0.00, 0.04]. Furthermore, under both ELBMR and NBRM, the true values for both β_0 and β_1 were recovered successfully. The standard deviation, and accordingly the BCI, decrease with increasing sample sizes in all the experiments. Comparing the results for the asymmetric error example, for which all the three methods were tested, it can be concluded that the PRMR works best in recovering the true values of the regression parameters, as both the ELBMR and the

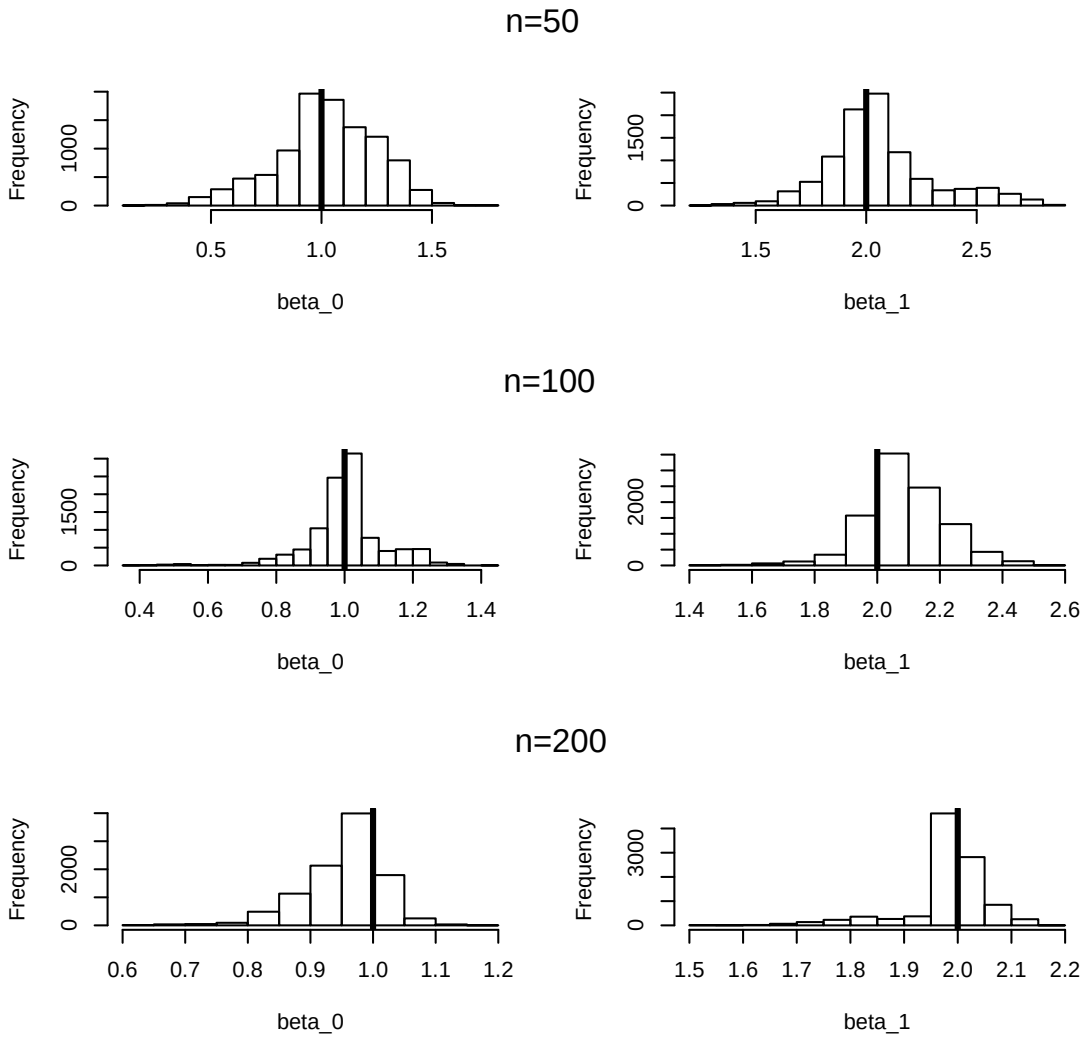


Figure 2.4: Posterior Histograms - Symmetric Error Distribution

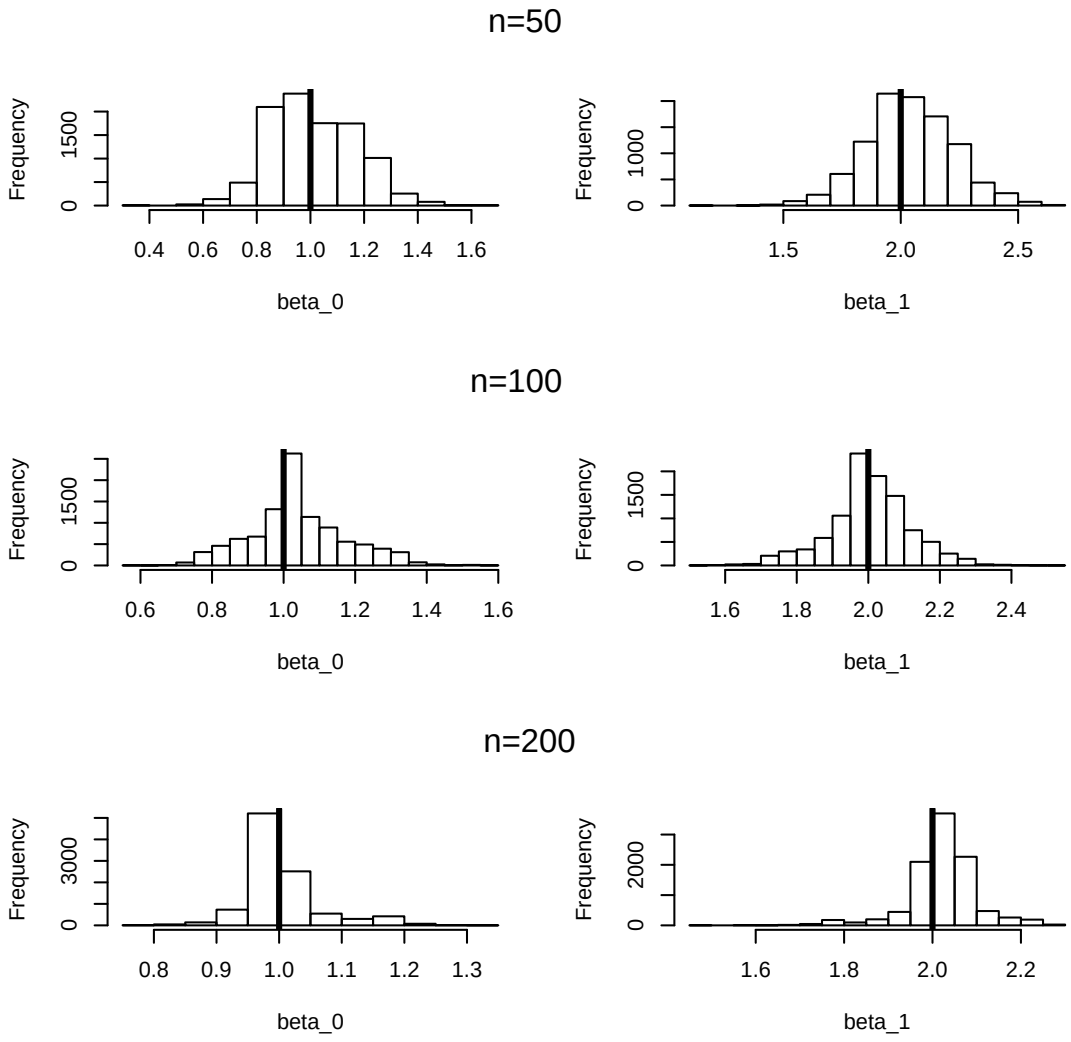


Figure 2.5: Posterior Histograms - Skewed Error Distribution

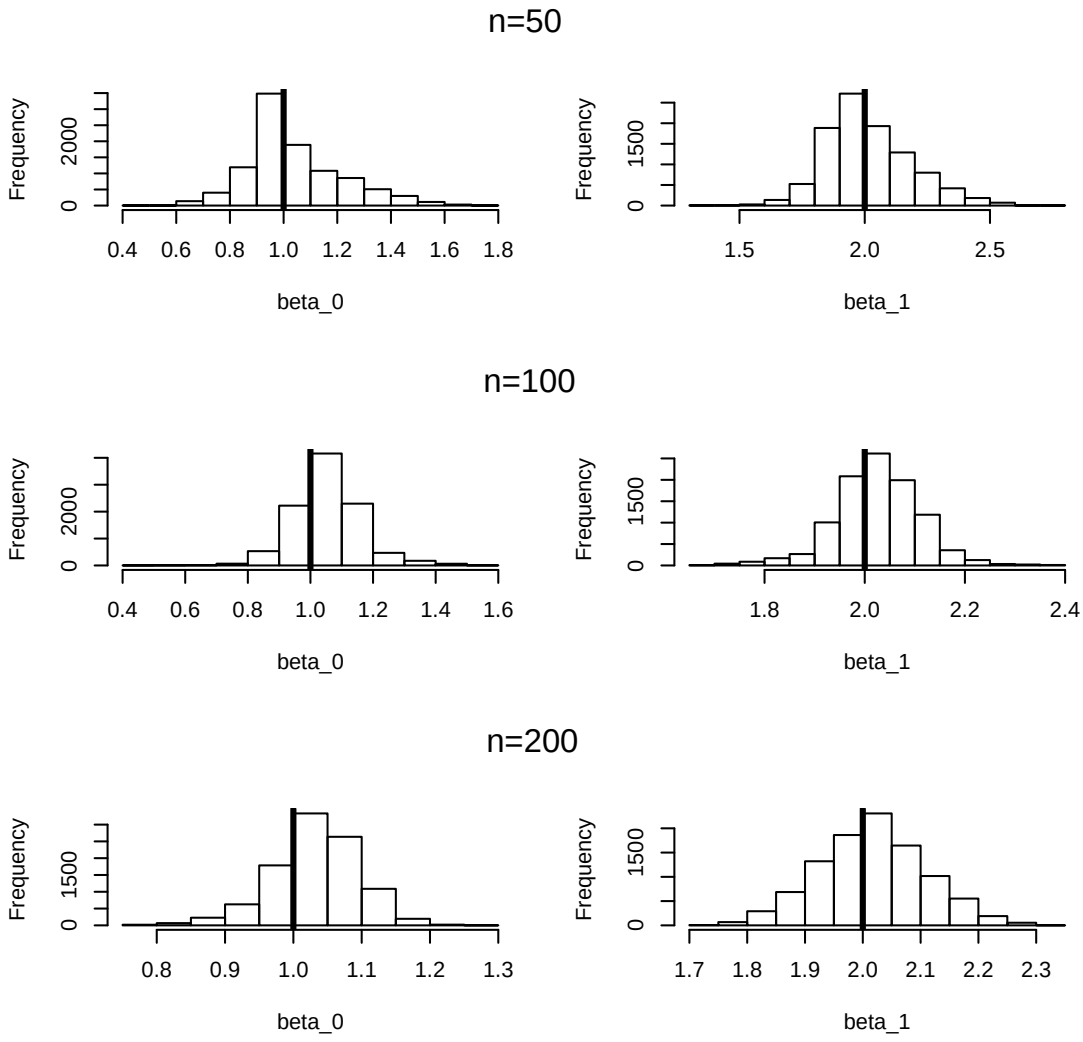


Figure 2.6: Posterior Histograms - Asymmetric Error Distribution

NPBMR demonstrate larger absolute biases. Regarding the estimated SD and BCI, the ELBMR demonstrates the best performance as the estimated SD is the very low, even in the smallest dataset. The NPBMR demonstrates similar results to the PBMR in terms of SD and BCI. In conclusion, the results indicate that all the proposed methodologies work well for mode regression in finite samples, with the PBMR to outperform in terms of recovering the true values and the ELBMR in terms of the magnitude of the estimated SD.

Figure 2.7 exhibits the empirical samples from the joint posterior distributions of the PBMR parameters, which were obtained using the output of the MCMC sampler for the regression parameters $\widehat{\beta}_0$ and $\widehat{\beta}_1$. These samples can be used to obtain a consistent estimator of the covariance or correlation structure of the parameter estimators, which is difficult to estimate under the classical approach. For example in case (a), with sample size n=100, the estimate is:

$$\widehat{Cov}\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = \begin{pmatrix}3 & -1\\ -1 & 6\end{pmatrix}.$$

2.4.3 Simulation Example 3

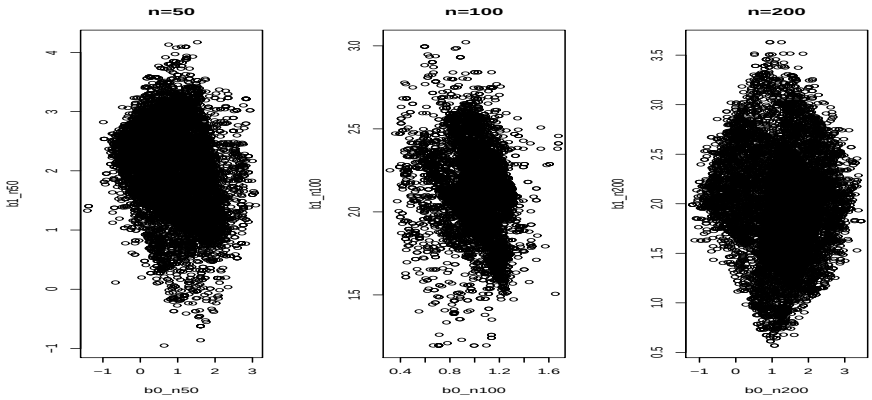
In this sub-section the results of a third simulation example are presented which was performed with the aim of comparing the performance of the proposed approach with the classical mode regression approach. Specifically, the simulation study in Kemp and Silva (2012) was replicated, but only for a sample of size 250, to give the opportunity to compare their results with the results obtained under the proposed Bayesian mode regression approach.

Simulation data was generated by the simple linear model:

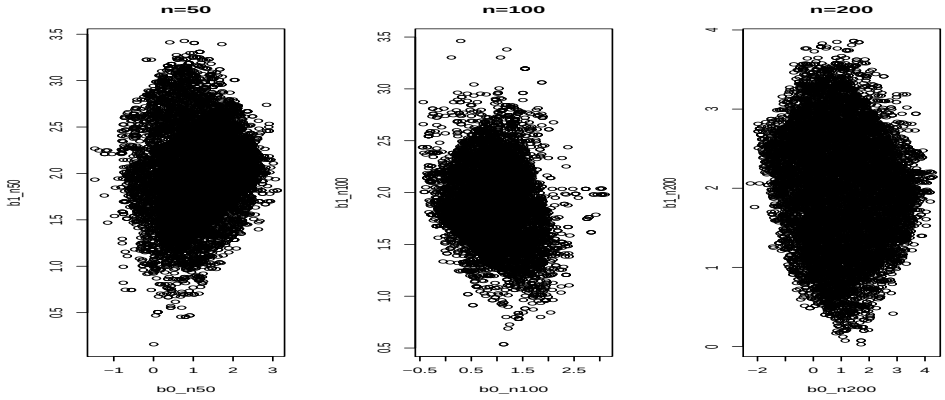
$$y_i = \beta_0 + \beta_1 x_i + (1 + v x_i)\epsilon_i, \tag{2.4.20}$$

where $\beta_0 = 0$ and $\beta_1 = 1$, $x_i \sim \chi^2_{(3)}$ distribution, scaled to have variance 1, and ϵ_i were generated as independent draws from a re-scaled log-Gamma random variable,

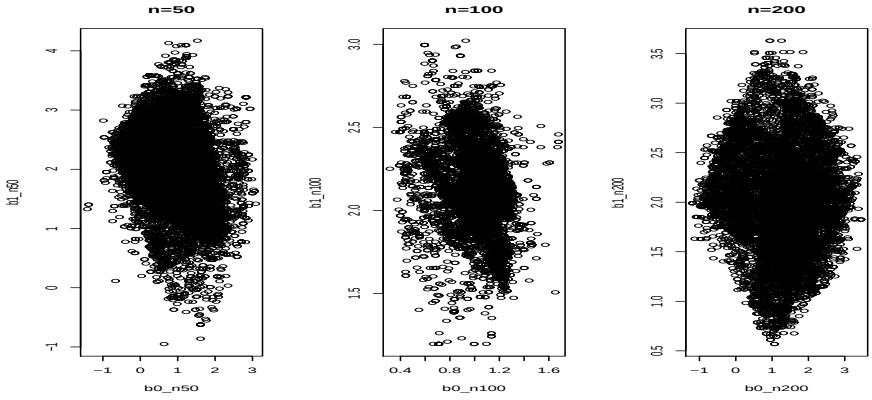
$$\epsilon_i = -\lambda\, ln(Z_i), \tag{2.4.21}$$



(a) Symmetric error



(b) Skewed error



(c) Asymmetric error

Figure 2.7: Plots showing the Empirical Samples from the Joint Distributions of Mode Regression Parameters

where Z follows a Gamma distribution with mean 1 and scale parameter $\frac{1}{\alpha}$, to ensure that ϵ_i has zero mode. Furthermore, $\lambda = [(1 + 2E(x_i)v + E(x_i^2)v^2)\psi(\alpha)]$, where $\psi(\cdot)$ is the trigamma function, to ensure that the unconditional variance of the error $(1 + vx_i)$ is equal to one.

The study was performed for $\alpha \in \{0.05, 5\}$ and for $v \in \{0, 2\}$. Each simulation experiment was replicated 100 times. Table 2.4 demonstrates the estimated values (E.V.), the 95% Bayesian credible intervals (BCI) and Root mean square error (RMSE) for the estimates obtained under PBMR and NPBMR (convergence was assessed using time series plots and convergence diagnostics measures contained in the R package boa (Smith (2007)))and provides a comparison with the E.V., the 95% classical confidence intervals (CI) and RMSE obtained under the two classical mode regression models: Mode 1.6 and Mode 0.8. Mode 1.6 and Mode 0.8 correspond to $k = 1.6$ and $k = 0.8$ respectively in the bandwidth selection rule, bandwidth=$k\, mad\, n^{-0.143}$, where mad is the median of the absolute deviation from the median of ordinary least squares regression residuals.

The results of the analysis suggest that the Bayesian mode regression estimates are strong competitors of the classical mode regression estimates. This is evident from the precision of the estimated parameter values, the length of the BCI and the values of the estimated root mean square (RMSE). In all the scenarios, both the PBMR and the NPBMR successfully recover the true values of the model parameters and also, in most of the cases provide shorter BCI as compared to the CI estimated for the parametric approaches. In addition, the RMSE indicates a high goodness of fit for all models fitted by the Bayesian approaches, with the PBMR method to demonstrate a comparable or lower RMSE than the parametric approach in all the scenarios considered and the NPBMR approach to outperform in terms of goodness of fit the classical approach in the heteroscedastic scenarios.

Finally, it should be noted that the selection of the value/prior for σ plays an important role on the precision of the parameters, something also evident from Kemp and Silva (2012).

Table 2.4: Simulation Example 3 - Comparison between Classical and Bayesian Approach for Mode Regression

			PBMR			NPBMR			Mode 1.6			Mode 0.8		
α	v		E.V.	95% BCI	RMSE	E.V	95% BCI	RMSE	E.V.	95% CI	RMSE	E.V.	95% CI	RMSE
5.00	0	β_0	-0.03	(-0.11,0.12)	0.03	0.06	(-0.21,0.36)	0.17	0.05	(-0.31, 0.41)	0.05	0.03	(-0.69, 0.75)	0.04
		β_1	1.03	(0.94,1.09)		1.09	(0.89,1.32)		1.00	(0.77, 1.24)		1.01	(0.56,1.45)	
	2	β_0	0.06	(-0.06,0.12)	0.09	0.07	(-0.03,0.21)	0.17	0.04	(-0.15,0.23)	0.04	0.02	(-0.25,0.29)	0.03
		β_1	1.02	(0.94,1.16)		1.06	(0.80,1.22)		1.00	(0.63,1.37)		1.01	(0.48,1.53)	
0.05	0	β_0	0.07	(0.00,0.21)	0.09	0.01	(-0.03,0.07)	0.02	0.28	(0.12,0.42)	0.36	0.13	(-0.09,0.35)	0.15
		β_1	1.01	(0.95,1.05)		0.98	(0.95,1.06)		1.06	(0.90,1.11)		1.02	(0.87,1.17)	
	2	β_0	0.03	(-0.001,0.04)	0.03	0.06	(0.04,0.09)	0.07	0.19	(0.09,0.29)	0.26	0.03	(0.01,0.21)	0.19
		β_1	1.00	(0.98,1.02)		1.002	(0.97,1.04)		1.05	(0.91,1.19)		1.11	(0.85,1.19)	

2.4.4 The Body Mass Index (BMI) Data Example

Following the introduction of the BMI example in Section 2.1, the proposed methodology was applied to investigate the research question: "What is the effect of factors such as gender, age, consumption of alcohol, consumption of fruit and vegetables and smoking on the typical body mass index (BMI)?"

