

Logistic Quantile Regression to Evaluate Bounded Outcomes

Vivi Wong

Kandidatuppsats 2018:16
Matematisk statistik
Juni 2018

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Logistic Quantile Regression to Evaluate Bounded Outcomes

Vivi Wong*

June 2018

Abstract

Lower urinary tract symptoms in men are common when men get older, and these symptoms can be measured with I-PSS (International Prostate Symptom Score), a scale between 0-35. The density function of the bounded outcome variable, I-PSS, is highly skewed to the right. It can therefore be difficult to analyze this type of variables with standard regression methods such as OLS, since these methods give us the effect of the explanatory variables on the mean of the response variable.

Epidemiological studies commonly study how lifestyle and several other factors affect health-related problems. We will therefore study the effect physical activity has on lower urinary tract symptoms by using logistic quantile regression, which is an appropriate method to use when we have bounded outcomes. The method works well because instead of the mean, it focuses on quantiles and it takes the bounded interval into account.

The results show a negative relationship between total physical activity and lower urinary tract symptoms, so men who are more physical active will more likely have lower and milder symptoms.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden.
E-mail: viviwong.96@gmail.com. Supervisor: Ola Hössjer Disa Hansson.

Sammanfattning

Hos äldre män är nedre urinvägssymptom ett vanligt förekommande problem. Dessa symptom kan mätas med hjälp av I-PSS (International Prostate Symptom Score), en skala mellan 0-35. I-PSS, som är begränsad inom ett intervall, är väldigt skev åt höger. Det kan därmed leda till svårigheter och felaktiga resultat om man använder sig av regressionsmetoder så som OLS, då dessa undersöker effekten de förklarande variablerna har på medelvärdet av responsvariabeln.

Epidemiologiska studier undersöker hur livsstilsfaktorer och en rad andra faktorer påverkar hälsorelaterade problem. I denna uppsats undersöker vi vilken effekt fysisk aktivitet har på nedre urinvägssymptom hos män genom att använda oss av logistisk kvantilregression, vilket är en bra metod att använda när responsvariabeln ligger inom ett begränsat intervall. Metoden använder istället för medelvärdet kvantiler och tar hänsyn till det begränsade intervallet.

Resultaten visar ett negativt förhållande mellan total fysisk aktivitet och nedre urinvägssymptomer, vilket säger att personer som är mer fysiskt aktiva kommer med högre sannolikhet att visa mindre omfattande och mildare symptom.

Preface

This is a bachelor's degree thesis worth 13.5 credits at Stockholm University.

I am very grateful for the support I received from my external supervisor Matteo Bottai. I want to thank you for giving me this project, giving me access to the data, taking time to explain theory and more. I really appreciate that.

Secondly, I need to thank the Unit of Biostatistics at the Institute of Environmental Medicine (IMM), Karolinska Institutet for offering a desk place so I could work with the thesis over there.

Lastly, I would like to thank my supervisors Ola Hössjer and Disa Hansson for their guidance of the project.

Contents

1	Introduction	1
1.1	Quantile regression	1
1.2	Bootstrap	2
1.3	Bounded outcomes	3
1.3.1	Bounded outcomes in medicine	3
1.3.2	Bounded outcomes in epidemiology	3
1.3.3	Bounded outcomes in finance	4
2	Background	5
2.1	Quantiles and transformations	5
2.1.1	Expectation	5
2.2	Logistic regression	6
2.3	Logistic quantile regression	7
2.4	Bootstrap	8
3	Data analysis	11
3.1	Research question	11
3.2	Previous research	11
3.3	Description of data	11
3.3.1	Variables	12
3.3.2	Missing data	12
3.4	Models	13
3.4.1	Categorical explanatory variables	14
3.4.2	Continuous explanatory variables	15
3.4.3	The best fitting epsilon	15
3.5	Software	17
3.6	Results	17
3.6.1	Regression results of models with interval variables	18
3.6.2	Regression results of models with continuous variables	20
3.6.3	Pseudo- R^2 and average of the absolute differences of $\hat{Q}_{y_i}(p \epsilon)$	22
3.7	Interpretation	23
3.7.1	First model	23
3.7.2	Second model	24
3.7.3	Third model	25
3.7.4	Fourth model	25
3.7.5	The model with the best fitted epsilon	26
4	Discussion	28
4.1	Limitations	28
4.2	Possible future work	28
4.3	Personal Considerations	29

5	References	30
6	Appendixes	33
A1	Plots and a table	33
A2	Results when $\epsilon = 10^{-11}$	40
A3	Results when $\epsilon = 1$	46
7	Abbreviations	52
	List of Figures	53
	List of Tables	54

1 Introduction

In this study we estimate psychological outcomes. The characteristics of these outcomes can lead to some issues when estimated with standard regression methods, where the focus is on the mean. It has been quite difficult during a long time to find articles about some simple and well performing models for such type of outcome variables, until in 2010 when Bottai et al. [4] published a paper about logistic quantile regression. This method has been proved to work very well and has been used in many research papers since [10, 13, 15, 27, 29].

In the following sections, we introduce some terms and methods, such as quantile regression, that are used in the rest of this thesis.

1.1 Quantile regression

Using standard linear regression, which analyzes the relationship between the covariates and the mean of the response variable, does not give the whole picture of the distribution of the response variable conditional on the covariates [6, 19, 29]. Quantile regression, on the other hand, models the relationship between the independent variables quantiles and the response variable instead [29]. This makes it easier to evaluate location, skewness and other features [29], which can lead to a more complete picture of the relationship between the explanatory variables and the outcome [6, 19].

