Quantile Regression

Reference range of Thyroid function test in pregnancy

Kevin Brosnan

Final Year Project: Interim Report

Mathematical Sciences

Department Of Mathematics and Statistics

University of Limerick

Limerick

Ireland

November, 2013

A final year project submitted in partial fulfillment of the B.Sc degree in Mathematical Sciences

Supervisor: Dr. Kevin Hayes

Second Reader: Dr. Norma Bargary

Abstract

Acknowledgements

I wish to thank the academic staff of the Maths and Statistics Department of University of Limerick for all knowledge they have instilled on me over the course of my undergraduate programme. I would like to express my deep gratitude to Dr. Kevin Hayes, my supervisor, for his patient guidance, enthusiastic encouragement and useful critiques of my final year project.

Finally, I wish to thank my parents for their support and encouragement throughout my study.

Contents

1 Introduction **1**

2 Analytical Background **2**
2.1 Quantiles 2
2.2 Quantile Regression 4
2.2.1 Frequentist Approach 5
2.2.2 Bayesian Approach 7
2.2.3 Linear Mixed Models Approach 8
2.3 Implementation 10
2.3.1 Frequentist Approach 10
2.3.2 Bayesian Approach 13
2.3.3 Linear Mixed Models Approach 16

3 Next Steps **19**
3.1 Thyroid Data 19
3.2 Non-overlapping Quantile Estimates 20

References **22**

Appendix **23**

A R code for Engel Data Example **23**
A.1 Quantreg Code 23
A.2 BayesQR Code 24
A.3 Lqmm Code 26

Chapter 1

Introduction

Thyroid disorders are the second most common endocrinologic disorders found in pregnancy. Overt hypothyroidism is estimated to occur in $0.3 - 0.5\%$ of pregnancies. Subclinical hypothyroidism tends to occur in $2 - 3\%$, and hyperthyroidism is present in $0.1 - 0.4\%$.(Abalovich et al., 2007) Physiological changes of pregnancy, including 50% increase in plasma volume, increased thyroid binding globulin production and a relative iodine deficiency, means that thyroid hormone reference ranges for non-pregnant women may not be appropriate in pregnancy.

One acceptable approach for establishing legitimate reference ranges requires that a Box-Cox transformation be applied to the data and prediction ranges calculated using classical polynomial regression. Alternatively, non-parametric smoothing such as quantile regression can be used to estimate the 2.5% and 97.5% percentiles. While this approach provides an estimate of the reference range, there are problems with the method. The main problem is the issue of crossing quantiles which makes it difficult to assign each patient to a single range. This is the issue that will be focused on throughout this project.

Chapter 2

Analytical Background

The objective of regression analysis is to establish a relationship between a response variable, Y, and the predictor variables, $\{x_1, \ldots, x_p\}$. In real world applications, Y cannot be calculated perfectly from the X variables. For modelling purposes we formulate Y, for a fixed value of each x_i as a random variable. We generally proceed by summarising the relationship of the response variable for fixed values of the predictors using measures of centrality, specifically the mean, median and mode.

Quantile regression uses the median as its central tendency and is the method of interest in this project. In this chapter we outline the theory behind quantile regression, focusing attention on the frequentist approach to quantile regression, Bayesian quantile regression and linear quantiles for mixed models. Finally, we will examine the computational implementation required for each of the quantile regression methods outlined above.

2.1 Quantiles

A quantile τ of the dependant variable Y is defined such that $100\tau\%$ of the population have values less than the τ^{th} quantile and $100(1-\tau)\%$ of the population have values greater than the τ^{th} quantile (see Figure 2.1).

The median is defined such that 50% of your population have a value above this value and 50% of your population have a value below this value. The median can therefore be interrupted as the 0.5^{th} quantile. The quantile or percentile refers to the general case of this.

More formally, let Y be a continuous real valued random variable, it may be characterised by its distribution function as

$$F_Y(y) = \mathbb{P}(Y \leq y),$$

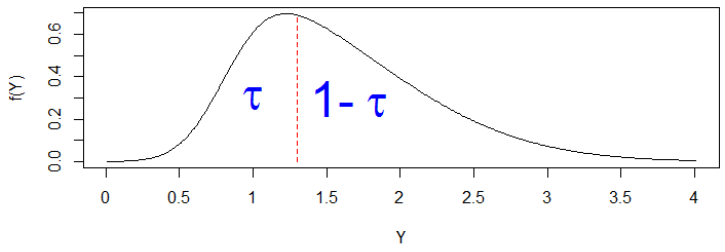

Figure 2.1: Graphical illustration of the τ^{th} quantile.

while for any $0 < \tau < 1$,

$$Q(\tau) = \inf \{y : F_Y(y) \geq \tau\}$$

is called the τ^{th} quantile of Y. When estimating quantiles, we want to determine the value of y in the sample data corresponding to a given probability τ. The τ^{th} quantile in a sample of data refers to the probability of τ for a value y, such that

$$F_Y(y_\tau) = \tau.$$

Another form of expressing the τ^{th} quantile mathematically is

$$y_\tau = F_Y^{-1}(\tau).$$

y_τ is such that it constitutes the inverse of the function $F_Y(\tau)$ for a probability τ.

If the distribution function $F_Y(y)$ is monotonically increasing, quantiles are well defined for every $\tau \in (0, 1)$. However, if a distribution function $F_Y(y)$ is not strictly monotonically increasing, there are some τ's for which a unique quantile can not be defined. In the latter case one must use the smallest value that y can take on for a given probability τ. In both cases the problem can be defined mathematically as seeking the value of y satisfying

$$y_\tau = F_Y^{-1}(\tau) = \inf \{y : F_Y(y) \geq \tau\}. \tag{2.1}$$

Therefore, y_τ is equal to the inverse of the function $F_Y(\tau)$ which in turn is equal to the infimum of y such that the distribution function $F_Y(y)$ is greater or equal to a given probability τ which in turn is the τ^{th} quantile.