A person's typical BMI was modelled as a function of the person's age, age_i, the total units of alcohol consumed per week, $alcohol_i$, the portion of fruit and vegetables consumed the previous day, $fruit\&veg_i$ the person's cigarette smoking status, $smoking_i$ (0= Non-smoker, 1= Light smokers, under 10 a day, 2= Moderate smokers, 10 to under 20 a day, 3=Heavy smokers, 20 or more a day), and of a gender indicator, $male_i$ (1=male, 0=female):

$$bmi_i = \beta_0 + \beta_1 age_i + \beta_2 alcohol_i + \beta_3 fruit\&veg_i + \beta_4 smoking_i + \beta_5 gender_i + \epsilon_i \quad (2.4.22)$$

The BMI range is from 15.9 to 56.0 (range =40.1) indicating a significant disparity between high and low BMI scores. The average BMI is 27.75 with standard deviation of 5.13 (Table 2.1 in section 2.1). The high levels for range and standard deviation suggest the presence of outliers which cause the mean to be pulled in the direction of the tail. As a consequence, the mean, median, and mode do not coincide and it can be easily concluded that the distribution of the data is positively skewed. Figure 2.1 in section 2.1 demonstrates the density of BMI for the total, males and females, verifying that all three distributions are positively skewed. The mode represents the most typical value and is the value at the peak of the distribution. Even though, mean regression and quantile regression could have been applied to model BMI these methods cannot reveal any information about the mode, or about the effect of the covariates on the most typical case.

Table 2.5 presents the estimation results obtained with the traditional mean, quantile and the parametric bayesian mode regressions. The analysis was performed for the total of responders but also for males and females separately. For the PBMR and ELBMR models, an independent improper uniform prior was chosen for all the components of $\boldsymbol{\beta}$ and a gamma prior with mean $3sd(\mathbf{bmi})$ for σ. Realisations

Table 2.5: BMI dataset - Estimation results for mean, quantile and mode regression

	Mean Regression		Quantile Regression						Parametric Bayesian Mode Regression	
			0.25		0.50		0.75			
Total (n=4,138)										
Variable	Estimate	95%CI.	Estimate	95%CI.	Estimate	95%CI.	Estimate	95%BCI	Estimate	95%BCI
const	25.4	(24.8,25.9)	20.9	(20.4,21.4)	23.7	(23.1,24.3)	28.1	(27.3,28.9)	21.8	(21.6,21.9)
age	0.05	(0.04,0.06)	0.05	(0.04,0.06)	0.06	(0.05,0.07)	0.05	(0.04,0.07)	0.09	(0.093,0.98)
alcohol	-0.006	(-0.014,0.003)	0.004	(-0.004,0.012)	0.001	(-0.008,0.011)	-0.005	(-0.017,0.008)	0.01	(0.001,0.011)
fruit&veg	-0.06	(-0.13,-0.01)	-0.002	(-0.06,0.06)	-0.06	(-0.13,-0.001)	-0.11	(-0.19,-0.02)	-0.06	(-0.08,-0.04)
smoking	-0.37	(-0.64,-0.10)	-0.40	(-0.66,-0.14)	-0.34	(-0.63,-0.04)	-0.36	(-0.76,0.04)	-0.16	(-0.49,0.06)
male	0.42	(0.11,0.74)	1.36	(1.06,1.66)	0.75	(0.40,1.09)	-0.29	(-0.76,0.17)	0.83	(0.65,0.95)
Males (n=1,814)										
const	25.3	(24.6,26.1)	22.2	(21.5,23.0)	24.4	(23.6,25.1)	27.5	(26.5,28.5)	25.0	(24.2,25.4)
age	0.05	(0.04,0.06)	0.05	(0.04,0.07)	0.06	(0.04,0.07)	0.05	(0.03,0.07)	0.03	(0.03,0.04)
alcohol	0.007	(-0.002,0.017)	0.008	(-0.001,0.02)	0.01	(0.001,0.02)	0.007	(-0.007,0.02)	0.01	(0.003,0.02)
fruit&veg	-0.03	(-0.11,0.05)	-0.02	(-0.10,0.06)	-0.03	(-0.11,0.05)	-0.02	(-0.13,0.10)	0.03	(-0.01,0.07)
smoking	-0.43	(-0.80,-0.07)	-0.57	(-0.94,-0.21)	-0.48	(-0.86,-0.11)	-0.51	(-1.02,0.01)	-0.96	(-1.30,0.14)
Females (n=2,324)										
const	25.8	(25.0,26.5)	21.1	(20.3,21.8)	23.6	(22.7,24.5)	28.2	(27.1,29.4)	23.3	(22.8,23.7)
age	0.04	(0.03,0.06)	0.05	(0.04,0.06)	0.06	(0.05,0.08)	0.06	(0.04,0.08)	0.08	(0.07,0.08)
alcohol	-0.03	(-0.04,-0.01)	-0.01	(-0.02,0.01)	-0.01	(-0.03,0.01)	-0.02	(-0.05,-0.001)	-0.04	(-0.06,-0.03)
fruit&veg	-0.10	(-0.18,-0.01)	0.01	(-0.07,0.09)	-0.07	(-0.17,0.03)	-0.15	(-0.28,-0.02)	-0.12	(-0.19,-0.04)
smoking	-0.27	(-0.66,0.12)	-0.29	(-0.67,0.08)	-0.19	(-0.66,0.27)	-0.29	(-0.89,0.31)	-0.82	(-1.11,-0.59)

were simulated from the posterior distributions by means of a single-component Metropolis-Hastings algorithm. Each of the parameters was updated using a random-walk Metropolis algorithm with a Gaussian proposal density centred at the current state of the chain. The estimates for the NPBMR model were obtained by fitting a truncated Dirichlet Process (DP) mixture model, with independent independent improper Normal priors for all the components of $\boldsymbol{\beta}$. Two parallel chains of equal length with different initial values were run for the model.

The results indicate that gender has a statistically significant effect on the BMI both on the mean and median, but also for the mode. On average, the BMI is 0.42 units lower for women than for men. However, as indicated by quantile regression, the effect of gender differs at different quantile levels. More specifically, at the 25% level, the BMI of women is around 1.36 units lower than the corresponding BMI for men but this gap is smaller for the median case (0.75). Mode regression reveals that the gender differential on the most typical BMI is higher than both the mean and the median. According to the results, the typical BMI for women is 0.83 units lower than the corresponding BMI for men.

Age has a positive and statistically significant effect on the BMI, both on the mean and on the estimated quantiles, for the total, but also for men and women separately. In the case of mode regression age has also a positive significant effect, but the effect is stronger, as compared to the mean and the estimated quantiles, in the case of the total population and for women, whereas it is weaker in the case of men.

Furthermore, additional consumption of fruits and vegetables has a negative and statistically significant effect on the BMI on the mean, median and mode, as well as on the 75% quantile level for the total population. The results indicate that the negative effect on the mean, median and mode is similar (-0.06), whereas the effect at the 75% level is higher. The results for females are not much different, although the effect of additional consumption of fruits and vegetables is more pronounced, and in this case it is not statistically significant for the median level. The results for men indicate that the effect of additional consumption of fruits and vegetables is not statistically significant at any estimated statistic.

In addition, the results of the analysis suggest that, for the total population and for men, smoking has also a negative and statistically significant effect on the BMI on the mean, the median and the 25% quantile level, but not on the mode, whereas, for women, smoking has a negative statistically significant effect on the BMI only on the mode. Finally, the effect of alcohol is very small to be reported even for when it is statistically significant.

In conclusion, the results of the analysis indicate that mode regression is a useful statistical technique, especially when analysing data with outliers. In this example, even thought in many cases the overall effect of covariates on the response variable was similar under the three regression methods, this was not always the case and in addition often the marginal effects of the covariates were different under different regression methods. This justifies the usefulness of mode regression as an alternative analysis tool.

2.5 Conclusions

Identifying the typical value or pattern could be one of the most efficient statistical methods of data analysis, in particular, for big data analysis. In this chapter a novel Bayesian mode regression framework has been presented which includes three approaches: a parametric method, a nonparametric method and an empirical likelihood-based method. It should be noted that, in the area of mode regression, there is no literature from a Bayesian perspective. The chapter demonstrates that the estimates are consistent and asymptotically Normal under fairly standard conditions and even under misspecification of the likelihood function. The numerical studies suggest that the proposed Bayesian mode regression estimates are strong competitors to the classical mode regression estimates.

Chapter 3

Fully Parametric Classical Mode Regression: An illustration via Big Data Analysis

3.1 Introduction

As it has been mentioned before, despite its advantages, limited work exists in the literature in the area of mode regression. In the classical literature work on mode regression was carried out by Lee (1989,1993), Kemp and Silva (2012) and Yao and Li (2014) whereas no work exists in this area from the Bayesian perspective.

Motivated by the latter, in Chapter 2 a novel Bayesian mode regression framework has been presented which includes three approaches: a parametric method, a nonparametric method and an empirical likelihood-based method.

On the other hand, research from the classical perspective involves either semiparametric or nonparametric mode regression methods, which have a slow rate of convergence and are subject to bandwidth selection; thus have little, if any, practical use. To this end in this Chapter a fully parametric mode regression method, based on the Gamma density is introduced with good theoretical properties and finite sample results, as well as easy and fast implementation.

In addition, this chapter demonstrates a quick and effective methodology for identifying patterns in big data and for exploring the effect of different factors on

the typical value. As it is always beneficial to demonstrate the applicability of a new approach within a valid domain, the approach is demonstrated through the analysis of an almost a decade-long dataset from the Health Survey for England. The aim of the analysis is to explore the effect of socio-economic characteristics and behavioural habits of adults in England on the typical Body Mass Index (BMI).

The proposed method is a 2-step approach. In the first step, mode estimation is used to identify the mode BMI for each of the years considered in the analysis and accordingly the intervals containing the most typical BMI observations are selected. This first step is easy and quick to carry out. Although data-mining pattern-finding algorithms are already available, the mode could be a quick and effective alternative for pattern-finding, and at the same time it is statistically meaningful, as it facilitates selecting the most typical observations in a dataset. Then, all the collected data for the typical BMI intervals and the associated factors are merged to construct a new (smaller) dataset which will be used for the second step of the analysis: the mode regression. In the case of multi-modal distribution, the data corresponding to all identified modes will be collected. This will increase the size of the resulting dataset to be used for mode regression.

It should be noted that mode estimation has already been used in modern science for data analysis (Hedges and Shah (2003), Heckman et al. (2001), Kumar and Hedges (1998), Markov et al. (1997)) and mode-based clustering techniques have also been developed (Li et al. (2007)).

The proposed methodology includes a data reduction step. In general, as data reduction is accomplished by throwing away some data, such techniques reduce the richness and quality of the data and may lead to a reduction of the information content in the data. However, even though such techniques are often criticised by many practitioners and researchers, the proposed methodology retains data that explain much of the variance and omits data that explain little of the variance, as the methodology ensures that the after data reduction the remaining sample contains the most typical observation.

The chapter is structured as follows. Section 3.2 details the fully parametric mode regression method. Section 3.3 introduces the concept of big data and presents

the analysis steps of The Health Survey for England data and explores the dependency of BMI on other covariates. Concluding remarks are provided in Section 3.4.

3.2 Fully Parametric Mode Regression

The Gamma distribution, which covers a wide range of skewed, even heavily skewed distributions, has been widely used and successfully applied to parametric quantile regression (Noufaily and Jones (2013)). Moreover, the expression of the mode of the Gamma distribution is fairly tractable and provides an ideal method to develop a fully parametric mode regression.

Let Y be a positive response variable according to the Gamma distribution with a density function as follows:

$$f(y|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y}, y > 0, \tag{3.2.1}$$

where, $\Gamma(\alpha)$ is the Gamma function, $\alpha > 0$ determines the shape of the distribution and $\beta > 0$ is the rate parameter.

The Gamma distribution is very flexible and its density can have different shapes depending on the values of the two distribution parameters, including the exponential distribution with rate λ when $\alpha = 1$ and $\beta = \frac{1}{\lambda}$, the $\chi^2(\kappa)$ distribution when $\alpha = \frac{\kappa}{2}$ and $\beta = \frac{1}{2}$, while it attends a Normal distribution at the limit as $\alpha \to \infty$. This evident flexibility makes the Gamma distribution an attractive candidate for data-driven statistical modelling. Figure 3.1 shows a few different Gamma densities corresponding to different values of (α, β).

The mode of the Gamma distribution with $\alpha > 1$ is given by:

$$\mu = mode(y) = \frac{\alpha - 1}{\beta}. \tag{3.2.2}$$

The fully parametric mode regression (PMR) is developed by first re-parameterising the Gamma density in equation (3.2.1) in terms of the mode (μ) of Y and then introducing a regression-based functional form. To obtain a regression structure for the mode of the response variable, let $\mu = \frac{\alpha-1}{\beta}$ and set $\phi = \alpha - 1$, i.e. $\alpha = 1 + \phi$

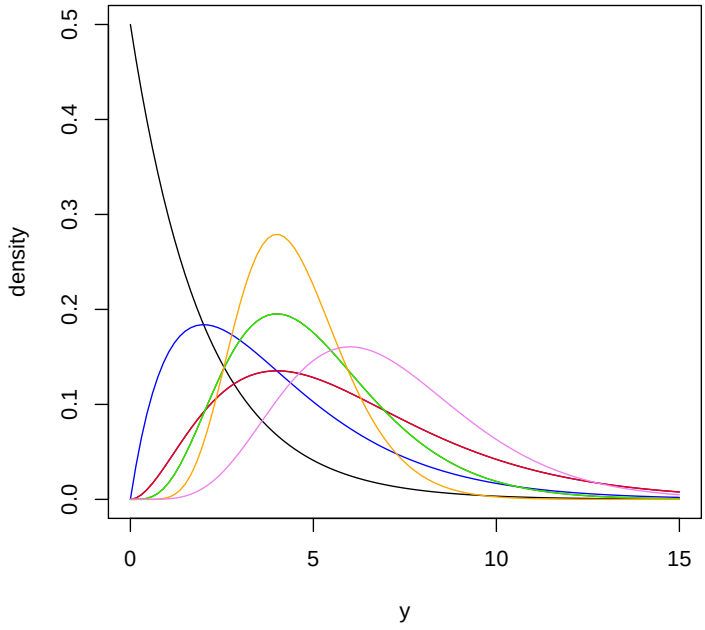


Figure 3.1: Gamma Densities for Different Combinations of α and β

and $\beta = \frac{\phi}{\mu}$. It follows from equation (3.2.1) that

$$f(y|\mu,\phi) = \frac{(\frac{\phi}{\mu})^{(1+\phi)}}{\Gamma(\phi)} y^{\phi} e^{\left(-\frac{\phi}{\mu}y\right)}, y > 0, \tag{3.2.3}$$

where, $\mu > 0$ and $\phi > 0$.

3.2.1 Regression and Model Fitting

Estimation of the unknown parameters is obtained via maximum likelihood, in a similar manner as for generalised linear models, and like regular maximum likelihood estimators they feature standard asymptotic properties.

Let $y_1, ..., y_n$ be a random sample such that $y_i \sim \Gamma(\mu_i, \phi)$, $i = 1...n$. The Gamma regression model is introduced through a link function which defines the mode of y_i as:

$$g(\mu_i) = \boldsymbol{x}'\underset{\sim}{b} = \eta_i, \tag{3.2.4}$$

where, $\underset{\sim}{b} = (b_1, ..., b_k)$, is a $k \times 1$ vector of unknown regression parameters, $\boldsymbol{x}' =$

$(x_1, ..., x_k)^T$ is the vector of k regressors and η_i is a linear predictor. Finally, $g(\cdot)$ is a strictly monotonic and twice differentiable link function.

There are several choices for the link function $g(\cdot)$, however a particularly useful link function is the logarithmic link function, $g(\mu) = log(\mu)$, in which case,

$$\mu_i = e^{\boldsymbol{x}'\underset{\sim}{b}}. \tag{3.2.5}$$

Thus, the density function of y conditional on $\underset{\sim}{b}$ and ϕ is given by

$$f(y|\underset{\sim}{b}, \phi) = \frac{\phi^{(1+\phi)} e^{-(1+\phi)(\boldsymbol{x}'\underset{\sim}{b})}}{\Gamma(\phi)} y^{\phi} e^{-\phi e^{(-\boldsymbol{x}'\underset{\sim}{b})} y}.$$

Given a sample of n independent observations $(x_i, y_i), i = 1...n$, the likelihood function is given by

$$L(\underset{\sim}{b}, \phi|x_i, y_i) = \frac{\phi^{n(1+\phi)}}{\Gamma^n(\phi)} \prod_{i=1}^{n} y_i^{\phi} \quad e^{-(1+\phi)\sum_{i=1}^{n} x_i'\underset{\sim}{b}} \quad e^{-\phi \sum_{i=1}^{n} y_i e^{-x_i'\underset{\sim}{b}}}, \tag{3.2.6}$$

and the corresponding log-likelihood function is defined as

$$l(\underset{\sim}{b}, \phi) = n(1+\phi)log(\phi) - nlog(\Gamma(\phi)) + \phi \sum_{i=1}^{n} log(y_i) - (1+\phi) \sum_{i=1}^{n} x_i'\underset{\sim}{b} - \phi \sum_{i=1}^{n} y_i e^{-x_i'\underset{\sim}{b}}. \tag{3.2.7}$$

This model is a standard maximum likelihood problem for which there is no closed-form solution. Maximum likelihood estimates of $\underset{\sim}{b}$ and ϕ can be obtained by direct numerical optimisation of the log-likelihood function in equation (3.2.7), which can be easily computed using any statistical software for linear programming, for example, the `optim` function in R.

The optimisation algorithm requires the specification of initial values to be used in the iterative scheme. The initial values are set as the estimates of $\underset{\sim}{b}$ obtained from a linear regression of the transformed response $(g(y_1), ..., g(y_n))$ on $\boldsymbol{x}$. A number of randomly chosen initial values for the parameter ϕ were used and the one that gave the maximum log-likelihood value was selected.

3.2.2 Asymptotic Properties

Given that parameter estimation is performed by maximum likelihood, the estimators enjoy standard asymptotic properties. In this sub-section the score function

and the Fisher information matrix for $(\underset{\sim}{b}, \phi)$ are derived. Details of the derivations are given in Appendix A.2.

The score function, obtained by differentiating the log-likelihood function with respect to the unknown parameters is given by (S_β, S_ϕ), where,

$$\begin{aligned} S_\beta &= \textstyle\sum_{i=1}^{n} \left(-\frac{(1+\phi)}{\mu_i} + \frac{\phi y_i}{\mu_i^2}\right) \frac{\partial \mu_i}{\partial \eta} x_{ik}, \\ S_\phi &= \textstyle\sum_{i=1}^{n} \frac{1}{\phi} + log(\phi) + 1 - log(\mu_i) - \frac{\psi(\phi)}{\Gamma(\phi)} + log(y_i) - \frac{y_i}{\mu_i}, \end{aligned} \tag{3.2.8}$$

where $\psi(\phi)$ is the digamma function, and, from equation (3.2.5), $\eta = \log(\mu)$ and $\frac{\partial \mu_i}{\partial \eta} = e^\eta$.

Under standard regularity conditions for maximum likelihood estimation as $n \to \infty$,

$$\sqrt{n}\begin{pmatrix} \widehat{\underset{\sim}{b}} - \underset{\sim}{b}_0 \\ \widehat{\phi} - \phi_0 \end{pmatrix} \sim N_{k+1}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, I^{-1}\right), \tag{3.2.9}$$

where, I is the Fisher information matrix, $\underset{\sim}{b}_0$ and ϕ_0 are the true values of $\underset{\sim}{b}$ and ϕ respectively and

$$I = \begin{pmatrix} I_{bb} & I_{b\phi} \\ I_{\phi b} & I_{\phi\phi}, \end{pmatrix} \tag{3.2.10}$$

where, $I_{bb} = X^T W X$, $I_{b\phi} = X^T T w_{b\phi}$, $I_{\phi b} = I_{b\phi}^T$ and $I_{\phi\phi} = tr(D)$ and,

$$\begin{aligned} &W = diag(w_{bb_1}, ..., w_{bb_n}) \quad \text{with} \quad w_{bb_i} = \left(\tfrac{(1+\phi)}{\mu_i^2} - 2\tfrac{\phi y_i}{\mu_i^3}\right)\left(\tfrac{d\mu_i}{d\eta}\right)^2, \\ &T = diag\left(\tfrac{d\mu_i}{d\eta}\right), w_{b\phi} = (w_{b\phi_1}, ..., w_{b\phi_n}) \quad \text{with} \quad w_{b\phi_i} = -\tfrac{1}{\mu_i} + \tfrac{y_i}{\mu_i^2}, \\ &D = diag(w_{\phi\phi_1}, ..., w_{\phi\phi_n}) \quad \text{with} \quad w_{\phi\phi_i} = -\tfrac{1}{\phi^2} + \tfrac{1}{\phi} - \tfrac{\psi'(\phi)\Gamma(\phi) - (\psi'(\phi))^2}{(\Gamma(\phi))^2}, \end{aligned} \tag{3.2.11}$$

where, $\psi'(\phi)$ is the trigamma function and $\frac{d\mu_i}{d\eta} = e^\eta$ (from equation (3.2.5)).

