Bayesian Parameter Estimation and Variable Selection for Quantile Regression

A thesis submitted for the degree of
Doctor of Philosophy

by

Craig Reed

Department of Mathematics

School of Information Systems,
Computing and Mathematics

July 2011

Abstract

The principal goal of this work is to provide efficient algorithms for implementing the Bayesian approach to quantile regression. There are two major obstacles to overcome in order to achieve this. Firstly, it is necessary to specify a suitable likelihood given that the frequentist approach generally avoids such specifications. Secondly, sampling methods are usually required as analytical expressions for posterior summaries are generally unavailable in closed form regardless of the prior used.

The asymmetric Laplace (AL) likelihood is a popular choice and has a direct link to the frequentist procedure of minimising a weighted absolute value loss function that is known to yield the conditional quantile estimates. For any given prior, the Metropolis Hastings algorithm is always available to sample the posterior distribution. However, it requires the specification of a suitable proposal density, limiting it's potential to be used more widely in applications.

It is shown that the Bayesian quantile regression model with the AL likelihood can be converted into a normal regression model conditional on latent parameters. This makes it possible to use a Gibbs sampler on the augmented parameter space and thus avoids the need to choose proposal densities. Using this approach of introducing latent variables allows more complex Bayesian quantile regression models to be treated in much the same way. This is illustrated with examples varying from using robust priors and non parametric regression using splines to allowing model uncertainty in parameter estimation. This work is applied to comparing various measures of smoking and which measure is most suited to predicting low birthweight infants. This thesis also offers a short tutorial on the `R` functions that are used to produce the analysis.

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Craig Reed*)

Acknowledgements

I would like to thank Professor David B. Dunson for his suggestions regarding Bayesian variable selection for quantile regression, Dr Paul Thompson for giving me the `R` code to implement the Metropolis Hastings algorithm for the splines example, Professor Janet Peacock for providing the birthweight data and my supervisors Dr Keming Yu and Dr Veronica Vinciotti for their supervision of this work. My thanks also go to all fellow students and staff who made my time at Brunel a thoroughly enjoyable experience. Finally, my thanks go to my parents and friends who provided unwavering support throughout my studies.

This project was funded by an EPSRC doctoral grant.

Table of Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

1.1 Why use quantile regression?

Since the introduction of quantile regression in a paper by Koenker and Bassett (1978), there has been much interest in the field. Quantile regression is used when an estimate of the various quantiles (such as the median) of a conditional distribution is desired. It can be seen as a natural analogue in regression analysis to the practice of using different measures of central tendency and statistical dispersion to obtain a more comprehensive and robust analysis (Koenker, 2005).

To get an idea of the usefulness of quantile regression, note the identity linking the conditional quantiles to the conditional mean:

$$E(y|x) = \int_0^1 Q_\tau(y|x)d\tau, \tag{1.1}$$

where $E(y|x)$ denotes the conditional expectation of y given x and $Q_\tau(y|x)$ denotes the conditional τth quantile of y given x. In essence, this result implies that traditional mean regression is a summary of all possible quantile regressions. Hence, a simple mean regression analysis can be insufficient to describe the complete relationship between y and x. This is demonstrated empirically by Min and Kim (2004), who simulate data based on a wide-class of non Gaussian error distributions. They conclude that simple mean regression cannot satisfactorily capture the key properties of the data and that even the conditional mean estimate can be misleading.

The robustness property of quantile regression is also important. It is widely

known that the mean is not a robust estimate when the underlying distribution is asymmetric or has non-negligible probabilities of extreme outcomes (the distribution has long tails). In such cases, the median (central quantile) offers a more robust estimate of the centre of the distribution. Situations like these are commonly encountered in real datasets from a number of disciplines such as social sciences, economics, medicine, public health, financial return, environment and engineering. Examples are presented in the next section.

This thesis focuses solely on the problem of parameter estimation in Bayesian quantile regression models, firstly assuming the regression model is fixed and then later relaxing this assumption. In addition, this thesis focuses on quantile regression models specified linearly in terms of the regression parameters. However, the ideas presented in the next chapter can be extended to handle nonlinear quantile regression under a Bayesian framework. For a discussion about using these models for prediction and other extensions, see Chapter 5.

1.2 Examples of Cases where Quantile Regression is Useful

Quantile regression enjoys some wide ranging applications. Here are some of them.

- Many asymmetric and long-tailed distributions have been used to model the innovation in autoregressive conditional heteroscedasticity (ARCH) models in finance. Specifically, the conditional autoregressive value at risk (CAViaR) model introduced by Engel and Manganelli (2004) is a very popular time series model for estimating the value at risk in finance.

- In ecology, there exist complex interactions between different factors affecting organisms that cannot all be measured and accounted for in statistical models. This leads to data which often exhibit heteroscedasticity and as such, quantile regression can give a more complete picture about the underlying data generating mechanism (Cade and Noon, 2003).

- In the study of maternal and child health and occupational and environmental risk factors, Abrevaya (2001) investigates the impact of various demographic characteristics and maternal behaviour on the birthweight of infants born in the U.S. Low birthweight is known to be associated with a wide range of subsequent health problems and developmental markers.

- Based on a panel survey of the performance of Dutch school children, Levin (2001) found some evidence that for those individuals within the lower portion of the achievement distribution, there is a larger benefit of being placed in classes with individuals of similar ability. This benefit decreases monotonically as the quantile of interest is increased.

- Chamberlain (1994) infers that for manufacturing workers, the union wage premium, which is at 28 percent at the first decile, declines continuously to 0.3 percent at the upper decile. The author suggests that the location shift model estimate (least squares estimate) which is 15.8 percent, gives a misleading impression of the union effect. In fact, this mean union premium of 15.8 percent is captured primarily by the lower tail of the conditional distribution.

These examples demonstrate the fact that quantile regression can be an important part of any statistician's toolbox. For more details and examples, see Yu et al. (2003).

1.3 The Frequentist Approach

The linear parametric model specifies the conditional quantiles as

$$Q_\tau(y_i|\mathbf{x_i}) = \mathbf{x_i}^T\boldsymbol{\beta}(\tau),\ i = 1, \ldots, n, \tag{1.2}$$

where $\mathbf{x_i}$ denotes the ith column of the $n \times (p+1)$ design matrix $\mathbf{X}$ made up of p predictors and the intercept and $\boldsymbol{\beta}(\tau)$ denotes the $(p+1) \times 1$ vector of associated regression parameters for a fixed value of τ.

In classical (frequentist) quantile regression estimation, the aim is to find an estimator $\hat{\beta}(\tau)$ of $\beta(\tau)$. This is often done without relying on a specification of the form of the residual distribution such as the assumption that the residuals are normally distributed with mean 0 and variance σ^2. Analysis may focus on one value of τ, say $\tau = 0.5$ for the conditional median, or a set of values for τ. If just one value of τ is of interest, Koenker and Bassett (1978) show that minimising the loss function given by

$$\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x_i}^T \boldsymbol{\beta}), \tag{1.3}$$

where

$$\rho_\tau(u) := \begin{cases} \tau u & \text{if} \quad u \geq 0, \\ (1-\tau)|u| & \text{if} \quad u < 0. \end{cases} \tag{1.4}$$

leads to the τth regression quantile. In the case of multiple quantile regressions, the procedure of minimising (1.3) could be repeated with different values of τ. More generally, the entire path of $\beta(\tau)$ could be modelled through the *quantile regression process* in which τ becomes a continuous variable in $(0, 1)$. Koenker and Bassett (1978) show that the problem of minimising (1.3) can be converted into a linear program and give details on how to solve it efficiently for any or all $\tau \in (0, 1)$. This procedure now comes as standard in the `quantreg` package (Koenker, 2009) for R (R Development Core Team, 2010).

Without specifying the form of the residual distribution, frequentist inference for quantile regression focuses on asymptotic theory (see Koenker and Bassett (1978)). In particular, for the linear location shift model $y_i = \mathbf{x_i}^T \boldsymbol{\beta} + \epsilon_i$, where ϵ_i are i.i.d from a density with distribution function F_ϵ, Koenker and Bassett (1978) show that the quantity $\sqrt{n}(\hat{\boldsymbol{\beta}}(\tau) - \boldsymbol{\beta}(\tau))$ converges in distribution to a normal distribution with mean 0 and variance given by $\tau(1-\tau)\boldsymbol{\Omega}^{-1}/s^2(\tau)$, where $\boldsymbol{\Omega} = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{x_i}\, \mathbf{x_i}^T$ and $s(t)$ is the *sparsity function*, the derivative of the quantile function F_ϵ^{-1}. The dependence of the asymptotic covariance matrix on the sparsity function makes this approach unreliable (Billias et al., 2000) and it is very sensitive to the assumption of i.i.d errors (Chen and Wei, 2005).

An alternative approach, suggested by Koenker (1994) is to make use of the theory of rank tests. The advantage of this approach is twofold. Firstly, it avoids the need to estimate the sparsity function and secondly, it is more robust to the model assumptions (Chen and Wei, 2005). The origins of this procedure is related to testing the hypothesis $\boldsymbol{\beta}_2 = \boldsymbol{\nu}$ in the regression model $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}$, where $\boldsymbol{\nu}$ is a pre-specified vector. The idea is to calculate the vector of regression rank score functions $\hat{\mathbf{a}}(t, \boldsymbol{\nu})$ by solving the linear programming problem

$$\max_{\mathbf{a}}\{(\mathbf{y} - \mathbf{X}_2\boldsymbol{\nu})^T\mathbf{a} | \mathbf{X}_1^T\mathbf{a} = (1-t)\mathbf{X}_1^T\mathbf{1_n},\ \mathbf{a} \in [0,1]^n\}, \tag{1.5}$$

where $\mathbf{1_n}$ denotes an $n \times 1$ vector of ones. Then, the vector of τth quantile scores $\hat{\mathbf{b}}_\tau(\boldsymbol{\nu})$ can be defined as

$$\hat{\mathbf{b}}_\tau(\boldsymbol{\nu}) := \hat{\mathbf{a}}(\tau, \boldsymbol{\nu}) - (1-\tau)\mathbf{1}_n. \tag{1.6}$$

Under the null hypothesis $\boldsymbol{\beta}_2 = \boldsymbol{\nu}$, it can be shown that the test statistic

$$T_n(\boldsymbol{\nu}) := \frac{n^{-1/2}\mathbf{X}_2^T\hat{\mathbf{b}}_\tau(\nu)\boldsymbol{\Theta}^{-1/2}}{\sqrt{\tau(1-\tau)}} \tag{1.7}$$

converges in distribution to a standard normal, where

$$\boldsymbol{\Theta} = n^{-1}\mathbf{X}_2^T(\mathbf{I} - \mathbf{X}_1(\mathbf{X}_1^T\mathbf{X}_1)^{-1}\mathbf{X}_1^T)\mathbf{X}_2. \tag{1.8}$$

This test can be inverted to yield confidence intervals, as explained in Koenker (1994).

The disadvantage of this approach, as pointed out by Chen and Wei (2005), is that the computing complexity is exponential in both n, the number of observations and p, the number of regression parameters. This makes it extremely expensive for medium to large sized datasets. For such datasets, a third option is the bootstrap. There are many versions of this. The package `quantreg` (Koenker, 2009) for `R` (R Development Core Team, 2010) offers 4 methods for bootstrapping. These are the xy-pair method, the method of Parzen et al. (1994), the Markov chain marginal bootstrap (MCMB) of He and Hu (2002) and Kocherginsky et al.

(2005) and a generalised bootstrap of Bose and Chatterjee (2003) and Chamberlain and Imbens (2003). These methods are not recommended for $p < 20$ or $n < 20$ due to stability issues (Chen and Wei, 2005).

1.4 The Bayesian Approach

For Bayesians, the missing specification of the residual distributions poses a problem as there is consequently no likelihood specified and learning about any unknown parameters is not possible. A simple solution to this problem has been suggested by Yu and Moyeed (2001), among others, who employ a "pseudo" likelihood $l(\mathbf{y}|\boldsymbol{\beta})$ given by

$$\tau^n(1-\tau)^n \exp\left\{-\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x_i}^T\boldsymbol{\beta})\right\}. \tag{1.9}$$

This is a "pseudo" likelihood in the sense that it is only used to link the Bayesian approach of estimation to the frequentist approach through the property that maximising the log-likelihood is equivalent to minimising (1.3). It is not based on the belief that it is the true data generating mechanism. The likelihood $l(\mathbf{y}|\boldsymbol{\beta})$ can alternatively be viewed as arising from the model $y_i = \mathbf{x_i}^T\boldsymbol{\beta} + \epsilon_i$, where ϵ_i is i.i.d. from the standard asymmetric Laplace (AL) distribution with skewness parameter τ and density function

$$f_\tau(z) = \tau\,(1-\tau)\,\exp\{-\rho_\tau(z)\}. \tag{1.10}$$

Yu and Moyeed (2001) place an improper prior $\pi(\boldsymbol{\beta}) \propto 1$ on the regression parameters $\boldsymbol{\beta}$. Under the improper prior, the posterior mode also corresponds to the minimisation of (1.3). In this sense, priors on $\boldsymbol{\beta}$ can be used to impose regularisation. For example, setting the prior to be independent double exponential distributions (or AL with $\tau = 0.5$) with a common shape parameter λ results in a posterior mode that corresponds to the L1 norm quantile regression studied by Li and Zhu (2008). Li et al. (2010) extended this idea to obtain Bayesian regularised estimates based on other forms of penalty such as the elastic net.

A further appealing feature of the AL distribution is that it is a member of the tick exponential family introduced by Komunjer (2005). It was illustrated in Komunjer (2005) that for likelihood based inference, using this family of distributions was necessary to achieve consistency of the maximum likelihood estimators.

Under a flat prior, Yu and Moyeed (2001) form the posterior distribution $\pi(\boldsymbol{\beta}|\mathbf{y})$ and show that it is proper. This posterior distribution cannot be sampled directly, so Yu and Moyeed (2001) resort to the Metropolis Hastings (MH) algorithm to provide joint samples from $\pi(\boldsymbol{\beta}|\mathbf{y})$.

Yu and Stander (2007) extend this work to analysing a Tobit quantile regression model, a form of censored model in which $y_i = y_i*$ is observed if $y_i* > 0$ and $y_i = 0$ is observed otherwise. A regression model then relates the unobserved y_i* to the covariates $\mathbf{x_i}$. Geraci and Bottai (2007) use the AL likelihood and combine Markov Chain Monte Carlo (MCMC) with the expectation maximising (EM) algorithm to carry out inference on quantile regression for longitudinal data. Chen and Yu (2009) use the AL likelihood combined with non-parametric regression modelling using piecewise polynomials to implement automatic curve fitting for quantile regression and Thompson et al. (2010) use the same approach but using natural cubic splines.

Tsionas (2003) employs a different approach to sampling from the joint posterior of Yu and Moyeed (2001) by using data augmentation. His approach relies on a representation of the AL distribution as a mixture of skewed normal distributions where the mixing density is exponential. He then implements a Metropolis within Gibbs algorithm to simulate from the augmented joint posterior distribution.

Alternatives to the AL likelihood have been suggested by Dunson and Taylor (2005), who use Jeffreys' (Jeffreys, 1961) substitution likelihood and Lancaster and Jun (2010), who use an approach based on the Bayesian exponentially tilted empirical likelihood introduced by Schennach (2005).

Kottas and Krnjajić (2009) point out that the value of τ not only controls the quantile but also the skewness of the AL distribution resulting in limited flexibility.

In particular, the residual distribution is symmetric when modelling the median. This motivated Kottas and Gelfand (2001) and Kottas and Krnjajić (2009) to consider a more flexible residual distribution constructed using a Dirichlet process prior but still having τth quantile equal to 0. Kottas and Krnjajić (2009) include a general scale mixture of AL densities with skewness τ in their analysis, but conclude that in terms of ability to predict new observations, a general mixture of uniform distributions performs the best.

Despite these concerns, the AL distribution is easy to work with for applied researchers if the key aim is parameter estimation. In particular, as will be shown in the next section, the AL distribution can be represented in terms of the symmetric double exponential distribution. This is well known to have a representation as a scale mixture of normals. By augmenting the data with latent variables, it is possible to implement the Gibbs sampler to sample from the resulting augmented posterior distribution under a normal prior. Gibbs sampling, where possible, has the advantage of being "automatic", in the sense that the researcher does not have to specify a candidate distribution necessary for MH sampling. Perhaps more importantly, this approach is easily extended to allow for more complex models such as random effect models. In addition, since the marginal likelihoods conditional on the latent parameters are available in closed form under a normal prior, it is possible to compute approximate Bayes factors to compare models. More generally, it is possible to incorporate covariate set uncertainty into the analysis. Such analysis would be computationally very expensive for the approach of Kottas and Krnjajić (2009). Finally, it is possible to use Rao-Blackwellisation to approximate the marginal density $\pi(\boldsymbol{\beta}|\mathbf{y})$. This may be useful for obtaining simultaneous credible intervals using the method of Held (2004).

This particular strategy of data augmentation differs from Tsionas (2003) in that the resulting full conditionals are available to sample from directly using standard algorithms. However, at the time of writing the manuscript, it was soon realised that work by Kozumi and Kobayashi (2009) essentially used the same

approach as the one demonstrated in the next chapter. The approach of Kozumi and Kobayashi (2009) differs only in the parameterisation used in the mixture of normals representation and was obtained by using results about a different parameterisation of the AL distribution appearing in Kotz et al. (2001). As will be shown in Chapter 2, there are key differences that make this approach more efficient than that of Kozumi and Kobayashi (2009). Firstly, in the Gibbs sampler developed in Chapter 2, the entire set of latent variables can be sampled efficiently using the algorithm described in Michael et al. (1976). Secondly, when adding a scale parameter as discussed in Chapter 5, it is demonstrated that a Gibbs sampler can be designed that still only requires two blocks, unlike Kozumi and Kobayashi (2009) who implement a three block sampler when considering a scale parameter. Results in Liu et al. (1994) suggest that the Gibbs sampler described in this thesis is likely to be more efficient than that of Kozumi and Kobayashi (2009).

It is nevertheless important to emphasise that this work was done independently and that it was only through an associate referee's observation that anything was known about this new manuscript Kozumi and Kobayashi (2009). As a result, the authors were invited to become joint authors of the manuscript Reed et al. (2010) which is awaiting a small revision. See Chapter 5 for further details.