2.2 Quantile Regression

Quantile regression is a statistical technique used to estimate conditional quantile functions. Classic linear regression methods are based on minimising sum-of-squares residuals and can be used to estimate models for conditional mean functions. Quantile regression methods however offer a way of estimating models for the conditional median function and all other conditional quantile functions. As quantile regression can estimate the entire family of conditional quantile functions it provides a much more powerful statistical analysis of relationships among random variables (Koenker, 2000). The need for something beyond linear regression was first advocated by Mosteller and Tukey (1977):

What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of of x's. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions. (Mosteller and Tukey, 1977)

Features that characterise quantile regression and distinguish it from other regression methods are the following:

1. quantile regression can characterise the entire conditional distribution of Y through different values of τ;
2. heteroscedasticity can be detected;
3. median regression estimators can be more efficient then mean regression estimators if heteroscedasticity is detected;
4. the minimisation problem as illustrated in equation 2.2 can be solved efficiently by linear programming methods, making estimation easy;
5. quantiles are robust in regards to outliers.

The technical details in the remainder of this section will help to explain how to implement quantile regression contrasting three fitting methods. Quantile regression can be seen as one statistical method which can be used to complete the regression picture.

2.2.1 Frequentist Approach

Quantile regression transforms a conditional distribution function into a conditional quantile function by slicing it in to segments. These segments describe the cumulative distribution of a conditional variable Y given the explanatory variables x_i with the use of quantiles as defined in equation (2.1). For a dependant variable Y given the explanatory variable $X = x$ and fixed τ, $0 < \tau < 1$, the conditional quantile function is defined as the τ^{th} quantile $Q_{Y|X}(\tau|x)$ of the conditional distribution function $F_{Y|X}(y|x)$. For the estimation of the location of the conditional distribution function, the conditional median $Q_{Y|X}(0.5|x)$ can be used as an alternative to the conditional mean.

In ordinary least squares, modelling a conditional distribution function of a random sample $(y_1, \ldots, y_n)$ with a parametric function $\mu(x_i, \beta)$ where x_i represents the independent variables, β the corresponding estimates and μ the conditional mean, one addresses the minimisation problem

$$\min_{\beta \in \Re} \sum_{i=1}^{n} (y_i - \mu(x_i, \beta))^2.$$

We therefore obtain the conditional expectation function $\mathbb{E}[Y|x_i]$. The conditional expectation function is the best predictor of Y given x_i in the sense that it solves a minimum mean squared error prediction problem. It can be simply evaluated as

$$\mathbb{E}[Y|x_i] = \underset{m(x_i)}{\operatorname{argmin}}\, \mathbb{E}[(Y - m(x_i))^2]$$

where $m(x_i)$ is any function of x_i. This equation is minimised at $m(x_i) = \mathbb{E}[Y|x_i]$. While the approach is similar for quantile regression the central feature now becomes ρ_τ, which acts as a check function. Define $\rho_\tau(x)$ by

$$\rho_\tau(x) = \begin{cases} \tau * x, & \text{if } x \geq 0 \\ (\tau - 1) * x, & \text{if } x < 0 \end{cases}$$

The check function, $\rho_\tau(x)$, ensures that:

(i) all ρ_τ are positive;

(ii) the scale is according to the probability τ.

In quantile regression, the τ^{th} *sample* quantile may be found by solving:

$$\min_{\xi \in \Re} \sum_{i=1}^{n} \rho_\tau(y_i - \xi) \tag{2.2}$$

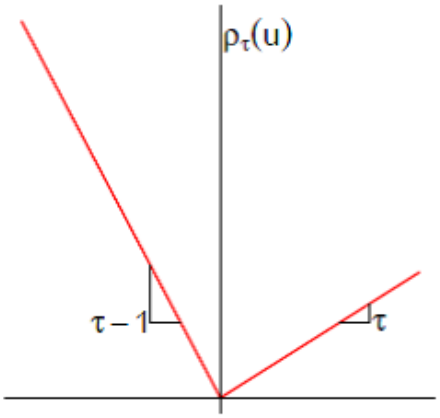

Figure 2.2: Quantile Regression ρ Function

While it is more common to define the sample quantiles in terms of the order statistics, $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n))}$, which results in a sorted arrangement of the original sample. Their formulation as a minimisation problem has the advantage that it yields a natural generalisation of the quantiles to the regression context. The idea of estimating the unconditional mean is simply

$$\hat{\mu} = \underset{\mu \in \Re}{\operatorname{argmin}} \sum (y_i - \mu)^2.$$

This estimation can be extended to the estimation of the linear conditional mean function $E(Y|X = x) = x'\beta$ by solving

$$\hat{\beta} = \underset{\beta \in \Re^p}{\operatorname{argmin}} \sum (y_i - x'\beta)^2.$$

Similarly, the linear conditional quantile function, $Q_Y(\tau|X = x) = x_i'\beta(\tau)$, can be estimated by solving the minimiser

$$\hat{\beta} = \underset{\beta \in \Re^p}{\operatorname{argmin}} \sum \rho_\tau(y_i - x'\beta).$$

In contrast to ordinary least squares, the minimisation in quantile regression is done for each subset defined by ρ_τ. The τ^{th} quantile is estimated with the parametric function $\xi(x_i, \beta)$. The requirement for the square in the unconditional mean case is to ensure each calculated term is positive, the check function ρ_τ ensures this in the quantile regression case and so the square of difference is not required.

2.2.2 Bayesian Approach

Unlike the frequentist approach to statistics, the Bayesian approach provides us with the entire posterior distribution of the parameter of interest. Additionally, it allows for uncertainty of a parameter to be taken into account when making a prediction. Irrespective of the true distribution of the data, Bayesian inference for quantile regression produces the likelihood function based on the asymmetric Laplace distribution.

A random variable ω follows the asymmetric Laplace distribution if its probability density function is given by

$$f_\tau(\omega;\mu,\sigma) = \frac{\tau(1-\tau)}{\sigma}\exp\left\{-\rho_\tau\left(\frac{\omega-\mu}{\sigma}\right)\right\},$$

where $0<\tau<1$, μ is the location parameter, σ is the scale parameter and $\rho_\tau(u)$ is the loss function defined as

$$\rho_\tau(x) = \begin{cases} \tau * x, & \text{if } x \geq 0 \\ (\tau-1)*x, & \text{if } x < 0, \end{cases}$$

or in a simpler form,

$$\rho_\tau(u) = \frac{|u| + (2\tau-1)u}{2}. \tag{2.3}$$

Using general modelling techniques like ordinary least squares the estimates of the regression parameters β are computed by assuming that

(i) conditional on x, the random variables Y_i, are mutually independent with distributions $f(y;\mu_i)$ specified by the values of $\mu_i = E[Y_i|x_i]$;

(ii) for some known link function g, $g(u_i) = x_i'\beta$.