3.2.3 Estimation of Confidence Intervals

Having obtained the maximum likelihood parameter estimates and using their asymptotic properties, it is possible to construct confidence intervals for the estimated parameters $\widehat{\underset{\sim}{b}}$. The expected Fisher informatics matrix, I can be transformed into the asymptotic variance of $\widehat{\underset{\sim}{b}}$, $\Sigma(\underset{\sim}{b})$. Then the estimated asymptotic variance matrix $\widehat{\Sigma}(\widehat{\underset{\sim}{b}})$ can be used to obtain a $100(1-\alpha)\%$ confidence interval for the estimated parameters:

$$\widehat{\underset{\sim}{b}} \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{\Sigma}(\widehat{\underset{\sim}{b}})}, \tag{3.2.12}$$

where $z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ quantile of the Normal distribution.

3.2.4 Finite Sample Experiments

In this sub-section the accuracy and robustness of the fully parametric mode regression is demonstrated through a Monte Carlo simulation.

The experiment is designed in such a way to examine the performance of the proposed approach under a series of different underlying density functions, including both monotone and unimodal density shapes, with both light and heavy tails.

Specifically, $n = 100$ observations for the response variable y were generated from the following three different density functions:

1. Gamma distribution with $y \sim Gamma(\phi + 1, \frac{\phi}{\mu})$ and $\phi = 8$,
2. Log-normal with $y \sim LogN(log(\mu) + \sigma^2, \sigma)$ and $\sigma = 0.25$,
3. Chi-square distribution with $y \sim \chi^2(\mu + 2)$,

where, $x \sim Uniform(0, 2)$ and

$$\mu = e^{\eta} \text{ with } \eta = b_0 + b_1 x.$$

Furthermore, for each density function, the following two sets of parameter values were considered:

a) $b_0 > 0$, $b_1 > 0$: $b_0 = 3$, $b_1 = 1$,

b) $b_0 > 0$, $b_1 < 0$: $b_0 = 3$, $b_1 = -1$

Each simulation experiment was replicated 100 times. In the analysis the estimated parameters were compared to the true parameter values. For each dataset two statistics were computed: the bias for each regression parameter and the root mean squared error for η, which were averaged over the 100 data sets from each scenario.

Table 3.1 reports the mean biases and the mean root mean squared errors taken over the 100 simulations. Figure 3.2 presents a series of boxplots which summarise the parameter estimates for the three error distributions in the simulation example.

Table 3.1: Simulation Example - Mean Biases for b_0 and b_1 and Root Mean Squared Errors for η

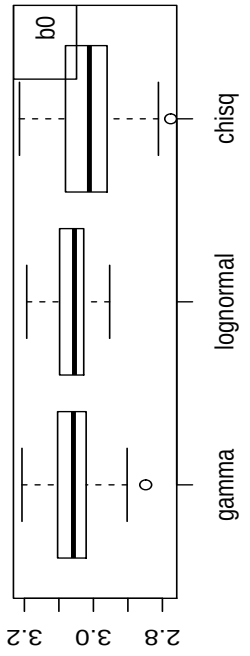
$f(\cdot)$	b_0	b_1	η
$b_0 > 0$, $b_1 > 0$			
Gamma	0.06	0.003	0.07
LogN	0.06	0.002	0.07
χ^2	0.02	-0.008	0.03
$b_0 > 0$, $b_1 < 0$			
Gamma	0.06	0.007	0.07
LogN	-0.003	0.03	0.04
χ^2	0.01	0.05	0.10

Examining these results it can be concluded that generally the mean biases and the root mean squared errors are quite small, which implies that, even under the small sample size of $n = 100$ the proposed method performs well. For most of the cases the bias is less than 0.1 and only in a couple of cases (both for b_0) it increases to 0.2. The simulation experiment was repeated with a larger dataset $n = 500$ but the differences in the estimated parameters were not significant.

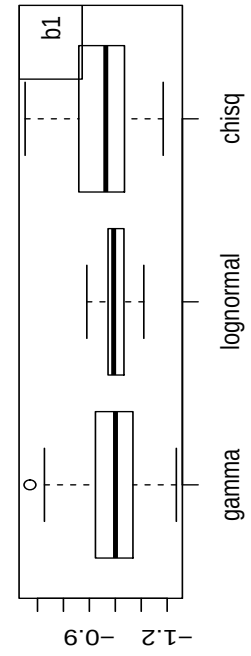
3.3 Big Data

According to IBM, every day, 2.5 quintillion bytes of data are being created[1]. These data come from different sources: sensors that gather climate information, social media sites, digital pictures and videos, purchase transaction records, and mobile phone GPS signals, among others. These data can be structured, semi-structured or unstructured. New big data technologies and tools (big data analytics) have been developing during the last years. Big data analytics assist in understanding the information contained within the data and in identifying the most important

[1] $http://www.ibm.com/software/data/bigdata/what-is-big-data.html$



(a) $b_0 > 0, b_1 > 0$



(b) $b_0 > 0, b_1 < 0$

Figure 3.2: Simulation Example - Boxplots of Parameter Estimates for the three Error Distributions

information. Sectors in which big data has received increasing attention include, the financial markets, medicine, meteorology, biology and physics. However, understanding and utilising big data is a daunting task. Also, there is a high cost associated with the capture, storage, processing, and analysis of such data.

Data could be the realisations of a random variable or variables, or the signals or symbols of events. Statistically, patterns in data are commonly described in terms of their centre, spread, shape, and unobserved features. If the random variables follow a symmetric distribution, the mean and variance are good measures of central tendency and variability. However, often data and variables do not follow a symmetrical distribution. In particular, big data often contains asymmetric variables and complex correlations. Furthermore, uncovering meaningful patterns in an efficient way is often required in big data analysis.

For example, the Body Mass Index (BMI) is a measure of body fat based on individual height and weight which indicates obesity. Given that more people are dying in England due to being overweight or obese than anywhere else in Europe[2] analysis of this data is of particular importance. Data on the BMI and other relevant variables is available from the Health Survey for England (HSE) which is an annual survey since 1991 designed to measure health and health related behaviours in adults and children living in private households in England. Measured height and weight data are recorded as part of a core data set. It can be assumed that this type of data is available from many different sources, e.g. from hospitals for different periods of time. Finding patterns over the population and time could be the first step in analysing such data.

Moreover, while identifying the typical value or pattern is an important part in big data analysis, this is not the only scientific objective of interest. Usually, quickly identifying the unknown correlations and/or complex relationships among variables is desirable. For example, the BMI data from the HSE also includes data on general health, smoking, drinking, fruit and vegetable consumption, blood pressure measurements, blood and saliva samples and other topic-specific health indica-

[2] $www.noo.org.uk/NOO_about_obesity/mortality$

tors. It is interesting to examine how different environmental exposures and lifestyle choices (smoking, alcohol status, medication status, fruit and vegetables consumed and region individuals live in) as well as genetic factors (gender, age, and ethnicity) influence the BMI. Data-mining pattern-finding algorithms may not be suitable for this purpose. Mode regression, which models the relationship between the pattern and other covariates could achieve this objective.

3.3.1 Big BMI Data Analysis

The dataset used for the analysis is taken from the Health Survey for England for the years 1997-2011, excluding the years 2000 and 2001 as data were not available from the survey for all the variables considered in the analysis. Figure 3.3 displays the histograms of the BMI density for each of the years considered in the analysis. This dataset consists of data from independent cross-sectional surveys and covers observations for 13 years. The dataset contains data on 195,173 individuals, hence it can be classified as an example of a big data dataset. Based on the data the aim is to identify BMI patterns, unknown correlations and other useful information efficiently. Following the proposed methodology, the analysis consists of two steps.

The first step involves identifying the pattern in the data. For each year, mode estimation was used to identify the range of modes (typical values) of the BMI variable and it was found that all BMIs follow a unimodal pattern; although the method can also be used for multimodal cases. Several methods for mode estimation exist in the literature. This thesis proposes the use of the Parzen's kernel mode estimation method, in which the mode is obtained by maximising the kernel density estimate.

Parzen (1962) discussed the problem of estimating a probability density function and estimating the mode of this density: Let $X_1, X_2, ...X_n$ be iid random variables with an absolutely continuous distribution function $F(x) = P(X \leq x)$, then $F(x) = \int_{-\infty}^{x} f(x)dx$, where $f(x)$ is the probability density function. An estimate of the distribution function $F(x)$ can be obtained by taking the sample distribution function, $F(x) = \frac{1}{n}$(no of observations $\leq x$ among $X_1, X_2, ...X_n$), which is a binomially distributed random variable. An estimate of the probability density function

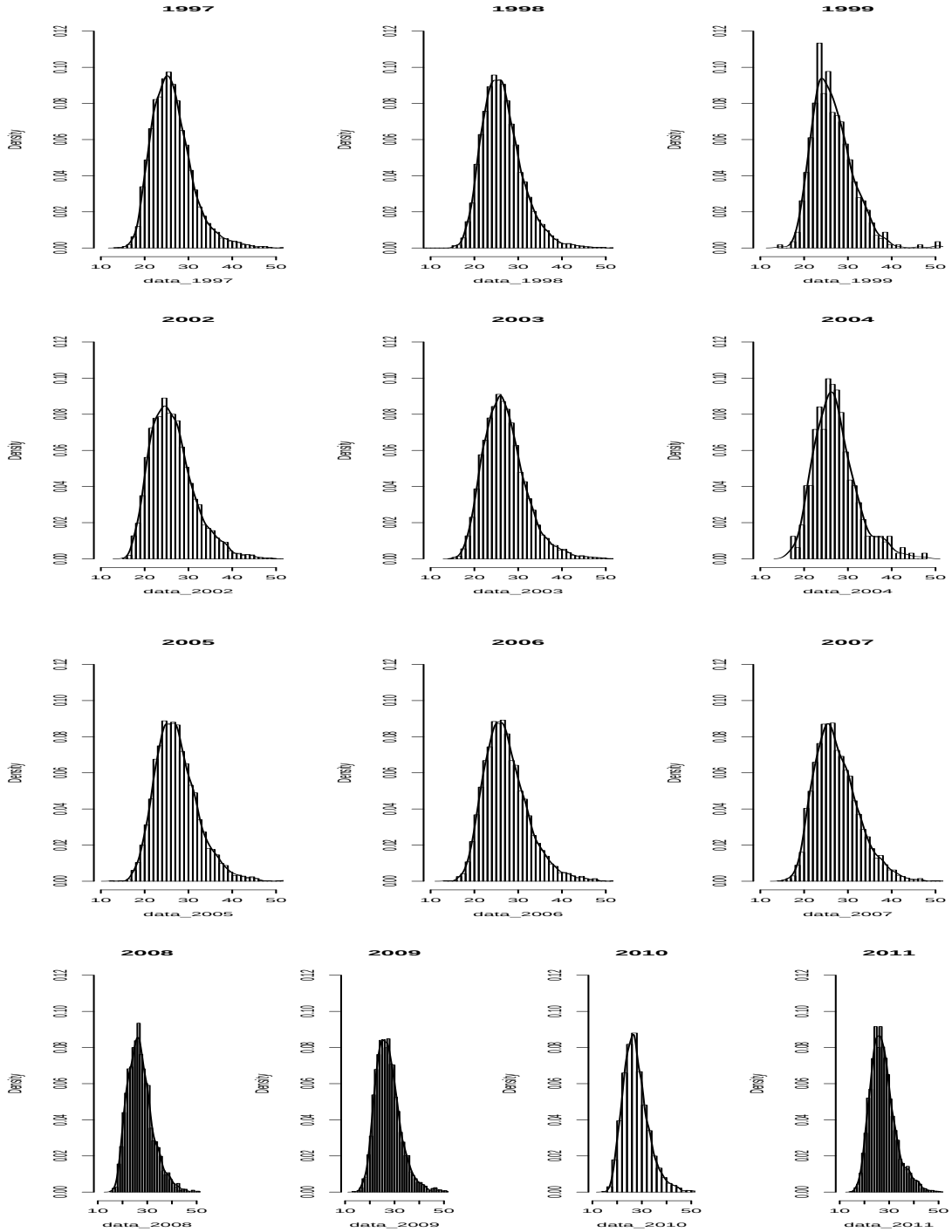


Figure 3.3: BMI Densities for the 13 years considered in the analysis

$f(x)$ can be obtained by

$$f(x) = (F_n(x+h) - F_n(x-h))/(2h), \text{ where } 0 < h < 1. \tag{3.3.13}$$

The estimate in equation (3.3.13) can be written as a weighted average over the

sample distribution function:

$$\widehat{f}(x) = \int_{-\infty}^{\infty} \frac{1}{h} K\left(\frac{x-y}{h}\right) dF_n(y) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x-X_j}{h}\right) \tag{3.3.14}$$

where, $K(y)$ can be any function which satisfies the following properties

$$\begin{aligned} &K1 : sup_{-\infty<y<-\infty}|K(y)| < \infty \\ &K2 : \lim_{y\to\infty} |yK(y)| = 0 \\ &K3 : \textstyle\int_{-\infty}^{\infty} K(y)dy = 1 \end{aligned} \tag{3.3.15}$$

and $\lim_{n\to\infty} h_n = 0$.

This estimator of $f(x)$ is consistent and asymptotically normal. An estimate of the mode can be obtained by maximising the estimated probability density function $f(x)$

$$mode = \arg \max_{-\infty<x<-\infty} f(x)$$

This estimator can be obtained using the `parzen` function in the "modeest" package in R.

When applying mode regression for the analysis of big data, in addition to the estimation of the mode, it is also necessary to identify the intervals containing most of the observations. Given that the mode can be also defined as the centre of an interval of a given length containing the majority of observations, and given that the mode estimator is asymptotical normal, having obtained an estimate of the mode it is possible to obtain the required intervals by applying a simple rule of thumb, based on the empirical rule. The rule states that, given a symmetric distribution, approximately 99.7% of the data values fall within three standard deviations (sd) of the mode, therefore, $interval = 3sd$. It should be noted here, that there is an element of subjectivity in choosing the criterion for identifying the ranges containing most of the observations, which is also applies to the identified ranges themselves.

Having identified these intervals, the next step is to collect the associated values of the covariates corresponding to these BMI values. Clearly, this first step is easy and quick to carry out. Then, all the collected data for the BMI and the associated

factors are merged to construct a new (smaller) dataset which will be used for the second step of the analysis: the mode regression.

Table 3.2 demonstrates the identified BMI mode intervals for each of the years considered in the analysis. The resulting dataset contained 14,272 observations for the BMI and the corresponding 11 covariates. The analysis was performed to provide a quick answer to the scientific question "what is the effect of factors such as gender, age, ethnic origin, income, waist hip ratio, number alcohol units consumed, consumption of fruit and vegetables as well as smoking on the typical BMI?"

Table 3.2: Typical BMI Values for the years 1997-2011

Year	Typical BMI range
1997	24-26
1998	24-26
1999	23-26
2002	22-27
2003	25-27
2004	23-28
2005	24-28
2006	24-27
2007	24-27
2008	24-27
2009	24-28
2010	22-24
2011	24-28

3.3.2 Regression Analysis

To answer the scientific question, a person's typical BMI is modelled as a function of the person's gender, sex_i (1=men, 0 =women), age, age_i (divided by 10), ethnic origin, $origin_i$ (1=white, 2=mixed, 3=asian or asian-british, 4=black or black-

british, 5=other), the total household income, $income_i$, the person's waist-hip ratio, $waisthip_i$, the frequency of drinking alcohol in the past 12 months, $alcohol_i$ (0=non drinker, 1=once or twice per year, 2=once every couple of months, 3=once or twice per month, 4=once or twice per week, 5=three of four days per week, 6=five or six days per week, 7=almost every day), the portion of fruit and vegetables consumed the previous day, $fruit\&veg_i$, and the number of cigarettes smoked per day, $cigs_i$.

The logarithmic link function based mode regression is given by:

$$\begin{aligned}\log(bmi_i) &= \beta_0 + \beta_1 sex_i + \beta_2 age_i + \beta_3 mixed_i + \beta_4 asian_i \\ &+\beta_5 black_i + beta_6 other_i + \beta_7 income_i + \beta_8 waisthip_i + \beta_9 alcohol_i \\ &+\beta_{10} fruit\&veg_i + \beta_{11} cigs_i + \epsilon_i\end{aligned} \tag{3.3.16}$$

The analysis was carried out for the total sample and for men and women separately. Parameter estimates were obtained via the parametric mode regression method described in section 3.2. Table 3.3 presents parameter estimates and the corresponding 95% confidence intervals.

The results of the analysis indicate that (in terms of the logarithmic BMI) in the case of the total sample, the gender, the age, the total household income, the waist-to-hip ratio and smoking are the variables that have a statistically significant effect on the typical BMI at the 95% level. Specifically, typically men have a 1% lower BMI compared to women and the BMI increases with age (a year increase in age causes a 0.14% increase in the typical BMI). The waist-to-hip ratio, has a strong positive effect on the typical BMI, as a unit increase in the waist-to-hip ratio would result in a 19% increase in the typical BMI. Furthermore, smoking is negatively correlated with the typical BMI, as typically 1 extra cigarette smoked per day would result in a 0.03% decrease in the typical BMI. Finally, the total household income is positively correlated with typical BMI, as a unit increase in the total household income would result in an increase of 0.03% in the typical BMI. Similar results were obtained both for men and women, although, in the case of men *age*, and in the case of women *income*, did not appear to have a statistically significant effect on the typical BMI. Furthermore, the effect of the waist-to-hip ratio on the typical BMI is more pronounced in the case of men, as a unit change in the waist-to-hip ratio of men would result in a 25% increase in the typical BMI, compared to a 15% increase

in the case of women.

Table 3.3: BMI Big Data (Dataset 1) - Parameter Estimates and 95%Corresponding Confidence Intervals (CI)

parameter	Total	95% CI	Men	95% CI	Women	95% CI
const	3.07	(3.04,3.09)	3.00	(2.96,3.05)	3.10	(3.07,3.13)
men	-0.01	(-0.014, -0.005)	-	-	-	-
age	0.0014	(0.0003,0.003)	0.0002	(-0.001,0.002)	0.002	(0.001, 0.004)
mixed	0.00	(-0.02,0.02)	0.001	(-0.03,0.03)	0.0002	(-0.03, 0.03)
asian	-0.002	(-0.01, 0.009)	-0.003	(-0.02,0.01)	0.0006	(-0.02,0.02)
black	0.004	(-0.01,0.02)	0.002	(-0.02,0.02)	0.009	(-0.01,0.03)
other	-0.01	(-0.03,0.002)	-0.01	(-0.03,0.008)	-0.01	(-0.03,0.008)
income	0.0003	(0.00004,0.0006)	0.001	(0.0002, 0.001)	0.00005	(-0.0003,0.0005)
waisthip	0.19	(0.16,0.22)	0.25	(0.20,0.30)	0.15	(0.11,0.19)
alcohol	-0.001	(-0.001,0.0003)	-0.0004	(-0.002,0.001)	-0.0008	(-0.002,0.0005)
fruit&veg	0.0002	(-0.0004,0.0009)	-0.0002	(-0.001,0.0008)	0.0007	(-0.0002,0.002)
smoking	-0.0003	(-0.0006,-0.0001)	-0.0004	(-0.0007, -0.0001)	-0.0002	(-0.0006,0.0002)

3.3.3 Effect of Physical Activity

The second step in the analysis focused on examining the effect of physical activity on the typical BMI. To perform this analysis a subset of the dataset was used, consisting of 8 years for which data on physical activity was available. The analysis involved the investigation of the effect of physical activity, *p.activity* (1=no exercise, 2=light exercise, 3=moderate exercise, 4=vigorous exercise), together with the above covariates on the typical BMI. In the first example the variable physical activity was treated as a numerical variable according to the following scale:1=no exercise, 2=light exercise, 4=moderate exercise, 9=vigorous exercise. The resulting dataset contained 7,130 observations for the BMI and the corresponding covariates.