1.5 Thesis Outline

The outline of this thesis is as follows. In Chapter 2, working with the AL likelihood and using data augmetation, a simple Gibbs sampler is developed to sample the augmented posterior distribution. This is compared to the MH algorithm of Yu and Moyeed (2001). The approach is extended to non parametric Bayesian quantile regression using natural cubic splines and the resulting Gibbs sampler is compared to the MH algorithm of Thompson et al. (2010). In Chapter 3, the method introduced in the previous chapter is used to analyse the dataset obtained from 1,254 women booking for antenatal care at St. George's hospital between August 1982 and March 1984. Gibbs sampling the posterior under a more robust

prior on $\boldsymbol{\beta}$ is considered. Chib's method (Chib, 1995) is used to calculate approximate Bayes factors for comparing competing models. In Chapter 4, the ideas of the previous chapters are extended to deal with model uncertainty and model selection in more detail. Stochastic search variable selection for quantile regression (QR-SSVS) similar in spirit to George and McCulloch (1997) is introduced and applied to a simulated dataset and the Boston Housing data. Finally, Chapter 5 concludes the thesis and offers suggestions of future work.

The Appendix provides details about the `R` functions that have been written to implement the Gibbs sampler described in Chapter 2 with a normal prior and the QR-SSVS algorithm described in Chapter 4. These have been used to obtain all analyses reported in this thesis. A short tutorial is provided on how to use these `R` functions.

Chapter 2

Bayesian Quantile Regression using Data Augmentation

2.1 Introducing the Latent Variables

From the literature review in the previous chapter, it is now clear that relying on a fully parametric model specified using the AL likelihood may be restrictive. Nevertheless, it remains the most straightforward and easily extended approach to obtain Bayesian estimates in quantile regression models. Results from this chapter will allow a regression model with the AL likelihood to be converted into a normal regression model with latent variables.

Firstly, note that the check function (1.4) can equivalently be defined as

$$\rho_\tau(u) := \tfrac{1}{2}|u| + (\tau - \tfrac{1}{2})u. \tag{2.1}$$

Using the definition of the check function (2.1), we can write the AL likelihood $l(\mathbf{y}|\boldsymbol{\beta})$ given in (1.9) as

$$\prod_{i=1}^{n} \exp\{-\tfrac{1}{2}|y_i - \mathbf{x_i}^T\boldsymbol{\beta}|\} \prod_{i=1}^{n} \tau(1-\tau) \exp\{-(\tau - \tfrac{1}{2})(y_i - \mathbf{x_i}^T\boldsymbol{\beta})\}. \tag{2.2}$$

Notice that the first product in (2.2) is proportional to the product of n double exponential densities (or AL densities with $\tau = 0.5$). Well known results from Andrews and Mallows (1974) and West (1987) show that the double exponential distribution admits a representation as a scale mixture of normals. In particular,

$$f_{\tau=0.5}(z) = \frac{1}{4} \exp\left\{-\frac{|z|}{2}\right\} = \int_0^\infty \frac{1}{\sqrt{2\pi v}} \exp\left\{-\frac{z^2}{2v}\right\} \frac{1}{8} \exp\left\{-\frac{v}{8}\right\} \mathrm{d}v. \tag{2.3}$$

It is in fact more convenient to parameterise in terms of $w = v^{-1}$ in (2.3), giving the representation

$$f_{\tau=0.5}(z) = \frac{1}{4}\exp\left\{-\frac{|z|}{2}\right\} = \int_0^\infty \sqrt{\frac{w}{2\pi}}\exp\left\{-\frac{z^2 w}{2}\right\}\frac{1}{8w^2}\exp\left\{-\frac{1}{8w}\right\}\mathrm{d}w. \tag{2.4}$$

This means that a double exponential distribution can be obtained by marginalising over w, where $z|w$ is normal with mean 0 and precision w and w has an inverse Gamma distribution with parameters $(1, \frac{1}{8})$. Likelihood (2.2) can therefore be obtained by marginalising over the entire $n \times 1$ vector of latent parameters $\mathbf{w}$ from the augmented likelihood $l(\mathbf{y}|\boldsymbol{\beta}, \mathbf{w})$ proportional to

$$\prod_{i=1}^{n}\left\{\sqrt{w_i}\exp\{-\tfrac{1}{2}w_i(y_i - \mathbf{x_i}^T\boldsymbol{\beta})^2 - (\tau - \tfrac{1}{2})(y_i - \mathbf{x_i}^T\boldsymbol{\beta})\}\right\}, \tag{2.5}$$

under the prior $\pi(\mathbf{w}) = \prod_{i=1}^{n}\pi(w_i)$, where

$$\pi(w_i) \propto w_i^{-2}\exp(-\tfrac{1}{8}w_i^{-1}). \tag{2.6}$$

The full Bayesian specification is completed by a prior on the unknown regression parameters $\boldsymbol{\beta}$. The multivariate normal prior

$$\pi(\boldsymbol{\beta}) \propto \exp\{-\tfrac{1}{2}(\boldsymbol{\beta} - \mathbf{b_0})^T\mathbf{B_0}(\boldsymbol{\beta} - \mathbf{b_0})\}, \tag{2.7}$$

is semi-conjugate. For now, it is assumed that the prior mean vector $\mathbf{b_0}$ and the prior precision matrix $\mathbf{B_0}$ are fixed although this will be relaxed later. An improper prior is obtained by setting $\mathbf{B_0} = c\mathbf{I}$, and letting $c \to 0$.

The joint posterior distribution $\pi(\boldsymbol{\beta}, \mathbf{w}|\mathbf{y})$ is given by

$$\pi(\boldsymbol{\beta}, \mathbf{w}|\mathbf{y}) \propto l(\mathbf{y}|\boldsymbol{\beta}, \mathbf{w})\pi(\mathbf{w})\pi(\boldsymbol{\beta}). \tag{2.8}$$

Given the result that the marginal posterior distribution $\pi(\boldsymbol{\beta}|\mathbf{y}) = \int \pi(\boldsymbol{\beta}, \mathbf{w}|\mathbf{y})\mathrm{d}\mathbf{w}$ remains proper if an improper prior is used for $\pi(\boldsymbol{\beta})$ (Yu and Moyeed, 2001), this is also the case for the augmented posterior distribution (2.8).

Sampling directly from $\pi(\boldsymbol{\beta}, \mathbf{w}|\mathbf{y})$ and the marginal $\pi(\boldsymbol{\beta}|\mathbf{y})$ remains difficult. However, the conditional posterior distributions $\pi(\boldsymbol{\beta}|\mathbf{w}, \mathbf{y})$ and $\pi(\mathbf{w}|\boldsymbol{\beta}, \mathbf{y})$ can

be sampled easily and efficiently. This motivates the Gibbs sampler to produce approximate samples from $\pi(\boldsymbol{\beta}, \mathbf{w}|\mathbf{y})$ and using the sampled values of $\boldsymbol{\beta}$ as samples from $\pi(\boldsymbol{\beta}|\mathbf{y})$.

Combining (2.5) with (2.7) reveals that $\pi(\boldsymbol{\beta}|\mathbf{w}, \mathbf{y})$ is multivariate normal with precision matrix

$$\mathbf{B_1} = \mathbf{X}^T\mathbf{W}\mathbf{X} + \mathbf{B_0}, \tag{2.9}$$

and mean

$$\mathbf{b_1} = \mathbf{B_1}^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{y} + (\tau - \tfrac{1}{2})\mathbf{X}^T\mathbf{1_n} + \mathbf{B_0}\mathbf{b_0}). \tag{2.10}$$

Here, $\mathbf{W}$ denotes an $n \times n$ diagonal matrix with the weights w_i forming the diagonal and $\mathbf{1_n}$ denotes an $n \times 1$ vector of ones. Note that if $\tau = 0.5$, then the posterior mean $\mathbf{b_1}$ becomes

$$\mathbf{b_1} = \mathbf{B_1}^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{y} + \mathbf{B_0}\mathbf{b_0}). \tag{2.11}$$

If the predictors are centered and $\tau \neq 0.5$, then

$$\mathbf{b_1} = \mathbf{B_1}^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{y} + \mathbf{B_0}\mathbf{b_0} + \boldsymbol{\xi}), \tag{2.12}$$

where $\boldsymbol{\xi}$ is a $(p+1) \times 1$ vector with the first element equal to $n(\tau - \frac{1}{2})$ and the remaining elements equal to 0. Sampling this normal distribution is most efficiently done using a Cholesky decomposition of $\mathbf{B_1}$.

To obtain $\pi(\mathbf{w}|\boldsymbol{\beta}, \mathbf{y})$, first note that w_i is conditionally independent of all remaining elements of $\mathbf{w}$ given $\boldsymbol{\beta}$ and $\mathbf{y}$. For a particular value of i, combining (2.5) with (2.6), the density function is proportional to

$$w_i^{-3/2} \exp\left\{-\frac{w_i(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2} - \frac{1}{8w_i}\right\}. \tag{2.13}$$

This density function can be compared to the kernel of an inverse Gaussian (IG) density function with pdf

$$f(w_i|\lambda_i, \mu_i) \propto w_i^{-3/2} \exp\left\{-\frac{\lambda_i(w_i - \mu_i)^2}{2\mu_i^2 w_i}\right\}, \qquad \mu_i, \lambda_i > 0, \tag{2.14}$$

using the parameterisation of Chhikara and Folks (1989). The parameters of (2.14) are the scale parameter $\lambda_i = \lambda = \frac{1}{4}$ and the location parameter $\mu_i =$

Algorithm 2.1 Gibbs sampler for augmented quantile regression model with initial values for $\boldsymbol{\beta}$. Draws M burn in samples followed by an additional N samples for inference.

Given: Prior mean vector $\mathbf{b_0}$, prior precision matrix $\mathbf{B_0}$ and initial values $\boldsymbol{\beta}^{(0)}$.

for $k = 1$ **to** $M + N$ **do**

- Sample $\mathbf{w}^{(k)}|\boldsymbol{\beta}^{(k-1)}, \mathbf{y}$ by sampling the ith component of $\mathbf{w}$ $(i = 1, \ldots, n)$ from the inverse Gaussian distribution with shape parameter $\frac{1}{4}$ and location $\frac{1}{2}|y_i - \mathbf{x_i}^T\boldsymbol{\beta}^{(k-1)}|^{-1}$.
- Sample $\boldsymbol{\beta}^{(k)}|\mathbf{w}^{(k)}, \mathbf{y}$ from the multivariate normal distribution with precision matrix

$$\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X} + \mathbf{B_0}$$

and mean vector

$$(\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X} + \mathbf{B_0})^{-1}(\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{y} + (\tau - \tfrac{1}{2})\mathbf{X}^T\mathbf{1} + \mathbf{B_0}\mathbf{b_0}),$$

where $\mathbf{W}^{(k)}$ is a diagonal matrix with the elements of $\mathbf{w}^{(k)}$ forming the diagonal.

end for.

$\frac{1}{2}|y_i - \mathbf{x_i}^T\boldsymbol{\beta}|^{-1}$. Consequently, sampling from $\pi(\mathbf{w}|\boldsymbol{\beta}, \mathbf{y})$ requires n samples from the inverse Gaussian distribution with the same scale parameter $\frac{1}{4}$ but different location parameters μ_i. The inverse Gaussian distribution can be sampled efficiently using the algorithm of Michael et al. (1976). Note the difference here between Kozumi and Kobayashi (2009) who obtain a generalised inverse Gaussian density for each of their latent variables, which cannot be sampled as efficiently.

The Gibbs sampler is summarised in Algorithm 2.1. Of course, as this is an ordinary Gibbs sampler, it is possible to rearrange the steps without altering the target distribution to which the Gibbs sampler converges. This leads to Algorithm 2.2. In this case, starting values $\mathbf{w}^{(0)}$ are required. Given that the prior on $\mathbf{w}$ is proper and that it is more difficult to make a sensible first guess at these initial values, they could be drawn at random from the prior.

Algorithm 2.2 Gibbs sampler for augmented quantile regression model with initial values for $\mathbf{w}$. Draws M burn in samples followed by an additional N samples for inference.

Given: Prior mean vector $\mathbf{b_0}$, prior precision matrix $\mathbf{B_0}$ and initial values $\mathbf{w}^{(0)}$.

for $k = 1$ **to** $M + N$ **do**

• Sample $\boldsymbol{\beta}^{(k)}|\mathbf{w}^{(k-1)}, \mathbf{y}$ from the multivariate normal distribution with precision matrix

$$\mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{X} + \mathbf{B_0}$$

and mean vector

$$(\mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{X} + \mathbf{B_0})^{-1}(\mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{y} + (\tau - \tfrac{1}{2})\mathbf{X}^T\mathbf{1} + \mathbf{B_0}\mathbf{b_0}),$$

where $\mathbf{W}^{(k-1)}$ is a diagonal matrix with the elements of $\mathbf{w}^{(k-1)}$ forming the diagonal.

• Sample $\mathbf{w}^{(k)}|\boldsymbol{\beta}^{(k)}, \mathbf{y}$ by sampling the ith component of $\mathbf{w}$ $(i = 1, \ldots, n)$ from the inverse Gaussian distribution with shape parameter $\frac{1}{4}$ and location $\frac{1}{2}|y_i - \mathbf{x_i}^T\boldsymbol{\beta}^{(k)}|^{-1}$.

end for.

2.2 Engel data: Comparing Augmented Posterior Summaries with Frequentist Estimate/ Marginal Posterior Mode

The first application is to assess the accuracy of the point estimates obtained by retaining only the sampled values of $\boldsymbol{\beta}$ from the Gibbs sampler described in the previous section. Whilst the marginal posterior mode in this case corresponds to finding the classical quantile regression estimate, it is in general unreliable to estimate the marginal posterior mode purely from MCMC output. This is because it is difficult to know whether all islands of high probability have been visited by the MCMC algorithm.

To illustrate, consider Engel's data in Koenker and Bassett (1982) and also available in the `quantreg` (Koenker, 2009) package. The dataset consists of 235 observations on the annual household food expenditure in Belgian francs. There is one predictor which is annual household income. Both the intercept and the coefficient of the predictor were assigned improper priors. Inference was based on 10,000 samples following a burn in of 1,000.

$\tau = 0.1$			
	Frequentist Estimate	Posterior Mean	Posterior Median
Intercept	110.142	111.398	111.189
Foodexp	0.402	0.398	0.399
$\tau = 0.25$			
	Frequentist Estimate	Posterior Mean	Posterior Median
Intercept	95.484	94.709	94.827
Foodexp	0.474	0.475	0.475
$\tau = 0.5$			
	Frequentist Estimate	Posterior Mean	Posterior Median
Intercept	81.482	82.625	82.556
Foodexp	0.560	0.559	0.559
$\tau = 0.75$			
	Frequentist Estimate	Posterior Mean	Posterior Median
Intercept	62.397	60.467	60.338
Foodexp	0.644	0.646	0.646
$\tau = 0.9$			
	Frequentist Estimate	Posterior Mean	Posterior Median
Intercept	67.351	66.164	66.141
Foodexp	0.686	0.687	0.687

Table 2.1: Comparison of frequentist estimate (also marginal posterior mode) and posterior mean and median, estimated from the Gibbs sample by retaining only the $\boldsymbol{\beta}$ values. The summary statistics are calculated from 11,000 iterations with the first 1,000 discarded.

As can be seen in Table 2.1, both the estimated posterior mean and median under the augmented model are good approximations to the marginal posterior mode, the median being slightly closer in more cases than the mean. The Rao-Blackwellised estimate of the mean was also calculated but was almost identical to the mean calculated directly from the samples and is therefore not reproduced in the table.

2.3 Stackloss data: Comparison of Gibbs sampler and Metropolis-Hastings algorithm

The second example uses Brownlee's stack loss data (Brownlee, 1960). The data originates from 21 days of operation of a plant for the oxidation of ammonia to nitric acid. The nitric oxides produced are absorbed in a countercurrent absorption tower. The response variable is 10 times the percentage of the ingoing ammonia to the plant that escapes from the absorption column unabsorbed and is a measure of the "efficiency" of a plant. There are 3 covariates in this dataset which are air flow (x_1), which represents the rate of operation of the plant, water temperature (x_2), the temperature of cooling water circulated through coils in the absorption tower and acid concentration (x_3), which is the concentration of the acid circulating, minus 50, times 10.

This section compares the posterior estimates obtained using the Gibbs sampler to those obtained using the univariate random walk Metropolis Hastings (MH) within Gibbs algorithm of Yu and Moyeed (2001). This involves updating each of the 4 parameters including the intercept one by one using an MH step based on a proposed value that is the current value plus random noise whilst conditioning on all remaining parameters. For this comparison, improper priors on all unknown regression parameters are again used. 11,000 samples are drawn from each Markov chain, 1,000 of which discarded as burn in. The analysis uses $\tau = \{0.05, 0.25, 0.5, 0.75, 0.95\}$. Table 2.2 presents the posterior mean together with the 95% highest posterior density (HPD) region in parentheses for each

chain. A word of caution: the HPD intervals are conditional on the choice of likelihood. They are presented here as a way to compare how well both algorithms explore the posterior distribution.

It can be seen from Table 2.2 that both the estimates and the HPD intervals are very similar across all quantiles, indicating that for small dimensional problems, Gibbs and MH perform similarly well at exploring the posterior distribution despite the $n = 21$ additional parameters for the Gibbs sampler to update. The MH algorithm was run using the optimal settings recommended by Yu and Moyeed (2001).

It should be noted that the Gibbs sampler for $\tau = 0.5$ produces samples that have significantly lower autocorrelation than MH (see Figure 2.1). Whilst it is true that any MH algorithm is always likely to have higher autocorrelation than a Gibbs sampler given that all candidate values are accepted in Gibbs sampling, another factor may be that the Gibbs sampler updates the regression parameters in one block and the latent parameters in another block. In contrast, the advantage associated with being able to choose univariate candidate densities for MH by holding the remaining parameters constant is negated by the fact that the predictors are correlated. The correlation between water temperature and acid concentration is about 0.39, between acid concentration and air flow is 0.5 and between water temperature and air flow is 0.78. As a result, any sampler that updates each parameter one by one conditional on the other parameters being held fixed is less likely to be able to make large moves and fully explore the posterior distribution in a reasonable amount of time. On this basis, the Gibbs sampler is more efficient when $\tau = 0.5$.