However, in this project we are interested in the conditional quantile, $q_\tau(y_i|x_i)$, in contrast to the conditional mean, $E[Y_i|x_i]$. Simple assumptions can be made so that regardless of the distribution of the data it is possible to solve for the quantiles in the framework of the general linear model. Assuming the following makes this possible,

(i) $f(y;\mu_i)$ is asymmetric Laplace;

(ii) $g(\mu_i) = x_i'\beta(\tau) = q_\tau(y_i|x_i)$ for any $0<\tau<1$.

An issue with using Bayesian statistics is the requirement of a conjugate prior distribution for quantile regression formulation. While this is generally not known Markov Chain Monte Carlo (MCMC) methods can extract posterior distributions

of the unknown parameters which allows the use of any prior distribution. While giving us the marginal and joint posterior distributions of all the unknown parameters, the Bayesian approach, also provides us with a very practical way of including parameter uncertainty in predictive inferences. Given the observations, $y = (y_1, \ldots, y_n)$, the posterior distribution of β, $\pi(\beta|y)$ is given by

$$\pi(\beta|y) \propto L(y|\beta)p(\beta),$$

where $p(\beta)$ is the prior distribution of β and $L(y|\beta)$ is the likelihood function written as

$$L(y|\beta) = \tau^n(1-\tau)^n \exp\left\{-\sum_i \rho_\tau(y_i - x_i'\beta)\right\} \tag{2.4}$$

which is using equation 2.3 with a location parameter $\mu_i = x_i'\beta$.

The optimum strategy is to choose β such that the resulting joint posterior distribution will be proper. It can be shown that the best choice for the prior of β is for it to be improper uniform.

2.2.3 Linear Mixed Models Approach

In statistics it is sometimes necessary to take into account the correlation of observations which belong to the same unit or cluster of the data being analysed. Mixed effects models represent an efficient, flexible and popular way of analysing this complex data. The modelling technique attempts to model and estimate the variability between clusters by using cluster specific random effects. The fact that mixed models can estimate the between cluster variability is a significant advantage over standard modelling techniques as they can provide conditional inferences. In mixed models, both fixed and random effects are assumed to be location-shift effects.

The general idea of linear mixed models approach for quantile regression came from Marco Geraci and Matteo Bottai's asymmetric Laplace approach. A generalisation of this model was further developed by Geraci and Bottai (2013) and is the foundation of the `lqmm R` package which will be discussed later.

A continuous random variable $\omega \in \Re$ is said to follow an asymmetric Laplace distribution with parameters (μ, σ, τ), $\omega \sim AL(\mu, \sigma, \tau)$, if its density can be expressed as

$$p(\omega|\mu, \sigma, \tau) = \frac{\tau(1-\tau)}{\sigma} \exp\left\{-\frac{1}{\sigma}\rho_\tau(\omega - \mu)\right\}$$

where $-\infty < \mu < \infty$ is the location parameter, $\sigma > 0$ is the scale parameter,

$0 < \tau < 1$ is the skew parameter and $\rho_\tau(\upsilon)$ is

$$\rho_\tau(x) = \begin{cases} \tau * x, & \text{if } x \geq 0 \\ (\tau - 1) * x, & \text{if } x < 0 \end{cases}$$

This is the general loss function which is used in each method of quantile regression described in this chapter. In our case, the parameter μ is of great interest as this is the τ^{th} quantile of ω, that is that $Pr(\omega \leq \mu) = \tau$.

If the random variable ω is comprised of n independent $\omega_i's$ with common skew (τ) and scale (σ) parameters and different location (μ), then $\omega_i \sim AL(\mu_i, \sigma, \tau)$ for $i = 1, \ldots, n$. This leads to a simplified expression for $\omega's$ density function

$$p(\omega|\mu, \sigma, \tau) = \sigma_n(\tau) \exp\left\{-\frac{1}{\sigma}\rho_\tau(\omega - \mu)\right\}$$

where $\sigma_n(\tau) = \frac{\tau^n(1-\tau)^n}{\sigma^n}$ and $\rho_\tau(y - \mu) = \sum_{i=1}^n \rho_\tau(\omega_i - \mu_i)$.

Geraci and Bottai (2013) proposed a random-intercepts quantile regression model for longitudinal data using the asymmetric Laplace to model the τ^{th} conditional quantile of a continuous response variable. In particular, they assumed the following regression function

$$Q_{y|u}(\tau|x, u) = X\beta^{(\tau)} + u,$$

where (y, X) represents the longitudinal data, u a vector of subject-specific random effects and $Q_{y|u}$ denotes the inverse of the unknown distribution $F_{y|u}$. The τ^{th} regression quantile of $y|u$ was then estimated under the convenient assumption $y|u \sim AL(X\beta^{(\tau)} + u, \sigma^{(\tau)}, \tau)$, where the τ-dependant parameters $\beta^{(\tau)}$ and $\sigma^{(\tau)}$ have a frequentist interpretation. There exists a link between the L_1 norm regression problem and the asymmetric Laplace based estimation of the coefficients $\beta^{(\tau)}$ which I will not get into here.

2.3 Implementation

As R (R Core Team, 2013) will be used as the development environment for the solution to this problem, we will review the current packages that implement the above quantile regression methods. The main package available in R for each of the methods described above are outlined in the list below. We will discuss each of these methods in the following pages. Advantages, disadvantages and inadequacies of each of the approaches will be highlighted. Included in the following is output from R using the specified approach to compute the quantiles. The sample data used is the Engel data set which is available with the `quantreg` package available on CRAN. The data consists of income and food expenditure values for 235 Belgian working class households.