First, the effect of 12 independent variables on the typical BMI was estimated

according to the model below:

$$\begin{aligned}\log(bmi_i) &= \beta_0 + \beta_1 sex_i + \beta_2 age_i + \beta_3 mixed_i + \beta_4 asian_i \\ &+\beta_5 black_i + \beta_6 other_i + \beta_7 income_i + \beta_8 waisthip_i + \beta_9 alcohol_i \\ &+\beta_{10} fruit\&veg_i + \beta_{11} cigs_i + \beta_{12} p_a ctivity_i + \epsilon_i\end{aligned} \tag{3.3.17}$$

Table 3.4: BMI Big Data (Dataset 2) - Parameter Estimates and 95% Corresponding Confidence Intervals (CI) - Regression Analysis (1)

parameter	Total	95% CI	Men	95% CI	Women	95% CI
const	2.99	(2.96,3.03)	2.91	(2.85,2.97)	3.03	(2.98,3.07)
men	-0.016	(-0.022,-0.010)	-	-	-	-
age	0.01	(0.07,0.011)	0.01	(0.004,0.01)	0.01	(0.009,0.013)
mixed	-0.003	(-0.03,0.03)	0.001	(-0.04,0.04)	-0.004	(-0.04,0.04)
asian	0.004	(-0.01,0.02)	0.0004	(-0.02,0.02)	0.01	(-0.01,0.03)
black	0.01	(-0.01,0.03)	0.005	(-0.02,0.03)	0.01	(-0.01,0.04)
other	-0.01	(-0.03,0.01)	-0.01	(-0.03,0.02)	-0.01	(-0.03,0.02)
income	0.001	(0.0004,0.001)	0.001	(0.0002,0.001)	0.001	(0.0003,0.002)
waisthip	0.21	(0.17,0.26)	0.30	(0.23,0.37)	0.15	(0.10,0.21)
alcohol	0.0001	(-0.001,0.001)	0.0002	(-0.002,0.002)	0.00001	(-0.002,0.002)
fruit&veg	0.001	(0.0004,0.002)	0.001	(-0.0002,0.002)	0.002	(0.0002,0.003)
smoking	-0.0001	(-0.0004,0.0002)	-0.0002	(-0.001,0.0002)	0.0001	(-0.0004,0.001)
p.activity	0.0002	(-0.001,0.001)	0.0003	(-0.001,0.002)	0.0002	(-0.001,0.002)

Table 3.4 presents the parameter estimates and the corresponding 95% confidence intervals. Again, the analysis was performed for the total sample and separately for men and women. The results of the analysis indicated that in the case of the total sample, but also for men and women separately, the gender, the age, the total household income and the waist-to-hip ratio are the variables that have a statistically significant effect on the typical BMI at the 95% level. Specifically, typically men have a lower BMI than women, the BMI increases with age and the waist-to-hip ratio has a strong positive effect on the BMI. In addition, the results also indicate that in the case of women, the consumption of additional fruits and vegetables also

has a positive statistically significant effect on the typical BMI. Finally, given the results, it can be concluded that the effect of the physical activity (treated as a numeric variable) does not have a significant effect on the typical BMI.

Evidence in the literature suggests that habitual physical activity plays a bigger role in attenuating age-related weight gain, rather than in promoting weight loss (Dipietro (1999)) and that increased physical activity reduces the magnitude of the age-related increase in the BMI and has an important and protective effect against weight gain (Bottai et al. (2014)).

Table 3.5: BMI Big Data (Dataset 2) - Parameter Estimates and Corresponding 90% Confidence Intervals (CI) - Regression Analysis (2)

parameter	Total	95% CI	Men	95% CI	Women	95% CI
const	2.96	(2.93,2.98)	2.87	(2.82,2.91)	3.00	(2.96,3.04)
men	-0.03	(-0.03,-0.02)	-	-	-	-
mixed	-0.01	(-0.03,0.01)	-0.002	(-0.04,0.03)	-0.01	(-0.05,0.02)
asian	-0.003	(-0.02,0.01)	-0.003	(-0.02,0.01)	-0.002	(-0.02,0.02)
black	0.003	(-0.01,0.02)	0.01	(-0.02,0.03)	0.003	(-0.02,0.03)
other	-0.02	(-0.03,-0.001)	-0.01	(-0.03,0.009)	-0.020	(-0.04,0.002)
income	0.0004	(0.0001,0.0008)	0.0005	(0.00003,0.001)	0.0004	(-0.0001,0.0008)
waisthip	0.32	(0.29,0.35)	0.39	(0.34,0.44)	0.26	(0.22,0.31)
alcohol	0.0009	(-0.0002,0.002)	0.0007	(-0.0009,0.002)	0.0009	(-0.0007,0.002)
fruit&veg	0.002	(0.0007,0.002)	0.001	(0.0001,0.002)	0.002	(0.0008,0.003)
smoking	-0.0005	(-0.0007,-0.0002)	-0.0005	(-0.0008,-0.0001)	-0.0004	(-0.0008,-0.00001)
p.activity	-0.0008	(-0.002,-0.00004)	-0.0004	(-0.001,0.0007)	-0.001	(-0.002,0.00004)

To examine the relationship between the typical BMI and age-related weight gain, first the regression in equation (3.3.17) was re-run, excluding the variable *age* from the model. The results of the analysis are shown in Table 3.5. According to these results, in the case of the total sample, the variable physical activity has a negative and statistical significant effect on the typical BMI. Specifically, a unit change in physical activity would result in a 0.08% decrease in the typical BMI. This

indicates a strong relationship between age and physical activity on the typical BMI and suggests further investigation.

To examine the combined effect of physical activity (treated as a numerical variable) and age an interaction term (between age and physical activity) was added to the model 3.3.17:

$$\begin{aligned}\log(bmi_i) &= \beta_0 + \beta_1 sex_i + \beta_2 age_i + \beta_3 mixed_i + \beta_4 asian_i + \beta_5 black_i \\ &+\beta_6 other_i + \beta_7 income_i + \beta_8 waisthip_i + \beta_9 alcohol_i + \beta_{10} fruit\&veg_i \\ &+\beta_{11} cigs_i + \beta_{12} p_a ctivity_i + \beta_{13} age_i * p_a ctivity_i \epsilon_i\end{aligned} \tag{3.3.18}$$

Table 3.6: BMI Big Data (Dataset 2) - Interaction between Age and Physical Activity (1)

parameter	Total	95% CI	Men	95% CI	Women	95% CI
const	3.03	(2.99,3.06)	2.97	(2.91,3.03)	3.05	(3.00,3.10)
men	-0.01	(-0.02,-0.01)	-	-	-	-
age	0.002	(-0.001,0.01)	-0.003	(-0.01,0.002)	0.01	(0.002,0.01)
mixed	-0.001	(-0.03,0.03)	0.004	(-0.04,0.05)	-0.003	(-0.04,0.04)
asian	0.004	(-0.01,0.02)	0.001	(-0.02,0.02)	0.01	(-0.01,0.03)
black	0.01	(-0.01,0.03)	0.01	(-0.02,0.03)	0.01	(-0.02,0.04)
other	-0.01	(-0.03,0.01)	-0.01	(-0.03,0.02)	-0.01	(-0.04,0.02)
income	0.001	(0.0003,0.001)	0.001	(0.0000.0.001)	0.001	(0.0003,0.001)
waisthip	0.21	(0.16,0.25)	0.28	(0.21,0.35)	0.16	(0.10,0.21)
alcohol	-0.0001	(-0.001,0.001)	-0.0002	(-0.002,0.002)	-0.0001	(-0.002,0.002)
fruit&veg	0.001	(0.0002,0.002)	0.001	(-0.001,0.002)	0.001	(0.0001,0.003)
smoking	-0.0001	(-0.0005,0.0002)	-0.0003	(-0.001,0.0001)	0.00004	(-0.0005,0.001)
p.activity	-0.01	(-0.01,-0.003)	-0.01	(-0.01,-0.004)	-0.004	(-0.01,-0.0001)
$p.activity * age$	0.001	(0.001,0.002)	0.002	(0.001,0.003)	0.001	(0.0001,0.002)

Table 3.6 presents parameter estimates and the corresponding 95% confidence intervals. Again, the analysis was performed for the total sample and separately for men and women. The results of the analysis indicated that in the case of the total

sample, but also for men and women separately, the gender, the total household income, the waist-to-hip ratio and smoking have a statistically significant effect on the typical BMI at the 95% level. As before, typically men have a lower BMI than women, the BMI increases with age, the waist-to-hip ratio has a strong positive effect on the typical BMI and smoking is negatively correlated with the typical BMI. In addition, the results again indicate that in the case of women, the consumption of additional fruits and vegetables also has a positive statistically significant effect on the typical BMI. Concerning the effect of the age and the physical activity, the results indicate that when the two variables are taken separately, the variable physical activity has a negative statistically significant effect on the typical BMI, for the total sample, but also for men and women separately, whereas, age has a positive statistical significant effect on the typical BMI for women. In addition, treated jointly it can be concluded that they have a positive, statistically significant effect on the typical BMI, indicating that the marginal effect of physical activity on the typical BMI in not the same for all the individuals. As indicated by the coefficient of the interaction term there is a positive heterogeneous effect of a unit increase in physical activity across age.

In last two examples, the categorical variable physical activity is assumed to be and treated as a numeric variable. This involves making the assumption that distances between each consecutive pair of points on the observed variable can be quantified by a number, i.e. either they are assumed equidistant or a different ratio is being chosen. This is often a reasonable but simplifying assumption, however, the chosen scale does not necessary lead to an optimal interpretation and information about the ordering is being lost or is based on an unrealistic assumptions.

To avoid this criticism, in the next example the variable physical activity is treated as a categorical variable with four levels (1=no exercise, 2=light exercise, 3=moderate exercise, 4=vigorous exercise). Thus, in equation (3.3.17), the variable *p.activity* was replaced by three dummy variables.

$$\log(bmi_i) = \beta_0 + \beta_1 sex_i + \beta_2 age_i + \beta_3 mixed_i + \beta_4 asian_i$$
$$+\beta_5 black_i + \beta_6 other_i + \beta_7 income_i + \beta_8 waisthip_i + \beta_9 alcohol_i$$
$$+\beta_{10} fruit\&veg_i + \beta_{11} cigs_i + \beta_{12} ligth_i + \beta_{13} moderate_i + \beta_{14} vigorous_i + \epsilon_i \tag{3.3.19}$$

The results of the analysis (Table 3.7) suggest that the effect of the variable physical activity (treated as a categorical variable) is not statistically significant for the typical BMI.

Table 3.7: BMI Big Data (Dataset 2) - Parameter Estimates and 95%Corresponding Confidence Intervals (CI) - Regression Analysis (3)

parameter	Total	95% CI	Men	95% CI	Women	95% CI
const	2.99	(2.95,3.02)	2.91	(2.85,2.97)	3.02	(2.97,3.07)
men	-0.016	(-0.022,-0.010)	-	-	-	-
age	0.009	(0.007,0.011)	0.006	(0.004,0.009)	0.011	(0.009,0.013)
mixed	-0.003	(-0.032,0.026)	0.001	(-0.042,0.044)	-0.004	(-0.043,0.036)
asian	0.004	(-0.011,0.019)	0.0003	(-0.020,0.021)	0.009	(-0.012,0.030)
black	0.007	(-0.01,0.03)	0.004	(-0.02, 0.03)	0.01	(-0.014,0.041)
other	-0.007	(-0.03,0.01)	-0.006	(-0.03,0.02)	-0.009	(-0.04,0.02)
income	0.001	(0.0004,0.001)	0.001	(0.0002,0.001)	0.001	(0.0003,0.002)
waisthip	0.21	(0.17,0.26)	0.30	(0.23,0.37)	0.16	0.10,0.21)
alcohol	0.000	(-0.001, 0.001)	0.0002	(-0.002,0.002)	0.0002	(-0.002,0.002)
fruit&veg	0.001	(0.0004,0.002)	0.001	(-0.0002,0.002)	0.002	(0.0002,0.003)
smoking	-0.0001	(-0.0004, 0.0002)	0.001	(-0.0002,-0.001)	0.0002	(0.0001,0.001)
light	0.005	(-0.008,0.018)	-0.0004	(-0.018,0.017)	0.011	(-0.008,0.03)
moderate	0.004	(-0.008,0.015)	-0.001	(-0.016,0.015)	0.009	(-0.008,0.03)
vigorous	0.005	(-0.007,0.017)	0.001	(-0.014,0.017)	0.009	(-0.008,0.03)

Next, the combined effect of physical activity and age on the typical BMI was investigated by adding 3 interaction terms between physical activity and age to the regression model in 3.3.19. This enabled an investigation of whether the age-related increase in the typical BMI is different for people with different habitual physical

activities according to the model below:

$$\begin{aligned}\log(bmi_i) &= \beta_0 + \beta_1 sex_i + \beta_2 age_i + \beta_3 mixed_i + \beta_4 asian_i \\ &+\beta_5 black_i + \beta_6 other_i + \beta_7 income_i + \beta_8 waisthip_i + \beta_9 alcohol_i \\ &+\beta_{10} fruit\&veg_i + \beta_{11} cigs_i + \beta_{12} ligth_i + \beta_{13} moderate_i + \beta_{14} vigorous_i + \\ &\beta_{15} age * ligth_i + \beta_{16} age * moderate_i + \beta_{17} age * vigorous_i + \epsilon_i\end{aligned} \tag{3.3.20}$$

Table 3.8: BMI Big Data (Dataset 2) - Interaction between Age and Physical Activity (2)

parameter	Total	95% CI	Men	95% CI	Women	95% CI
const	3.00	(2.96,3.05)	2.95	(2.87,3.02)	3.03	(2.97,3.09)
men	-0.014	(-0.02,-0.008)	-	-	-	-
age	0.008	(0.002,0.013)	0.005	(-0.003,0.013)	0.010	(0.002,0.018)
mixed	-0.001	(-0.03,0.03)	0.004	(-0.04,0.05)	-0.003	(-0.04,0.04)
asian	0.004	(-0.01,0.02)	0.001	(-0.02,0.02)	0.009	(-0.01,0.03)
black	0.006	(-0.01,0.02)	0.004	(-0.02,0.03)	0.01	(-0.02,0.04)
other	-0.008	(-0.03,0.01)	-0.008	(-0.03,0.02)	-0.009	(-0.04,0.02)
income	0.001	(0.0003,0.001)	0.001	(0.00001,0.001)	0.001	(0.0002,0.001)
waisthip	0.207	(0.16, 0.25)	0.28	(0.21,0.35)	0.16	(0.10,0.21)
alcohol	-0.0001	(-0.001,0.001)	-0.0002	(-0.002,0.002)	-0.0001	(-0.002,0.002)
fruit&veg	0.001	(0.0002,0.002)	0.001	(-0.001,0.002)	0.001	(0.0001,0.003)
smoking	-0.0001	(-0.0005,0.0002)	-0.0003	(-0.0007,0.0001)	0.00002	(-0.0005,0.001)
light	0.007	(-0.029,0.04)	0.01	(-0.04,0.07)	0.002	(-0.05,0.05)
moderate	0.009	(-0.02,0.04)	0.008	(-0.04,0.05)	0.01	(-0.03,0.06)
vigorous	-0.028	(-0.06,0.004)	-0.04	(-0.08,0.006)	-0.015	(-0.062,0.03)
$light * age$	-0.001	(-0.008,0.006)	-0.003	(-0.013,0.007)	0.002	(-0.008,0.01)
$moderate * age$	-0.001	(-0.007,0.005)	-0.002	(-0.01,0.007)	-0.001	(-0.009,0.008)
$vigorous * age$	0.008	(0.002,0.02)	0.010	(0.001,0.02)	0.006	(-0.003,0.015)

The results of the analysis (Table 3.8) suggest that, as in previous examples, gender, age, income, waist-to-hip ratio and consumption of fruits and vegetables are the variables that have a statistically significant effect on the typical BMI. Concerning

the effect of the physical activity, the results indicate that when the variables are taken separately, age has a positive significant effect on the typical BMI, whereas, none of the 3 levels of physical activity can be considered as significantly different from zero at the 95% level. However, treated jointly, it can be concluded that vigorous activity has a positive, statistically significant effect on the typical BMI, for the total sample and for men and women separately, indicating that the vigorous activity has a heterogeneous effect on the typical BMI across age.

The last part of the analysis involved running separate single variable models (1 covariate: age) for people performing no physical activity, light physical activity, moderate physical activity and vigorous physical activity, respectively. According to the results of the analysis, for people who performed no physical activity a unit change in age causes a 0.11% increase in the typical BMI. The increase in the typical BMI is lower for people who perform light or moderate physical activity (0.098% and 0.093% respectively), whereas the respective effect for people who perform vigorous physical activity is larger (0.20%).

This result verifies the results often found in the literature, which suggest an inverse association between physical activity and age-related weight gain. A possible explanation for the positive coefficient of the interaction term between age and vigorous physical activity in model (3.3.20) as well as the larger coefficient of age in the single variable model for the vigorous activity dataset could be attributed to the nature of the physical activity variable, which only captures the intensity. Current physical activity recommendations refer to frequency, intensity and duration of physical activity as factors influencing healthy weight. Including only the intensity may capture gain in muscle mass associated with more intense physical activity rather than weight loss, which is more likely to occur with longer durations.

In summary, the analysis suggests that regular physical activity plays a role in attenuating age-related weight gain and that increasing physical activity may be necessary to effectively maintain a constant body weight with increasing age.

3.3.4 Data Reduction step: Out-of-sample Validation

As it has been mentioned before the proposed methodology involved a data reduction step which is accomplished by throwing away some data. Such techniques have been criticised for reducing the richness and quality of the data and may lead to a reduction of the information content of the data. In this section a comparison of the predictive power of the reduced-data model to the predictive power of the full-data model is performed, using an independent validation data set. The validation data set contained 1.316 observations from the 2012 Health Survey for England on the following 11 variables: the body mass index, bmi_i, gender, sex_i, age, age_i, ethnic origin, $origin_i$ (1=white, 2=mixed, 3=asian or asian-british, 4=black or black-british, 5=other), the total household income, $income_i$, the person's waist-hip ratio, $waisthip_i$, the frequency of drinking alcohol in the past 12 months, $alcohol_i$ (0=non drinker, 1=once or twice per year, 2=once every couple of months, 3=once or twice per month, 4=once or twice per week, 5=three of four days per week, 6=five or six days per week, 7=almost every day) and the number of cigarettes smoked per day, $cigs_i$.

$$\begin{aligned}\log(bmi_i) &= \beta_0 + \beta_1 sex_i + \beta_2 age_i + \beta_3 mixed_i + \beta_4 asian_i \\ &+\beta_5 black_i + beta_6 other_i + \beta_7 income_i + \beta_8 waisthip_i \\ &+\beta_9 alcohol_i + \beta_{10} cigs_i + \epsilon_i\end{aligned} \tag{3.3.21}$$

The model in 3.3.21 was fitted both under the reduced-data dataset and the full-data dataset. The predictive power of each model was accessed using the root mean square error (rmse) which obtained by

$$rmse = \sqrt{\frac{1}{n}\sum_{1=1}^{n}\left(\widehat{bmi}_i - bmi(validation)_i\right)} \tag{3.3.22}$$

In addition for each of the datasets the computational time was recorded.

The results indicate that the predictive power of the reduced-data model (rmse=0.19) is very similar to the predictive power of the full-data model (rmse=0.16), indicating that the proposed methodology retains data that explain much of the variance and omits data that explain little of the variance. However, the results for the com-

putation time indicate that the reduced model performs much better in terms of computation time (0.47 sec) as compared to the computation time of the full-data model (1.75 sec)

Combining the results it can be concluded that using the reduced-data model contributes to a 73% reduction in the computational time at the expense of a small decrease in predictive power.

3.4 Conclusions

In this chapter a fully parametric mode regression methodology, based on the Gamma distribution, is developed and a simple and quick 2-step methodology for the analysis of big data is proposed. The method is demonstrated through the analysis of the BMI big data dataset. Initially mode estimation is used for uncovering the typical pattern of a decade-long BMI dataset and then mode regression is applied for exploring the effect of a number of factors on the typical BMI. A fully parametric mode regression method is proposed which provides a quick and meaningful tool for big data analysis. The method demonstrates both good finite sample and asymptotic results.

Chapter 4

Binary Quantile Regression and Variable Selection

4.1 Introduction

Applications of regression models for binary response variables are quite common and models such as logistic regression and probit regression, are widely used in many fields and applications. However, these conventional binary regression models, focus on the estimation of the conditional mean function, which is not always the prime interest for a researcher. Also, they assume that the errors are independent of the regressors, which is rarely the case in practice. Quantile regression extends the mean regression model to conditional quantiles of the response variable and can provide estimation for a family of quantile functions that describe the entire underlining distribution of the response variable. Furthermore, quantile regression parameter estimates are not biased by a location-scale shift of the conditional distribution of the dependent variable. Quantile regression has been used by many researchers in different fields and has also been extended to censored data, count data and proportions.

The potential benefits of binary quantile regression have been recognised by several authors (e.g. Manski (1975), Horowitz (1992), Kordas (2006) and Benoit and Van den Poel (2010)) who developed different estimation techniques for the binary quantile regression model.