2.4 Bayesian Quantile Regression with Natural Cubic Splines

Recently, Thompson et al. (2010) used the approach of Yu and Moyeed (2001) to implement non parametric Bayesian quantile regression using natural cubic

$\tau = 0.05$		
	Gibbs	MH
Intercept	-41.099(-89.951, -2.328)	-41.873(-92.301,-3.637)
x_1	0.398(-0.165,0.900)	0.387(-0.181,0.901)
x_2	1.545(-0.176,3.163)	1.519(-0.150,3.149)
x_3	-0.050(-0.648,0.607)	-0.028(-0.603,0.615)
$\tau = 0.25$		
	Gibbs	MH
Intercept	-37.749(-54.170,-21.421)	-37.133(-54.318, -21.008)
x_1	0.654(0.350,0.933)	0.672(0.368,0.970)
x_2	1.013(0.363,1.770)	1.004(0.328, 1.743)
x_3	-0.092(-0.361, 0.147)	-0.106(-0.387, 0.133)
$\tau = 0.5$		
	Gibbs	MH
Intercept	-38.613(-53.419,-23.587)	-38.660(-54.983,-22.687)
x_1	0.839(0.613,1.072)	0.838(0.620, 1.058)
x_2	0.725(0.222,1.352)	0.732(0.186,1.386)
x_3	-0.115(-0.322,0.078)	-0.116(-0.334,0.083)
$\tau = 0.75$		
	Gibbs	MH
Intercept	-48.528(-68.976, -23.097)	-48.872(-68.624,-22.496)
x_1	0.862(0.589,1.131)	0.849(0.562,1.132)
x_2	1.033(0.263,1.810)	1.068(0.261,1.910)
x_3	-0.065(-0.380,0.186)	-0.060(-0.389,0.189)
$\tau = 0.95$		
	Gibbs	MH
Intercept	-41.236(-90.369,45.144)	-43.047(-91.716,41.351)
x_1	0.766(0.191,1.428)	0.786(0.186,1.477)
x_2	1.485(-0.194,2.806)	1.401(-0.224,2.755)
x_3	-0.144(-1.124,0.538)	-0.118(-1.114,0.561))

Table 2.2: Comparison of Gibbs sampler and Metropolis-Hastings(MH). The posterior means were recorded together with the 95% highest posterior density (HPD) intervals (in parentheses).

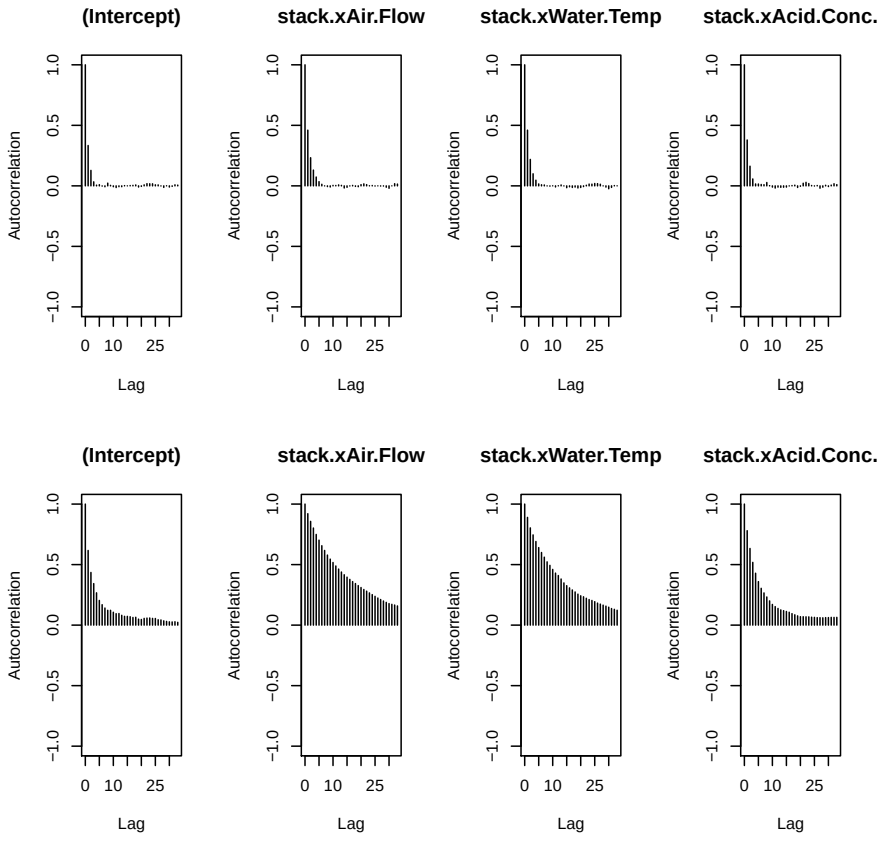

Figure 2.1: Stackloss data: Comparing autocorrelation of Gibbs and MH at $\tau = 0.5$. The top row corresponds to the Gibbs sampler, the bottom row to MH with optimal settings.

splines (NCS). In this section, a Gibbs sampler designed using the same approach as the previous sections is compared to the MH algorithm used by Thompson et al. (2010). Following the authors, artificial data was simulated based on the motorcycle data obtained in an experiment to test crash helmets and discussed in Silverman (1985). This dataset is a classic example of where polynomial regression is inappropriate. The response variable $\mathbf{y}$ of the motorcycle data is a record of the head acceleration, measured in multiples of the acceleration due to gravity g. The explanatory variable $\mathbf{x}$ is the time, measured in milliseconds, after a simulated motorcycle accident. An artificial dataset was formed by simulating 100 observations at 30 evenly spaced time points from a normal distribution with mean equal to the value of the smoothing spline fitted to the motorcycle data at each time point and standard deviation equal to 20.

Using the notation of Green and Silverman (1994), let t_i be the ordered set of knots that are in the range of $\mathbf{x}$ and let $g_i = g(t_i)$ denote the value of the NCS at the knot points t_i. The AL likelihood $l(\mathbf{y}|\mathbf{g})$, where $\mathbf{g}$ is a 30×1 vector with elements g_i, is then proportional to

$$\prod_{i,j} \exp\{-\tfrac{1}{2}|y_{ij} - g_i|\} \prod_{i,j} \exp\{-(\tau - \tfrac{1}{2})(y_{ij} - g_i)\}, \tag{2.15}$$

where i runs from 1 to 30 and j runs from 1 to 100. The model of Thompson et al. (2010) assumes a multivariate normal prior for $\mathbf{g}$,

$$\pi(\mathbf{g}|\lambda) \propto \lambda^{n/2} \exp(-\tfrac{\lambda}{2}\mathbf{g}^T\mathbf{K}\mathbf{g}). \tag{2.16}$$

This choice was motivated by the fact that the log density is proportional to the roughness penalty $\int_a^b g''(x)^2 dx$ as a consequence of theorem 2.1 in Green and Silverman (1994). The matrix $\mathbf{K}$ is a fixed symmetric matrix of rank 28 defined as $\mathbf{Q}\mathbf{R}^{-1}\mathbf{Q}^T$. Defining $h_i = t_{i+1} - t_i$ for $i = 1, \ldots, 29$, the matrix $\mathbf{Q}$ is 30×28 with entries q_{ij}, $i = 1, \ldots, 30$, $j = 2, \ldots, 29$, with $q_{j-1,j} = h_{j-1}^{-1}$, $q_{jj} = -h_{j-1}^{-1} - h_j^{-1}$, $q_{j+1,j} = h_j^{-1}$ for $j = 2, \ldots, 29$ and $q_{ij} = 0$ for $|i - j| \geq 2$. The matrix $\mathbf{R}$ is a symmetric 28×28 matrix and has elements r_{ij}, $i = 2, \ldots, 29$, $j = 2, \ldots, 29$, with

$r_{ii} = \frac{1}{3}(h_{i-1} + h_i)$, $i = 2, \ldots, 29$, $r_{i,i+1} = r_{i+1,i} = \frac{1}{6}h_i$, $i = 2, \ldots, 28$ and $r_{ij} = 0$ for $|i - j| \geq 2$. The parameter λ denotes the smoothing parameter that acts as a compromise between smoothness and fidelity to the data. Finally, the model of Thompson et al. (2010) treats λ as unknown and gives it a Gamma hyperprior with parameters c_0 and d_0,

$$\pi(\lambda) \propto \lambda^{c_0 - 1} \exp(-d_0 \lambda). \tag{2.17}$$

Instead of using the random walk Metropolis within Gibbs as in Yu and Moyeed (2001), Thompson et al. (2010) opt for an MH algorithm that updates the entire vector $\mathbf{g}$ in one block followed by an update for λ. The main disadvantage with this approach is the "curse of dimensionality" - to find a suitable candidate density that performs well and gives good mixing is extremely hard in high dimensional problems.

In order to implement a Gibbs sampler, an additional 3,000 latent variables in a 30×100 matrix $\mathbf{W}$ need to be introduced into the model. The augmented likelihood $l(\mathbf{y}|\mathbf{g}, \mathbf{W}, \mathbf{y})$ is proportional to

$$\prod_{i,j} \left\{ \sqrt{w_{ij}} \exp\{-\tfrac{1}{2} w_{ij} (y_{ij} - g_i)^2\} \right\} \prod_{i,j} \exp\{-(\tau - \tfrac{1}{2})(y_{ij} - g_i)\}. \tag{2.18}$$

The final component of this augmented model is the independent and identically distributed inverse Gamma priors on each w_{ij} with parameters $(1, \frac{1}{8})$. Just as in section 2, the likelihood of Thompson et al. (2010) can be recovered by marginalising over $\mathbf{W}$.

Routine calculations reveal that the conditional posterior distribution of $\mathbf{g}$ is multivariate normal with precision matrix

$$\mathbf{\Omega} + \lambda \mathbf{K} \tag{2.19}$$

and mean vector

$$(\mathbf{\Omega} + \lambda \mathbf{K})^{-1} \mathbf{u}, \tag{2.20}$$

where $\mathbf{\Omega}$ denotes a 30×30 diagonal matrix with $\mathbf{\Omega}_{i,i} = \sum_{j=1}^{100} w_{ij}$ and $\mathbf{u}$ is a 30×1 vector with elements $u_i = \sum_{j=1}^{100} w_{ij} y_{ij} + 100(\tau - \frac{1}{2})$. The full conditional of each

Algorithm 2.3 Gibbs sampler for quantile regression with natural cubic splines. Draws M burn in samples and N samples for inference.

Given: Precision matrix $\mathbf{K}$ and initial values $\mathbf{g}^{(0)}$.

for $k = 1$ **to** $M + N$ **do**

• Sample each component $w_{ij}^{(k)}|\mathbf{g}^{(k-1)}, \mathbf{y}$ by sampling from the inverse Gaussian distribution with shape parameter $\frac{1}{4}$ and location $\frac{1}{2}|y_{ij} - g_i^{(k-1)}|^{-1}$.

• Sample $\lambda^{(k)}|\mathbf{g}^{(k-1)}, \mathbf{y}$ from the gamma distribution with location $14 + c_0$ and scale $\frac{1}{2}\{\mathbf{g}^{(k-1)}\}^T\mathbf{K}\mathbf{g}^{(k-1)}$.

• Sample $\mathbf{g}^{(k)}|\mathbf{W}^{(k)}, \lambda^{(k)}, \mathbf{y}$ from the multivariate normal distribution with precision matrix

$$\mathbf{\Omega}^{(k)} + \lambda^{(k)}\mathbf{K}$$

and mean vector

$$(\mathbf{\Omega}^{(k)} + \lambda^{(k)}\mathbf{K})^{-1}\mathbf{u}^{(k)},$$

where $\mathbf{\Omega}^{(g)}$ is a 30×30 diagonal matrix with $\mathbf{\Omega}_{i,i}^{(g)} = \sum_{j=1}^{100} w_{ij}^{(g)}$ and $\mathbf{u}^{(k)}$ is a 30×1 vector with elements $u_i^{(k)} = \sum_{j=1}^{100} w_{ij}^{(k)} y_{ij} + 100(\tau - \frac{1}{2})$.

end for.

w_{ij} is inverse Gaussian with parameters $(\frac{1}{2}|y_{ij} - g_i|^{-1}, \frac{1}{4})$, with w_{ij} conditionally independent of each other given $\mathbf{g}$ and the data $\mathbf{y}$. Finally, the conditional posterior for λ is gamma with parameters $(14 + c_0, \frac{1}{2}\mathbf{g}^T\mathbf{K}\mathbf{g} + d_0)$. This Gibbs sampler can be summarised in Algorithm (2.3)

Thompson et al. (2010) analysed $\tau = 0.95$ and ran the MH algorithm for 250,000 iterations discarding 50,000 as burn in and retaining every 10th iteration to reduce autocorrelation and for storage purposes. Figure 2.2 plots the NCS obtained by the MH algorithm and that obtained by the Gibbs sampler using 11,000 iterations, 1,000 of which were discarded as burn in.

At first glance, all seems fine. Figure 2.2 show that both the MH algorithm and the Gibbs sampler produce curves that can accurately reconstruct the true underlying curve and are very similar to each other.

Thompson et al. (2010) assess the rate of convergence by running 3 separate MH samplers initialised with wildly different starting values. These same starting values were used to start 3 additional Gibbs samplers. Figure 2.3 shows the first 10,000 iterations of the 3 chains under MH sampling for the first knot g_1. Figure 2.4 shows the first 100 iterations of the 3 chains under Gibbs sampling for g_1.

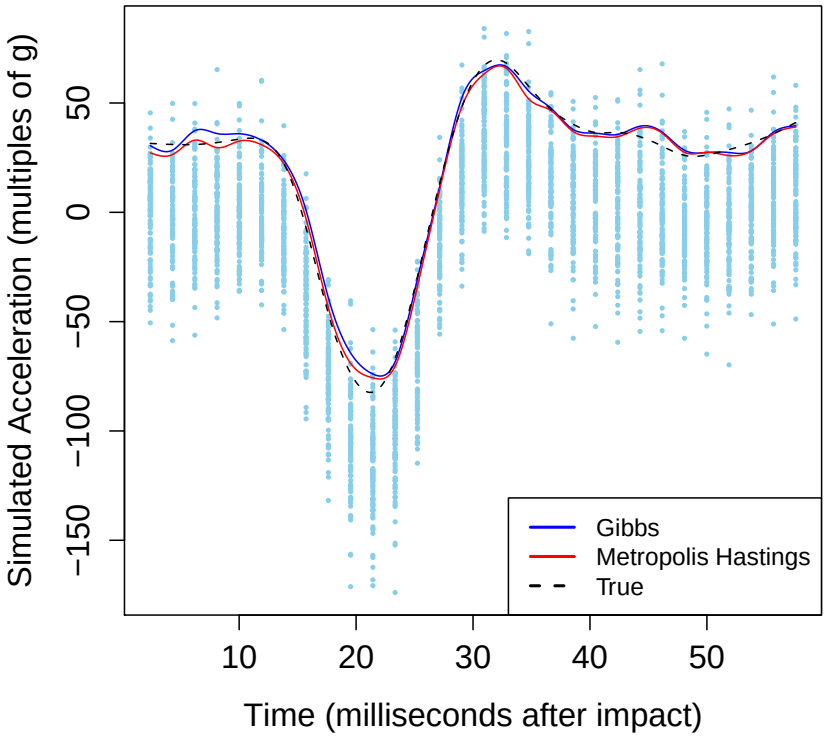

Figure 2.2: Bayesian nonparametric regression using NCS for $\tau = 0.95$. The blue curve is obtained from our Gibbs sampler, the red curve is obtained from MH. The dashed black curve is the true underlying curve.

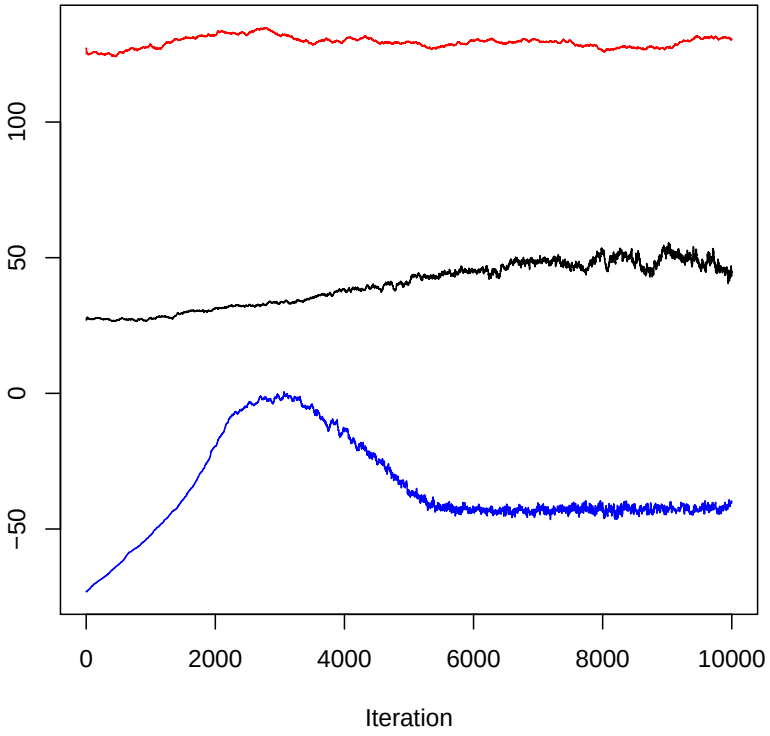

Figure 2.3: Plot of the first 10,000 iterations of the Metropolis Hastings chains, plotted for g_1.

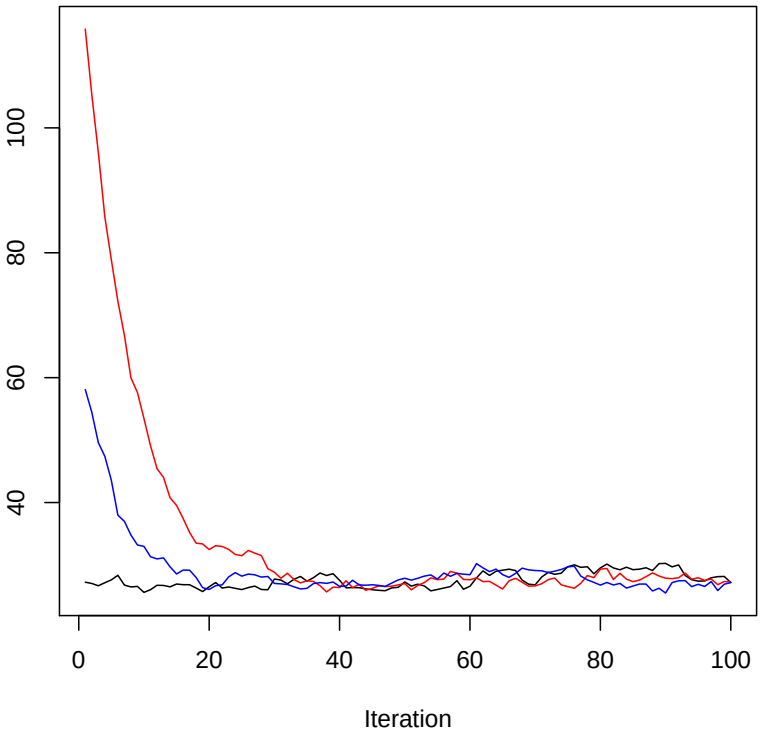

Figure 2.4: Plot of the first 100 iterations of the Gibbs chains, plotted for g_1.