- The frequentist method is available in the `quantreg` package developed by Koenker (2013)
- The Bayesian method is available in the `bayesQR` package developed by Benoit et al. (2013)
- The linear mixed models method is available in the `lqmm` package developed by Geraci (2012)

2.3.1 Frequentist Approach

The frequentist approach to quantile regression is implemented in R through Roger Koenker's extensive `quantreg` package. The package is currently in version 5.05. The package contains an array of functions for calculating regression quantiles, plotting the results, formatting the results table and a multitude of data sets which can be used as test data for the package. The basic fitting routine is used as follows `rq(formula, tau=.5, data, method='br')`. The function can accept more parameters than shown here however these are the parameters of interest here. The `formula` argument specifies the model that is desired. In the Engel data example below I fitted a simple bivariate linear model so the formula was simply `foodexp~income`, if we had two explanatory variables it would simply be `foodexp~income + something else`. The parameter `tau` is defaulted to calculate the median regression line, however it will accept a single quantile of interest or a vector of quantiles which is used in the example below. The `data` argument requires the name of the `data.frame` which contains the variables named in the `formula` argument. In the case of our example `data=engel` was passed to the function. The `method` argument specifies the calculation method which the package should use to calculate the regression quantiles of interest.

The `rq` function will automatically use `method='br'` if no method is specified. The `br` method calculates the regression quantiles using exterior point methods. It controls the quantile regression fitting by the simplex approach embodied in the algorithm of Koenker and d'Orey (1987) based on the median regression algorithm of Barrodale and Roberts (1974). If all values of `tau` lie in $(0, 1)$ then the regression values are returned for the single or multiple quantiles requested. On the other hand, if `tau` lies outside $[0, 1]$ parametric programming methods are used to find all the solutions to the quantile regression problem for `tau` in $(0, 1)$. This method is efficient for problems containing up to several thousand observations and has the advantage of being able to calculate the full quantile regression process. It also implements a scheme for computing confidence intervals for the estimated parameters based on an inversion of a rank test described in Koenker (1994).

Two other methods to compute the regression quantiles are `method='fn'` and `method='pfn'`. These methods both use the Frisch-Newton algorithm to compute the regression quantiles. The algorithms full detail's are explained in Koenker and Portnoy (1997). In brief, the approach is the reverse of the simplex method, rather than travelling around the exterior of the constraint set it starts from within the constraint set and moves towards the exterior. Instead of taking steepest decent steps at each intersection of exterior edges, it takes Newton steps based on a log-barrier Lagrangian form of the objective function. For exceptionally large problems `method='pfn'` further adds a pre-processing step to the algorithm which can help to speed up the process considerably.

There are also two methods for penalised quantile regression available in the package are `method='lasso'` and `method='scad'`. These methods implement the lasso penalty and Fan and Li's smoothly clipped absolute deviation penalty, respectively. A parameter `lambda` is passed to both functions and is key to the calculations made. In the `lasso` case, if `lambda` is a scalar quantity the penalty function is the l1 norm of the last coefficients, under the assumption that the first coefficient is an intercept parameter that should not be subject to the penalty. When `lambda` is a vector it defines a coordinatewise specific vector of lasso penalty parameters. Similarly, for the `scad` method, if `lambda` is a scalar quantity the penalty function is the scad modified l1 norm of the last coefficients, under the assumption that the first coefficient is an intercept parameter that should not be subject to the penalty. When `lambda` is a vector it defines a coordinatewise specific vector of scad penalty parameters. It should be noted that while these methods are available, Koenker himself states that "These methods should probably be regarded as experimental".

To provide a somewhat more visual explanation of the `quantreg` package I will illustrate an example of its usage on the Engel data set available with the

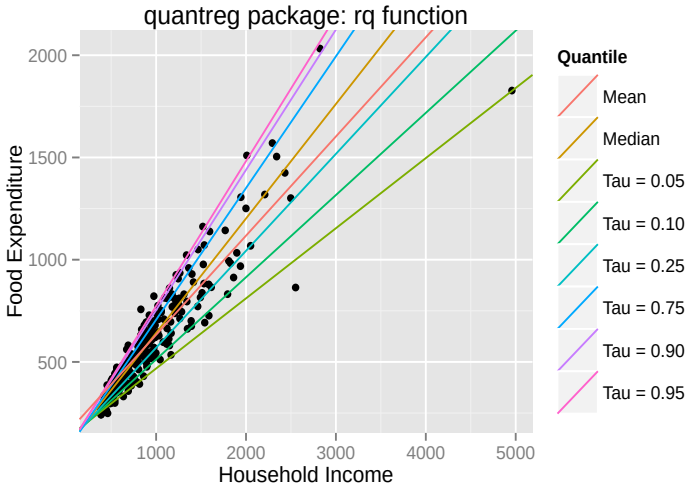

Figure 2.3: Quantreg Estimation of Quantiles on Engel Data

package. The data contains 235 observations of food expenditure and household income for 19^{th} century working class Belgian households. Figure 2.3 shows the $\{0.05_{th}, 0.10_{th}, 0.25_{th}, 0.75_{th}, 0.90_{th}, 0.95_{th}\}$ quantile regression lines, the median fit and the least squares estimate of the conditional mean function. The quantiles were calculated using the `br` method as described above. The least squares estimate was computed using the `lm` function available in `R` for linear regression. It can be seen at the bottom left of the graph the regression lines intersect with one another which is the issue I hope to address in this paper.

Quantile	Intercept	Slope
$\tau = 0.05$	124.880 (98.302,130.517)	0.343 (0.343, 0.390)
$\tau = 0.10$	110.142 (79.888,146.189)	0.402 (0.342, 0.451)
$\tau = 0.25$	95.484 (73.786,120.098)	0.474 (0.420, 0.494)
Median	81.482 (53.259,114.012)	0.560 (0.487, 0.602)
$\tau = 0.75$	62.397 (32.745,107.314)	0.644 (0.580, 0.690)
$\tau = 0.90$	67.351 (37.118,103.174)	0.686 (0.649, 0.742)
$\tau = 0.95$	64.104 (46.265, 83.579)	0.709 (0.674, 0.734)
Mean	147.4754	0.4852

Table 2.1: Regression line coefficients for figure 2.3

The quantiles were also calculated after taking a log transformation of the data. This was used as an attempt to stop the quantiles crossing. It can be seen in figure 2.4 that the quantiles are not crossing at a point close to the origin, however the 0.95^{th} and 0.90^{th} quantile lines are still crossing. A log transformation

therefore does not appear to solve the crossing quantiles problem but does improve on the previous example. The R code used to produce this graphic is available in appendix A.1.