The general binary regression model is defined as:

$$\begin{aligned} y^* &= \boldsymbol{x}'\boldsymbol{\beta} + \epsilon_i, \\ y &= I\{y^* \geq 0\}, \end{aligned} \tag{4.1.1}$$

where, y_i^* is a continuous, scalar latent variable, y is the observed binary outcome of this latent variable, $I(\cdot)$ is the indicator function, $\boldsymbol{x}$ is a $p \times 1$ vector of explanatory variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of parameters and ϵ is a scalar random error term. If the distribution of ϵ conditional on x is known up to a finite set of parameters, $\boldsymbol{\beta}$ can be estimated by different techniques, including maximum likelihood. If it is assumed that ϵ has a Normal distribution then the binary probit model arises, whereas, if a logistic distribution is assumed then the model (4.1.1) becomes the binary logit model. Specifying the distribution of ϵ a priori, will yield inconsistent estimators if the distribution of ϵ is misspecified. A more flexible model is obtained by imposing only one assumption on ϵ, the quantile restriction $Q_\tau(\epsilon_i|x_i) = 0$.

Let $Q_\tau\left(y^*|\boldsymbol{x}\right)$ denote the conditional quantile of the latent variable y^* given $\boldsymbol{x}$, defined as:

$$Q_\tau\left(y^*|\boldsymbol{x}\right) \equiv F_{y^*}^{-1}(\tau|\boldsymbol{x}) \equiv \boldsymbol{x}'\boldsymbol{\beta}(\tau),$$

where $F(\cdot)$ is the distribution function of the latent variable y^* and $\tau \in [0, 1]$.

By the equivalence property to monotone transformations of the conditional quantile function (Powell (1986)), the τ^{th} conditional quantile function of the observed variable y_i in the model (4.1.1) can be expressed as:

$$Q_\tau\left(y|x\right) = I\{\boldsymbol{x}'\boldsymbol{\beta}(\tau) \geq 0\}. \tag{4.1.2}$$

Binary quantile regression was first introduced by Manski (1975, 1985). In these papers he introduced the Maximum Score Estimator (MSE), which requires very weak assumptions on the relation of errors to regression variables and can accommodate for heteroscedasticity of unknown form. Estimates of the regression parameters in model (4.1.1) can be obtained by:

$$\widehat{\boldsymbol{\beta}}(\tau) = \arg \max_{\{\boldsymbol{\beta}:\|\boldsymbol{\beta}\|=1\}} \sum_{i=1}^{n} [y_i - (1-\tau)] I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0\}, \tag{4.1.3}$$

where, $(x_i, y_i, i = 1, ..., n)$ is a random sample of observation and $0 < \tau < 1$ is the τ^{th} regression quantile. Identification of $\boldsymbol{\beta}$ is only possible up to a scale, thus to make estimation possible a scale normalisation is necessary. Manski (1975, 1985) used the normalisation $||\boldsymbol{\beta}|| = 1$, where $||\cdot||$ denotes the Euclidean norm.

Manski (1985) provided the conditions under which the maximum score and binary quantile regression estimators are consistent. However, this work faces important technical drawbacks in both optimising the objective function and inferring the regression parameters. The rate of convergence of $\widehat{\boldsymbol{\beta}}(\tau)$ and its asymptotic distribution were derived by Cavanagh (1987). Kim and Pollard (1990) showed that it is not asymptotically normal, but the estimator converges in distribution to the maximum of a complicated multidimensional stochastic process. Furthermore, the model is nonlinear in parameters thus its estimation is computationally more demanding than conventional linear quantile regression models. Delgado et al. (2001) attempted to solve the problem by using sub-sampling methods to form confidence intervals. They provided simulation evidence that suggests inconsistency of the bootstrap, a result that was later proved by Abrevaya and Huang (2005).

The maximum score estimator has a slow rate of convergence and a complicated asymptotic distribution because it is obtained by maximising a step function. To remedy some of these shortcomings Horowitz (1992) developed a smoothed maximum score estimator (SMSE) under a linear median regression specification for the latent variable in the binary model, which can be computed using standard optimisation routines. Kordas (2006) extended this estimator to a family of conditional quantile functions giving the opportunity for a complete understanding of the conditional distribution of the latent response variable given covariates:

$$\widehat{\boldsymbol{\beta}}_{smse}(\tau) = \arg \max_{\{\boldsymbol{\beta}:|\beta_1|=1\}} \sum_{i=1}^{n} [y_i - (1-\tau)] K\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau)}{h_n}\right) \tag{4.1.4}$$

where K is a smooth continuous function and h_n is a sequence of real positive constants converging to zero as the sample size increases. Identification of $\boldsymbol{\beta}$ up to scale requires that $\boldsymbol{x}$ has at least one component whose probability distribution conditional on the remaining components is absolutely continuous with respect to the Lebesgue measure (Manski (1985)). To make estimation possible Horowitz (1992)

imposes the normalisation, $|\beta_1| = 1$. This requires to arrange the components of $\boldsymbol{x}$ appropriately, so that x_1, satisfies this condition and accordingly, to re-arrange the components of $\boldsymbol{\beta}$ so that β_1 is the coefficient corresponding to x_1. Kordas (2006) discusses two possible normalisation methods $||\boldsymbol{\beta}|| = 1$ or $|\beta_p| = 1$. In this work the latter normalisation method was chosen.

Horowitz's approach is computationally simpler than the maximum score estimator. Also, under stronger conditions than in Manski (1975, 1985), Horowitz's estimator converges at a faster rate and is asymptotically normally distributed.

Benoit and Van den Poel (2010) provided numerical evidence for the usefulness of Bayesian quantile regression for binary response models based on the Asymmetric Laplace distribution.

Although both the maximum score and smoothed maximum score estimators have desirable asymptotic properties, they are difficult to implement in practice, and most importantly, they do not necessarily guarantee convergence and a unique solution. Specifically, the objective function in the maximum score estimator is discontinuous (step-function) therefore it cannot be solved using a gradient-based optimisation method, whereas, the objective function of the smoothed maximum score estimator can have several local maxima, therefore stochastic search algorithms are necessary to identify the global maximum (e.g. the simulated annealing algorithm suggested by Horowitz (1992)). Even though algorithms for solving both the MSE and the SMSE are readily available these are not included in standard software packages. Furthermore, the non-standard structure of their objective functions cannot always guarantee global convergence. These practical limitations motivate the development of the estimator described in this chapter. An alternative estimation approach is proposed, based on a nonlinear asymmetrical weighted loss function, which can be implemented by an iteratively reweighted least square algorithm (IRLS). The IRLS algorithm is computationally simple and guarantees convergence to a unique solution (Kokic et al. (1997)).

The remainder of the chapter is organised as follows. Section 4.2 introduces the Binary quantile regression, provides the asymptotic properties of the estimator and describes the proposed estimation approach and the corresponding algorithm

for binary quantile regression. Section 4.3 introduces the method of variable selection via the modern adaptive lasso technique and describes how this method can be implemented in the framework of the binary quantile regression. An estimation approach and the algorithm for variable selection using a penalised binary quantile regression objective function are provided. Section 4.4 illustrates the proposed methods through a Monte Carlo study and a real example. Concluding remarks are provided in Section 4.5. Technical proofs can be found in Appendix A.3.

4.2 Binary Quantile Regression

The estimator in equation (4.1.3) can be viewed as a $\tau - quantile$ version of the general linear binary quantile regression problem (Koenker and Bassett (1978)), which is obtained by solving:

$$\widehat{\boldsymbol{\beta}}(\tau) = \arg \min_{\{\boldsymbol{\beta}:|\beta_1|=1\}} \mathcal{R}_u(x) \tag{4.2.5}$$

where,

$$\mathcal{R}_u(x) = \sum_{i=1}^{n} w_i(\tau)|y_i - I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0\}|$$

and

$$w_i(\tau) = \begin{cases} \tau & \text{if } y_i - I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0\} \geq 0; \\ (1-\tau) & \text{if } y_i - I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0\} < 0. \end{cases}$$

A smoothed version of the model (4.2.5) can be contracted by replacing the indicator function with a smooth cumulative distribution function (cdf), $K(\cdot)$ (Horowitz (1992)), such as:

$$\widehat{\boldsymbol{\beta}}_{smse}(\tau) = \arg \min_{\{\boldsymbol{\beta}:|\beta_1|=1\}} \sum_{i=1}^{n} w_i(\tau) \left| y_i - K\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau)}{h_n}\right)\right| \tag{4.2.6}$$

where,

$$w_i(\tau) = \begin{cases} \tau & \text{if } y_i - K\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau)}{h_n}\right) \geq 0; \\ (1-\tau) & \text{if } y_i - K\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau)}{h_n}\right) < 0. \end{cases}$$

and $K(\cdot)$ satisfies the following properties,

$$\begin{aligned} &K1 : |K(v) < M| \text{ for some finite M and v } \in (-\infty, \infty) \\ &K2 : \lim_{v\to-\infty} K(v) = 0 \text{ and } \lim_{v\to\infty} K(v) = 1. \end{aligned} \tag{4.2.7}$$

4.2.1 Estimation of the Smoothed Binary Quantile Regression Model

In this sub-section an alternative estimation approach for estimating binary quantile regression models is developed, which is simple, is guaranteed to converge to a unique solution and can be implemented with standard software packages.

In a recent paper, Blevins and Khan (2013) demonstrated that for binary data the maximum score objective function in equation (4.2.5) is equivalent to the quadratic loss objective function under the median restriction, i.e for $\mathbf{w} = 0.5$. Since quantile regression can be viewed as a generalisation of median regression, in this chapter this work is extended to the estimation of binary regression quantiles using a nonlinear least asymmetric weighted squares (LAWS) approach. For any given quantile the estimator in model (4.2.5) is mathematically equivalent to the nonlinear LAWS estimator. Hence, the binary quantile regression objective function in equation (4.2.5), under Kordas (2006) normalisation can be written as:

$$\widehat{\boldsymbol{\beta}}_{laws}(\tau) = \arg \min_{\{\boldsymbol{\beta}:|\beta_p|=1\}} \sum_{i=1}^{n} w_i(\tau) \left(y_i - I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0\}\right)^2 \tag{4.2.8}$$

where, $\widehat{\boldsymbol{\beta}}_{laws}(\tau) = (\widehat{\beta}', 1)'$ and

$$w_i(\tau) = \frac{\mathcal{R}_u \left(y_i - I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0\}\right)}{\left(y_i - I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0\}\right)^2} \tag{4.2.9}$$

In the case of binary data it can be shown that equation (4.2.9) is equal to

$$w_i(\tau) = \begin{cases} \tau & \text{if } y_i - I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0\} \geq 0; \\ (1-\tau) & \text{if } y_i - I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0\} < 0. \end{cases} \tag{4.2.10}$$

The concept of LAWS was first introduced by Newey and Powell (1987), who used the so-called regression expectiles to investigate the underlying conditional distribution. Recently LAWS re-gained interest in the context of semiparametric or

geoadditive regression (see for example Schnabel and Eilers (2009) and Sobotka and Kneib (2010)). Breckling and Chambers (1988) proposed a M-quantile regression based on an asymmetric loss function and Jones (1994) showed that expectiles are quantiles of a transformation of the original distribution. Nonparametric estimation of regression expectiles was considered by Yao and Tong (1996) who used a kernel method based on a locally linear fit. Compared to quantile regression, the LAWS is reasonably efficient under normality conditions (Efron (1991)). Confidence intervals for expectiles based on an asymptotic Normal distribution were introduced by Sobotka et al. (2013).

4.2.2 Estimation Algorithm

The algorithm to estimate the model (4.2.8) is a nonlinear weighted least squares algorithm. However, since the weights are determined by the residuals that vary from iteration to iteration, a nonlinear IRLS approach is implemented.

To enable estimation, following Horowitz (1992), the standard Normal distribution, with cdf $\Phi(\cdot)$ is taken as the Kernel density and a customary normalisation $\boldsymbol{\beta}_n = 1$ is imposed. Then, the nonlinear binary regression estimator is obtained by minimising the nonlinear smoothed LAWS function (slaws):

$$\widehat{\boldsymbol{\beta}}_{slaws}(\tau) = \arg \min_{\{\boldsymbol{\beta}:|\beta_p|=1\}} \sum_{i=1}^{n} w_i(\tau) \left(y_i - \Phi \left(\frac{\boldsymbol{x}_i' \boldsymbol{\beta}(\tau)}{h_n} \right) \right)^2 \tag{4.2.11}$$

where, $\widehat{\boldsymbol{\beta}}_{slaws}(\tau) = (\widehat{\beta}', 1)'$ and

$$w_i(\tau) = \begin{cases} \tau & \text{if } y_i - \Phi \left(\frac{\boldsymbol{x}_i' \boldsymbol{\beta}(\tau)}{h_n} \right) \geq 0; \\ (1-\tau) & \text{if } y_i - \Phi \left(\frac{\boldsymbol{x}_i' \boldsymbol{\beta}(\tau)}{h_n} \right) < 0. \end{cases} \tag{4.2.12}$$

The steps of the algorithm for fitting the binary quantile regression model are described in Algorithm 2. These steps can be easily implemented using standard software packages such as R or Stata.

Algorithm 2 Binary quantile regression via nonlinear LAWS

1: Obtain an initial estimate of β by running standard nonlinear OLS regression.
2: Obtain an initial estimate of the residuals $\epsilon_i^0 = y_i - \Phi\left(\frac{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}(\tau)}{h_n}\right)$.
3: Construct the weights, $w_i^0(\tau)$ using equation (4.2.12) and estimate equation (4.2.11) via nonlinear WLS regression.
4: Obtain new estimates of the residuals, $\epsilon_i^1 = y_i - \Phi\left(\frac{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{slaws}(\tau)}{h_n}\right)$.
5: Update the weights to obtain $w_i^1(\tau)$ using equation (4.2.12).
6: Estimate equation (4.2.11) by nonlinear WLS regression.
7: Repeat steps 4 to 6 until convergence.

4.2.3 Asymptotic Properties

Regarding the asymptotic properties of the estimator, it can be shown that, under the following assumptions, Theorem 4.2.1 can be established.

Assumption 1. The vectors (x_i', ϵ_i') are identically and independently distributed random variables.

Assumption 2. $F_{\epsilon_i}(\cdot)$ is a distribution function with $F(0) = \tau$ and $Q_\tau(\epsilon_i|x_i) = 0$ for $\tau \in (0, 1)$.

Assumption 3. $\beta_n \in \mathbb{B}$, the closure of an open convex set of $\Re^{p-1}$.

Assumption 4. The support of x_i is not contained in any proper linear subspace of $\Re^p$.

Assumption 5. The density function, $f_{\epsilon_i|x_i}(\cdot)$ is positive in a neighborhood of 0.

Assumption 6. The weights $w_i(\tau)$ are independent of the regression parameters.

Assumption 7. The n vectors $x_j, j = 1...p-1$ are independently distributed with the first component of $x_{i1} \equiv 1$ for all i almost surely.

Assumption 8. $0 < P(y_i = 1|x_i) < 1$ for almost every x_i.

Theorem 4.2.1. *(proof is provided in Appendix A.3)*
If $h_n \to 0$, *then* $\widehat{\beta}(\tau) - \beta_0(\tau) \xrightarrow{p} 0$.

Furthermore, under regularity conditions identical to the ones in Horowitz (1992), the estimator enjoys asymptotic properties similar to those of the maximum score estimator Manski (1975, 1985). In particular, the rate of convergence can be as fast as the $O(n^{-1/3})$ and it has a non-Gaussian limiting distribution.

The slower rate of convergence relative to the smoothed maximum score estimator in Horowitz (1992) is due to a bias condition, where the bias of the estimator converges at the rate of h_n. This is in contrast to the rate of h_n^2 for the smoothed maximum score estimator. However, according to Blevins and Khan (2013) this bias condition can be easily corrected, e.g. by using a different kernel function to the Normal cdf, or via other bias-reducing mechanisms, such as jackknifing.

4.3 Variable Selection via Penalised Binary Quantile Regression

Variable selection plays an important role in the model-building process. A common problem when constructing a predictive model is the large number of candidate predictor variables. Identifying the smallest set of relevant variables has many advantages: (i) the process is cost-effective, usually simpler, and potentially faster, (ii) it improves the prediction performance of the predictors (iii) knowledge about the relevant variables can enhance the understanding of the underlying problem. Furthermore, multicollinearity and overfitting are areas of concern when a large number of independent variables are incorporated in a regression model.

The problem of overfitting also arises in quantile regression models. First, Koenker (2004) developed a L1-regularisation quantile regression method to shrink individual effects in longitudinal data towards a common value and Li and Zhu (2008) considered the L1-norm (LASSO) regularised quantile regression. The lasso is a regularised technique for simultaneous estimation and variable selection (Sobotka et al. (2013)). Even though the lasso is generally able to provide consistent variable selection and optimal prediction, scenarios exist in which the lasso selection cannot be consistent.

To solve this problem Zou (2006) developed a new version of the lasso, the adaptive lasso. This is a weighted L1 penalty which allows different penalisation parameters for different regression coefficients. The weights are determined by an initial estimator, $\widehat{\boldsymbol{\beta}}(\tau)$, e.g. the classical quantile regression estimator, and are used to construct weights based on the importance of each predictor. The most important advantage of the adaptive lasso is its oracle property, which estimators based on the

classical lasso do not enjoy. The oracle property requires that as the sample size increases the coefficient of non-relevant terms approaches zero and the probability of selecting the correct model goes to 1. Also, it requires that consistent model selection does not come at the expense of efficiency: the asymptotic distribution of the non-zero components of $\widehat{\boldsymbol{\beta}}$ must be the same as the "oracle model", when $\boldsymbol{y}$ is regressed only on the relevant variables. Wu and Liu (2009) considered variable selection through penalised quantile regression with adaptive lasso penalties in the framework of a linear model.

It should be noted that in Bayesian terms, the lasso procedure can be interpreted as a posterior mode estimate under independent Laplace priors for the regression coefficients (Tibshirani (1996), Park (2008)). Based on this principle Li (2010) proposed a Bayesian regularized quantile regression model by assuming that the model residuals come from the skewed Laplace distribution. The Laplace distribution has the attractive property that it can be represented as a scale mixture of normals with an exponential mixing density which leads to the development of a hierarchical Bayesian interpretation of the Lasso, which can be easily estimate by a Gibbs sampling algorithm. Benoit (2013) extended this work to bayesian lasso binary quantile regression.

In this section the modern adaptive lasso variable selection technique is extended to Binary quantile regression, in the framework of the nonlinear LAWS approach. Suppose that $\widehat{\boldsymbol{\beta}}(\tau)$ is a consistent estimator of $\boldsymbol{\beta}(\tau)$, the binary quantile regression estimator in equation (4.2.5). Then the $\tau - quantile$ version of the adaptive lasso binary quantile regression estimator, $\widehat{\boldsymbol{\beta}}^*$, is given by:

$$\widehat{\boldsymbol{\beta}}^*(\tau) = \arg \min_{\{\boldsymbol{\beta}:|\beta_1|=1\}} \sum_{i=1}^{n} w_i(\tau) \left| y_i - I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0\} \right| + \lambda_n \sum_{j=1}^{p} w_j^{lasso} \left|\boldsymbol{\beta}_j\right| \quad (4.3.13)$$

where, $w_i(\tau)$ is defined in equation (4.2.10), $\mathbf{w}^{lasso} = \frac{1}{|\widehat{\boldsymbol{\beta}}(\tau)|}$ is a known weights vector (Zou (2006)) and λ is a nonnegative regularisation parameter which controls the level of penalisation, with greater values implying more aggressive model selection. The second term in equation (4.3.13) is the adaptive lasso binary quantile regression penalty function, that is crucial for the success of the lasso.

4.3.1 Estimation Algorithm

In this sub-section the estimation approach to obtain the penalised binary quantile regression estimator in equation (4.3.13) is presented. The approach is simple and has the advantage of being implementable in standard software packages such as R or Stata.

Like the estimator for non-penalised binary quantile regression, developed in section 4.2, the estimator of the adaptive lasso binary quantile regression in equation (4.3.13) is mathematically equivalent to the penalised nonlinear LAWS estimator given:

$$\widehat{\boldsymbol{\beta}}^{*}_{adapt.lasso_{laws}}(\tau) = \arg \min_{\{\boldsymbol{\beta}:|\beta_p|=1\}} \sum_{i=1}^{n} w_i(\tau) \left(y_i - I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0\}\right)^2 + \lambda_n \sum_{j=1}^{p} w_j^{lasso} \left|\boldsymbol{\beta}_j\right| \tag{4.3.14}$$

where, $\widehat{\boldsymbol{\beta}}_{laws}(\tau)$ is a consistent estimator of $\boldsymbol{\beta}(\tau)$ in equation (4.2.8), $w_i(\tau)$ is defined as before, $\mathbf{w}^{lasso} = \frac{1}{|\widehat{\boldsymbol{\beta}}_{laws}(\tau)|}$ and λ is a nonnegative regularisation parameter.

