Quantile Regression

Reference range of Thyroid function test in pregnancy

Kevin Brosnan

Final Year Project
Mathematical Sciences

Department Of Mathematics and Statistics
University of Limerick
Limerick
Ireland
May, 2014

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

Supervisor: Dr. Kevin Hayes

Abstract

Contents

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 appears to occur in $2-3\%$, and hyperthyroidism is present in $0.1-0.4\%$.[1] 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 aim of this project is to address the problem of the method not routinely quantifying the precision of the end points of the reference range.

Chapter 2

Analytical Background

Regression analysis is used to expose a relationship between a response variable and predictor variables. That is that, a researcher is interested in analyzing the behaviour of a dependant variable y_i given the information contained in a set of explanatory variables x_i. In real world applications, the response variable cannot be calculated perfectly from the predictor variables and so the response variable for a fixed value of each predictor variable is a random variable. Due to this issue in the real world, it is generally best to summarise the relationship of the response variable for fixed values of the predictors using measures of centrality. Common measures of centrality are the mean (average value), median (middle value) or the mode (most common value).

Quantile regression uses the median as its central tendency and is the method of interest in this project. In this chapter I will outline the theory behind Quantile regression, focusing my attention on the Frequentist approach to Quantile Regression, Bayesian Quantile Regression and Linear Quantiles for Mixed Models. Finally, I will examine the computational implementation required for each of the Quantile regression methods outlined above.

2.1 What are Quantiles?

A quantile τ 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 2.1). The median is defined such that 50% of your population have a value above the median and 50% of your population have a value below the median. The median can therefore be interrupted as the 0.5^{th} quantile. Similarly, the quartiles divide the population into 4 segments each with an equal proportion of the population. The quintiles divide the population into 5 equal segments while the deciles divide it into 10 equal segments. The quantile or percentile refers to the general case of this.

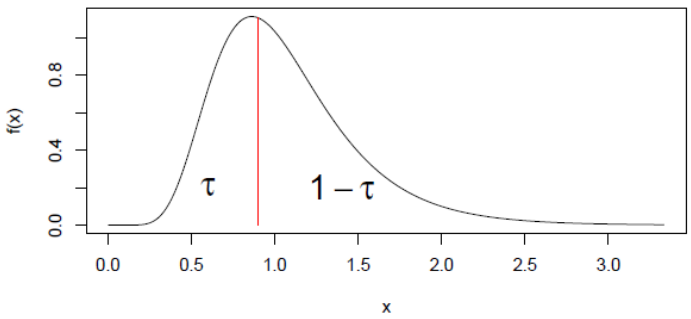

Figure 2.1: Graphical explanation of a quantile

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

$$F_Y(y) = Prob(Y \leq y)$$

while for any $0 < \tau < 1$

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

is called the τ^{th} quantile of Y. As stated earlier the median $Q(0.5)$ plays the central role. This means that for a distribution function $F_Y(y)$ one can determine for a given value of y the probability τ of occurrence. For quantiles, we wish to do entirely the opposite, that is we want to determine for a given probability τ of the sample data set the corresponding y value. A τ^{th} quantile refers, in a sample of data, to the probability of τ for a value y.

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

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.

Note that there are two scenarios. On the one hand, 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 this case one uses the smallest value that y can take on for a given probability τ. In both cases, with and without a monotonically increasing function, the problem can be defined mathematically as:

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

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 is in turn the τ^{th} quantile.

2.2 Theory of Quantile Regression

Quantile Regression is a statistical technique used to estimate conditional quantile functions. Classic linear regression methods are based on minimising sums of squared 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 in comparison with linear regression just estimating the conditional mean, quantile regression provides a much more powerful statistical analysis of relationships among random variables.[2] This need for something beyond linear regression was first noted by Mosteller and Tukey in their influential text (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.[3]

Features that characterise Quantile Regression and differentiate 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 "go further". Quantile regression is a 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 into segments. These segments describe the cumulative distribution of a conditional variable Y given the explanatory variable 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 OLS, 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 gets the following minimisation problem:

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

We therefore obtain the conditional expectation function $E[Y|x_i]$. While the approach is similar for Quantile regression the central feature now becomes ρ_τ, which acts as a check function.