The disadvantage of the MH algorithm is immediately evident from Figure 2.3 and Figure 2.4. The chains obtained using MH have not forgotten their starting values after 10,000 iterations. In fact, even after running the 3 chains for the full 250,000 iterations, discarding the first 50,000 iterations and thinning, the posterior mean for $\mathbf{g}$ (not shown here) was different for each chain. In contrast, observe that the 3 chains obtained by running the Gibbs sampler converged on each other very quickly, despite having 3,000 additional parameters to update. This is likely to be due to the fact that the Gibbs sampler blocks all 3,000 latent parameters together, thus reducing the negative effect alluded to in Liu et al. (1994) of having 3,000 additional parameters to update. Running the Gibbs samplers with 3 different starting values each for 11,000 iterations discarding the first 1,000 gave values of the Gelman-Rubin (Gelman and Rubin, 1992) diagnostic of between 1.000 and 1.013. The posterior mean for $\mathbf{g}$ was virtually identical for each chain. These results demonstrate the apparent superiority of the Gibbs sampler in these higher dimensional cases.

The posterior mean for λ was about 0.03. This value of λ indicates that those curves that fit the data well but are fairly "wiggly" are preferred for this example.

2.5 Summary

This chapter has provided the framework for allowing Bayesian parameter estimation to be implemented on more complex quantile regression models in a relatively straightforward manner. Despite the observations by Liu et al. (1994) suggesting that adding latent variables will slow convergence, no evidence of this has been observed. In fact, it is particularly evident when analysing the nonparametric quantile regression model with natural cubic splines that the Gibbs sampler is a much more efficient MCMC sampler. Although it was possible to accurately reconstruct the underlying curve using both the Gibbs sampler and the MH algorithm of Thompson et al. (2010), the MH sampler requires good prior knowledge about starting values whereas the Gibbs sampler appears not to be affected by

the starting values. These observations could be due to a number of factors, most notably that it is extremely difficult to choose a sensible proposal density in high dimensional problems and that the Gibbs sampler can update the latent parameters jointly whilst sampling directly from the conditional $\pi(\mathbf{g}|\mathbf{W}, \lambda, \mathbf{y})$.

Chapter 3

An Application in Epidemiology: Is Maternal Cotinine a Better Predictor of Low Birthweight Infants than the Reported Number of Cigarettes?

3.1 Introduction and Method

Many previous studies analysing infant birthweight have analysed how various factors have affected *average* birthweight (Peacock et al. (1998) and references therein). However, as Abrevaya and Dahl (2008) have pointed out, there are greater costs associated with low birthweight (LBW) infants. Moreover, it has been observed that it is more likely for LBW infants to have a greater mortality rate, in addition to likely problems in development and education (LBW infants are more likely to repeat a year) and ultimately, are more likely to be unemployed (see Abrevaya and Dahl (2008) and references therein).

The aim of this study is to use quantile regression on the St George's Birthweight study data to explore whether results presented in Peacock et al. (1998) hold for the lower quantiles of the conditional distribution. In using quantile regression, it will be possible to analyse the median and hence investigate the robustness of the original results.

To answer some of the questions related to this study, model comparisons will

be required. The Bayesian approach to comparing competing models is to use the Bayes factor, defined as $\pi(\mathbf{y}|\text{Model 1})/\pi(\mathbf{y}|\text{Model 2})$. If there is no reason to suspect that model 1 is more likely to have generated the data than model 2 *a priori*, then the Bayes factor is equivalent to the posterior odds $\pi(\text{Model 1}|\mathbf{y})/\pi(\text{Model 2}|\mathbf{y})$. Using the Bayesian approach compares the likelihoods averaged over the parameters of the models rather than the maximum likelihoods used in frequentist statistics. The averaging of the likelihood over the parameters naturally penalises a model for the size of its parameter space, hence offering a model comparison tool that trades off goodness of fit against model complexity.

Combined with the AL likelihood (1.9), the prior

$$\boldsymbol{\beta}|\boldsymbol{\lambda} \sim N(\mathbf{0}, \boldsymbol{\Lambda}^{-1}) \tag{3.1}$$

is used throughout this chapter. Here, $\boldsymbol{\lambda}$ is a $(p+1)\times 1$ vector of hyperparameters λ_j and $\boldsymbol{\Lambda}$ is a $(p+1)\times(p+1)$ diagonal matrix with $\boldsymbol{\Lambda}_{j,j} = \lambda_j$. To let the data influence the results as much as possible, λ_j can be set to a constant c and then letting c tend to 0. This results in a joint improper uniform prior for $\boldsymbol{\beta}$. However, this improper prior leads to indeterminate Bayes factors. A compromise between robustness and avoiding indeterminate Bayes factors is to give each λ_j a gamma hyperprior with parameters $(\frac{1}{2}, \frac{1}{2})$. Marginalised over $\boldsymbol{\lambda}$, this gives a product of standard Cauchy$(0, 1)$ distributions as the joint prior on $\boldsymbol{\beta}$. These have more probability mass in the tails than the normal.

Just as before, data augmentation plays a key role in designing more efficient Gibbs samplers. Under the improper prior on $\boldsymbol{\beta}$, the resulting Gibbs sampler is the same as that in Chapter 2 using $\mathbf{b_0} = \mathbf{0}$ and $\mathbf{B_0} = c\mathbf{I}$ and $c \to 0$. With the Cauchy priors, an additional update of each of the latent λ_j is required, similar in spirit to the Gibbs sampler for NCS. These can be updated independently of each other and are exponentially distributed with rate parameter $\frac{1}{2}(1+\beta_j^2)$. The procedure is described in Algorithm 3.1.

In using the Gibbs sampler to analyse this dataset, no indications of lack of

Algorithm 3.1 Gibbs sampler for augmented quantile regression model under independent Cauchy priors on $\boldsymbol{\beta}$. Draws M burn in samples followed by an additional N samples for inference.

Given: Initial values $\boldsymbol{\beta}^{(0)}$.

for $k = 1$ **to** $M + N$ **do**

- Sample $\mathbf{w}^{(k)}|\boldsymbol{\beta}^{(k-1)}, \mathbf{y}$ by sampling the ith component of $\mathbf{w}$ $(i = 1, \ldots, n)$ from the inverse Gaussian distribution with shape parameter $\frac{1}{4}$ and location $\frac{1}{2}|y_i - \mathbf{x_i}^T\boldsymbol{\beta}^{(k-1)}|^{-1}$.
- Update each component of $\boldsymbol{\lambda}^{(k)}$ from an exponential distribution with rate $\frac{1}{2}\{1 + \{\beta_j^{(k-1)}\}^2\}$.
- Sample $\boldsymbol{\beta}^{(k)}|\mathbf{w}^{(k)}, \boldsymbol{\lambda}^{(k)}, \mathbf{y}$ from the multivariate normal distribution with precision matrix

$$\mathbf{X}^T\mathbf{W^{(k)}}\mathbf{X} + \boldsymbol{\Lambda}^{(k)}$$

and mean vector

$$(\mathbf{X}^T\mathbf{W^{(k)}}\mathbf{X} + \boldsymbol{\Lambda}^{(k)})^{-1}(\mathbf{X}^T\mathbf{W^{(k)}}\mathbf{y} + (\tau - \tfrac{1}{2})\mathbf{X}^T\mathbf{1},$$

where $\mathbf{W}^{(k)}$ and $\boldsymbol{\Lambda}^{(k)}$ are diagonal matrices. Each element of $\mathbf{w}^{(k)}$ forms the diagonal of $\mathbf{W}^{(k)}$ and the diagonal elements of $\boldsymbol{\Lambda}^{(k)}$ are $\{\lambda_j^{(k)}\}$.

end for.

convergence were found after 1,000 iterations. Once these first 1,000 iterations had been discarded, inference was based on 10,000 further iterations.

For the analysis under the Cauchy priors, the Bayes factors were calculated from the samples using Chib's method (Chib, 1995). This approach uses a rearrangement of Bayes theorem. In this case, conditional on a model, Bayes theorem gives

$$l(\mathbf{y}) = \frac{l(\mathbf{y}|\boldsymbol{\beta})\pi(\boldsymbol{\beta})}{\pi(\boldsymbol{\beta}|\mathbf{y})}. \tag{3.2}$$

Equation (3.2) remains true if $\boldsymbol{\beta}$ is substituted by it's posterior mean $\hat{\boldsymbol{\beta}}$. Thus, an algorithm can be developed making use of the Gibbs samples to calculate an approximate marginal for $\mathbf{y}$ conditional on a fixed model. This is summarised in Algorithm 3.2.

Algorithm 3.2 can be repeated with other models to form approximate Bayes factors. The most computationally intense part in this calculation is evaluating the posterior marginal density ordinate $\pi(\hat{\boldsymbol{\beta}}|\mathbf{y})$, which requires additional iterations from the Gibbs sampler to get a good approximation.

Algorithm 3.2 Chib's method for calculating approximate marginal likelihood $l(\mathbf{y})$.

Given: Gibbs sample obtained using Algorithm 3.1.

- From the Gibbs sample, discard the burn in and average over the remaining samples to produce an estimate of the marginal posterior mean $\hat{\boldsymbol{\beta}}$.
- Evaluate $l(\mathbf{y}|\hat{\boldsymbol{\beta}})$ and $\pi(\hat{\boldsymbol{\beta}})$. These are both available in closed form, the first being the AL likelihood and the second a product of Cauchy density ordinates.
- Using the current state of the Gibbs sampler, obtain N addition samples using Algorithm 3.1. Record the density ordinate at each sampled value of $\mathbf{w}$ and $\boldsymbol{\lambda}$, $\pi(\hat{\boldsymbol{\beta}}|\mathbf{w}^{(k)}, \boldsymbol{\lambda}^{(k)})$ for $k = M + N + 1, \ldots, M + 2N$.
- Approximate $\pi(\hat{\boldsymbol{\beta}}|\mathbf{y})$ with

$$\hat{\pi}(\hat{\boldsymbol{\beta}}|\mathbf{y}) = \frac{1}{N} \sum_{k=M+N+1}^{M+2N} \pi(\hat{\boldsymbol{\beta}}|\mathbf{w}^{(k)}, \boldsymbol{\lambda}^{(k)}). \tag{3.3}$$

- Plug these values into (3.2) to obtain an approximation of $\hat{\pi}(\mathbf{y})$.

3.2 Results

The original research by Peacock et al. (1998), investigated two main questions, namely i) whether maternal serum cotinine level, a metabolite of nicotine, is a better predictor of infant's birthweight than the reported number of cigarettes smoked by the mother and ii) what the effect of passive smoke exposure on birthweight among women who do not smoke is. This dataset was analysed again to investigate relationships at the median and the lower tails of the conditional birthweight distribution. Following the original analysis, quantile regression was used with the response variable being adjusted birthweight (birthweight adjusted for gestational age, maternal height, sex of infant and parity, where the adjusted birthweight is effectively a ratio of observed to expected values and can be interpreted as percentage differences from expected values (Bland et al., 1990)). The models investigated included one or more of the following covariates:

- cotinine
- number of cigarettes

- nicotine yield of cigarette

For each model, data recorded at three time points was analysed: at booking clinic (approximately 14 weeks gestation), at 28 weeks gestation and at 36 weeks gestation. The difference between cotinine measured at each of the different time points was also modelled to help identify whether a change in smoking habit has an effect on the adjusted birthweight. As a final anaysis, data for women who were not active smokers (determined by a cotinine measurement level less than 15ng/ml) were analysed to see if there are any effects of passive smoke on the adjusted birthweight.

A quick glance at the scatterplot of cotinine level at booking against adjusted birthweight (Figure 3.1) suggests that the linear regression model used by Peacock et al. (1998) seems sensible for the majority of conditional quantiles of the adjusted birthweight distribution. Here, the term "linear" is referring both to regression linear in the parameters and a regression equation that is a straight line. The plots look similar for cotinine measured at 28 weeks and 36 weeks so are not presented here. This analysis is therefore based on simple linear quantile regression.

Figure 3.2 plots the posterior mean cotinine as a function of τ for $\tau \in \{0.05, 0.1, \ldots, 0.45, 0.5\}$. The shaded region is the associated 95% HPD interval. Just as in the previous chapter, the HPD interval should be interpreted with caution as it is subject to the likelihood representing the true underlying data generating mechanism, something that is never assumed in the analysis. It does however serve as a rough guide for exploratory analysis.

Figure 3.2 shows that at all 3 timepoints the posterior mean birthweight gradually decreases until $\tau = 0.15$, then it appears to decreases significantly faster as τ gets smaller. The implications are that smoking has a much larger effect on the more severely underweight infants.

Turning to the question of whether cotinine is a better predictor than the reported number of cigarettes, the Bayes factors were calculated at $\tau \in \{0.03, 0.1, 0.25, 0.5\}$.

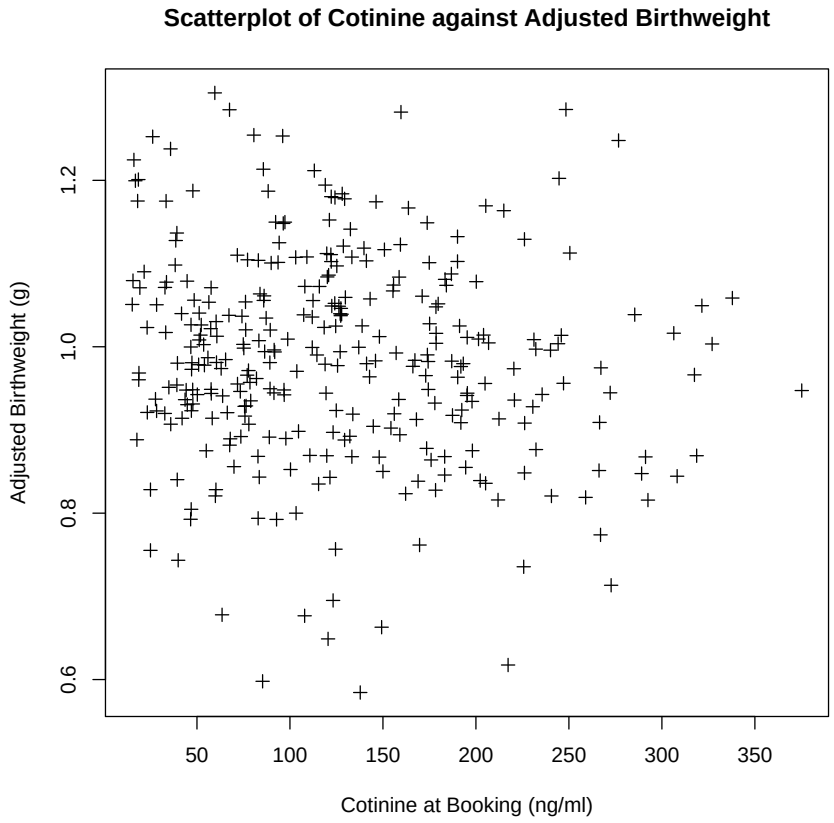

Figure 3.1: Scatterplot of cotinine level at booking against adjusted birthweight. The points include all those classed as smokers with a cotinine level above 15ng/ml.

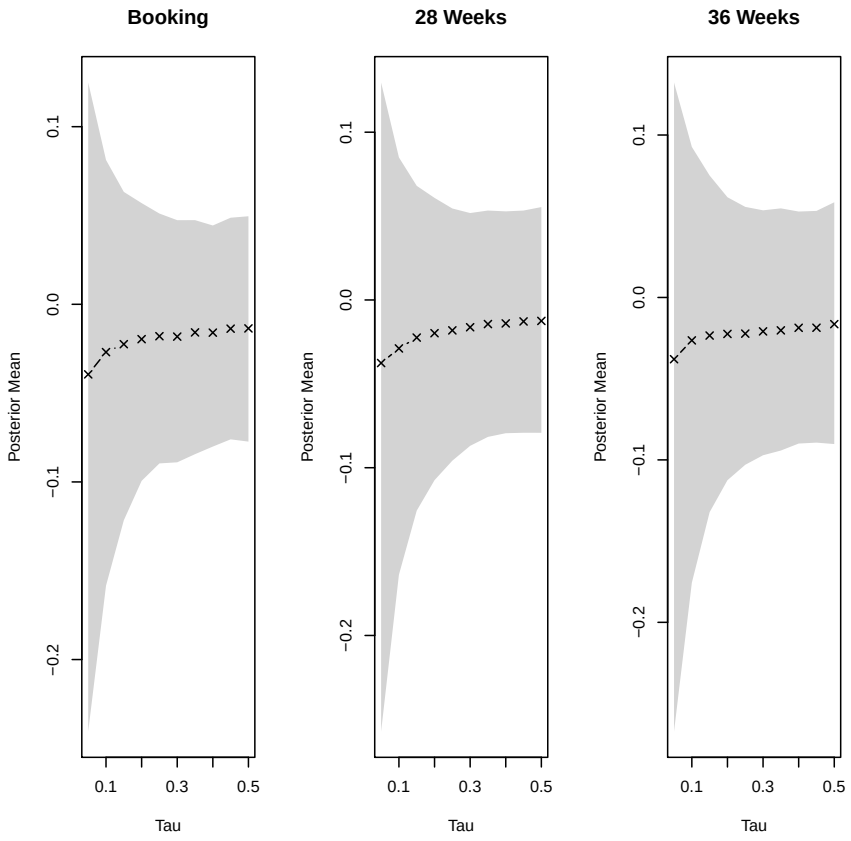

Figure 3.2: Plot of posterior mean cotinine against quantile. The shaded area is the 95% HPD interval.

τ	Booking	28 Weeks	36 Weeks
0.03	0.958	1.054	0.964
0.1	1.006	1.041	0.9129
0.25	1.072	1.059	1.070
0.5	1.074	1.032	1.162

Table 3.1: Bayes factors against Model 1: Model 1 number of cigarettes vs. Model 2 cotinine.

Bayes factor against Model	Evidence against Model
1:3	Weak
3:20	Positive
20:150	Strong
>150	Very strong

Table 3.2: Interpretation of Bayes factors from Kass and Raftery (1995).