Again, as in the non-penalised binary quantile regression estimator, to enable estimation the Indicator function is replaced by the standard Normal kernel density, $\Phi(\cdot)$. Then, the nonlinear adaptive lasso smoothed binary quantile regression estimator is defined as:

$$\widehat{\boldsymbol{\beta}}^{*}_{adapt.lasso_{slaws}}(\tau) = \arg \min_{\{\boldsymbol{\beta}:\|\beta_p\|=1\}} \sum_{i=1}^{n} w_i(\tau) \left(y_i - \Phi\left(\frac{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau)}{h_n}\right)\right)^2 + \lambda_n \sum_{j=1}^{p} w_j^{lasso} \left|\boldsymbol{\beta}_j\right| \tag{4.3.15}$$

where, $w_i(\tau)$ is defined in equation (4.2.12), $\widehat{\boldsymbol{\beta}}_{slaws}(\tau)$, is a consistent estimator of the binary quantile regression estimator in equation (4.2.11), $\mathbf{w}^{lasso} = \frac{1}{|\widehat{\boldsymbol{\beta}}_{slaws}(\tau)|}$ and λ is a nonnegative regularisation parameter.

The estimator can be obtained by an iteratively re-weighted least square algorithm (IRLS). The steps of the algorithm for fitting the adaptive lasso binary quantile regression model are described in Algorithm 3.

Choice of λ

The selection of the tuning parameters λ should be based on a data-driven approach to allow for increasing flexibility with the sample size. The most common

Algorithm 3 Variable Selection via Penalised Binary quantile regression

1: Obtain an initial estimate for non-penalised binary quantile regression, $\widehat{\boldsymbol{\beta}}_{slaws}(\tau)$, via Algorithm 2.
2: Calculate $\mathbf{w}^{lasso} = \frac{1}{|\widehat{\boldsymbol{\beta}}_{slaws}(\tau)|}$.
3: Use the initial estimates $\widehat{\boldsymbol{\beta}}_{slaws}(\tau)$ to obtain an initial estimate of the residuals $\epsilon_i^0 = y_i - \Phi\left(\frac{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{slaws}(\tau)}{h_n}\right)$.
4: Construct the initial weights, $w_i^0(\tau)$ using equation (4.2.12).
5: Use $\mathbf{w}^{lasso}$ and $w_i^0(\tau)$ to optimise the objective function in equation (4.3.15) via direct numerical optimisation.
6: Obtain new estimates of the residuals, $\epsilon_i^1 = y_i - \Phi\left(\frac{\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}_{salaslaws}(\tau)}{h_n}\right)$.
7: Update the weights to obtain $w_i^1(\tau)$ using equation (4.2.12).
8: Re-estimate equation (4.3.15) via direct numerical optimisation.
9: Repeat steps 6 to 8 until convergence.

way for its selection is the method of K-fold cross-validation. This is a measure of the out-of-sample estimation error under different configurations for tuning parameters, without collecting additional data.

The first step of the approach involves selecting a grid of candidate values for λ and dividing the data into K roughly equal folds. For each candidate value of λ the model is fitted K-1 times, each time leaving out one of the folds and the model prediction error of computed using the Kth fold by:

$$E_k(\lambda) = \sum_{i \in K^{th} fold} \left(y_i - \widehat{y}_{(-i)}(\lambda)\right)^2, \tag{4.3.16}$$

where, $\widehat{y}_{(-i)}(\lambda)$ is the fitted value from the model that excludes the fold containing i.

This gives the cross-validation error

$$CV(\lambda) = \frac{1}{K}\sum_{k=1}^{K} E_k(\lambda) \tag{4.3.17}$$

The selected tuning parameter is the one that minimises the cross-validation error.

4.4 Numerical Experiments

In this section the proposed approach for binary quantile regression and variable selection is demonstrated through two simulated and one real examples. The first simulation example is carried out to examine the performance of the proposed binary quantile regression estimator, using a nonlinear least asymmetric weighted squares (LAWS) approach. The second simulation example demonstrates the proposed approach for variable selection in binary quantile regression models. The real example is based on the widely studied transport-choice dataset described in Horowitz (1993). All programs were written and executed in the free statistical package R.

4.4.1 Simulation Example 1 - Binary Quantile Regression

In the first simulation experiment the following model was considered for simulating data:

$$y_i^* = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \tag{4.4.18}$$

where $x_{pi} \sim N(0,1)$, $i = 1, ..., n$ and $n = 500$ and $\boldsymbol{\beta} = (-0.1, -1, 1)$.

For the model error ϵ_i the following three specifications were considered:

- a homoscedastic symmetric error specification: $\epsilon_i \sim N(0,1)$.
- a homoscedastic asymmetric error distribution: $\epsilon_i \sim \chi^2(1)$, minus its median.
- a heteroscedastic error distribution: $\epsilon_i \sim (2 + x_{1i})N(0,1)$.

The model parameters were estimated using the proposed binary quantile regression approach. For each case 150 Monte Carlo simulations were run. Table 4.1 summarises the estimated parameters and the standard errors for β_0 and β_1 under all three error specifications[1]. The results of the analysis indicate that even in a relatively small sample size the estimator works relatively well, especially in the homoscedastic cases. Therefore, it can be concluded that the proposed binary quantile

[1]The value of β_2 has been normalised to 1.

Table 4.1: Simulation Example 1 - Estimated Parameters and (Standard Deviations)

	Normal		Heteroscedastic		Asymmetric	
τ	β_0	β_1	β_0	β_1	β_0	β_1
0.10	-1.21	-0.97	-2.09	-1.90	-0.52	-1.01
	(0.05)	(0.05)	(0.11)	(0.12)	(0.03)	(0.04)
0.25	-0.66	-0.91	-1.1	-1.36	-0.33	-0.99
	(0.04)	(0.05)	(0.06)	(0.09)	(0.03)	(0.04)
0.50	-0.09	-0.89	0.01	-0.83	-0.02	-0.94
	(0.03)	(0.04)	(0.04)	(0.05)	(0.03)	(0.04)
0.75	0.48	-0.90	0.96	-0.49	0.61	-0.86
	(0.04)	(0.04)	(0.05)	(0.05)	(0.04)	(0.05)
0.90	1.01	-0.94	1.87	-0.27	1.54	-0.87
	(0.05)	(0.05)	(0.08)	(0.07)	(0.07)	(0.06)

regression estimator is a viable alternative to the smoothed maximum score estimator given that its implementation simplicity does not come at the expense of finite sample performance.

4.4.2 Simulation Example 2 - Variable Selection

In this sub-section the performance of the proposed penalised binary quantile regression approach is investigated through a simulated example.

In this example data was simulated from the following regression model:

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta}(\tau) + \epsilon_i, \tag{4.4.19}$$

where $x_i \sim N(0,1)$, $i = 1, ..., n$, $n = 200$ and

$$\boldsymbol{\beta} = (0.5, 1.5, 0, 0, 2, 0, -1, 1)$$

.

20 validation and 20 training and 200 testing observations were simulated from

the model and three homoscedastic and one heteroscedastic specifications for the model error ϵ_i were considered,

- a homoscedastic symmetric error specification: $\epsilon_i \sim N(0,1)$
- a Laplace distribution: $\epsilon_i \sim Laplace(0,1)$
- a mixture of two Normal distributions: $\epsilon_i \sim 0.1N(0,1) + 0.9N(0,9)$
- a heteroscedastic error distribution: $\epsilon_i \sim (2 + x_{1i})N(0,1)$

The model was fitted using the generated data set. The experiment was repeated 100 times. All the penalised quantile regression estimates were obtained via direct numerical optimisation using the R function `optim`. The penalty parameter in lasso λ was chosen using the a cross-validation method.

In the analysis the estimated parameters were compared to the true parameter values. For every data generating process the bias was calculated, which was averaged over the 100 generated datasets from each scenario.

The results of the simulations are summarised in Table 4.2. It can be observed that, in general, the proposed method performs well when comparing the estimates $\widehat{\beta_j}$ with the true values β_j as the majority of the estimated biases are around or smaller than $|0.1|$.

4.4.3 Work-trip Mode-Choice Data Example

In order to assess the practical applicability of the proposed approach the method was tested on a previously published maximum score dataset (Horowitz (1993)). Mode choice modelling and prediction relate closely to transportation policies and can be useful for estimating travel demand and for mitigating traffic congestion. The dataset contains 842 observations sampled randomly from the Washington, D.C. area transportation study for each of the following four dependent variables: (i) the number of cars owned by traveller households, CARS, measured in car units; (ii) the transit out-of-vehicle travel time minus automobile out-of-vehicle travel time, DOVTT, measured in minutes; (iii) the transit in-vehicle travel time minus automobile in-vehicle travel time, DIVTT, also measured in minutes; and (iv) the transit

Table 4.2: Simulation Example 2 - Estimated Bias for Model Parameters

τ	$\widehat{\beta_0}$	$\widehat{\beta_1}$	$\widehat{\beta_2}$	$\widehat{\beta_3}$	$\widehat{\beta_4}$	$\widehat{\beta_5}$	$\widehat{\beta_6}$
Normal (0,1)							
0.10	0.30,	0.09	0.03	-0.004	0.05	-0.02	-0.03
0.25	0.08	0.03	-0.0009	-0.02	0.04	-0.01	-0.04
0.5	-0.05	0.01	-0.02	-0.03	0.008	-0.02	0.009
0.75	-0.09	-0.02	-0.01	0.007	0.07	-0.03	-0.04
0.90	-0.26	0.099	0.008	0.003	0.11	-0.007	-0.06
$Laplace(0,1)$							
0.10	0.08	0.08	0.04	0.05	0.09	-0.004	-0.04
0.25	-0.01	-0.08	-0.07	-0.01	-0.08	0.009	0.001
0.5	-0.003	-0.03	-0.04	-0.04	-0.02	-0.02	-0.03
0.75	0.06	0.02	-0.07	-0.06	0.04	-0.1	-0.11
0.90	-0.07	-0.12	-0.08	-0.11	-0.09	-0.03	-0. 13
Normal mixture							
0.10	0.34	0.09	-0.01	-0.04	0.20	-0.009	-0.09
0.25	0.18	0.06	-0.01	0.02	0.09	-0.004	-0.06
0.5	-0.04	0.0008	-0.04	-0.01	0.02	-0.03	-0.04
0.75	-0.18	0.04	-0.03	-0.01	0.04	-0.03	-0.08
0.90	-0.35	0.02	-0.04	-0.04	0.09	-0.02	-0.06
Heteroscedastic model							
0.10	0.05	0.40	0.06	-0.08	0.09	-0.05	-0.06
0.25	0.12	0.05	0.02	0.005	-0.22	-0.01	0.08
0.50	-0.29	-0.22	-0.03	-0.06	-0.17	-0.04	-0.02
0.75	0.03	0.03	-0.05	-0.07	0.01	-0.09	0.10
0.90	-0.10	-0.03	-0.09	0.16	0.13	-0.0002	-0.17

fare minus automobile travel cost, DCOST, measured in US dollars. The dependent variable of the resulting binary choice model was CHOOSE, which equals to 1 if the car is used and 0 otherwise, representing the latent variable "willingness to use a car". All continuous variables were standardised to have zero mean and unit

standard deviation for better comparison with results in the literature. Scale normalisation is achieved by setting the coefficient of DCOST equal to 1, as in Horowitz (1993), to enable the comparison of the obtained results to previous research.

Table 4.3 provides estimates of the model parameters for the median case ($\tau = 0.5$) as well as a comparison with the results obtained by three different estimation approaches, namely the smoothed maximum score estimator (Horowitz (1993)), a mixed integer optimisation (MIP) method (Florios and Skouras (2008)) and a Bayesian binary quantile regression (BBQR) approach based on the asymmetric Laplace distribution (Benoit and Van den Poel (2010)).

Table 4.3: Mode-Choice Data: Model Parameters Estimates

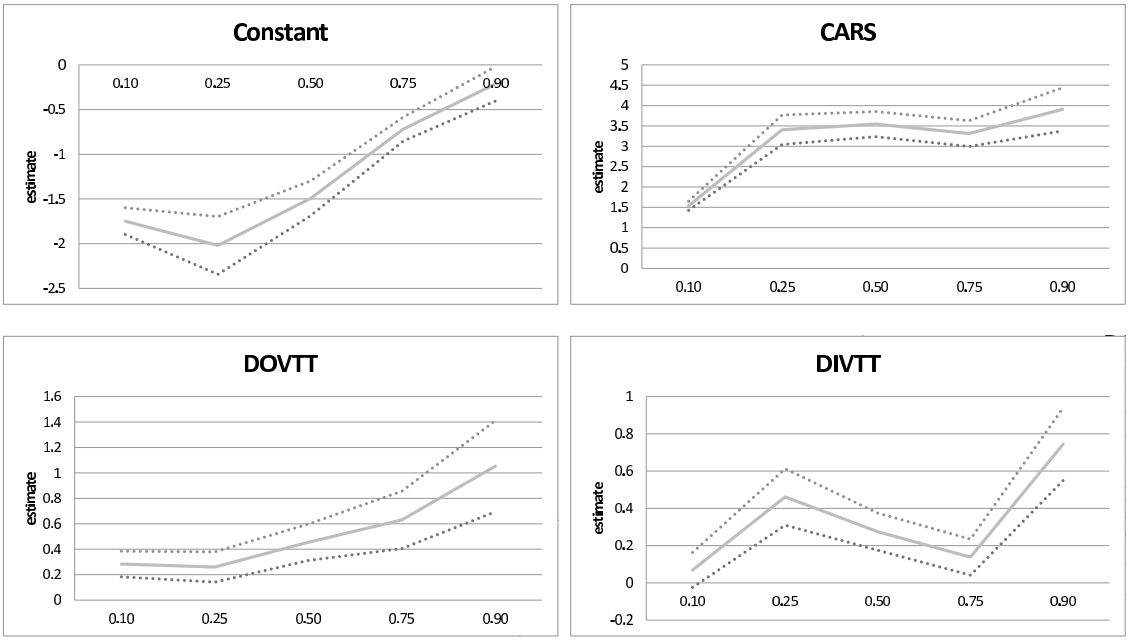
AUTHOR	INTERCEPT	CARS	DOVTT	DIVTT	DCOST	Method
Horowitz (1993)	-0.276	0.052	0.011	0.005	1	MSCORE
Florios and Skouras (2008)	5.122	3.916	0.962	0.401	1	MIP
Benoit and Van den Poel(2010)	4.825	3.375	1.018	0.282	1	BBQR
Current study	-1.493	3.545	0.455	0.274	1	LAWS

The analysis suggests that the results obtained by Horowitz (1993) are quite different from the ones obtained by Florios and Skouras (2008), and Benoit and Van den Poel (2010). According to Horowitz (1993), DCOST and CARS are the most important variables influencing the work-trip mode choice, with DCOST being by far the most important variable. In contrast, the results obtained by the other two methods, which are very similar between them, show that the variable CARS is by far the most important variable with the other variables having a small impact. The difficulty in computing maximum score estimates, discussed in Section 4.1, has been identified by many authors. In the context of computing estimators such algorithms are problematic because the statistical properties of such procedures can differ from those of exact estimates, e.g. as the ones provided by (Florios and Skouras (2008)).

The proposed LAWS approach delivers very similar estimates to the ones obtained both under MIP and BBQR. Furthermore, the technique is able to provide a more in-depth view of the relationship of the dependent variable and the covariates, as it allows to estimate the relationships at different parts of the distribution of

the response variable. Figure 4.1 illustrates the effect of covariates on the response variable at 0.10, 0.25, 0.50 , 0.75 and 0.90 quantile levels. The solid line represents the point estimates of the regression coefficients for the different quantiles and the dotted lines represent the upper and lower levels of a 95% confidence interval.

Figure 4.1: Mode-choice Dataset: Quantile Curves for Model Parameters



These results indicate that the effect of CARS and DOVTT on the unobserved willingness to take the car become stronger for higher conditional quantiles. This means that the effect of these variables is not constant across various quantiles of the latent variable. Specifically, commuters who have a low willingness to use the car are less affected by the number of cars whereas commuters with high willingness to use a car are more affected by the number of cars. Furthermore, commuters with increasing willingness to use a car are more affected by increasing out-ofvehicle trasportation time. In addition the results indicate that CARS is the most important variable as it has three times higher effect than the second variable, followed by the variable DCOST. The effect of DOVTT on the unobserved willingness to take the car is much lower than both CARS and DCOST, whereas, the respective effect of DIVTT is very small as compared to all the other variables.

4.5 Conclusions

In this chapter an alternative estimation approach to binary quantile regression and variable selection is proposed. The approach is based on a nonlinear asymmetrical weighted loss function which can be implemented by an iteratively reweighted least square algorithm (IRLS). Existing algorithms for fitting quantile regression models are not computational straight forward, hence they do not necessarily guarantee convergence and a unique solution. Also, due to their non-standard objective functions they cannot be computed using standard software packages. The main advantage of the proposed approach is that the IRLS algorithm is guaranteed to converge to a unique solution, whereas its computational simplicity makes it an attractive alternative to conventional methods. The results of the simulation study indicate that the ease of implementation does not come at the expense of finite sample performance.

Chapter 5

Conclusions and Future Work

5.1 Summary

Despite being an important measure of central tendency with potential benefits in data analysis, mode, and specifically mode regression, has been neglected in the statistical literature. This is mainly due to the lack of tools for implementing the existing mode regression methods, but also due to the limitations of the proposed estimators in terms of consistency and accuracy. A similar phenomenon is observed in the area of binary quantile regression, where, despite the popularity of binary models, there are no simple estimation techniques available that can be implemented with standard statistical packages.

This thesis presented a number of new regression methods for mode and binary quantile regression. The main objective of this work was to develop models which are simple, perform well in finite samples, have good large sample properties and can be implemented using standard statistical software. Furthermore, the thesis demonstrated the applicability of mode estimation and mode regression in big data analysis, which is currently a topic of increasing interest and importance in many fields of the global economy, for example, medicine, market research, finance, meteorology, environment and biology.

A Bayesian approach to mode regression was described in Chapter 2, where three distinct methods of estimation were presented. The first method involved a parametric Bayesian mode regression method, which was based on a uniform likelihood

function. The other two methods approached the problem in a nonparametric way with the aim of increasing flexibility and addressing the possibility of misspecification. The first method aimed at relaxing the distributional assumption on the prior of σ by employing a Dirichlet process prior. The second method aimed at avoiding the critical dependence on the parametric uniform distribution using the method of empirical likelihood, which combines the reliability of nonparametric methods with the flexibility and effectiveness of the likelihood approach. A fully parametric mode regression method, based on the Gamma density was introduced in Chapter 3. In addition, as it is always beneficial to demonstrate the applicability of a new approach within a valid domain, Chapter 3 also demonstrated how mode estimation and mode regression can be used for big data analysis through the analysis of the Health Survey for England data for the years 1997-2010. The aim of the analysis was to explore the effect of socio-economic characteristics and behavioural habits of adults in England on the typical Body Mass Index (BMI).

The proposed method for binary quantile regression was presented in Chapter 4. Although binary quantile regression has been previously studied in the literature, the existing methods involve complex estimation techniques. In contrast, the proposed method is simple and can be implemented with standard statistical packages. Furthermore, the method has been extended to accommodate variable selection via the modern adaptive lasso technique.

5.2 Discussion and Future Research Directions

The work presented in Chapters 2 and 3 paves the way for future research in the area of mode regression, and especially towards its application to big data analysis.

The Bayesian mode regression approach described in Chapter 2 was based on a parametric mode regression, which may lead to inconsistent estimators due to misspecification. Even though two new nonparametric approaches were presented in this chapter, there is room for further work in this area. Thompson et al. (2010) presented a nonparametric alternative to the Bayesian parametric quantile regression model of Yu and Moyeed (2001), using natural cubic splines, which provides

more flexible modelling. Similarly, developing spline-based nonparametric mode regression could be an extension of the proposed Bayesian inference of parametric mode regression.

Furthermore, an additional limitation of the proposed Bayesian inference method is the dependence on prior selection. Prior selection is very important in Bayesian modelling; the appropriate choice of priors, however, is a challenging task. Chapter 2 provided a number of suggestions for suitable priors for the model parameters $\boldsymbol{\beta}$ and σ, however, alternative options could further improve the performance of the model.

In Chapter 3 the proposed new fully parametric mode regression model is based on the Gamma distribution. However, the choice of the Gamma distribution is not binding for mode regression modelling. Extensions of the model can investigate the exploitation of other distributions for the response variable y which may allow increased flexibility and improved applicability. A natural first choice is the flexible generalised Gamma distribution, which is a generalisation of the two-parameter Gamma distribution. However, flexible mixtures of Gamma distributions are also worth exploring.

In addition, the inference approach described in Chapter 3 can be extended to a Bayesian framework. Put in a Bayesian framework, this approach will inherit the merits of mode regression in a modelling approach that takes into consideration uncertainty when making predictions.

Finally, an additional area that has not been addressed in the existing literature is variable selection for mode regression models. The proposed inference methods, both from the classical and the Bayesian perspective, can be extended to incorporate variable selection techniques. The application of modern adaptive lasso techniques, but also conventional methods based on either the Bayesian or the Akaike Information Criterion are available options for further research.