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

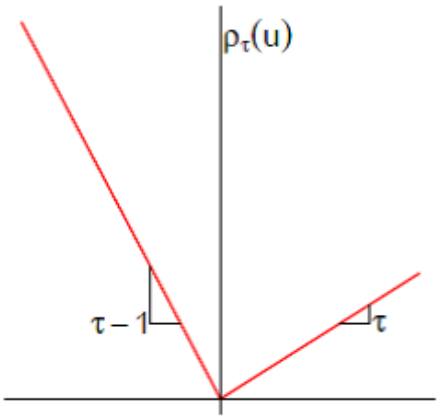

Figure 2.2: Quantile Regression ρ Function

This check function 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:

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

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. Just as the idea of estimating the unconditional mean, viewed as the minimiser,

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

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

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

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

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

In contrast to OLS, the minimisation in Quantile regression is done for each subset defined by ρ_τ. The τ^{th} quantile is estimated with the parametric function $\xi(x_i, \beta)$.

2.2.2 Bayesian Approach

Unlike the frequentist approach to statistics, the Bayesian approach provides one 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 U follows the asymmetric Laplace distribution if its probability density function is given by

$$f_\tau(u) = \tau(1-\tau)\exp\{-\rho_\tau(u)\}$$

where $0 < \tau < 1$ 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

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

This is the same loss function used in the frequentist approach to Quantile regression. Location (μ) and scale (σ) parameters can be incorporated to obtain

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

In the conventional linear model, the unknown regression parameters β estimates are produced 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, $g(u_i) = x_i'\beta$.

In this well-known framework, a Gaussian distribution for the Y_i's with an identity link function yields the quadratic loss function estimator of β.

However, in this paper we are interested in the conditional quantile, $q_\tau(y_i|x_i)$, in contrast to the conditional mean, $E[Y_i|x_i]$. It is possible to solve the problem in the framework of the linear model, regardless of the distribution of the data, by assuming that

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

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

In general a conjugate prior distribution is not available for Quantile regression formulation, however Markov Chain Monte Carlo (MCMC) methods may be used to extract the posterior distributions of unknown parameters. This allows for 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.

Theorem 1. If the likelihood function is given by equation 2.4 and $p(\beta) \propto 1$, then the posterior distribution of β, $\pi(\beta|y)$, will have a proper distribution. In other words

$$0 < \int \pi(\beta|y)d\beta < \infty$$

or equivalently,

$$0 < \int L(y|\beta)p(\beta)d\beta < \infty$$

In practice the assumption is made that the components of β have independent improper uniform prior distributions which is a special case of the above theorem.[4]

2.2.3 Linear Mixed Models Approach

A number of sampling designs such as multilevel, longitudinal and cluster sampling typically require the application of statistical methods that allow for the correlation between observations that belong to the same unit or cluster. Mixed effects models represent highly popular and flexible models to analyze complex data. They model and estimate between-cluster variability by means of cluster-specific random effects. These, in turn, provide a modelling structure for estimating the intraclass correlation coefficient (ICC). Due to the presence of an estimate of the ICC it is now possible to conduct inference at the population or cluster level. This is a huge advantage of mixed models over standard modelling techniques.

2.3 Implementation

As I will be using R [5] as my development environment for my solution to this problem, I will review the current packages that implement the above Quantile regression methods. The main package in R for each of the methods described above are outlined in the list below. I will discuss each of these methods in the following pages. I will highlight the advantages, disadvantages and inadequacies of each of the methods. Included in the following is output from R using the specific method to calculate the quantiles. The data used was the Engel data set which is available with the Quantreg package available on CRAN. The data consists of 235 observations on income and expenditure on food for Belgian working class households.