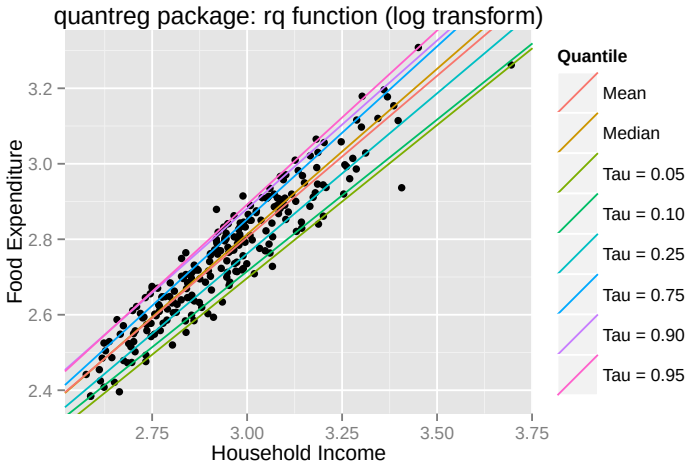

Figure 2.4: Quantreg Estimation of Quantiles on log transformation of Engel Data

Finally, the coefficients of the quantile regression calculations for figure 2.3 are shown in table 2.1 while the log transformation coefficients are shown in table 2.2. The values in parenthesis below the actual value are the confidence bands for that value. Note that the intercept and the slope differ for each of the quantiles and the linear regression model.

Quantile	Intercept	Slope
$\tau = 0.05$	0.2638 (0.2283,0.4607)	0.8113 (0.7398,0.8213)
$\tau = 0.10$	0.3033 (0.1130,0.3601)	0.8041 (0.7866,0.8645)
$\tau = 0.25$	0.2151 (0.1249,0.4095)	0.8495 (0.7822,0.8799)
Median	0.1817 (0.0303,0.3911)	0.8766 (0.8051,0.9302)
$\tau = 0.75$	0.1048 (-0.0057,0.2805)	0.9156 (0.8552,0.9535)
$\tau = 0.90$	0.2075 (0.0335,0.2759)	0.8909 (0.8674,0.9493)
$\tau = 0.95$	0.1258 (0.0804,0.2771)	0.9222 (0.8720,0.9381)
Mean	0.2368	0.8559

Table 2.2: Regression line coefficients for figure 2.4

2.3.2 Bayesian Approach

The implementation of the Bayesian approach to quantile regression is available in the bayesQR package developed by Benoit et al.. The implementation as expected

is much less efficient with regard to processing time than the `quantreg` package due to the iteration process required for Bayesian calculations. The package currently resides in version 2.1 and was updated in 2013. The quantile regression function is called by `bayesQR(formula, data, quantile=0.5, ndraw, prior)`.

The `formula` argument specifies the model that is desired. It follows the same format as in the `rq` function discussed earlier. The `data` parameter is an optional parameter to specify the data object from which the dependant and independent variables are taken. The parameter `quantile` is defaulted to calculate the median regression line, however it will accept a single quantile of interest and will also accept a vector of quantiles as was an option in the `quantreg` package. The `ndraw` parameter specifies how many Markov Chain Monte Carlo draws are to be taken when estimating each quantile required. The `prior` argument allows the user to pass a prior distribution to the model if known otherwise the prior distribution is calculated based on the model type being used.

The package can compute Bayesian quantiles for four types of models; continuous dependant variable without adaptive lasso variable selection, continuous dependant variable with adaptive lasso variable selection, binary dependant variable without adaptive lasso variable selection and binary dependant variable with adaptive lasso variable selection. The computational effort required for each of these methods is similar and is in general extremely computationally intense given a relatively large data set.

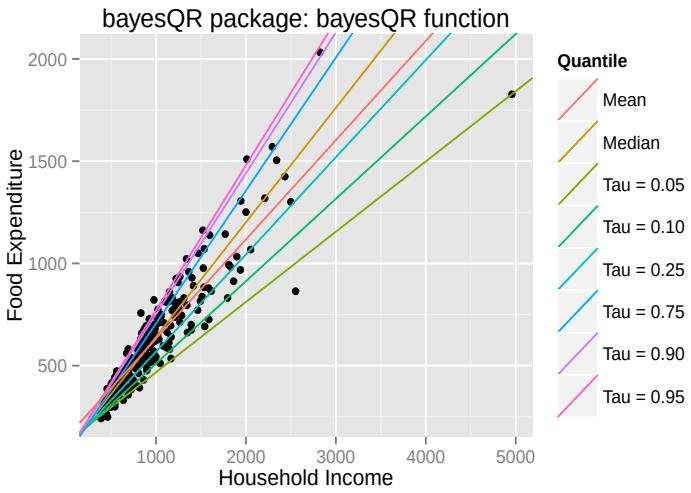

Figure 2.5: bayesQR Estimation of Quantiles on Engel Data

Figure 2.5 shows the regression quantiles produced by the `bayesQR` method on the same data used in the `quantreg` example above. The regression lines are close but not identical to those produced by the frequentist approach in figure 2.3. The

problem of crossing quantiles is still present when using the Bayesian approach to quantile regression.

Quantile	Intercept	Slope
Tau = 0.05	124.561	0.344
Tau = 0.10	107.495	0.403
Tau = 0.25	94.078	0.475
Median	81.227	0.560
Tau = 0.75	57.229	0.650
Tau = 0.90	64.690	0.689
Tau = 0.95	64.605	0.708
Mean	147.475	0.485

Table 2.3: Regression line coefficients for figure 2.5

As before the quantiles were also calculated after taking a log transformation of the data. It can be seen in figure 2.6 that the quantiles appear not to cross anywhere in the domain in which it is plotted. The `R` code used for the two graphics can be found in appendix A.2.