Chapter 4 described a new estimation technique for binary quantile regression for modelling a single quantile. Koenker (2005) notes that, in the case of binary response variables, the conditional probabilities cannot be estimated from a single binary quantile regression, thus the estimation of multiple conditional quantile functions is

of particular importance. Empirical likelihood provides the means for simultaneous estimation of multiple binary regression quantiles. Yang and He (2012) presented a method for empirical likelihood estimation for quantile regression models, which can form the basis of further research in the area of binary quantile regression. In this paper, Yang and He present the advantages of simultaneous estimation of multiple quantiles: the approach avoids the problem of crossing quantiles but also allows quantiles to share strength between them, thus leading to more accurate estimation. Furthermore, approaching the problem from the Bayesian perspective has the additional advantage of exploring commonality across quantiles through the use of informative priors.

Bibliography

Abrevaya, J. and Huang, J. (2005). On the bootstrap of the maximum score estimator. *Econometrica*, 73(4):1175–1204.

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422), 669-679.

Amemiya, T. (1985). *Advanced econometrics*. Harvard university press.

Benoit, D. F. and Van den Poel, D. (2010). Binary quantile regression: a bayesian approach based on the asymmetric laplace distribution. *Journal of Applied Econometrics*, 27(7):1174–1188.

Berlinet, A., Vajda, I., and Van der Meulen, E. (1998). About the asymptotic accuracy of barron density estimates. *Information Theory, IEEE Transactions*, 44(3):999–1009.

Bickel, P. and Fan, J. (1996). Some problems on the estimation of unimodal densities. *Statistica Sinica*, 6:23–46.

Birgé, L. (1997). Estimation of unimodal densities without smoothness assumptions. *The Annals of Statistics*, 25(3):970–981.

Blevins, J. R. and Khan, S. (2013). Local nlls estimation of semi-parametric binary choice models. *The Econometrics Journal*, 16(2):135–160.

Bottai, M., Frongillo, E. A., Sui, X., O'Neill, J. R., McKeown, R. E., Burns, T. L.and Liese, A. D. B. S. N., and Pate, R. R. (2014). Use of quantile regression to investigate the longitudinal association between physical activity and body mass index. *Obesity*, 2(5): 149–156.

Breckling, J. and Chambers, R. (1988). M-quantiles. *Biometrika*, 75(4):761–771.

Brunner, L. (1992). Bayesian nonparametric methods for data from a unimodal density. *Statistics & Probability letters*, 14(3):195–199.

Cavanagh, C. (1987). Limiting behavior of estimators defined by optimization. *unpublished manuscript (Department of Economics, Harvard University)*.

Chernozhukov, V. and Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2):293–346.

Delgado, M., Rodrıguez-Poo, J. and Wolf, M. (2001). Subsampling inference in cube root asymptotics with an application to Manski's maximum score estimator. *Economics Letters*, 73(2):241–250.

Dipietro, L. (1999). Physical activity in the prevention of obesity: current evidence and research issues. *Medicine and science in sports and exercise*, 31(11):542–546.

Dunson, D., Pillai, N. and Park, J. (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):163–183.

Eddy, W. (1980). Optimum kernel estimators of the mode. *The Annals of Statistics*, 8(4):870–882.

Efron, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, 1(93):125.

Feller, W. (1971). *An introduction to probability theory and its applications*, volume 2. Wiley-New york.

Ferguson, T. S.(1973). A Bayesian analysis of some nonparametric problems. *The annals of statistics*, 1(2): 209–230.

Florios, K. and Skouras, S. (2008). Exact computation of max weighted score estimators. *Journal of Econometrics*, 146(1):86–91.

Fogelholm, M. and Kukkonen-Harjula, K. (2000). Does physical activity prevent weight gain–a systematic review. *Obesity reviews*, 1(2):95–111.

Gasser, T., Hall, P., and Presnell, B. (1998). Nonparametric estimation of the mode of a distribution of random curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):681–691.

Grenander, U. (1965). Some direct estimates of the mode. *The Annals of Mathematical Statistics*, 36(1):131–138.

Hall, P. and Huang, L. (2001). Nonparametric kernel regression subject to monotonicity constraints. *The Annals of Statistics*, 29(3):624–647.

Hall, P., Peng, L., and Rau, C. (2001). Local likelihood tracking of fault lines and boundaries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):569–582.

Hastings, W. K.(1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Heckman, D., Geiser, D., Eidell, B., Stauffer, R., Kardos, N. and Hedges, S. (2001). Molecular evidence for the early colonization of land by fungi and plants. *Science*, 293(5532):1129.

Hedges, S. and Shah, P. (2003). Comparison of mode estimation methods and application in molecular clock analysis. *BMC Bioinformatics*, 4(1):31.

Ho, M.-W. (2006). Bayes estimation of a symmetric unimodal density via s-paths. *Journal of Computational and Graphical Statistics*, 15(4):848–860.

Horowitz, J. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica: Journal of the Econometric Society*, 60(3):505–531.

Horowitz, J. (1993). Semiparametric estimation of a work-trip mode choice model. *Journal of Econometrics*, 58(1):49–70.

Huber, P. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1(5):799–821.

Jones, M. (1994). Expectiles and m-quantiles are quantiles. *Statistics & Probability Letters*, 20(2):149–153.

Kemp, G. C. and Santos Silva, J. (2012). Regression towards the mode. *Journal of Econometrics*, 170(1):92–101.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *The Annals of Statistics*, 18(1):191–219.

Koenker, R. (2000). Quantile regression. *International Encyclopedia of the Social Sciences*.

Koenker, R. (2004). Quantile regression for longitudinal data. *Journal of Multivariate Analysis*, 91(1):74–89.

Koenker, R. (2005). *Quantile regression*, volume 38. Cambridge Univ Pr.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 46(1):33–50.

Kokic, P., Chambers, R., Breckling, J. and Beare, S. (1997). A measure of production performance. *Journal of Business & Economic Statistics*, 15(4):445–451.

Kordas, G. (2006). Smoothed binary regression quantiles. *Journal of Applied Econometrics*, 21(3):387–407.

Kottas, A. and Fellingham, W. (2012). Bayesian semiparametric modeling and inference with mixtures of symmetric distributions. *Statistics and Computing*, 22(1):93–106.

Kumar, S. and Hedges, S. (1998). A molecular timescale for vertebrate evolution. *Nature*, 392(6679):917–920.

Lazar, N. (2003). Bayesian empirical likelihood. *Biometrika*, 90(2):319–326.

Lee, M. (1989). Mode regression. *Journal of Econometrics*, 42(3):337–349.

Lee, M. (1993). Quadratic mode regression. *Journal of Econometrics*, 57(1-3):1–19.

Li, J., Ray, S., and Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8(8):1687-1723.

Li, Y. and Zhu, J. (2008). L1-norm quantile regression. *Journal of Computational and Graphical Statistics*, 17(1):163-185.

Lu, Z., Tjstheim, D. and Yao, Q. (2007). Adaptive varying-coefficient linear models for stochastic processes: asymptotic theory. *Statistica Sinica*, 17:177–197.

Manski, C. (1975). Maximum score estimation of the stochastic utility model of choice. *Journal of Econometrics*, 3(3):205–228.

Manski, C. (1985). Semiparametric analysis of discrete response:: Asymptotic properties of the maximum score estimator. *Journal of Econometrics*, 27(3):313–333.

Manski, C. (1991). Regression. *Journal of Economic Literature*, 29(1):34–50.

Markov, H., Valtchev, T., Borissova, J. and Golev, V. (1997). An algorithm to "clean" close stellar companions. *Astronomy and Astrophysics Supplement Series*, 122(1):193–199.

Meyer, M. (2001). An alternative unimodal density estimator with a consistent estimate of the mode. *Statistica Sinica*, 11(4):1159–1174.

Molanes Lopez, E., Keilegom, I. and Veraverbeke, N. (2009). Empirical likelihood for non-smooth criterion functions. *Scandinavian Journal of Statistics*, 36(3):413–432.

Monahan, J. and Boos, D. (1992). Proper likelihoods for Bayesian analysis. *Biometrika*, 79(2):271–278.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Newey, W. K. and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica: Journal of the Econometric Society*, 55(4):819–847.

Noufaily, A. and Jones, M. (2013). Parametric quantile regression based on the generalized gamma distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(5):723–740.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75(2):237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *The Annals of Statistics*, 18(1):90–120.

Owen, A. (1991). Empirical likelihood for linear models. *The Annals of Statistics*, 19(4):1725–1747.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Powell, J. (1986). Censored regression quantiles. *Journal of econometrics*, 32(1):143–155.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22:300–325.

Roberts, G. and Rosenthal, J. (2001). Optimal scaling for various metropolis-hastings algorithms. *Statistical Science*, 16(4):351–367.

Schnabel, S. K. and Eilers, P. H. (2009). Optimal expectile smoothing. *Computational Statistics & Data Analysis*, 53(12):4168–4177.

Sethuraman, J. (1994). A Constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639-650.

Silverman, B. (1986). *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall/CRC.

Smith, B. (2007). boa: an R package for MCMC output convergence assessment and posterior inference. *Journal of Statistical Software*, 21(11):1–37.

Sobotka, F., Kauermann, G., Schulze Waltrup, L. and Kneib, T. (2013). On confidence intervals for semiparametric expectile regression. *Statistics and Computing*, 23:135–148.

Sobotka, F. and Kneib, T. (2010). Geoadditive expectile regression. *Computational Statistics & Data Analysis*.

Thompson, P. and Cai, Y. and Moyeed, R. and Reeve, D. and Stander, J. (2010). Bayesian nonparametric quantile regression using splines. *Computational Statistics & Data Analysis*, 54(4):1138-1150.

Tremblay, M. S., Inman, J. W. and Willms, J. D. (2000). The relationship between physical activity, self-esteem, and academic achievement in 12-year-old children. *Pediatric Exercise Science*, 12(3):312–323.

Van der Vaart, A. (1998). *Asymptotic statistics*. Cambridge Univ Pr.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 50(1):1–25.

Wu, Y. and Liu, Y. (2009). Variable selection in quantile regression. *Statistica Sinica*, 19(2):801.

Yang, Y. and He, X. (2012). Bayesian empirical likelihood for quantile regression. *Annals of Statistics*, 40(12):1102–1131.

Yao, Q. and Tong, H. (1996). Asymmetric least squares regression estimation: A nonparametric approach. *Journal of Nonparametric Statistics*, 6(2-3):273–292.

Yao, W. and Li, L. (2014). A new regression model: modal linear regression. *Scandinavian Journal of Statistics*, 41(3):656671.

Yasukawa, K. (1926). On the probable error of the mode of skew frequency distributions. *Biometrika*, 18(3/4):263–292.

Yu, K and Moyeed, R. (2001). Bayesian quantile regression. *Statistics & Probability Letters*, 54(4):437–447.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Benoit, D. and Alhamzawi, R. and Yu, K. (2013). Bayesian lasso binary quantile regression. *Computational Statistics*, 28(6): 2861-2873.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological*, 58(1): 267-288.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482): 681–686.

Li, Q. and Ruibin, X. and Nan L. (2010). Bayesian regularized quantile regression. *Bayesian Analysis*, 5(3), 533-556.

Appendix A

Proofs of Theoretical Results

A.1 Proofs of Main Results: Chapter 2

Proof of theorem 2.2.1

The γth moments of marginal posterior distribution of $\boldsymbol{\beta}$ is given by

$$E[|\boldsymbol{\beta}|^{\gamma}|\sigma, \boldsymbol{y}] = \int \frac{1}{(2\sigma)^n} \prod_{i=1}^{n} I[|y_i - x_i'\boldsymbol{\beta}| < \sigma]\, \pi(\boldsymbol{\beta}, \sigma) d\boldsymbol{\beta}\, d\sigma.$$

Note that $\prod_{i=1}^{n} I[|y_i - x_i'\boldsymbol{\beta}| < \sigma]$ provides joint bands for all components β_j $(j = 0, 1, ..., p)$ of $\boldsymbol{\beta}$. Assume $0 < |\beta_j| < B_j < \infty$ $(j = 0, 1, ..., p)$, even if some of $|y_i - x_i'\boldsymbol{\beta}| < \sigma$ are true and some are not.

Therefore,

$$E[|\boldsymbol{\beta}|^{\gamma}|\sigma, \boldsymbol{y}] = \int \frac{1}{(2\sigma)^n} d\sigma \int_{-B_0}^{B_0} \int_{-B_1}^{B_1} ... \int_{-B_p}^{B_p} \prod_{j=0}^{p} |\beta_j|^{r_j}\, \pi(\boldsymbol{\beta}, \sigma)\, d\boldsymbol{\beta},$$

which is clearly finite.

Similarly, for the γth moment of marginal posterior of σ with $\gamma < n$ is defined as $E[|\sigma|^{\gamma}|\boldsymbol{\beta}, \boldsymbol{y}]$, and can be provided finite in the same way.

Proof of theorem 2.3.1

We will show Theorem 2.3.1 by applying a generic consistency lemma, Lemma 4.1, of Lu et al. (2007). For convenience of statement, we define $R_n(\lambda, \boldsymbol{\beta}) \equiv n^{-1} \sum_{i=1}^{n} \log(1+ \lambda' g(X_i, Y_i, \boldsymbol{\beta}))$ and $R(\lambda, \boldsymbol{\beta}) \equiv E\{\log(1+\lambda' g(X_i, Y_i, \boldsymbol{\beta}))\}$. Then note that $R_n(\hat{\lambda}(\boldsymbol{\beta}), \boldsymbol{\beta}) = -\Gamma_n(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{B}} R_n(\hat{\lambda}(\boldsymbol{\beta}), \boldsymbol{\beta})$, where $\hat{\lambda}(\boldsymbol{\beta})$ and $\Gamma_n(\boldsymbol{\beta})$ are defined in (2.3.15) and (2.3.16), respectively, and $\mathbb{B}$ is a compact subset of $\mathbb{R}^p$ containing the

true parameter vector $\boldsymbol{\beta}_0$ as an interior point. Further, we denote by $H_n(\lambda, \boldsymbol{\beta})$ for the left-hand side of (2.3.15) divided by n, that is $H_n(\lambda, \boldsymbol{\beta}) \equiv n^{-1} \sum_{i=1}^{n} \{g(X_i, Y_i, \boldsymbol{\beta})/[1+ \lambda' g(X_i, Y_i, \boldsymbol{\beta})]\}$, and hence for any $\boldsymbol{\beta} \in \mathbb{B}$, $\hat{\lambda}(\boldsymbol{\beta})$ is the solution of λ to the equation $H_n(\lambda, \boldsymbol{\beta}) = 0$.

We will need the following lemma on the continuity for the quantities related.

Lemma A.1.1. *Under Assumptions 2 and 3, we have the following results:*

(L1) $E\{g(X, Y, \boldsymbol{\beta})\}$ *and* $E\{g(X, Y, \boldsymbol{\beta})g(X, Y, \boldsymbol{\beta})'\}$ *are twice continuously differentiable with respect to* $\boldsymbol{\beta}$*.*

(L2) There exist p *dimensional compact neighborhoods* C_λ *and* $C_{\boldsymbol{\beta}}$ *around* 0*, in which* $H_0(\lambda, \boldsymbol{\beta}) = E[g(X, Y, \boldsymbol{\beta})/\{1 + \lambda' g(X, Y, \boldsymbol{\beta})\}]$ *is twice continuously differentiable in* $\boldsymbol{\beta} \in C_{\boldsymbol{\beta}}$ *and* $\lambda \in C_\lambda$*, and* $E[g(X, Y, \boldsymbol{\beta})g(X, Y, \boldsymbol{\beta})'/\{1 + \lambda' g(X, Y, \boldsymbol{\beta})\}]$ *is uniformly continuous with respect to* $\boldsymbol{\beta} \in C_{\boldsymbol{\beta}}$ *and* $\lambda \in C_\lambda$*.*

The proof of this lemma is similar to that of Lemma A.1 of Yang and He (2012, pp. 1121). We only need to notice $g(X_i, Y_i, \boldsymbol{\beta}) = (Y_i - \boldsymbol{\beta}' X_i) I_{\{|Y_i - \boldsymbol{\beta}' X_i| < \sigma\}} X_i$ and apply Assumptions 2 and 3. As an illustration, we provide the proof for $E[g(X_i, Y_i, \boldsymbol{\beta})]$ here. Note that

$$\begin{aligned} Eg(X_i, Y_i, \boldsymbol{\beta}) &= E_X \int (y - \boldsymbol{\beta}' X) I_{\{|y - \boldsymbol{\beta}' X| < \sigma\}} X f_X(y) dy \\ &= E_X \int_{\boldsymbol{\beta}' X - \sigma}^{\boldsymbol{\beta}' X + \sigma} (y - \boldsymbol{\beta}' X) X f_X(y) dy, \end{aligned}$$

where E_X stands for the expectation with respect to the distribution G_X of the random variable X. Then the first order derivative of $Eg(X_i, Y_i, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, through simple algebraic calculations, is

$$\frac{\partial Eg(X_i, Y_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = E_X\{\sigma X(f_X(\boldsymbol{\beta}' X + \sigma) - f_X(\boldsymbol{\beta}' X - \sigma)) - XX'(F_X(\boldsymbol{\beta}' X + \sigma) - F_X(\boldsymbol{\beta}' X - \sigma))\}.$$

Now by Assumptions 2 and 3, clearly $\frac{\partial Eg(X_i, Y_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ is further differentiable with respect to $\boldsymbol{\beta}$. The remaining parts of this lemma can be proven similarly with details omitted. ‡

We further define $\lambda_0(\boldsymbol{\beta})$ to be the solution of λ to the equation $H(\lambda, \boldsymbol{\beta}) \equiv E\{g(X_i, Y_i, \boldsymbol{\beta})/[1 + \lambda' g(X_i, Y_i, \boldsymbol{\beta})]\} = 0$. By Lemma A.1.1, Assumption 5 and the

implicit function theorem, $\lambda_0(\boldsymbol{\beta})$ uniquely exists in the neighbourhood C_λ of $\mathbf{0} \in \mathbb{R}^p$. By this uniqueness, as $Eg(X, Y, \boldsymbol{\beta}_0) = 0$, we have $\lambda_0(\boldsymbol{\beta}_0) = 0$. Therefore it follows that $R(\lambda_0(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0) = E\{\log(1 + (\lambda_0(\boldsymbol{\beta}_0))' g(X_i, Y_i, \boldsymbol{\beta}_0))\} = 0$. Note that under Assumptions 1–5, $\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta} \in \mathbb{B}} R(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta})$.

To show the consistency of $\hat{\boldsymbol{\beta}}$ to $\boldsymbol{\beta}_0$, we will apply a lemma below that is a special case of Lemma 4.1 of Lu et al. (2007). Here we need to define a uniform metric $\|\cdot\|_{\mathbb{B}}$ for the distance of any continuous function $\lambda : \mathbb{B} \mapsto \mathbb{R}^p$ from $\lambda_0(\cdot)$, that is $\|\lambda(\cdot) - \lambda_0(\cdot)\|_{\mathbb{B}} = \sup_{\boldsymbol{\beta} \in \mathbb{B}} \|\lambda(\boldsymbol{\beta}) - \lambda_0(\boldsymbol{\beta})\|$ with $\|\cdot\|$ standing for the Euclidean norm of $\mathbb{R}^p$.

Lemma A.1.2. *Suppose $\boldsymbol{\beta}_0 \in \mathbb{B}$ (a compact subset of $\mathbb{R}^p$) satisfies $R(\lambda_0(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0) = \inf_{\boldsymbol{\beta} \in \mathbb{B}} R(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta})$, and that the following hold.*

(i) $R_n\left(\hat{\lambda}(\hat{\boldsymbol{\beta}}), \hat{\boldsymbol{\beta}}\right) \leq \inf_{\boldsymbol{\beta} \in \mathbb{B}} R_n\left(\hat{\lambda}(\boldsymbol{\beta}), \boldsymbol{\beta}\right) + o_P(1)$.

(ii) For all $\delta > 0$, there exists $\epsilon(\delta) > 0$ such that

$$\inf_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| > \delta} R\left(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta}\right) \geq R\left(\lambda_0(\boldsymbol{\beta}_0), \boldsymbol{\beta}_0\right) + \epsilon(\delta).$$

(iii) Uniformly for all $\boldsymbol{\beta} \in \mathbb{B}$, $R\left(\lambda(\boldsymbol{\beta}), \boldsymbol{\beta}\right)$ is continuous [with respect to the uniform metric $\|\cdot\|_{\mathbb{B}}$] in $\lambda(\boldsymbol{\beta})$ at $\lambda_0(\boldsymbol{\beta})$.

(iv) $\|\hat{\lambda}(\cdot) - \lambda_0(\cdot)\|_{\mathbb{B}} = o_P(1)$.