- The frequentist method is available in the *quantreg* package developed by Roger Koenker [6]
- The Bayesian method is available in the *Brq* package developed by Rahim Alhamzawi [7]
- The Linear Mixed Models method is available in the *lqmm* package developed by Marco Geraci [8]

2.3.1 Frequentist Approach

The frequentist approach to quantile regression is implemented in R through Roger Koenker's extensive `quantreg` package. The package contains an array of functions for calculating regression quantiles, plotting the results, outputting the results to a latex file and a multitude of datasets 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 has more options available to it than can be seen here but the ones shown are the ones of most interest. 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 as was 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 QR 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 QR problem for `tau` in $(0, 1)$. This method is efficient for problems 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 calculate the regression quantiles are `method='fn'` and `method='pfn'`. These methods both use the Frisch-Newton algorithm to compute the regression quantiles. The algorithms detail's are explained in Koenker and Portnoy (1997). Rather than travelling around the exterior of the constraint set like the simplex method, the interior point approach starts from within the constraint set toward 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.

The final methods which are 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 dataset available with the package. The data contains 235 observations of food expenditure and household income for 19^{th} century working class Belgian households.

2.3.2 Bayesian Approach

2.3.3 Linear Mixed Models Approach

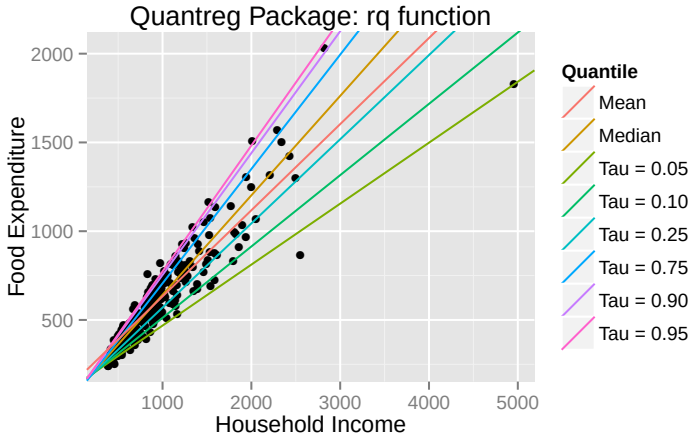

Figure 2.3: Quantreg Estimation of Quantiles on Engel Data

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 2.4

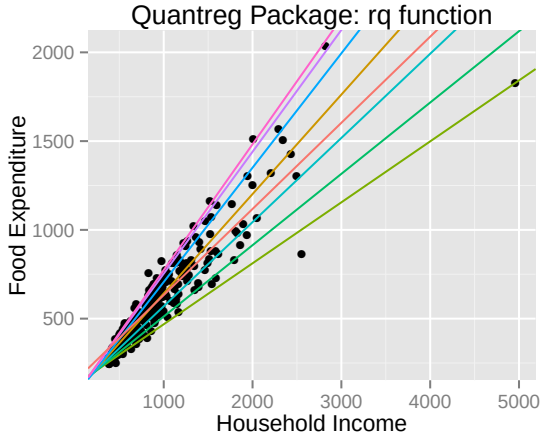

Figure 2.4: Quantreg Estimation of Quantiles on Engel Data

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 2.2: Regression line coefficients for figure 2.4