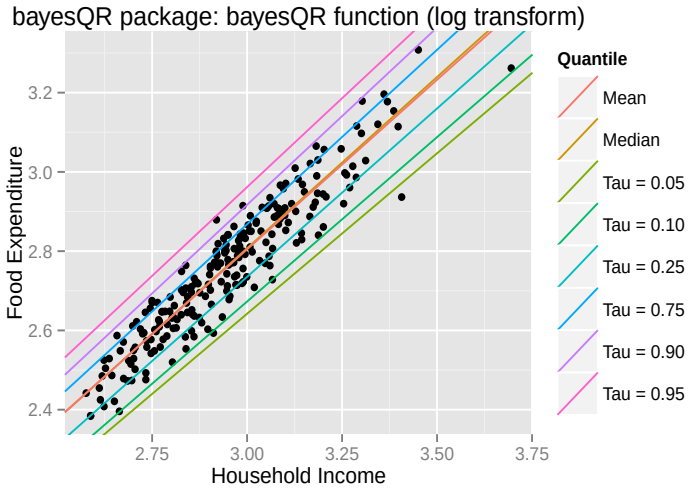

Figure 2.6: bayesQR Estimation of Quantiles on log transformation of Engel Data

Finally, the coefficients of the quantile regression calculations for figure 2.5 are shown in table 2.3 while the log transformation coefficients are shown in table 2.4. The `bayesQR` package does not provide a simple method for computing the confidence intervals for the quantiles. Note that the intercept and the slope differ for each of the quantiles and the linear regression model. The values are also different to those calculated by the `quantreg` package.

Quantile	Intercept	Slope
Tau = 0.05	0.2395	0.8084
Tau = 0.10	0.2194	0.8236
Tau = 0.25	0.1976	0.8465
Median	0.2184	0.8653
Tau = 0.75	0.2323	0.8798
Tau = 0.90	0.2312	0.8943
Tau = 0.95	0.2653	0.9031
Mean	0.2368	0.8559

Table 2.4: Regression line coefficients for figure 2.6

2.3.3 Linear Mixed Models Approach

The final package which will be presented here is the `lqmm` package developed by Marco Geraci. The package is the `R` implementation of linear mixed models approach for quantile regression. The package is currently in version 1.03 and was updated in 2013. The package contains a multitude of functions based around the implementation of the asymmetric Laplace solution to the regression quantiles. The functions of most interest are `lqm` and `lqmm` as these are the methods which calculate the regression quantiles.

The `lqm` function is called using `lqm(formula, data, iota=0.5)`. The `formula` and `data` arguments are specified exactly as they were in the `quantreg` packages `rq` function. The `iota` parameter is equivalent to the `tau` parameter in the previous two packages and can accept single or multiple tau values. The function computes an estimate of the τ^{th} quantile function of the response variable, conditional on the covariates, as specified by the `forumla` argument. The quantile predictor is assumed to be linear. The function maximises the likelihood of a Laplace regression which is equivalent to the minimisation of the weighted sum of absolute residuals. The optimisation algorithm is based on the gradient of the Laplace log-likelihood.

The `lqmm` function is called using `lqmm(fixed, random, iota=0.5)`. The function is similar to the `lqm` function but allows random effects to be specified as arguments. The `fixed` and `iota` parameters follow directly from the `formula` and `iota` arguments in the `lqm` function. The `random` argument allows the inclusion of random effects and should be specified as `~x1+ ... + xn`, where x_i is a random effect of interest. The function calculates an estimate of the τ^{th} quantile function of the response, conditional on the covariates, as specified by the `fixed` argument and on random effects, as specified by the `random` argument. The quantile predictor is again assumed to be linear and the function maximises the likelihood Laplace regression. The likelihood is numerically integrated via Gaus-

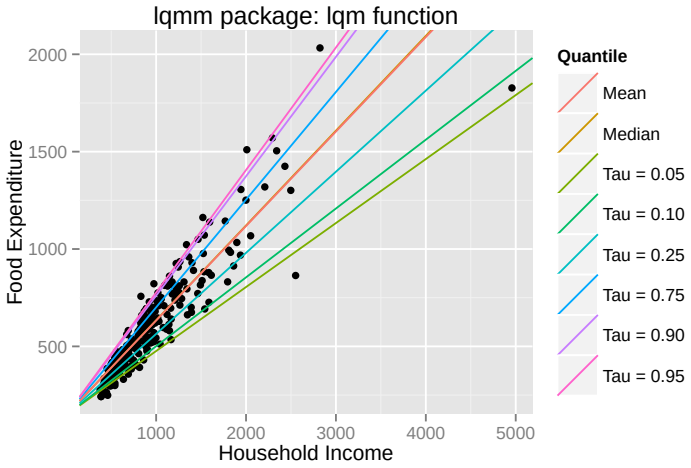

Figure 2.7: lqmm Estimation of Quantiles on Engel Data

sian quadrature techniques. The optimisation algorithm ia based on the gradient of the Laplace log-likelihood.

Quantile	Intercept	Slope
$\tau = 0.05$	147.475	0.329
$\tau = 0.10$	147.475	0.354
$\tau = 0.25$	147.475	0.417
Median	147.475	0.487
$\tau = 0.75$	147.475	0.553
$\tau = 0.90$	147.476	0.613
$\tau = 0.95$	147.476	0.629
Mean	147.475	0.485

Table 2.5: Regression line coefficients for figure 2.7

Again, I have used the Engel data set as the example data set to allow comparison between the three packages. I have used the `lqm` function to calculate the regression quantiles for this example. In figure 2.7 we can see that the regression quantiles appear to meet at a point on the y-axis. However, the median regression line and the mean regression line seem to be identical which is vastly different from what was experienced using the other two packages. The extreme quantile lines, $\tau = \{0.95, 0.05\}$, appear much closer to the median regression fit than they did previously.

When we look at table 2.5, which contains the coefficients for the graphic in figure 2.7, we can see that all the intercepts calculated by `lqm` are exactly equal to the intercept of the least squares estimate for the conditional mean function. This is vastly different from the other two packages where none of the intercepts were equal.