Table 3.1 presents the Bayes factor comparing cotinine and number of cigarettes at the 3 different time points. Table 3.2 is from Kass and Raftery (1995) giving the scale of evidence for the competing model. The Bayes factors in Table 3.1 range from 0.9129 to 1.162. Thus any evidence to support either cotinine or number of cigarettes is weak. The majority of Bayes factors do favour cotinine as a predictor ($\frac{3}{4}$ versus $\frac{1}{4}$).

The Bayes factors, although not presented here due to the fact that they were all bigger than 150, strongly suggest that the additional knowledge of the nicotine yield does not improve the predictive accuracy in using reported number of cigarettes relative to cotinine. Instead, the price is paid in additional model complexity.

Figure 3.3 shows how the difference in cotinine between the various time points varies with the quantile. The relationship appears fairly linear when comparing the cotinine recorded at booking with the cotinine recorded at 28 weeks and 36 weeks. However, this does not appear to hold when comparing the difference between cotinine recorded at 28 weeks to cotinine recorded at 36 weeks with a more noticeable effect at values of τ smaller than about 0.1. Overall, any difference in cotinine seems to have the greatest impact on the low birthweight infants.

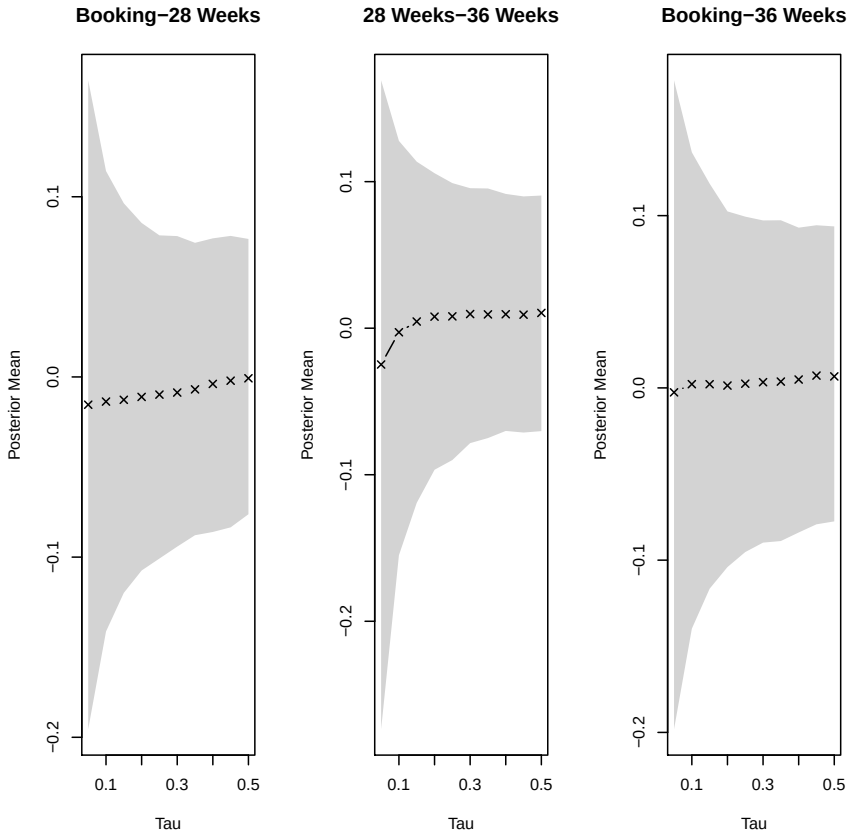

Figure 3.3: Plot of posterior mean difference in cotinine against quantiles. Shaded region is 95% HPD interval.

τ	Booking-28 Weeks	28 Weeks-36 Weeks	Booking-36 Weeks
0.03	11.316	9.399	10.013
0.1	19.407	18.874	18.830
0.25	26.745	27.764	25.375
0.5	33.526	30.801	29.663

Table 3.3: Bayes factors against Model 1: Model 1 null model vs. Model 2 cotinine difference.

τ	Booking	28 Weeks	36 Weeks
0.03	17.304	14.105	14.897
0.1	31.443	29.280	28.969
0.25	51.109	47.508	43.368
0.5	61.625	57.642	48.383

Table 3.4: Bayes factors against Model 1 for passive smokers: Model 1 null model vs. Model 2 cotinine.

Table 3.3 shows the Bayes factors comparing the null model with the model containing cotinine difference as a predictor. The Bayes factors this time range from 9.399 to 33.526, indicating evidence ranging from positive to strong that the difference in cotinine does not have any significant impact. Perhaps not surprising given Figure 3.3, the weakest evidence supporting the null model was observed at $\tau = 0.03$, with the evidence increasing steadily as τ increases and falling into the strong category at the median.

Finally, Figure 3.4 shows the effect of passive smoking on adjusted birthweight. There is again evidence of a linear relationship between the posterior means and τ from $\tau = 0.5$ down to about 0.2, then it appears to decrease at a quadratic rate for values of τ lower than 0.2.

The Bayes factors from Table 3.4 appear to suggest that there is evidence against passive smoking having a significant effect on birthweight. Just as for the cotinine difference, the evidence is strongest at the median and gets weaker as τ decreases. This is again in keeping with the relationship apparent in Figure 3.4.

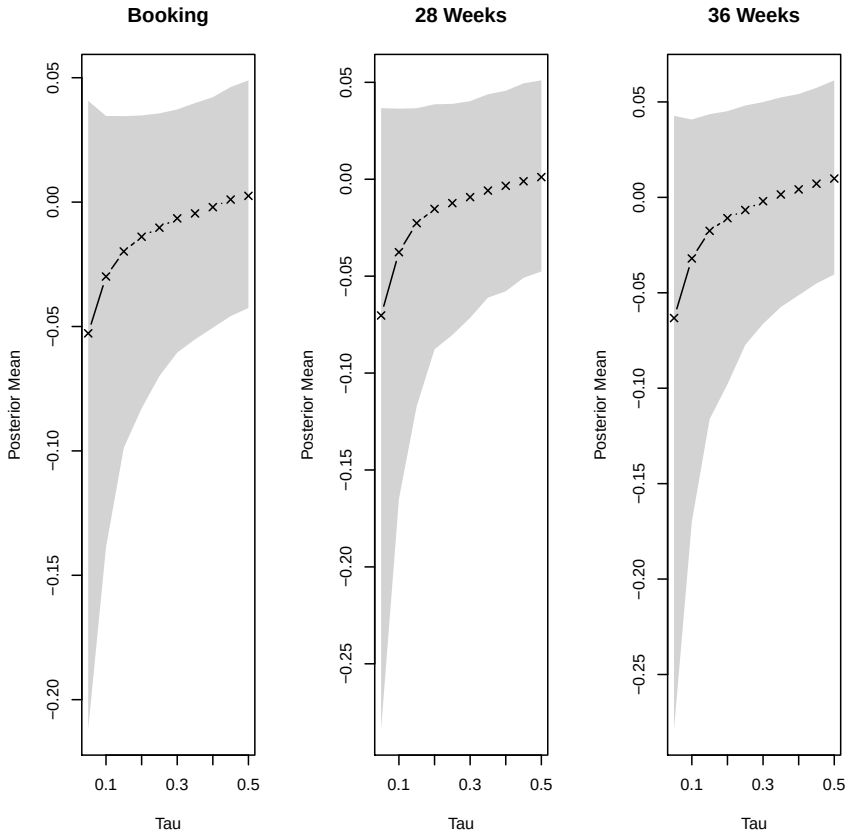

Figure 3.4: Posterior mean of cotinine for passive smokers against quantile. Shaded region is 95% HPD interval.

3.3 Conclusion of Study

The Bayesian approach to quantile regression has been successfully used in modelling the lower half of the conditional distribution of birthweight rather than the average birthweight carried out in the original study. Whilst many of the conclusions from median regression analysis are identical to the original research by Peacock et al. (1998), it is clear that the effects are significantly stronger at the extreme low quantiles. This study also gives further evidence, albeit weak, that cotinine is a better predictor of birthweight than the reported number of cigarettes at all quantiles except 3 percent. There is no significant evidence for both a change in cotinine level and passive smoking having any effect on birthweight.

Chapter 4

Bayesian Variable Selection for Quantile Regression

4.1 Introduction and Method

Up until now, the assumption has been that the underlying regression model is fixed. Proceeding in this way ignores model uncertainty. The cost of ignoring model uncertainty is now well known (see Hoeting et al. (1999) and references therein). The Bayesian answer to this problem is provided by Bayesian model averaging (BMA). A nice introductory tutorial to BMA is provided by Hoeting et al. (1999). Madigan and Raftery (1994) show that predictive accuracy, measured by a logarithmic scoring rule, is always higher if BMA is used compared to any single model.

The work in the previous two chapters can now be extended to handle model uncertainty in quantile regression. This chapter follows the manuscript of Reed et al. (2010) which has been submitted to the Journal of Computational and Graphical Statistics and is awaiting a revision.

The potential models considered are of the form

$$Q_\tau(y_i|\mathbf{z_i}, \mathbf{x_i}) = \mathbf{z_i}^T\boldsymbol{\alpha} + \mathbf{x_i}^T\boldsymbol{\beta}. \tag{4.1}$$

The $n \times q$ associated design matrix $\mathbf{Z}$ contains all predictors that should always appear in any model. In the majority of cases, this will be the intercept only so that $\mathbf{Z} = \mathbf{1_n}$ and $q = 1$. The $n \times (p - q)$ design matrix $\mathbf{X}$ contains the remaining

predictors for which variable selection is to be carried out on. The regression parameters are $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ which are distinguished deliberately to make notation clearer.

Following common practice for Bayesian variable selection, index each of the $p-q$ predictors in $\mathbf{X}$ by γ_j, $j = q+1, \ldots, p$ with $\gamma_j = 1$ if the jth predictor is present in the regression model and 0 otherwise. Additionally, denote the $(p-q)\times 1$ vector $\boldsymbol{\gamma}$ as having jth element γ_j. In this way, the vector $\boldsymbol{\gamma}$ uniquely defines each regression model $M_{\boldsymbol{\gamma}}$. For example, the model $M_{\boldsymbol{\gamma}}$ indexed by $\boldsymbol{\gamma} = [1,0,0,1,0]^T$ with $\mathbf{Z} = \mathbf{1_n}$ corresponds to the regression model

$$Q_\tau(y_i|\mathbf{z_i}, \mathbf{x_i}) = \alpha_0 + \beta_1 x_{1i} + \beta_4 x_{4i}.$$

In order to compare predictors in $\mathbf{X}$ in a meaningful way, it is necessary that the predictors are standardised, that is, they have sample mean equal to 0 and sample standard deviation equal to 1. The predictors in $\mathbf{Z}$ are not required to be standardised as they are common to all models. Given this, the AL likelihood (2.2) takes the form

$$\prod_{i=1}^{n} \exp\{-\tfrac{1}{2}|y_i - \mathbf{z_i}^T\boldsymbol{\alpha} - \mathbf{x_i}^T\boldsymbol{\beta}|\} \prod_{i=1}^{n} \tau(1-\tau)\exp\{-(\tau - \tfrac{1}{2})(y_i - \mathbf{z_i}^T\boldsymbol{\alpha})\}. \quad (4.2)$$

After augmenting the data with the latent vector $\mathbf{w}$ and assigning the independent inverse Gamma priors as in Chapter 2, the normal prior for the regression parameters is semi-conjugate. The assumed prior for $\boldsymbol{\alpha}$ takes the form

$$\pi(\boldsymbol{\alpha}) \propto \exp\left\{-\tfrac{1}{2}(\boldsymbol{\alpha} - \mu_{\mathbf{0}})^T \mathbf{A_0}(\boldsymbol{\alpha} - \mu_{\mathbf{0}})\right\}, \quad (4.3)$$

with $\mu_{\mathbf{0}}$ and $\mathbf{A_0}$ both fixed. The improper prior could be used here by setting $\mathbf{A_0} = c\mathbf{I}$, and letting $c \to 0$.

For regression parameters $\boldsymbol{\beta}$, the assumed prior is specified jointly with $\boldsymbol{\gamma}$ and depends on hyperparameters $\boldsymbol{\lambda}$ and π_0. The first component $\pi(\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{\lambda})$ is specified using

$$\beta_j|\gamma_j, \lambda_j \sim (1-\gamma_j)\delta_0 + \gamma_j N(0, \lambda_j^{-1}) \quad (4.4)$$

for $j = q+1, \ldots, p$. Here, δ_0 denotes a degenerate distribution with all its mass at zero. The joint prior $\pi(\boldsymbol{\gamma}, \boldsymbol{\beta}|\boldsymbol{\lambda}, \pi_0)$ is fully specified by letting

$$\gamma_j|\pi_0 \sim \text{Bernoulli}(\pi_0) \tag{4.5}$$

independently. The hyperparameter $\pi_0 = \Pr(\gamma_j = 1) = \Pr(\beta_j \neq 0)$ represents the prior probability of including a randomly selected predictor in the model.

Note that the prior (4.4) assigns prior probability $1 - \pi_0$ to the event that a randomly selected predictor is not present in the quantile regression model. This is in contrast to the specification given by George and Mcculloch (1993) where this prior probability is 0.

Given that the QR-SSVS procedure will effectively be computing multiple Bayes factors, a robust approach that avoids using improper priors leading to indeterminate Bayes factors uses a hyperprior on $\boldsymbol{\lambda}$. Following from the previous chapter, the prior specification

$$\lambda_j \sim \text{Gamma}(1/2, 1/2) \tag{4.6}$$

induces a heavy-tailed Cauchy prior marginally for the coefficients on the predictors selected to be in the model. Specifically, the marginal prior for $\beta_j|\gamma_j$ is given by

$$\beta_j|\gamma_j \sim (1 - \gamma_j)\delta_0 + \gamma_j \text{Cauchy}(0, 1). \tag{4.7}$$

Using a standard Cauchy prior for β_j given $\gamma_j = 1$ seems sensible given that the predictors are standardised.

To complete the full specification, it is necessary to specify π_0. A common choice here is to set $\pi_0 = \frac{1}{2}$ in an attempt to be non-informative (see e.g. George and McCulloch (1997)). However, this also assumes that the prior expected number of predictors in $\mathbf{X}$, $p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j$, is $(p - q)/2$. Alternatively, π_0 could be treated as unknown and given a conjugate beta hyperprior $\pi_0 \sim \text{Beta}(a_0, b_0)$, thus obtaining a more flexible prior on $p_{\boldsymbol{\gamma}}$. This allows the data to inform more strongly about the model size. The parameter π_0 can then be analytically integrated out to yield a beta-binomial prior marginally on $p_{\boldsymbol{\gamma}}$.

4.1.1 QR-SSVS algorithm

For computational convenience, define $\boldsymbol{X}_{\boldsymbol{\gamma}}$, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\lambda}_{\boldsymbol{\gamma}}$ as being the design matrix $\mathbf{X}$, the regression parameters $\boldsymbol{\beta}$ and the vector $\boldsymbol{\lambda}$ with the jth row or column deleted if $\gamma_j = 0$. Additionally define $\boldsymbol{\Lambda}_{\boldsymbol{\gamma}}$ to be a diagonal matrix containing only those elements of $\boldsymbol{\lambda}$ on the diagonal for which $\gamma_j = 1$. Now define the matrices $\mathbf{H}_{\boldsymbol{\gamma}}$ and $\mathbf{Q}_{\boldsymbol{\gamma}}$ as

$$\begin{aligned} \mathbf{H}_{\boldsymbol{\gamma}} &:= \mathbf{W}\boldsymbol{X}_{\boldsymbol{\gamma}}(\boldsymbol{X}_{\boldsymbol{\gamma}}{}^T\mathbf{W}\boldsymbol{X}_{\boldsymbol{\gamma}} + \boldsymbol{\Lambda}_{\boldsymbol{\gamma}})^{-1}\boldsymbol{X}_{\boldsymbol{\gamma}}{}^T\mathbf{W} \qquad (4.8) \\ \mathbf{Q}_{\boldsymbol{\gamma}} &:= \mathbf{A_0} + \mathbf{Z}^T(\mathbf{W} - \mathbf{H}_{\boldsymbol{\gamma}})\mathbf{Z}, \qquad (4.9) \end{aligned}$$

and the vector $\mathbf{r}$ as

$$\mathbf{r} := \mathbf{Z}^T\mathbf{W}\mathbf{y} + (\tau - \tfrac{1}{2})\mathbf{Z}^T\mathbf{1} + \mathbf{A_0}\mu_{\mathbf{0}}. \qquad (4.10)$$

Finally, define the function $\mathrm{f}(\boldsymbol{\gamma}) = \mathrm{f}(\gamma_{q+1}, \gamma_{q+2}, \ldots, \gamma_p)$, where

$$\begin{aligned} \mathrm{f}(\boldsymbol{\gamma}) &:= \Gamma(p\boldsymbol{\gamma} + a_0)\Gamma(p - q - p\boldsymbol{\gamma} + b_0)|\boldsymbol{\Lambda}_{\boldsymbol{\gamma}}|^{1/2}|\boldsymbol{X}_{\boldsymbol{\gamma}}{}^T\mathbf{W}\boldsymbol{X}_{\boldsymbol{\gamma}} + \boldsymbol{\Lambda}_{\boldsymbol{\gamma}}|^{-1/2}|\mathbf{Q}_{\boldsymbol{\gamma}}|^{-1/2} \\ &\times \exp\{\tfrac{1}{2}(\mathbf{y}^T\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y} + (\mathbf{r} - \mathbf{Z}^T\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y})^T\mathbf{Q}_{\boldsymbol{\gamma}}(\mathbf{r} - \mathbf{Z}^T\mathbf{H}_{\boldsymbol{\gamma}}\mathbf{y}))\}. \qquad (4.11) \end{aligned}$$

Equipped with these definitions, the QR-SSVS algorithm sampling from the joint marginal with π_0 integrated out is fully described in Algorithm 4.1 and Algorithm 4.2.

To prevent the QR-SSVS algorithm from getting stuck, it is necessary to marginalise out the vector of regression parameters $\boldsymbol{\beta}$ in updating the indicator γ_j. It is not necessary to marginalise out $\boldsymbol{\alpha}$ but it makes the calculations easier if it is integrated out when updating γ_j. Due to the use of these *reduced* conditionals, the algorithm no longer defines an ordinary Gibbs sampler. It instead defines a partially collapsed Gibbs sampler (van Dyk and Park, 2008), (Park and van Dyk, 2009). With such samplers, the order in which the parameters are updated necessarily affects the stationary distribution to which the algorithm converges.