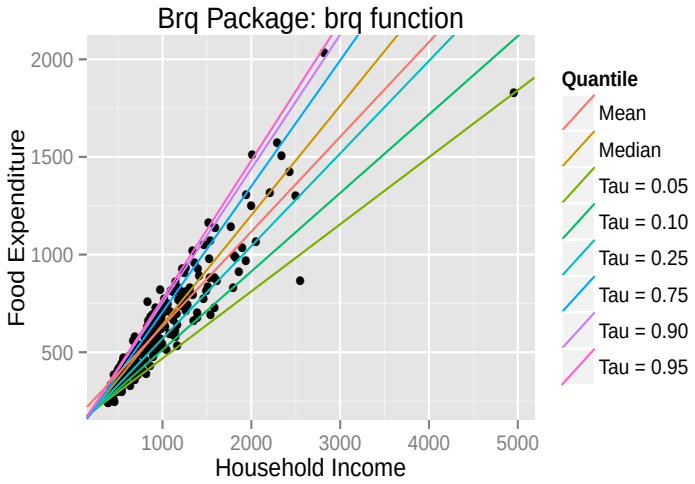

Figure 2.5: Brq Estimation of Quantiles on Engel Data

Quantile	Intercept	Slope
$\tau = 0.05$	124.880	0.343
$\tau = 0.10$	110.142	0.402
$\tau = 0.25$	95.484	0.474
Median	81.482	0.560
$\tau = 0.75$	62.397	0.644
$\tau = 0.90$	67.351	0.686
$\tau = 0.95$	64.104	0.709
Mean	147.475	0.485

Table 2.3: Regression line coefficients for figure 2.5

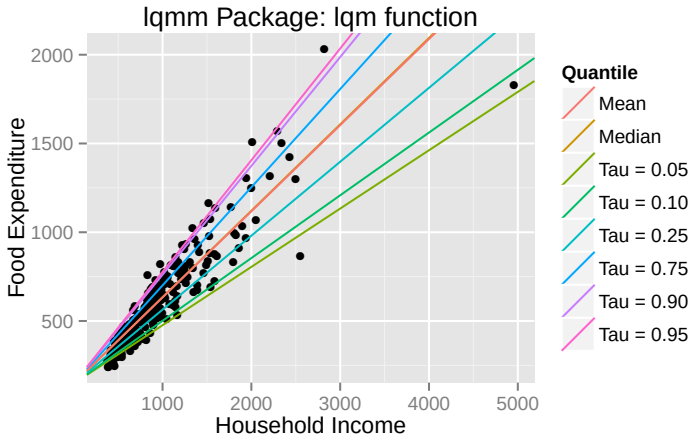

Figure 2.6: lqmm Estimation of Quantiles on Engel Data

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 2.4: Regression line coefficients for figure 2.6

References

[1] Abalovich M, Amino N, Barbour LA, Cobin RH, De Groot LJ, Glinoer D. *Management of thyroid dysfunction during pregnancy and postpartum: an Endocrine Society Clinical Practice Guideline*, J Clin Endocrinol Metab. Aug 2007

[2] Koenker R, *Quantile Regression*, International Encyclopedia of the Social Sciences, October 25^{th} 2000

[3] Mosteller F, Tukey JW, *Data Analysis and Regression: a second course in statistics*, Addison-Wesley, 1977

[4] Keming Yu, Rana A. Moyeed, *Bayesian Quantile Regression*, Statistics and Probability Letters, April 2001

[5] R Core Team (2013), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN $3 - 900051 - 07 - 0$, URL http://www.R-project.org/

[6] Roger Koenker (2013), *quantreg: Quantile Regression*, R package version 5.05, http://CRAN.R-project.org/package=quantreg

[7] Rahim Alhamzawi (2012), *Brq: Bayesian analysis of quantile regression models*, R package version 1.0, http://CRAN.R-project.org/package=Brq

[8] Marco Geraci (2012), *lqmm: Linear Quantile Mixed Models*, R package version 1.02, http://CRAN.R-project.org/package=lqmm

[9] Ramsey J, Silverman B, *Functional Data Analysis (2nd Edition)*, Springer, 2005.

[10] Ramsey J, Silverman B, *Functional Data Analysis with R*, Springer, 2009.

[11] Koenker R, *Quantile Regression (Econometric Society Monographs)*, Cambridge University Press, 2005

Acknowledgements

Appendices