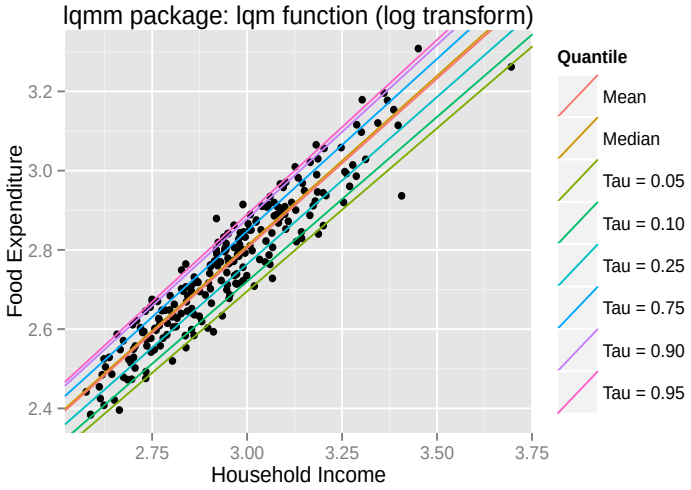

Figure 2.8: lqmm Estimation of Quantiles on log transformation of Engel Data

Finally, a plot of the quantiles calculated by `lqm` for the data after a log transformation is shown in figure 2.8. Again it seems to help in addressing the issue of crossing quantiles, however the upper quantiles, 0.90^{th} and 0.95^{th}, are again very close which could cause a problem. The coefficients are given in table 2.6. The `R` code required to reproduce the graphics and coefficients shown in this section can be found in appendix A.3.

Quantile	Intercept	Slope
$\tau = 0.05$	0.2259	0.8232
$\tau = 0.10$	0.2285	0.8306
$\tau = 0.25$	0.2327	0.8438
Median	0.2372	0.8577
$\tau = 0.75$	0.2409	0.8689
$\tau = 0.90$	0.2441	0.8780
$\tau = 0.95$	0.2452	0.8814
Mean	0.2368	0.8559

Table 2.6: Regression line coefficients for figure 2.8

Chapter 3

Next Steps

This chapter will outline the direction in which the future work in this project will take. The work yet to be completed falls under two key categories, specifically

(a) focusing on real Thyroid data as this is the motivation for this project;

(b) looking at non-parametric approaches that produce non-overlapping quantile estimates.

3.1 Thyroid Data

The data set used so far in this report is a very simple data set containing only 2 variables and has no relation to thyroid disease in pregnancy. The thyroid data currently available is split across 3 data sets, each of which contains 9 variables and a minimum of 106 observations. This is significantly larger than the sample data set and so the code used previously may not be applicable to the thyroid data. The remainder of this section will give an overview of the thyroid data currently available.

MRN	Age	Ethnicity	Smoking	T4	TSH	T3	TPO	Gestation
937395	30	Irish	no	14.60	0.37	4.60	5.30	12.00
1975207	28	Irish	no	13.80	2.07	5.30	0.00	13.00
464433	28	Irish	no	13.80	1.16	4.80	0.00	12.00
1123123	28	Irish	no	12.30	2.71	4.70	5.20	12.00
1855554	35	Irish	no	14.30	0.57	4.80	0.00	13.29
1068450	28	Irish	no	14.20	2.39	4.50	7.00	12.86

Table 3.1: Trimester 1 sample data

The 3 sample datasets available contain the same variables calculated during the 1^{st}, 2^{nd} and 3^{rd} trimesters of pregnancy. The 1^{st} trimester data was recorded during the $10^{th} - 14^{th}$ weeks of gestation, while the 2^{nd} and 3^{rd} trimesters were

recorded in the $14^{th} - 26^{th}$ week period of gestation and the $27^{th} - 42^{nd}$ week period of gestation respectively. A sample of the data from the first trimester can be seen in table 3.1. A data dictionary for the variables of the thyroid data is given in table 3.2.

Variable	Description
MRN	unique hospital identifier for each patient
Age	age of patient in years
Ethnicity	country/region of birth
Smoking	binomial response yes or no, if yes the number of cigarettes smoked per day is recorded
T4	level of thyroxine hormone in each patient
TSH	level of thyroid stimulating hormone in each patient
T3	level of triiodothyronine hormone in each patient
TPO	level of thyroid peroxidase enzyme in each patient
Gestation	number of weeks pregnant

Table 3.2: Data dictionary for thyroid data

The main objective is to develop a method which can calculate reference ranges for each of the variables with respect to the gestation week. This allows the medical practitioners to gain information on those patients who should be treated for thyroid disease. The thyroid data is of most relevance in this project and so the future work on this project will be largely focused around this data.

3.2 Non-overlapping Quantile Estimates

The major problem with quantile estimates for this data is the crossing of the quantile lines. Crossing quantile estimates make it impossible to assign an observation to a single reference range. For example, if we only treat the upper 5% of the population and if the 0.95^{th} and the 0.90^{th} quantile estimates are crossing then how do we decide who to treat? If we assign a patient to be above the 0.95^{th} quantile estimate than they get treated but may not actually need the treatment. The opposite could also occur, we assign a patient to be in the range between the 0.90^{th} and the 0.95^{th} quantile estimates, they do not receive treatment but yet they contract thyroid disease. A simple example like this shows just how important non-crossing quantile estimates really are.

To address the problem, the intention is to use non-parametric approaches to calculating the quantile estimates. While some work has been carried out by others on the topic there exists no general off-the-shelf solution to the problem of overlapping quantile estimates. The focus initially will be to use b-splines and

p-splines to attempt to eliminate the issue of overlapping quantiles. If this solution does not work I will then focus on more advanced ideas to rectify the problem.

References

Abalovich, M., N. Amino, L. Barbour, R. Cobin, L. De Groot, and D. Glinoer (2007, August). Management of thyroid dysfunction during pregnancy and postpartum: an endocrine society clinical practice guideline. *The Journal of Clinical Endocrinology and Metabolism*.

Barrodale, I. and F. Roberts (1974). Solution of an overdetermined system of equations in the $\ell 1$ norm. *Communications of the ACM* (17), 319–320.

Benoit, D. F., R. Al-Hamzawi, K. Yu, and D. Van den Poel (2013). *bayesQR: Bayesian quantile regression*. R package version 2.1.

Geraci, M. (2012). *lqmm: Linear Quantile Mixed Models*. R package version 1.02.

Geraci, M. and M. Bottai (2013). Linear quantile mixed models. *Statistics and Computing*.

Koenker, R. (1994). Confidence intervals for regression quantiles. *Asymptotic Statistics*, 349–359.

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

Koenker, R. (2013). *quantreg: Quantile Regression*. R package version 5.05.

Koenker, R. and V. d'Orey (1987). Computing regression quantiles. *Applied Statistics* (36), 383–393.

Koenker, R. and S. Portnoy (1997). The gaussian hare and the laplacian tortoise. *Statistical Science* (12), 279–300.