Although at first glance, the QR-SSVS algorithm seems computationally intensive, in practice the steps required to update both $\boldsymbol{\gamma}$ and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ can be implemented efficiently. The main computational overhead is computing the Cholesky

Algorithm 4.1 Component of QR-SSVS algorithm that updates the vector $\boldsymbol{\gamma}$.

Given: Index k, previous values of parameters $\mathbf{w}^{(k-1)}$, $\boldsymbol{\lambda}^{(k-1)}$ and previous model specified through $\boldsymbol{\gamma}^{(k-1)}$.

• Sample $\gamma_q^{(k)}|\boldsymbol{\gamma}_{-q}^{(k-1)}, \mathbf{w}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}, \mathbf{y}$, where $\boldsymbol{\gamma}_{-q}$ denotes the remaining elements of $\boldsymbol{\gamma}$ excluding the qth, from a Bernoulli trial with probability of success

$$\frac{\mathrm{f}(\gamma_q = 1, \boldsymbol{\gamma}_{-q}^{(k-1)})}{\mathrm{f}(\gamma_q = 1, \boldsymbol{\gamma}_{-q}^{(k-1)}) + \mathrm{f}(\gamma_q = 0, \boldsymbol{\gamma}_{-q}^{(k-1)})}.$$

for $j = q + 1$ **to** $(p - 1)$ **do**

• Sample $\gamma_j^{(k)}|\boldsymbol{\gamma}_{q:j-1}^{(k)}, \boldsymbol{\gamma}_{j+1:p}^{(k-1)}, \mathbf{w}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}, \mathbf{y}$ from a Bernoulli trial with probability of success

$$\frac{\mathrm{f}(\gamma_j = 1, \boldsymbol{\gamma}_{q:j-1}^{(k)}, \boldsymbol{\gamma}_{j+1:p}^{(k-1)})}{\mathrm{f}(\gamma_j = 1, \boldsymbol{\gamma}_{q:j-1}^{(k)}, \boldsymbol{\gamma}_{j+1:p}^{(k-1)}) + \mathrm{f}(\gamma_j = 0, \boldsymbol{\gamma}_{q:j-1}^{(k)}, \boldsymbol{\gamma}_{j+1:p}^{(k-1)})}.$$

end for

• Sample $\gamma_p^{(k)}|\boldsymbol{\gamma}_{-p}^{(k)}, \mathbf{w}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}, \mathbf{y}$ from a Bernoulli trial with probability of success

$$\frac{\mathrm{f}(\gamma_p = 1, \boldsymbol{\gamma}_{-p}^{(k)})}{\mathrm{f}(\gamma_p = 1, \boldsymbol{\gamma}_{-p}^{(k)}) + \mathrm{f}(\gamma_p = 0, \boldsymbol{\gamma}_{-p}^{(k)})}.$$

decomposition, that is, find the Cholesky factor $\mathbf{L}$ such that $\mathbf{L}\mathbf{L}^T = \boldsymbol{\Sigma}$ for some positive definite matrix $\boldsymbol{\Sigma}$. Once this is known, the new Cholesky factor can be recalculated when a predictor is added or removed from the design matrix $\boldsymbol{X}_{\boldsymbol{\gamma}}$ efficiently using techniques such as permuting the rows of the current Cholesky factor and applying orthogonal transformations. See Dongarra et al. (1979) for a more detailed explanation.

4.2 Revisiting the Stack Loss Data

To illustrate QR-SSVS, consider again the stackloss data in Section 2.3. If the intercept is to be included in all candidate models, then there are a total of $2^3 = 8$ potential models. This data was analysed using the hyperprior $\pi_0 \sim \mathrm{Beta}(1, 1)$. Each predictor was standardised before analysis except the intercept, which is assumed to appear in all models. The results are based on 10,000 iterations of

Algorithm 4.2 Stochastic search variable selection for quantile regression model. Draws M burn in samples followed by an additional N samples for inference.

Given: Initial values $\mathbf{w}^{(0)}$, $\boldsymbol{\lambda}^{(0)}$ and an initial model specified through $\boldsymbol{\gamma}^{(0)}$. Typically, the initial model is the full model, i.e. $\boldsymbol{\gamma}^{(0)} = \mathbf{1_{p-q}}$.

for $k = 1$ **to** $M + N$ **do**

• Sample $\boldsymbol{\gamma}^{(k)}$ using Algorithm 4.1.

• Sample $\boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta_\gamma}^{(k)} | \boldsymbol{\gamma}^{(k)}, \mathbf{w}^{(k-1)}, \boldsymbol{\lambda}^{(k-1)}, \mathbf{y}$ from the multivariate normal distribution with precision matrix

$$\begin{bmatrix} \mathbf{Z}^T\mathbf{W}^{(k-1)}\mathbf{Z} + \mathbf{A_0} & \mathbf{Z}^T\mathbf{W}^{(k-1)}\boldsymbol{X_\gamma} \\ \boldsymbol{X_\gamma}^T\mathbf{W}^{(k-1)}\mathbf{Z} & \boldsymbol{X_\gamma}^T\mathbf{W}^{(k-1)}\boldsymbol{X_\gamma} + \boldsymbol{\Lambda_\gamma} \end{bmatrix}$$

and mean

$$\begin{bmatrix} \mathbf{Z}^T\mathbf{W}^{(k-1)}\mathbf{Z} + \mathbf{A_0} & \mathbf{Z}^T\mathbf{W}^{(k-1)}\boldsymbol{X_\gamma} \\ \boldsymbol{X_\gamma}^T\mathbf{W}^{(k-1)}\mathbf{Z} & \boldsymbol{X_\gamma}^T\mathbf{W}^{(k-1)}\boldsymbol{X_\gamma} + \boldsymbol{\Lambda_\gamma} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{r}^{(k-1)} \\ \boldsymbol{X_\gamma}^T\mathbf{W}^{(k-1)}\mathbf{y} \end{bmatrix}.$$

The diagonal matrix $\mathbf{W}^{(k-1)}$ is as defined in previous chapters and $\mathbf{r}^{(k-1)} = \mathbf{Z}^T\mathbf{W}^{(k-1)}\mathbf{y} + (\tau - \frac{1}{2})\mathbf{Z}^T\mathbf{1} + \mathbf{A_0}\mu_\mathbf{0}$.

• Sample $\mathbf{w}^{(k)} | \boldsymbol{\alpha}^{(k)}, \boldsymbol{\beta}^{(k)}, \mathbf{y}$ by sampling the ith component of $\mathbf{w}$ $(i = 1, \ldots, n)$ from the inverse Gaussian distribution with shape parameter $\frac{1}{4}$ and location $\frac{1}{2}|y_i - \mathbf{z_i}^T\boldsymbol{\alpha}^{(k)} - \boldsymbol{x_{\gamma,i}}^T\boldsymbol{\beta_\gamma}^{(k)}|^{-1}$.

• Sample each component of $\boldsymbol{\lambda}^{(k)} | \boldsymbol{\beta}^{(k)}, \mathbf{y}$ from the exponential distribution with rate parameter $\frac{1}{2}(1 + \{\beta_j^{(k)}\}^2)$

end for.

$\tau = 0.05$	
Predictors	Probability
Intercept, x_1, x_2	0.283
Intercept, x_1, x_2, x_3	0.261
Intercept, x_2	0.144
$\tau = 0.25$	
Predictors	Probability
Intercept, x_1, x_2	0.630
Intercept, x_1, x_2, x_3	0.334
Intercept, x_1	0.029
$\tau = 0.5$	
Predictors	Probability
Intercept, x_1, x_2	0.564
Intercept, x_1, x_2, x_3	0.356
Intercept, x_1	0.065
$\tau = 0.75$	
Predictors	Probability
Intercept, x_1, x_2	0.595
Intercept, x_1, x_2, x_3	0.298
Intercept, x_1	0.084
$\tau = 0.95$	
Predictors	Probability
Intercept, x_1, x_2, x_3	0.394
Intercept, x_1, x_2	0.383
Intercept, x_1	0.123

Table 4.1: Models visited by QR-SSVS with their estimated posterior probability. The top 3 models are displayed for $\tau \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$.

QR-SSVS following a 1,000 iteration burn in. Table 4.1 presents the 3 most visited models at each quantile $\tau \in \{0.05, 0.25, 0.5, 0.75, 0.95\}$. Again, x_1 denotes air flow, x_2 denotes water temperature and x_3 is acid concentration.

The first observation is that the models selected when $\tau = 0.25$, $\tau = 0.5$ and $\tau = 0.75$ are ranked in the same order and have similar posterior probabilities. This might suggest that the data exhibit homoscedasticity in this region although this would have to be verified. One way to do this would be to examine posterior parameter estimates $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and see if they are roughly equal for these values of τ. At the extreme quantiles $\tau = 0.05$ and $\tau = 0.95$, the models ranked first and second have roughly equal posterior probabilities. This is typical for extreme

$a_0 = b_0 = 1$	
Predictors	Probability
Intercept, x_1, x_2	0.564
Intercept, x_1, x_2, x_3	0.356
Intercept, x_1	0.065
$a_0 = 3$, $b_0 = 6$	
Predictors	Probability
Intercept, x_1, x_2	0.710
Intercept, x_1	0.146
Intercept, x_1, x_2, x_3	0.125

Table 4.2: Models visited by QR-SSVS at $\tau = 0.5$ with their estimated posterior probability. The top 3 models are displayed for the hyperpriors $\pi_0 \sim \text{Beta}(1, 1)$ and $\pi_0 \sim \text{Beta}(3, 6)$ respectively.

quantiles and is due to the frequent lack of information provided in the tails of the conditional distribution. To see this, observe that the top 3 models account for over 97% of the total posterior probability when $\tau = 0.25$, $\tau = 0.5$, and $\tau = 0.75$ whereas when $\tau = 0.05$, the top 3 models account for only around 69% of the total posterior probability.

The effect of including prior information about the model size is now investigated. For this analysis, τ is set equal to 0.5 i.e. there is interest in discovering plausible models for the conditional median. This analysis has just been done using $\pi_0 \sim \text{Beta}(a_0, b_0)$ with $a_0 = b_0 = 1$. This is equivalent to a uniform prior on the prior probability of selecting a predictor. Suppose instead that $a_0 = 3$ and $b_0 = 6$ in the hyperprior for π_0. This density has a single mode at $\frac{2}{7}$. Alternatively, it is equivalent to saying that the expected model size $p_{\boldsymbol{\gamma}}$ is 1 with variance equal to 0.8. As a consequence, it can be viewed as including prior knowledge that models with a smaller number of predictors are more plausible. The results of using the 2 priors are compared in Table 4.2.

The effect of this prior knowledge has resulted in increasing the estimated posterior probability of the model that was ranked in first position under a non-informative prior from 0.564 to 0.710. This is likely to be due to the fact that using an informative prior on the model size reduces the number of alternative

models that are supported by the prior. Secondly, the effect of imposing sparsity in the range of plausible models can clearly be seen. The full model was ranked in second position under the non-informative prior with an estimated posterior probability of 0.356. However, when imposing sparsity, the full model was ranked in third place with an estimated posterior probability of 0.125. In contrast, the model containing only x_1 (air flow) was ranked in third place with an estimated posterior probability of 0.065 under the non-informative prior, but was ranked in second place with an estimated posterior probability of 0.146 when imposing sparsity. The conclusion is that an informative hyperprior on π_0 can be used to guide the QR-SSVS procedure depending on what models are thought to be most plausible before observing the data.

4.3 Application to Boston Housing data

To illustrate QR-SSVS on a larger dataset, consider the Boston housing data of Harrison and Rubinfeld (1978). The corrected data consists of $n = 506$ observations and $p = 16$ potential predictors of interest. These are the tract point latitudes/longitudes in decimal degrees (LAT/LON), the per capita crime (CRIM), the proportions of residential land zoned for lots over 25,000 square feet per town (ZN), the proportions of non-retail business acres per town (INDUS), whether or not the tract borders the Charles river (CHAS), nitric oxide concentration (parts per 10 million) per town (NOX), average number of rooms per dwelling (RM), the proportions of owner occupied units built prior to 1940 (AGE), the weighted distances to 5 Boston employment centres (DIS), the index of accessibility to radial highways per town (RAD), the full value property tax rate in 10,000s of US dollars per town (TAX), pupil to teacher ratios per town (PTRATIO), $1,000(\text{proportion of black people} - 0.63)^2$ (B) and the percentage values of lower status population (LSTAT). The response variable is CMEDV, the corrected median values of owner occupied housing in 1,000s of US dollars.

This data was analysed using the hyperprior representing ignorance, $\pi_0 \sim$

$\tau = 0.5$					
	MIP	Posterior median	95% credible interval	Frequentist estimate	95% rank interval
LON	0.996	-0.563	(-0.960, -0.163)	-0.495	(-0.989, -0.287)
LAT	0.867	0.194	(-0.057, 0.505)	0.249	(-0.007, 0.504)
CRIM	0.998	-0.953	(-1.399, -0.271)	-1.203	(-1.282, -0.222)
ZN	0.998	0.728	(0.224, 1.198)	0.835	(0.422, 1.150)
INDUS	0.748	0.000	(-0.566, 0.355)	0.006	(-0.405, 0.302)
CHAS	0.983	1.036	(-0.023, 2.259)	0.942	(0.367, 2.108)
NOX	0.978	-0.651	(-1.309, 0.000)	-0.682	(-1.213, -0.059)
RM	1.000	3.534	(2.893, 4.193)	3.516	(2.619, 4.335)
AGE	0.987	-0.617	(-1.163, -0.013)	-0.637	(-1.052, -0.173)
DIS	1.000	-1.784	(-2.406, -1.163)	-1.909	(-2.525, -1.379)
RAD	1.000	1.482	(0.592, 2.346)	1.733	(0.962, 2.426)
TAX	1.000	-1.917	(-2.721, -0.999)	-2.099	(-2.630, -1.312)
PTRATIO	1.000	-1.428	(-1.828, -1.011)	-1.485	(-1.742, -1.051)
B	1.000	1.096	(0.746, 1.445)	1.085	(0.848, 1.496)
LSTAT	1.000	-2.281	(-2.961, -1.607)	-2.254	(-2.904, -1.620)

Table 4.3: Marginal inclusion probabilities (MIPs), posterior summaries and corresponding frequentist estimates (based on the full model) of the Boston Housing data, presented for $\tau = 0.5$.

Beta$(1, 1)$. Just as before, each predictor was standardised before analysis except the intercept, which is assumed to appear in all models. The results this time are based on 50,000 iterations of QR-SSVS following a 5,000 iteration burn in due to the larger number of candidate predictors. Table 4.3 presents the marginal inclusion probabilities (MIPs), the posterior summaries and the frequentist results for comparison. The MIP for a predictor j is given by $\pi(\gamma_j = 1|\mathbf{y})$ and can be estimated by the proportion of occasions that $\gamma_j = 1$ during the QR-SSVS run. The 95% credible interval reported is the central 95% credible interval estimated by calculating the sample 2.5% and 97.5% quantiles and not the HPD as this is difficult to find given the enormity ($p = 2^{16} = 65,536$) of the model space. The frequentist analysis was based on fitting the full model to the data.

Note from Table 4.3 that many predictors appear in all models visited by QR-SSVS when $\tau = 0.5$. The median probability model, as defined by Barbieri and Berger (2004) includes all predictors. The posterior estimates and credible inter-

$\tau = 0.05$		
Predictors	Model Size	Probability
LON, LAT, CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT	15	0.268
LON, CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT	14	0.053
LON, LAT, CRIM, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT	14	0.052
LON, LAT, CRIM, ZN, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT	14	0.043
LON, LAT, CRIM, ZN, INDUS, CHAS, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT	14	0.025
$\tau = 0.5$		
Predictors	Model Size	Probability
LON, LAT, CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT	15	0.634
LON, LAT, CRIM, ZN, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT	14	0.186
LON, CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT	14	0.078
LON, CRIM, ZN, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT	13	0.044
LON, LAT, CRIM, ZN, INDUS, CHAS, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT	14	0.013

Table 4.4: The 5 models with the highest estimated posterior probability. The results are presented for $\tau = 0.05$ and $\tau = 0.5$.

vals are similar to the frequentist estimates and rank intervals. This is perhaps unsurprising given that the majority of predictors are important in the median regression model. In Table 4.4, the top 5 models for $\tau = 0.05$ and $\tau = 0.5$ are presented. In this case, the highest probability model for both values of τ is the same, but the associated probability changes for different quantiles. Observe again the larger model uncertainty when $\tau = 0.05$, with only 44.1% of the posterior probability accounted for by the top 5 models. In contrast, for median regression where $\tau = 0.5$, the top 5 models account for 95.5% of the total posterior probability.

4.4 Summary

The work done in this chapter has allowed researchers to fully take into account the model uncertainty inherent to regression analysis and apply it to quantile regression under a Bayesian framework. It is likely that posterior model probabilities may vary across quantiles and the models achieving the highest posterior probabilities may also vary. QR-SSVS can thus be used as a tool to rank models in order of posterior probabilities at each quantile. By examining MIPs, QR-SSVS can help uncover predictors that globally affect all quantiles and those that only affect quantiles locally.

Chapter 5

Conclusions and Future Research

5.1 A Summary of this Thesis

The principal aim of this work has been to make the Bayesian approach to quantile regression straightforward for applied researchers. The bonus of using the AL likelihood of Yu and Moyeed (2001) is that the resulting Bayesian quantile regression model can be converted into a normal regression model with latent variables. This has many advantages, including, for example, the ability to use Rao-Blackwellisation to approximate the marginal posterior density (see Held (2004) for an example of an application) and the ability to easily extend the model to handle random effects, non parametric regression using splines and covariate set uncertainty, to name but a few.

It has been shown that although the posterior mode under an improper prior has a direct correspondence to the frequentist procedure, the posterior mean and median are also close to the mode and are more readily available from the Gibbs sample. Alternatively, Rao-Blackwellised estimates are available for the posterior mean. Using Gibbs sampling on the augmented posterior distribution has been shown to be more efficient than MH on the marginal distribution, particularly when fitting natural cubic splines.

A variant of this Gibbs sampling approach was independently investigated by Kozumi and Kobayashi (2009) who used a different parameterisation than that appearing in this thesis. They also found an increase in efficiency using the

Gibbs sampler on the augmented space. On realising the similarity between the two approaches, a comparison was made between the progress of this work and of the work of Kozumi and Kobayashi (2009). It was felt that this work was at a more advanced stage and so Kozumi and Kobayashi (2009) were invited to combine their contribution to this work. This proposal was accepted, resulting in the joint manuscript (Reed et al., 2010) that has been submitted to the Journal of Computational Statistics and Data Analysis and is awaiting a small revision.