(v) For all $\{\delta_n\}$ with $\delta_n = o(1)$,

$$\sup_{\boldsymbol{\beta} \in \mathbb{B}} \sup_{\|\lambda(\boldsymbol{\beta}) - \lambda_0(\boldsymbol{\beta})\|_{\mathbb{B}} \leq \delta_n} \left|R_n\left(\lambda(\boldsymbol{\beta}), \boldsymbol{\beta}\right) - R\left(\lambda(\boldsymbol{\beta}), \boldsymbol{\beta}\right)\right| = o_P(1).$$

Then $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = o_P(1)$.

The proof of this lemma is omitted; see that of Lemma 4.1 of Lu et al. (2007, pp. 186).

The consistency of $\hat{\boldsymbol{\beta}}$ can be proven by checking the conditions in Lemma A.1.2 step by step: As $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ are the minimizers of $R_n(\hat{\lambda}(\boldsymbol{\beta}), \boldsymbol{\beta})$ and $R(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta})$,

respectively, (i) and (ii) hold obviously. By noting Lemma A.1.1, simple algebraic calculations lead to

$$R(\lambda, \boldsymbol{\beta}) = E_X \int_{\boldsymbol{\beta}'X-\sigma}^{\boldsymbol{\beta}'X+\sigma} \log\{1 + \lambda' X(y - \boldsymbol{\beta}'X)\} f_X(y)dy, \tag{A.1.1}$$

$$H(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta}) = E_X \int_{\boldsymbol{\beta}'X-\sigma}^{\boldsymbol{\beta}'X+\sigma} \frac{X(y - \boldsymbol{\beta}'X)}{1 + (\lambda_0(\boldsymbol{\beta}))' X(y - \boldsymbol{\beta}'X)} f_X(y)dy = 0, \tag{A.1.2}$$

and therefore (iii) also holds clearly by the following fact: as $\|\lambda(\cdot) - \lambda_0(\cdot)\|_{\mathbb{B}} \to 0$,

$$\begin{aligned}
&\sup_{\boldsymbol{\beta}\in\mathbb{B}} |R(\lambda(\boldsymbol{\beta}), \boldsymbol{\beta}) - R(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta})| \\
&\quad \le \sup_{\boldsymbol{\beta}\in\mathbb{B}} \left| E_X \int_{\boldsymbol{\beta}'X-\sigma}^{\boldsymbol{\beta}'X+\sigma} [\log\{1 + (\lambda(\boldsymbol{\beta}))' X(y - \boldsymbol{\beta}'X)\} - \log\{1 + (\lambda_0(\boldsymbol{\beta}))' X(y - \boldsymbol{\beta}'X)\}] \, f_X(y)dy \right| \\
&\quad \le \sup_{\boldsymbol{\beta}\in\mathbb{B}} \left| E_X \int_{\boldsymbol{\beta}'X-\sigma}^{\boldsymbol{\beta}'X+\sigma} \left[\frac{(\lambda(\boldsymbol{\beta}) - \lambda_0(\boldsymbol{\beta}))' X(y - \boldsymbol{\beta}'X)}{1 + (\lambda_0(\boldsymbol{\beta}))' X(y - \boldsymbol{\beta}'X)} \right.\right. \\
&\qquad\qquad \left.\left. - \frac{(\lambda(\boldsymbol{\beta}) - \lambda_0(\boldsymbol{\beta}))' XX' (y - \boldsymbol{\beta}'X)^2 (\lambda(\boldsymbol{\beta}) - \lambda_0(\boldsymbol{\beta}))}{[1 + (\lambda_0(\boldsymbol{\beta}) + \xi(\lambda(\boldsymbol{\beta}) - \lambda_0(\boldsymbol{\beta})))' X(y - \boldsymbol{\beta}'X)]^2} \right] f_X(y)dy \right| \\
&\quad \le \|\lambda(\cdot) - \lambda_0(\cdot)\|_{\mathbb{B}}^2 \sup_{\boldsymbol{\beta}\in\mathbb{B}} \left| E_X \int_{\boldsymbol{\beta}'X-\sigma}^{\boldsymbol{\beta}'X+\sigma} \left[\frac{\|XX'\|(y - \boldsymbol{\beta}'X)^2}{[1 + (\lambda_0(\boldsymbol{\beta}))' X(y - \boldsymbol{\beta}'X)]^2} \right] f_X(y)dy \right| \to 0,
\end{aligned} \tag{A.1.3}$$

where $|\xi| < 1$, the last inequality follows from equality of (A.1.2), and the last limit from the compactness of $\mathbb{B}$ together with the continuity of the integration part as a function of $\boldsymbol{\beta}$ on the RHS of the last inequality in (A.1.3). (iv) follows from a standard argument of the Z-estimator $\hat{\lambda}(\boldsymbol{\beta})$, which is the solution to $H_n(\lambda, \boldsymbol{\beta}) = 0$, uniformly converging to $\lambda_0(\boldsymbol{\beta})$, which is the solution to $H(\lambda, \boldsymbol{\beta}) = 0$, in Chapter 5.1 of Van der Vaart (1998); see also the argument on uniform convergence in the second paragraph on Yang and He (2012, pp. 1124). For (v), letting $\delta_n = o(1)$ and $\|\lambda - \lambda_0\|_{\mathbb{B}} \le \delta_n$, we notice that

$$\begin{aligned}
&R_n(\lambda(\boldsymbol{\beta}), \boldsymbol{\beta}) - R(\lambda(\boldsymbol{\beta}), \boldsymbol{\beta}) \\
&\quad = \{R_n(\lambda(\boldsymbol{\beta}), \boldsymbol{\beta}) - R_n(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta})\} + \{R_n(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta}) - R(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta})\} \\
&\qquad + \{R(\lambda_0(\boldsymbol{\beta}), \boldsymbol{\beta}) - R(\lambda(\boldsymbol{\beta}), \boldsymbol{\beta})\} \\
&\quad = I + II + III,
\end{aligned}$$

where by (A.1.3) III tends to 0, uniformly for $\boldsymbol{\beta} \in \mathbb{B}$ and with λ satisfying $\|\lambda - \lambda_0\|_{\mathbb{B}} \leq \delta_n$. That I tends to 0, uniformly for $\boldsymbol{\beta} \in \mathbb{B}$ and λ with $\|\lambda - \lambda_0\|_{\mathbb{B}} \leq \delta_n$, can be proven in the same way as for III, because in fact $E[I] = III$; II can also be proven easily to tend to zero.

Proof of theorem 2.3.2

Based on the consistency in Theorem 2.3.1, Theorem 2.3.2 can be proven similarly to Theorem 3.2 of Yang and He (2012) by noticing the difference of mode regression in this paper from quantile regression in Yang and He (2012). First, under Assumptions 2–4, it is easy to show as done in Lemma A.5 of Yang and He (2012) that

(C1)$\|\sum_{i=1}^{n}[g(X_i, Y_i, \boldsymbol{\beta}) - Eg(X_i, Y_i, \boldsymbol{\beta})]\| = O_p(n^{1/2})$, uniformly in $\boldsymbol{\beta}$ in a $o(1)$-neighborhood of $\boldsymbol{\beta}_0$.

(C2) $\|\sum_{i=1}^{n}[g(X_i, Y_i, \boldsymbol{\beta})g(X_i, Y_i, \boldsymbol{\beta})' - Eg(X_i, Y_i, \boldsymbol{\beta})g(X_i, Y_i, \boldsymbol{\beta})']\| = o_p(n)$, uniformly in $\boldsymbol{\beta}$ in a $o(1)$-neighborhood of $\boldsymbol{\beta}_0$.

(C3) $\|\sum_{i=1}^{n}[g(X_i, Y_i, \boldsymbol{\beta}) - Eg(X_i, Y_i, \boldsymbol{\beta}) - g(X_i, Y_i, \boldsymbol{\beta}_0) + Eg(X_i, Y_i, \boldsymbol{\beta}_0)]\| = o_p(n^{-1/2})$, uniformly in $\boldsymbol{\beta}$ for $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O_p(n^{-1/2})$.

These (C1)-(C3) together with Assumptions 1–5 ensure (2.3.17) holds true (c.f., Lemma 6 of Molanes Lopez et al. (2009)).

Further, maximizing the main terms on the RHS of (2.3.17) with respect to $\boldsymbol{\beta}$, we have

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = n^{-1/2}(V_{12}' V_{11}^{-1} V_{12})^{-1} V_{12}' V_{11}^{-1} W_n + o_P(n^{-1/2}), \tag{A.1.4}$$

where $\hat{\boldsymbol{\beta}}$ is the maximum empirical likelihood estimator of $\boldsymbol{\beta}_0$.

Then it follows from (2.3.17) and (A.1.4) that

$$\begin{aligned}
\pi(\boldsymbol{\beta}|data) &= \pi(\boldsymbol{\beta})\,\mathfrak{R}(\boldsymbol{\beta}) \\
&= \pi(\boldsymbol{\beta}) \times \exp\Big\{-\frac{n}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'V_{12}'V_{11}^{-1}V_{12}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + n^{1/2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'V_{12}'V_{11}^{-1}W_n \\
&\qquad -\frac{1}{2}W_n'V_{11}^{-1}\,W_n + o_P(1)\Big\} \\
&= \pi(\boldsymbol{\beta}) \times \exp\Big\{-\frac{n}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'V_{12}'V_{11}^{-1}V_{12}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) + n(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'V_{12}'V_{11}^{-1}V_{12}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \\
&\qquad -\frac{1}{2}W_n'V_{11}^{-1}\,W_n + o_P(1)\Big\} \\
&= \pi(\boldsymbol{\beta}) \times \exp\Big\{-\frac{n}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'V_{12}'V_{11}^{-1}V_{12}(\boldsymbol{\beta}-2\hat{\boldsymbol{\beta}}+\boldsymbol{\beta}_0) - \frac{1}{2}W_n'V_{11}^{-1}\,W_n + o_P(1)\Big\} \\
&= \pi(\boldsymbol{\beta})\exp\{-\frac{n}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'I_n(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + Q_n\}, \qquad \text{(A.1.5)}
\end{aligned}$$

where, by (A.1.4),

$$\begin{aligned}
Q_n &= -\frac{n}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'V_{12}'V_{11}^{-1}V_{12}(\boldsymbol{\beta}-2\hat{\boldsymbol{\beta}}+\boldsymbol{\beta}_0) + \frac{n}{2}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})'V_{12}'V_{11}^{-1}V_{12}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \\
&\qquad -\frac{1}{2}W_n'V_{11}^{-1}\,W_n + o_P(1) \\
&= \frac{n}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'V_{12}'V_{11}^{-1}V_{12}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) - \frac{1}{2}W_n'V_{11}^{-1}\,W_n + o_P(1) \\
&= \frac{n}{2}(n^{-1/2}(V_{12}'V_{11}^{-1}V_{12})^{-1}V_{12}'V_{11}^{-1}W_n + o_P(n^{-1/2}))'V_{12}'V_{11}^{-1}V_{12} \\
&\qquad \times (n^{-1/2}(V_{12}'V_{11}^{-1}V_{12})^{-1}V_{12}'V_{11}^{-1}W_n + o_P(n^{-1/2})) - \frac{1}{2}W_n'V_{11}^{-1}\,W_n + o_P(1) \\
&= \frac{1}{2}W_n'V_{11}^{-1}\,W_n + o_P(1) - \frac{1}{2}W_n'V_{11}^{-1}\,W_n + o_P(1) = o_P(1). \qquad \text{(A.1.6)}
\end{aligned}$$

Therefore (2.3.18) follows from (A.1.5) and (A.1.6) together with $\log(\pi(\boldsymbol{\beta})) = \log(\pi(\boldsymbol{\beta}_0)) + O(n^{-1/2})$ for $\boldsymbol{\beta}-\boldsymbol{\beta}_0 = O(n^{-1/2})$ owing to Assumption 6.

The remaining part of Theorem 2.3.2 can be proven, by using Assumption 6, as shown in the corresponding proof of Theorem 3.2 of Lu et al. (2007, pp. 186). The details are therefore omitted.

A.2 Proofs of Main Results: Chapter 3

Derivation of the Fisher Information

Note that, given a sample of n independent observations the log-likelihood sample

in (3.2.7) can also be obtained via

$$l(\underset{\sim}{b}, \phi) = \sum_{i=1}^{n} l(\mu_i, \phi), \tag{A.2.7}$$

where, $l(\mu_i, \phi) = (1+\phi)\{log(\phi) - log(\mu_i)\} - log(\Gamma(\phi)) + \phi log(y_i) - \frac{\phi}{\mu_i} y_i$.

Hence it follows that

$$\frac{\partial l(\underset{\sim}{b}, \phi)}{\partial \underset{\sim}{b}_k} = \sum_{i=1}^{n} \frac{\partial l(\mu_i, \phi)}{\partial \mu_i} \frac{d\mu_i}{d\eta} \frac{\partial \eta_i}{\partial \underset{\sim}{b}_k}, \tag{A.2.8}$$

where,

$$\begin{aligned} \tfrac{\partial l(\mu_i,\phi)}{\partial \mu_i} &= -\tfrac{(1+\phi)}{\mu_i} + \tfrac{\phi y_i}{\mu_i^2} \\ \tfrac{\partial \eta_i}{\partial \underset{\sim}{b}_k} &= x_{ik}. \end{aligned} \tag{A.2.9}$$

Thus the score function for $\underset{\sim}{b}$ is given by,

$$\frac{\partial l(\underset{\sim}{b}, \phi)}{\partial \underset{\sim}{b}_k} = \sum_{i=1}^{n} \left(-\frac{(1+\phi)}{\mu_i} + \frac{\phi y_i}{\mu_i^2} \right) \frac{\partial \mu_i}{\partial \eta} x_{ik}.$$

Similarly it can be shown that the score function for ϕ is given by

$$\frac{\partial l(\underset{\sim}{b}, \phi)}{\partial \phi} = \sum_{i=1}^{n} \frac{1}{\phi} + log(\phi) + 1 - log(\mu_i) - \frac{\psi(\phi)}{\Gamma(\phi)} + log(y_i) - \frac{y_i}{\mu_i}, \tag{A.2.10}$$

where, $\psi(\phi)$ is the digamma function.

Hence we arrive at the matrix expression (S_β, S_ϕ).

From (A.2.8), the second derivative of $l(\underset{\sim}{b}, \phi)$ with respect to $\underset{\sim}{b}s$ is given by

$$\begin{aligned} \tfrac{\partial^2 l(\underset{\sim}{b},\phi)}{\partial \underset{\sim}{b}_k \underset{\sim}{b}_l} &= \textstyle\sum_{i=1}^{n} \tfrac{\partial}{\partial \mu_i} \left(\tfrac{\partial l(\mu_i,\phi)}{\partial \mu_i} \tfrac{d\mu_i}{d\eta} \right) \tfrac{d\mu_i}{d\eta} \tfrac{\partial \eta_i}{\partial \underset{\sim}{b}_k} x_{ik} x_{il} \\ &= \textstyle\sum_{i=1}^{n} \left(\tfrac{\partial^2 l(\mu_i,\phi)}{\partial \mu_i^2} \tfrac{d\mu_i}{d\eta} + \tfrac{\partial l(\mu_i,\phi)}{\partial \mu_i} \tfrac{\partial}{\partial \mu_i} \tfrac{d\mu_i}{d\eta} \right) \tfrac{d\mu_i}{d\eta} x_{ik} x_{il}. \end{aligned} \tag{A.2.11}$$

Since $E\left(\frac{\partial l(\mu_i,\phi)}{\partial \mu_i}\right) = 0$, then

$$E\left(\frac{\partial^2 l(\underset{\sim}{b}, \phi)}{\partial \underset{\sim}{b}_k \underset{\sim}{b}_l} \right) = E\left(\frac{\partial^2 l(\mu_i, \phi)}{\partial \mu_i^2} \right) \left(\frac{d\mu_i}{d\eta} \right)^2 x_{ik} x_{il}. \tag{A.2.12}$$

From (A.2.9)-1,

$$\frac{\partial^2 l(\mu_i, \phi)}{\partial \mu_i^2} = \frac{(1+\phi)}{\mu_i^2} - 2\frac{\phi y_i}{\mu_i^3},$$

then

$$E\left(\frac{\partial^2 l(\underset{\sim}{b}, \phi)}{\partial \underset{\sim}{b}_k \underset{\sim}{b}_l} \right) = -E\left(\left(\frac{(1+\phi)}{\mu_i^2} - 2\frac{\phi y_i}{\mu_i^3} \right) \left(\frac{d\mu_i}{d\eta} \right)^2 x_{ik} x_{il} \right) = -X^T W X.$$

Similarly, from (A.2.8), the second derivative of $l(\underset{\sim}{b}, \phi)$ with respect to $\underset{\sim}{b}$ and ϕ can be written as

$$E\left(\frac{\partial^2 l(\underset{\sim}{b},\phi)}{\partial \underset{\sim}{b}\phi}\right) = -E\left(\left(-\frac{1}{\mu_i}+\frac{y_i}{\mu_i^2}\right)\left(\frac{d\mu_i}{d\eta}\right)x_{ik}\right) = -X^T T w_{b\phi}.$$

Finally, from (A.2.10),

$$E\left(\frac{\partial^2 l(\underset{\sim}{b},\phi)}{\partial \phi^2}\right) = -E\left(-\frac{1}{\phi^2}+\frac{1}{\phi}-\frac{\psi'(\phi)\Gamma(\phi)-(\psi'(\phi))^2}{(\Gamma(\phi))^2}\right) = -tr(D).$$

The Fisher information matrix is obtained from combining the results above.

A.3 Proofs of Main Results: Chapter 4

Proof of theorem 4.2.1

Proof. To establish consistency we use the results of Blevins and Khan (2013), who applied the standard consistency theorem of Newey and McFadden (1994) (Theorem 2.1). The proof is similar to those in Manski (1985) and Horowitz (1992).

Let $S_\tau(\boldsymbol{\beta}(\tau)) = [(2Pr(y=1|\boldsymbol{x}_i)-1)-(1-2\tau)]\, I(\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0)$ be the population score function. Under Assumptions 4 and 5, for any $0 < \tau < 1$, $S_\tau(\boldsymbol{\beta}(\tau)) \leq S_\tau(\boldsymbol{\beta}_0(\tau))$ with equality only if $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta}_0(\tau)$ (Manski (1985)'s Lemma 3 and Corollary 2).

As in Blevins and Khan (2013) the observations are iid by Assumption 1, compactness of $\mathbb{B}$ is established by Assumption 3 and the objective function is a sample average of bounded functions that are continuous in the parameters. Continuity of the objective function follows from Assumption 5.

To establish consistency it is necessary to show that as $n \to \infty$ the stochastic objective function $S_\tau(\boldsymbol{\beta}(\tau))$ converges in probability to a limit function $S_\tau(\boldsymbol{\beta}_0(\tau))$. Since $\widehat{\boldsymbol{\beta}}(\tau)$ maximises $S_\tau(\boldsymbol{\beta}(\tau))$ by definition it follows that $\widehat{\beta}(\tau) - \beta_0(\tau) \xrightarrow{p} 0$ (Amemiya (1985), Theorem 4.2.1).

Blevins and Khan (2013) proved that under the above assumptions $S_\tau(\boldsymbol{\beta}(\tau)) \xrightarrow{p} S_\tau(\boldsymbol{\beta}_0(\tau))$ by showing that, under the assumption $h_n \to 0$ the component of the limiting objective function that depends on $\beta(\tau)$ is

$$E\left[[1-2(Pr(y=1|\boldsymbol{x}_i))](I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau)\geq 0\} - I\{\boldsymbol{x}_i'\boldsymbol{\beta}_0(\tau)\geq 0\})\right],$$

which is clearly 0 for $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta}_0(\tau)$.

In a similar manner, under Assumption 6, the component of the limiting objective function that depends on $\beta(\tau)$ in this case is

$$E\left[[1 - 2(Pr(y = 1|\boldsymbol{x}_i)) + (1 - 2\tau)](I\{\boldsymbol{x}_i'\boldsymbol{\beta}(\tau) \geq 0\} - I\{\boldsymbol{x}_i'\boldsymbol{\beta}_0(\tau) \geq 0\})\right],$$

which is also clearly 0 for $\boldsymbol{\beta}(\tau) = \boldsymbol{\beta}_0(\tau)$. By the strict monotonicity of $K(\cdot)$ and Assumptions 2, 4 and 5, it follows that this component is also strictly positive if $\boldsymbol{\beta}(\tau) \neq \boldsymbol{\beta}_0(\tau)$ for all $0 < \tau < 1$. Therefore it is also minimised at $\boldsymbol{\beta}_0(\tau)$. Moreover, let $S^*_{n,\tau}$ denote the objective function in (4.2.8). Under Assumptions 3 and 7 by Lemma 4 of Horowitz (1992) $|S_{n,\tau} - S^*_{n,\tau}| \xrightarrow{p} 0$ a.s. uniformly. Thus, consistency is established. □