Mosteller, F. and J. Tukey (1977). *Data Analysis and Regression: a second course in statistics*. Addison-Wesley.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Appendix A

R code for Engel Data Example

A.1 Quantreg Code

```
# Quantreg Package: Simulation on Sample Data

 # log transformation of Data: TRUE/FALSE

  logTrans <- FALSE

 # Loading data and packages

 library(quantreg)
 library(ggplot2)
 data(engel)
 attach(engel)

 if(logTrans==TRUE){
    x <- log10(engel$income)
    y <- log10(engel$foodexp)
 }else{
    x <- engel$income
    y <- engel$foodexp
 }

 # Creating a data frame of quantile regression lines

 mdl.quant <- rq(y ~ x,tau=c(0.05,0.1,0.25,0.5,0.75,0.9,0.95))
 quant.results <- data.frame(t(coef(mdl.quant)))
 colnames(quant.results) <- c("Intercept", "Slope")

 # Creating a linear model to compare to quantiles (median)
```

```
linear.quant <- lm(y ~ x)

# Creating a data set to output to a latex table

quant.results[8,1] <- linear.quant$coefficients[1]
quant.results[8,2] <- linear.quant$coefficients[2]

quant.results[,"Names"] <- c("Tau = 0.05", "Tau = 0.10",
                             "Tau = 0.25", "Median", "Tau = 0.75",
                               "Tau = 0.90", "Tau = 0.95", "Mean")

quant.results <- data.frame(quant.results[,"Names"],
                            round(quant.results[,"Intercept"],3),
                                round(quant.results[,"Slope"],3))
colnames(quant.results) <- c("Quantile", "Intercept", "Slope")

# Plotting points and the quantile regression lines

if(logTrans==TRUE){
   title <-"quantreg package: rq function (log transform)"
}else{
   title <-"quantreg package: rq function"
}

plt.quant <- qplot(x=x, y=y, xlab="Household Income",
                      ylab="Food Expenditure", main=title) +
            geom_abline(aes(intercept=Intercept, slope=Slope,
            color=Quantile),show_guide=TRUE, data=quant.results)
```

A.2 BayesQR Code

```
# BayesQR Package: Simulation on Sample Data

  # log transformation of Data: TRUE/FALSE

  logTrans <- FALSE

  # Loading data and packages

  library(bayesQR)
  library(ggplot2)
```

```
library(quantreg)
data(engel)
attach(engel)

if(logTrans==TRUE){
  x <- log10(engel$income)
  y <- log10(engel$foodexp)
}else{
  x <- engel$income
  y <- engel$foodexp
}
# Creating a data frame of quantile regression lines

taus <- c(0.05,0.1,0.25,0.5,0.75,0.9,0.95)
intercept <- c(rep(0,7))
slope <- c(rep(0,7))

mdl.bays <- bayesQR(y ~ x, quantile=taus, ndraw=5000)

sum <- summary(mdl.bays, burnin=500)

for(i in 1:length(sum)){
  intercept[i] <- sum[[i]]$betadraw[1,1]
  slope[i] <- sum[[i]]$betadraw[2,1]
}

bays.results <- data.frame(cbind(intercept,slope))
colnames(bays.results) <- c("Intercept", "Slope")

# Creating a linear model to compare to quantiles (median)

linear.bays <- lm(y ~ x)
bays.results[8,1] <- linear.bays$coefficients[1]
bays.results[8,2] <- linear.bays$coefficients[2]

bays.results[,"Names"] <- c("Tau = 0.05","Tau = 0.10",
                        "Tau = 0.25","Median","Tau = 0.75",
                            "Tau = 0.90","Tau = 0.95","Mean")

bays.results <- data.frame(bays.results[,"Names"],
                        round(bays.results[,"Intercept"],4),
                            round(bays.results[,"Slope"],4))
colnames(bays.results) <- c("Quantile", "Intercept", "Slope")
```

```
  # Plotting points and the quantile regression lines

  if(logTrans==TRUE){
    title <-"bayesQR package: bayesQR function (log transform)"
  }else{
    title <-"bayesQR package: bayesQR function"
  }

  plt.bays <- qplot(x=x, y=y, xlab="Household Income",
                    ylab="Food Expenditure",main=title) +
    geom_abline(aes(intercept=Intercept, slope=Slope,
            color=Quantile),show_guide=TRUE, data=bays.results)
```

A.3 Lqmm Code

```
# lqmm Package: Simulation on Sample Data

 # log transformation of Data: TRUE/FALSE

 logTrans <- FALSE

 # Loading data and packages

 library(lqmm)
 library(ggplot2)
 library(quantreg)
 data(engel)
 attach(engel)

 if(logTrans==TRUE){
    x <- log10(engel$income)
    y <- log10(engel$foodexp)
 }else{
    x <- engel$income
    y <- engel$foodexp
 }
 # Creating a data frame of quantile regression lines

 mdl.mix <- lqm(y ~ x,iota=c(0.05,0.1,0.25,0.5,0.75,0.9,0.95))
 mix.results <- data.frame(t(coef(mdl.mix)),row.names=NULL)
 colnames(mix.results) <- c("Intercept", "Slope")
```

```
# Creating a linear model to compare to quantiles (median)

linear.mix <- lm(y ~ x)
mix.results[8,1] <- linear.mix$coefficients[1]
mix.results[8,2] <- linear.mix$coefficients[2]

mix.results[,"Names"] <- c("Tau = 0.05", "Tau = 0.10",
                           "Tau = 0.25", "Median", "Tau = 0.75",
                            "Tau = 0.90", "Tau = 0.95", "Mean")

mix.results <- data.frame(mix.results[,"Names"],
                          round(mix.results[,"Intercept"],4),
                           round(mix.results[,"Slope"],4))
colnames(mix.results) <- c("Quantile", "Intercept", "Slope")

# Plotting points and the quantile regression lines

if(logTrans==TRUE){
   title <- "lqmm package: lqm function (log transform)"
}else{
   title <- "lqmm package: lqm function"
}

plt.mix <- qplot(x=x, y=y, xlab="Household Income",
                 ylab="Food Expenditure", main=title) +
           geom_abline(aes(intercept=Intercept, slope=Slope,
           color=Quantile), show_guide=TRUE, data=mix.results)
```