Re-analysing the St George's birthweight study has confirmed the findings of the original study by Peacock et al. (1998), but additional insight has been gained into whether or not the metabolite cotinine is a more accurate predictor of low birthweigh infants than just the reported number of cigarettes and whether or not changing smoking habits during pregnancy or second hand smoke have an effect on the chances of having an underweight infant. It has been observed through quantile regression that these factors seem to have a stronger effect on the lower portion of the conditional birthweight distribution suggesting that any exposure to smoke, whether it is active or passive can increase the likelihood of having an underweight infant.

By taking account of model uncertainty in quantile regression, there is the possibility of choosing a single model as the "best" model, for example by taking the model with the highest estimated posterior probability or to take the model whose predictors have a marginal inclusion probability of greater than 0.5 (the median probability model, using the terminology of Barbieri and Berger (2004)). Alternatively, it allows the possibility of model averaging to fully take account of the model uncertainty inherent in all regression problems.

Finally, two of the main Gibbs sampling algorithms have been implemented in `R` for the package `MCMCpack`. These algorithms use an `R` interface that is very similar to what users would find using the `lm` command. As a result, anyone who is familiar with regression in `R` could use these functions easily. Whilst not having the flexibility of programs like WinBUGS, the samplers are hand crafted and are

consequently more efficient and faster. This is important when multiple quantile regressions are of interest.

It is important to again emphasise that the AL likelihood is a "pseudo" likelihood providing a bridge between the Bayesian and the classical approach to quantile regression. It is not thought to be an accurate representation of a true data generating likelihood. Topics that require further research include how accurate any credible intervals are when the assumption of the AL likelihood is violated. Posterior model probabilities and marginal inclusion probabilities may also be called into question in this situation. There are, however, extensions that may offer some improvements. These are discussed next.

5.2 Extensions

5.2.1 Shape parameter σ

A small improvement in flexibility can be made by incorporating a scale parameter σ. This now results in a likelihood $l(\mathbf{y}|\boldsymbol{\beta}, \sigma)$ given as

$$\tau^n(1-\tau)^n\sigma^{-n}\exp\left\{-\sigma^{-1}\sum_{i=1}^{n}\rho_\tau(y_i - \mathbf{x_i}^T\boldsymbol{\beta})\right\}, \tag{5.1}$$

and can be obtained from the model $y_i = \mathbf{x_i}^T\boldsymbol{\beta} + \sigma\epsilon_i$, where ϵ_i has the standard AL distribution with skewness τ.

The appropriate prior for $\mathbf{w}$ in the specification including the latent variables now depends on σ and is given as $\pi(\mathbf{w}|\sigma) = \prod_{i=1}^{n}\pi(w_i|\sigma)$, where

$$\pi(w_i|\sigma) \propto \sigma^{-1}w_i^{-2}\exp(-\tfrac{1}{8\sigma}w_i^{-1}), \tag{5.2}$$

so that $\pi(\mathbf{w}|\sigma)$ is the product of inverse Gamma densities with parameters $(1, (8\sigma)^{-1})$. The joint prior $\pi(\boldsymbol{\beta}, \sigma)$ can be specified by using $\pi(\boldsymbol{\beta}, \sigma) = \pi(\boldsymbol{\beta}|\sigma)\pi(\sigma)$. An improper prior for beta would yield a marginal posterior mode corresponding to the frequentist quantile regression estimate (if it is unique). Specifying a proper prior that is dependent on σ yields a marginal posterior mode that corresponds to the

regularised QR estimate. The normal prior is semi-conjugate

$$\pi(\boldsymbol{\beta}|\sigma) \propto \exp\left\{-\frac{1}{2\sigma}(\boldsymbol{\beta}-\mathbf{b_0})^T\mathbf{B_0}(\boldsymbol{\beta}-\mathbf{b_0})\right\}. \tag{5.3}$$

The inverse Gamma distribution is the semi conjugate prior for σ. Thus, with prior hyperparameters c_0 and d_0, this model is completed by specifying

$$\pi(\sigma) \propto \sigma^{-c_0-1}\exp(-d_0\sigma^{-1}). \tag{5.4}$$

This also includes the improper Jeffrey's prior $\pi(\sigma) \propto \sigma^{-1}$ obtained by letting c_0 and d_0 tend to 0 although whether or not this yields a proper posterior has yet to be verified.

Although there is now an additional parameter to update in the Gibbs sampler, it is still possible to construct a Gibbs sampler with only two steps. This is done by noting that the marginal conditional posterior $\pi(\sigma|\mathbf{w},\mathbf{y})$ is also inverse Gamma. This means that the joint conditional $\pi(\boldsymbol{\beta},\sigma|\mathbf{w},\mathbf{y})$ is available to sample from directly. This algorithm is summarised in Algorithm (5.1).

5.2.2 Multiple values of τ

For the purposes of using quantile regression to characterise the complete conditional distribution of $\mathbf{y}|\mathbf{x}$, analysis at more than one value of τ is required. The approach of Yu and Moyeed (2001) is to simply repeat the procedure fixing τ at different values. This has also been the approach adopted throughout this thesis. However, it is effectively fitting different likelihoods to the same data and it could be argued that at least one of the likelihoods is misspecified. One way to address this criticism is if τ is allowed to be a discrete random variable taking values τ_1, τ_2, $\cdots$, τ_S with probabilities p_1, p_2, ..., p_S respectively. The result is a marginal likelihood $l(\mathbf{y}|\boldsymbol{\beta},\sigma)$ given by

$$\sum_{s=1}^{S}\left\{\prod_{i=1}^{n} p_s \mathrm{AL}(\mathbf{x_i}^T\boldsymbol{\beta},\sigma,\tau_s)\right\}. \tag{5.5}$$

In this case, the approach of Yu and Moyeed (2001) can be viewed as finding conditional posterior distributions given $\tau=\tau_s$ for different values of s when the

Algorithm 5.1 Gibbs sampler for augmented quantile regression model with shape parameter σ. Draws M burn in samples followed by an additional N samples for inference.

Given: Prior mean vector $\mathbf{b_0}$, prior precision matrix $\mathbf{B_0}$ and initial values $\mathbf{w}^{(0)}$.

for $k = 1$ **to** $M + N$ **do**

• Sample $\sigma^{(k)}|\mathbf{w}^{(k-1)}, \mathbf{y}$ from an inverse Gamma distribution with shape $c_1 = c_0 + \frac{3n}{2}$ and scale

$$\begin{aligned} d_1 &= \mathbf{y}^T\mathbf{W}^{(k-1)}\mathbf{y} + \mathbf{b_0}^T\mathbf{B_0}\mathbf{b_0} + 2\sum_{i=1}^{n}\left(y_i - \frac{1}{8w_i^{(k-1)}}\right) \\ &- \{\mathbf{v}^{(k-1)}\}^T(\mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{X} + \mathbf{B_0})\mathbf{v}^{(k-1)}, \end{aligned}$$

where $\mathbf{W}^{(k-1)}$ is a diagonal matrix with the elements of $\mathbf{w}^{(k-1)}$ forming the diagonal and

$$\mathbf{v}^{(k-1)} = \mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{y} + (\tau - \tfrac{1}{2})\mathbf{X}^T\mathbf{1} + \mathbf{B_0}\mathbf{b_0}.$$

• Sample $\boldsymbol{\beta}^{(k)}|\sigma^{(k)}, \mathbf{w}^{(k-1)}, \mathbf{y}$ from the multivariate normal distribution with precision matrix

$$\frac{\mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{X} + \mathbf{B_0}}{\sigma^{(k)}}$$

and mean vector

$$(\mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{X} + \mathbf{B_0})^{-1}(\mathbf{X}^T\mathbf{W}^{(k-1)}\mathbf{y} + (\tau - \tfrac{1}{2})\mathbf{X}^T\mathbf{1} + \mathbf{B_0}\mathbf{b_0}).$$

• Sample $\mathbf{w}^{(k)}|\boldsymbol{\beta}^{(k)}, \sigma^{(k)}, \mathbf{y}$ by sampling the ith component of $\mathbf{w}$ $(i = 1, \ldots, n)$ from the inverse Gaussian distribution with shape parameter $\frac{1}{4}\{\sigma^{(k)}\}^{-1}$ and location $\frac{1}{2}|y_i - \mathbf{x_i}^T\boldsymbol{\beta}^{(k)}|^{-1}$.

end for.

underlying marginal likelihood is (5.5). However, likelihoods such as (5.5) that are discrete mixtures of densities can be notoriously tricky to analyse (see e.g. Diebolt and Robert (1994)).

Of course, it is not necessary for τ to be a discrete random variable. If τ is allowed to be continuous with a prior density $\pi(\tau)$ on $(0, 1)$, then this could be viewed as the Bayesian analogue of the quantile regression process. The marginal likelihood then generalises to

$$\int \left\{ \prod_{i=1}^{n} \mathrm{AL}(\mathbf{x_i}^T \boldsymbol{\beta}, \sigma, \tau) \right\} \pi(\tau) \mathrm{d}\tau. \tag{5.6}$$

Whether τ is discrete or continuous, there is the possibility of allowing the joint prior for $\boldsymbol{\beta}$ and σ to depend on τ. Priors could be used here to enforce the condition of monotonicity or in cases where there is more prior knowledge about the median than there is the tails of the distribution or vice versa.

5.2.3 Prediction

The advantage of prediction in the Bayesian framework over the frequentist approach is that it is possible for the predictive uncertainty to account for the uncertainty involved in the parameter estimation. Of course, predicting a new $y*$ given $\mathbf{x}*$ and the data $\mathbf{y}$ depends on the likelihood. Given that the AL likelihood is not really believed to have generated the data, predictive inference could be misleading. However, by allowing τ to be random, the marginal likelihood becomes much more flexible and predictive inference could potentially be implemented from this model taking into account the uncertainty involved in estimating all conditional quantiles. Prediction need not be at a new value of $\mathbf{x}*$, it could also be used in cross validation to assess model fidelity. An example of this is leave one out cross validation, in which a random data point is excluded from the data analysis and is then predicted. The mean square error is one possibility to compare the predicted value to the actual observed value. Given the criticisms of using the AL distribution, this will be a handy tool in what appears to be a key area of future

research.

5.2.4 Posterior mode using the EM algorithm

An interesting by-product of the work in Chapter 2 is that it allows an alternative to linear programming in order to find the posterior mode under an AL likelihood. As an illustration, consider the case of estimating the τth quantile from a sample $\mathbf{y}$ of size n, assuming it can be uniquely determined. Of course, in a simple case like this, the τth sample quantile can be deduced by ordering the data $\mathbf{y}$. An alternative approach would be to use the Expectation Maximising (EM) algorithm introduced by Dempster et al. (1977).

Proceeding as in the previous chapters, adopting the AL likelihood (1.9) for a fixed τ with location parameter μ and combining with the improper prior $\pi(\mu) \propto 1$ yields the posterior distribution with the mode at the τth sample quantile of $\mathbf{y}$. In Chapter 2, it was shown how to express the AL likelihood as a mixture of normals. As illustrated in Dempster et al. (1977), the EM algorithm is then equivalent to an iteratively reweighted least squares procedure.

Starting with an initial value μ_0, the expectation step is to find the posterior expectation q_i of each latent variable w_i conditional on the data $\mathbf{y}$ and μ_0. It was shown in Chapter 2 that the conditional posterior distribution of w_i is inverse Gaussian with scale parameter $\frac{1}{4}$ and location parameter $\frac{1}{2}|y_i - \mu|^{-1}$. The expectation of an inverse Gaussian random variable is equal to the location parameter so that

$$q_i = \tfrac{1}{2}|y_i - \mu|^{-1}. \tag{5.7}$$

Thus, using μ_0 in place of μ for the conditional expectations q_i in (5.7), a new estimate μ_1 can be obtained from the maximisation step. Due to the conjugacy of the prior, the maximum is available in closed form as

$$\mu_1 = \frac{\sum_{i=1}^{n} q_i y_i + n(\tau - \frac{1}{2})}{\sum_{i=1}^{n} q_i}. \tag{5.8}$$

The expectation and maximisation steps are then iterated to form μ_2, μ_3, ...

Algorithm 5.2 The EM algorithm for finding the posterior mode under a $\mathrm{N}(\mathbf{b_0}, \mathbf{B_0}^{-1})$ prior where each y_i has an AL distribution with skewness τ and location $\mathbf{x_i}^T\boldsymbol{\beta}$.

Given: Prior mean vector $\mathbf{b_0}$, prior precision matrix $\mathbf{B_0}$ and initial values $\boldsymbol{\beta}^{(0)}$. Set $k = 0$ and define a convergence criterion.

Repeat:

- Calculate $q_i^{(k)} = \frac{1}{2}|y_i - \mathbf{x_i}^T\boldsymbol{\beta}^{(k)}|$ for $i = 1, \ldots, n$ and form the $n \times n$ diagonal matrix $\mathbf{Q}^{(k)}$ such that $\mathbf{Q}_{i,i}^{(k)} = q_i^{(k)}$.
- Update $\boldsymbol{\beta}$ according to

$$\boldsymbol{\beta}^{(k+1)} = \{\mathbf{B_1}^{(k)}\}^{(-1)}(\mathbf{X}^T\mathbf{Q}^{(k)}\mathbf{y} + (\tau - \tfrac{1}{2})\mathbf{X}^T\mathbf{1_n} + \mathbf{B_0}\mathbf{b_0}),$$

where

$$\mathbf{B_1}^{(k)} = \mathbf{X}^T\mathbf{Q}^{(k)}\mathbf{X} + \mathbf{B_0}.$$

- Set $k \leftarrow k + 1$.

Until Convergence criterion is satisfied.

until some convergence criterion is satisfied. Interestingly, it turns out that this particular algorithm is exactly the same algorithm as that presented in Hunter and Lange (2000). Algorithm (5.2) describes the procedure for the general case where it is assumed that each y_i come from an AL distribution with location parameter $\mathbf{x_i}^T\boldsymbol{\beta}$ with the semi conjugate prior

$$\boldsymbol{\beta} \sim \mathrm{N}(\mathbf{b_0}, \mathbf{B_0}^{-1}).$$

The hyperparameters $\mathbf{b_0}$ and $\mathbf{B_0}$ are assumed fixed.

Provided that the prior is conjugate normal and that there is a unique posterior mode, Algorithm 5.2 can be used to find it. An example of where the mode is not unique is when n is even so that there are an even number of data points $\mathbf{y}$ and the sample median ($\tau = 0.5$) is of interest. In this case, the posterior distribution under the AL likelihood and an improper prior has a plateau in which there is a range of values $[\mu_{\min}, \mu_{\max}]$ for which the posterior density is maximised. The theory of the EM algorithm (e.g. Dempster et al. (1977)) suggests that for this example, it will converge to any value in $[\mu_{\min}, \mu_{\max}]$.

5.3 Recommendations

To conclude this thesis, some recommendations for future work are in order. The previous section has outlined a few of the potential extensions to the Bayesian quantile regression models. There are many other extensions not mentioned here such as Bayesian nonlinear quantile regression. The main drawback with any extension to the model is that in order to have an efficient MCMC algorithm for sampling the posterior distribution, it is usually necessary to code it or parts of it from scratch as was done in `MCMCpack`. In order to implement some of these extensions in practice, it would be useful to use the data augmentation approach in WinBUGS or JAGS and for them to recognise the efficient sampling strategies that exist for these models. This would give the Bayesian quantile regression models additional flexibility and would make it easier to investigate their performance in applied settings.

Bibliography

Abrevaya, J. (2001). The effects of demographics and maternal behavior on the distribution of birth outcomes. *Empirical Economics*, 26:247–257.

Abrevaya, J. and Dahl, C. M. (2008). The effects of birth inputs on birthweight: Evidence from quantile estimation on panel data. *Journal of Business and Economic Statistics*, 26:379–397.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society Series B (Methodological)*, 36(1):99–102.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *Annals of Statistics*, 32:870–897.

Billias, Y., Chen, S., and Ying, Z. (2000). Simple resampling methods for censored regression quantiles. *Journal of Econometrics*, 99(2):373–386.

Bland, J., Peacock, J. L., Anderson, H., Brooke, O., and Curtis, M. D. (1990). The adjustment of birthweight for very early gestational ages: Two related problems in statistical analysis. *Journal of the Royal Statistical Society Series C Applied Statistics*, 39(2):229–239.

Bose, A. and Chatterjee, S. (2003). Generalized bootstrap for estimators of minimizers of convex functions. *Journal of Statistical Planning and Inference*, 117:225–239.

Brownlee, K. A. (1960). *Statistical Methodology in Science and Engineering*, chapter 17. Wiley.

Cade, B. S. and Noon, B. R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1:412–420.

Chamberlain, G. (1994). *Quantile Regression, Censoring and the Structure of Wages*. Elsevier, New York.

Chamberlain, G. and Imbens, G. W. (2003). Nonparametric applications of Bayesian inference. *Journal of Business and Economic Statistics*, 21:12–18.

Chen, C. and Wei, Y. (2005). Computational issues for quantile regression. *Sankhyā*, 67:399–417.

Chen, C. and Yu, K. (2009). Automatic Bayesian quantile regression curve fitting. *Statistics and Computing*, 19(3):271–281.

Chhikara, R. S. and Folks, L. (1989). *The Inverse Gaussian Distribution: Theory, Methodology and Applications*. Marcel Dekker, New York.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90:1313–1321.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, 39(1):1–38.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society Series B*, 56(2):363–375.

Dongarra, J., Moler, C., Bunch, J., and Stewart, G. (1979). *Linpack Users' Guide*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

Dunson, D. B. and Taylor, J. A. (2005). Approximate Bayesian inference for quantiles. *Journal of Non-parametric Statistics*, 17:385–400.

Engel, R. F. and Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles. *Journal of Business and Economic Statistics*, 22:367–381.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulations using multiple sequences. *Statistical Science*, 7:457–511.

George, E. I. and Mcculloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 7:339–373.

Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using the asymmetric Laplace distribution. *Biostatistics*, 8:140–154.

Green, P. J. and Silverman, B. W. (1994). *Non Parametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman and Hall.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.

He, X. and Hu, F. (2002). Markov chain marginal bootstrap. *Journal of the American Statistical Association*, 97(459):783–795.

Held, L. (2004). Simutaneous posterior probability statements from Monte Carlo output. *Journal of Computational and Graphical Statistics*, 13(1).

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14:382–401.

Hunter, D. R. and Lange, K. (2000). Quantile regression via an MM algorithm. *Journal of Computational and Graphical Statistics*, 9(1):60–77.

Jeffreys, H. (1961). *Theory of Probability*. Clarendon Press, Oxford.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kocherginsky, M., He, X., and Mu, Y. (2005). Practical confidence intervals for regression quantiles. *Journal of Computational and Graphical Statistics*, 14:41–55.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, London.

Koenker, R. (2009). *quantreg: Quantile Regression*. R package version 4.44.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1):35–50.

Koenker, R. and Bassett, G. (1982). Robust tests of heteroscedasticity based on regression quantiles. *Econometrica*, pages 43–61.

Koenker, R. W. (1994). Confidence intervals for regression quantiles. In Mandl, P. and Huskova, M., editors, *Asymptotic Statistics*, pages 349–359. Springer-Verlag, New York.

Komunjer, I. (2005). Quasi-maximum likelihood estimation for conditional quantiles. *Journal of Econometrics*, 128:137–164.

Kottas, A. and Gelfand, A. E. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, 96:1458–1468.

Kottas, A. and Krnjajić, M. (2009). Bayesian semiparametric modeling in quantile regression. *Scandinavian Journal of Statistics*, 36:297–319.

Kotz, S., Kozubowski, T., and Podgórski, K. (2001). *The Laplace Distribution and Generalizations: A Revisit with Applications to Communications, Economics, Engineering and Finance*. Birkhäuser.

Kozumi, H. and Kobayashi, G. (2009). Gibbs sampling methods for Bayesian quantile regression. Technical report, Kobe University.

Lancaster, T. and Jun, S. J. (2010). Bayesian quantile regression methods. *Journal of Applied Econometrics*, 25(2):287–307.

Levin, J. (2001). For whom the reduction counts: A quantile regression analysis of class size on scholastic achievement. *Empirical Economics*, 26:221–246.

Li, Q., Xi, R., and Lin, N. (2010). Bayesian regularized quantile regression. *Bayesian Analysis*, 5:533–556.

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

Liu, J. S., Wong, W. H., and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40.

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: Concepts, structure and extensibility. *Statistics and Computing*, 10:325–337.

Madigan, D. M. and Raftery, A. E. (1994). Models selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association*, 89:1335–1346.

Martin, A. D., Quinn, K. M., and Park, J. H. (2010). *MCMCpack: Markov chain Monte Carlo (MCMC) Package*. R package version 1.0-6.

Michael, J. R., Schucany, W. R., and Haas, R. W. (1976). Generating random variates using transformations with multiple roots. *The American Statistician*, 30:88–90.

Min, I. and Kim, I. (2004). A monte carlo comparison of parametric and non-parametric quantile regressions. *Applied Economic Letters*, 11:71–74.

Park, T. and van Dyk, D. A. (2009). Partially collapsed Gibbs samplers: Illustrations and applications. *Journal of Computational and Graphical Statistics*, 18(2):283–305.

Parzen, M. I., Wei, L., and Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika*, 81:341–350.

Peacock, J. L., Cook, D. G., Carey, I. M., Jarvis, M. J., Bryant, A. E., Anderson, H. R., and Bland., J. M. (1998). Maternal cotinine level during pregnancy and birthweight for gestational age. *International Journal of Epidemiology*, 27(4):647–656.

Pemstein, D., Quinn, K. M., and Martin, A. D. (2007). *Scythe Statistical Library*. Version 1.0.2.

Plummer, M. (2004). *JAGS: Just Another Gibbs Sampler*.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2010). *coda: Output analysis and diagnostics for MCMC*. R package version 0.13-5.

R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Reed, C., Yu, K., Kozumi, H., and Kobayashi, G. (2010). Efficient Gibbs sampling for Bayesian quantile regression. Technical report, Brunel University.

Sarkar, D. (2010). *lattice: Lattice Graphics*. R package version 0.18-5.

Schennach, S. M. (2005). Bayesian exponentially tilted empirical likelihood. *Biometrika*, 92(1):31–46.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric curve fitting. *Journal of the Royal Statistical Society Series B*, 47.

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

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

Tsionas, E. G. (2003). Bayesian quantile inference. *Journal of Statistical Computation and Simulation*, 73:659–674.

van Dyk, D. A. and Park, T. (2008). Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103:790–796.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648.

Yu, K., Lu, Z., and Stander, J. (2003). Quantile regression: Applications and current research areas. *The Statistician*, 137:331–350.

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

Yu, K. and Stander, J. (2007). Bayesian analysis of a tobit quantile regression model. *Journal of Econometrics*, 137:260–276.

Appendix A

Practical Implementation in R

A.1 Introduction

Historically, Bayesian inference was only possible for a handful of simple models. With the power of today's computers and MCMC algorithms in abundance, it should be possible to conduct Bayesian inference on virtually all models regardless of their complexity. However, there remains little in the way of publicly available software for fitting Bayesian models using MCMC particularly compared to software for implementing frequentist methods.

For a researcher wishing to fit a Bayesian model, there are two options. The first of these is to use a variant of the BUGS software such as WinBUGS (Lunn et al., 2000), a program that is extremely versatile and can fit a range of models with only a small learning curve. However, this flexibility can come at a price. The researcher cannot guarantee that WinBUGS is using the most efficient MCMC algorithm to fit their model and it can be fairly slow. Time is required to organise the data into a format that can be recognised by WinBUGS. It also allocates memory to each parameter in the model, which can be inefficient in larger models.

Another program similar to BUGS is JAGS (Plummer, 2004). JAGS is based on compiled C++ code and is faster than WinBUGS. Additionally, it reads data in the same format as R. Additional modules can be added to JAGS to give it a greater array of samplers. This makes it more likely (although not guaranteed) that JAGS will use an efficient sampler. Just like WinBUGS, this program will

allocate memory to each parameter in a model, again making it potentially inefficient.

The second option is to code the algorithm in R (R Development Core Team, 2010). A main provider of MCMC algorithms for R is the package MCMCpack (Martin et al., 2010) which contains MCMC algorithms to fit a variety of different models including mean regression models, probit models, poisson change point and item response theory. The algorithms in MCMCpack use compiled C++ code using the Scythe statistical library (Pemstein et al., 2007) to do the bulk of the calculation and are hand crafted making them fast and efficient. This does mean that they lack the flexibility of WinBUGS and JAGS, often requiring conjugate prior distributions. Nevertheless, for quantile regression, particularly when several quantile regressions are of interest, the gains in speed and efficiency of using the functions in MCMCpack can be considerable.

In this section, a small tutorial is provided on the use of MCMCquantreg and SSVSquantreg that has been used to do all analysis reported in this thesis. Previous versions of these functions have been successfully included in MCMCpack.

A.2 Using Gibbs Sampling for Bayesian Quantile Regression in R

A.2.1 MCMCquantreg

Conducting Bayesian inference in R using the MCMCpack functions is much the same as using the lm function for frequentist linear mean regression as the syntax is very similar. The package MCMCpack can be installed by using the command

```
> install.packages("MCMCpack")
```

and choosing a mirror to download from. On completion of this, the package can be made available to the R session with the command

```
> library(MCMCpack)
```

Once the library has been loaded, it is possible to reproduce the various analyses in Chapter 2 and Chapter 4. In Section 2.2, Engel's data was used to compare augmented posterior summaries and the marginal posterior mode. The following commands implement Bayesian median regression on Engel's data with an improper prior on the the regression parameters $\boldsymbol{\beta}$ (assuming that the `quantreg` package is installed) and stores the results in `medfit`.

```
> library(quantreg)
> library(MCMCpack)
> data(engel)
> medfit <- MCMCquantreg(foodexp ~ income, data = engel)
```

Now suppose that regression at the 90th centile is of interest. It may be felt by the researcher that, given the frequent lack of data in the tails of the distribution, that a larger sample is needed to get an "appropriate" level of convergence. This is easily accomplished by altering the `mcmc` and the `burnin` options. For example, to get a sample of size 100,000 from the posterior ditribution following a burn in of 10,000 samples at $\tau = 0.9$, the command would be

```
> fit90pc <- MCMCquantreg(foodexp ~ income, data = engel, mcmc = 1e+05,
+     burnin = 10000, tau = 0.9)
```

By default, `MCMCquantreg` uses a different seed for each simulation. This is to ensure that the simulation will be different when the value of τ is changed. This is in contrast to many of the other MCMC functions in `MCMCpack` where a seed would need to be specified with the `seed` argument to ensure a different seed was used.

Unlike any object resulting from the `lm` command, objects resulting from `MCMCquantreg` or any other functions in `MCMCpack` such as `medfit` cannot be summarised by just typing it's name. Doing so will just display the exhaustive list of all values that were simulated during the MCMC run. The `summary` command, on the other hand, produces the following output:

```
> summary(medfit)

Iterations = 1001:11000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 10000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

              Mean       SD  Naive SE Time-series SE
(Intercept) 82.7984 2.238199 2.238e-02      7.435e-02
income       0.5584 0.002411 2.411e-05      7.421e-05

2. Quantiles for each variable:

              2.5%     25%     50%   75%   97.5%
(Intercept) 79.0821 81.3337 82.5426 83.92 88.1813
income       0.5524  0.5572  0.5589  0.56  0.5619
```

This is the summary produced from the `coda` package (Plummer et al., 2010). This particular form of summary is defined for an object with class `mcmc` such as `medfit`. Many other summary statistics and plots can be obtained from the `coda` package. For example,

```
> plot(medfit)
```

will plot a traceplot of each variable, and a corresponding estimated density plot. Another common plot is the autocorrelation function. The command used to produce an autocorrelation plot like Figure 2.1 used in Section 2.3 is

```
> autocorr.plot(medfit)
```

In Section 2.4, the Gelman-Rubin statistic was calculated. To obtain this, a second chain is needed and both chains need to be contained in an R list with an mcmc.list class attribute.

```
> medfit2 <- MCMCquantreg(foodexp ~ income, data = engel)
> medfitlist <- mcmc.list(medfit, medfit2)
> gelman.diag(medfitlist)

Potential scale reduction factors:


            Point est. 97.5% quantile
(Intercept)       1.00           1.01
income            1.01           1.03


Multivariate psrf


1.01
```

Many other convergence diagnostics are also implemented in coda. To see a more comprehensive list of all available functions, type

```
> help(package = "coda")
```

A.2.2 SSVSquantreg

To demonstrate SSVSquantreg in action, consider the model described in Tibshirani (1996). The model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\epsilon$ where $\boldsymbol{\beta} = [3, 1.5, 0, 0, 2, 0, 0, 0]^T$ and σ is set equal to 1. The design matrix $\mathbf{X}$ in constructed in such a way that the correlation between column i and column j is approximately $0.5^{|i-j|}$. In this example, the conditional quantiles are parallel. If there is strong prior knowledge about this fact, then QR-SSVS is not really needed. Nevertheless, this example serves to illustrate the commands that can be used. The following R commands will produce this data with $n = 101$ data points.

```
> rhomatrix <- matrix(0, 8, 8)
> for (i in 1:8) {
+     for (j in 1:8) {
+         rhomatrix[i, j] <- 0.5^(abs(i - j))
+     }
+ }
> set.seed(1)
> standnorm <- matrix(rnorm(808), 101, 8)
> U <- chol(rhomatrix)
> x <- standnorm %*% U
> beta <- c(3, 1.5, 0, 0, 2, 0, 0, 0)
> set.seed(2)
> y <- x %*% beta + rnorm(101)
> xs <- scale(x)
> models50pc <- SSVSquantreg(y ~ xs)
```

Note the second to last step, which standardises the predictors prior to analysis except the intercept, which is added implicitly unless otherwise specified in the formula. Future versions of `SSVSquantreg` will automate this process. A product of spike and slab priors with the slab corresponding to a Cauchy distribution (see Chapter 4) is placed on the regression parameters $\boldsymbol{\beta}$ and a beta-binomial prior is assumed for $p_{\boldsymbol{\gamma}}$. The default values of the hyperparameters a_0 and b_0 are set to 1 but can be chosen by the researcher with the options `pi0a0` and `pi0b0`.

The results from using `SSVSquantreg`, in this case `models50pc`, are in the form of a list. The first component is `gamma`, which contains all the models that were visited by the SSVS algorithm and `beta`, which contains the sampled values of the model specific regression parameters. The `beta` component of `models50pc` can be analysed in the same way as the output from `MCMCquantreg`. For this reason, the rest of this tutorial is devoted to analysing the `gamma` component.

The gamma component of models50pc has a qrssvs class attribute. There exists methods to handle this class for print, summary and plot, as well as a couple of additional functions.

```
> summary(models50pc$gamma)

Marginal inclusion probability of each predictor:

            Probability
(Intercept)      0.4386
xs1              1.0000
xs2              1.0000
xs3              0.4105
xs4              0.5436
xs5              1.0000
xs6              0.3079
xs7              0.3553
xs8              0.3613

For tau = 0.5, the median probability model
includes the following predictors:

xs1, xs2, xs4, xs5.
```

R can identify components of a list just by the first letter that makes them unique. In this case,

```
> summary(models50pc$g)
```

would also produce the desired output.

As can be seen from above, the summary command will produce a table listing all the candidate predictors together with the estimated marginal inclusion

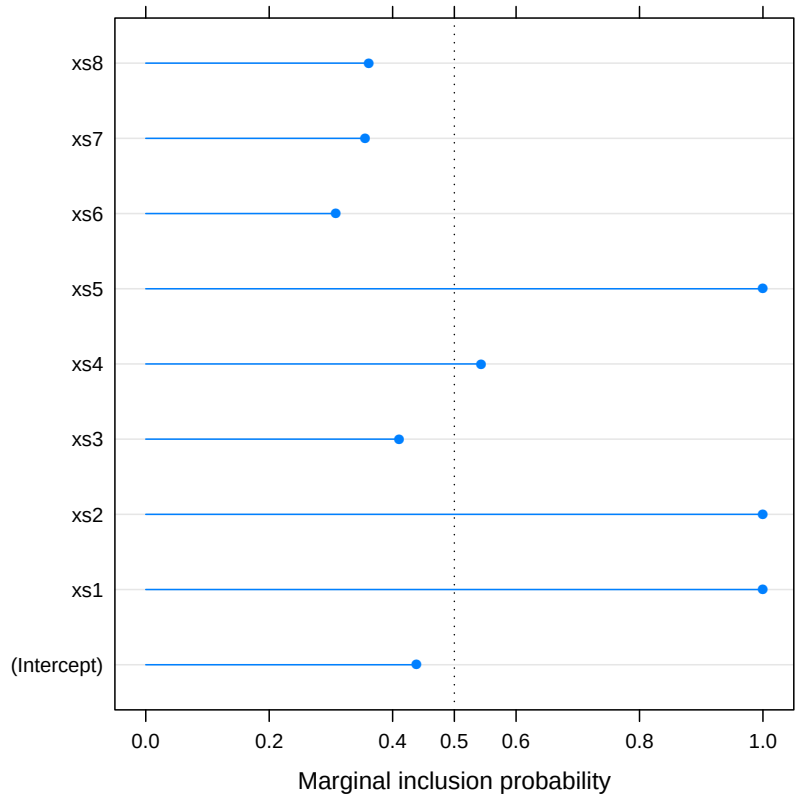

Figure A.1: R plot obtained by the `plot` function on an object of class `qrssvs`.

probabilities (MIPs). It will additionally provide the median probability model as defined in the previous section. The table can be extracted on it's own using the `mptable` command. It is also possible to plot these MIPs.

```
> print(plot(models50pc$g))
```

This plots the covariates on the y axis against the estimated MIPs on the x axis (see Figure A.1). It is produced as a trellis graph that can be manipulated in any way desired using the `lattice` package of Sarkar (2010).

The models that were visited most frequently can be displayed in a table using the `topmodels` command.

```
> topmodels(models50pc$g)
```

```
                                    Probability
xs1,xs2,xs5                              0.1058
```

```
xs1,xs2,xs4,xs5                                    0.0628
(Intercept),xs1,xs2,xs5                            0.0459
xs1,xs2,xs3,xs4,xs5                                0.0435
(Intercept),xs1,xs2,xs3,xs4,xs5,xs6,xs7,xs8        0.0343
```

The default behaviour is to produce the top 5 visited models although this can be changed using the `nmodels` option. Observe that even though there are only 9 predictors including the intercept, the model with the highest posterior probability only achieves an estimated probability of around 0.1. This illustrates that there is considerable model uncertainty in this problem and so inference based on one single fixed model does not capture this uncertainty.

Now consider a large problem in which there are $p = 60$ possible predictors giving a total of 2^{60} possible models. The design matrix is constructed in the same way as before, with the correlation between column i and column j being $0.5^{|i-j|}$. This simulation sets $\boldsymbol{\beta} = [2, 2, \ldots, 2, 1, 1, \ldots, 1, 0, 0, \ldots, 0]$ with 10 twos, 10 ones and 40 zeros.

```
> beta <- c(rep(2, 10), rep(1, 10), rep(0, 40))
> y <- x %*% beta + rnorm(101)
```

Now suppose that a priori it is certain that the variables x_1, x_2 through to x_{10} should appear in all models visited by QR-SSVS. This can be specified using the `include` option. The predictors can be specified either by name, or by the position that they appear in the formula containing all predictors. The second way is shown below. The algorithm is designed in such a way that forcing variables to be included in the model can improve computational speed.

```
> print(system.time(models50pc1 <- SSVSquantreg(y ~ xs)))

   user  system elapsed
 40.730   0.260  41.762
```

```
> print(system.time(models50pc2 <- SSVSquantreg(y ~ xs, include = 2:11)))

   user  system elapsed
 30.300   0.210  31.264
```

Note that the argument to `include` starts at 2. The intercept appears in position 1 of the formula and so if there is uncertainty about whether it appears, then the command above is how to specify it. Given that in general, researchers are not interested in whether the intercept appears or not and that it is likely that in general it does, future versions of `SSVSquantreg` will automatically include an intercept term. Regression parameters of any predictors that are a priori certain to appear are given an improper flat prior.

A.3 Summary

The previous section is not an exhaustive tutorial on the use of `MCMCquantreg` and `SSVSquantreg`. The help files of these functions will give further details on additional options that can be set. Most options are common to all functions in `MCMCpack` and give the researcher more direct control over the MCMC algorithms, such as which random number generator to use and whether the algorithm gives output to the `R` console while running. All of these options have default values and so as long as the researcher is familiar with the `lm` function for linear regression, in practice it should be easy to use `MCMCquantreg` and `SSVSquantreg`.