The main reference for this part is Koenker [18, 19]. Estimating the coefficients of quantile regression is similar to estimating OLS regression coefficients. For an OLS model, we know that we should solve for the following problem to get the sample mean

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^n (y_i - \mu)^2,$$

where y_1, \dots, y_n are the observations and μ the location parameter. For multiple linear regression with q explanatory variables $x_i = (x_{i1}, \dots, x_{iq})^T$, in order to get the estimated regression coefficients $\beta = (\beta_1, \dots, \beta_q)^T$ and intercept β_0 , we have to solve

$$\min_{(\beta_0, \beta) \in \mathbb{R}^{q+1}} \sum_{i=1}^n [y_i - (\beta_0 + \beta^T x_i)]^2.$$

For quantile regression we do not look at the sample mean, but instead we focus on the p th sample quantile, which is given by the following minimization problem

$$\min_{\xi \in \mathbb{R}} \sum_{i=1}^n l_p(y_i - \xi),$$

where $l_p(\cdot)$ is a tilted absolute function called the loss function, and ξ is a scalar. The loss function can be described in more detail by the following expression (see also Figure 5 in Appendix A1)

$$l_p(u) = \begin{cases} up, & \text{if } u \geq 0, \\ u(p-1), & \text{if } u < 0. \end{cases}$$

A general quantile regression model for the p th quantile can according to Koenker [19] be defined as

$$y_i = \beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq} + \varepsilon_i, \quad \text{where } i = 1, \dots, n,$$

where ε_i denotes the error term. Then the p th quantile of the conditional distribution of y_i given x_i is

$$Q_{y_i}(p) = \beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq}.$$

An advantage of using quantile regression is that quantiles have the following property for a non-decreasing function h :

$$Q_{h(y_i)}(p) = h[Q_{y_i}(p)].$$

A link function can then be written [29] as

$$h[Q_{y_i}(p)] = \beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq}. \quad (1)$$

To obtain the estimated regression coefficients for any quantile, we should solve the following minimization problem

$$\min_{\beta \in \mathbb{R}^{q+1}} \sum_{i=1}^n l_p[y_i - (\beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq})],$$

where $\beta = (\beta_{p,0}, \dots, \beta_{p,q})$. For simplicity, this can be rewritten as

$$\min_{\beta \in \mathbb{R}^{q+1}} \sum_{i=1}^n l_p[y_i - (\beta_{p,0} + \beta_p^T x_i)], \quad (2)$$

where $\beta_p = (\beta_{p,1}, \dots, \beta_{p,q})^T$.

1.2 Bootstrap

Collecting information of a whole population (e.g. every person in a city, or all people with a specific disease) can be extremely difficult or even impossible, instead a large sample that should be very accurate to the population is usually gathered. It is hard to get an efficient sample, but an easy way to obtain a good sample is to use the simple random sampling method. The method consists of drawing n observations from the population with

replacement, where every unit in the population has the same probability of being selected [8].

Bootstrap can be seen as a simple random sampling from the sample instead of the population. This method has been used in many different fields. When the desire is to estimate a certain population parameter from a sample but we do not make any parametric assumptions (for example the assumption that the observations are normally distributed), then the bootstrap often is a good option [7]. By applying bootstrap with replacement, we resample k new samples by taking random samples from the original dataset, each with same size as the original data. From each of these k samples we compute the estimates of the desired parameters. Note that k is recommended to be at least 100 for standard errors and 1000 for confidence intervals [7, 8].

1.3 Bounded outcomes

Outcome variables that take on values within a bounded interval are called bounded outcome variables [4, 29]. Problems can occur when analyzing bounded outcome variables with traditional statistical methods, since the frequency distributions of these variables can be unimodal, U-shaped, J-shaped, or a variety of other shapes. By instead using methods that constrain the inference within the bounded interval, we can draw more reliable conclusions [4].

Bounded response variables can be found in many different fields [29]. We continue by briefly give some examples of bounded outcomes.

1.3.1 Bounded outcomes in medicine

An example of medical bounded outcomes is the Glasgow Coma Scale, which measures consciousness in an interval between 3 and 15 [26]. Visual Analog Scale for Pain is another example, it takes values between 0 and 100 [4, 23]. A third example is another Visual Analog Scale (VAS), which lies between 0 and 100, which is a measurement for irritations in the eyes, nose, throat and airways. The statistical analysis method that was used for this bounded variable was logistic quantile regression [13].

1.3.2 Bounded outcomes in epidemiology

An example of epidemiological bounded response variables can be the Center for Epidemiologic Studies Depression scale (CES-D), which measures the level of depression within the interval 0 and 60 [4].

Another example is an ADHD symptoms counting. Parents report psychiatric disorder symptoms of their children by filling in Diagnostic Interview Schedule for Children Version IV (DISC-IV) modules. The total number of

symptoms is then calculated and the score can be seen as a bounded outcome [15].

Mini Mental State Examination (MMSE) is another example, which is bounded between 0 and 30. It measures the person’s cognitive status, where low values stand for cognitive impairment [10].

The method that was used in these studies to analyze the relationship between the bounded outcomes was logistic quantile regression [4, 10, 15].

1.3.3 Bounded outcomes in finance

Finance is another field with bounded outcomes, for example different types of recovery rates (e.g. credit recovery rate), which vary between 0 and 1 (0% and 100%). Here 1 indicates full recovery, i.e. zero loss.

Recovery rates have been studied with the logistic quantile regression model in [27], where they show that the logistic quantile regression method was a better alternative to use than other methods.

2 Background

When we have bounded outcomes, classical standard regression models may be inefficient and impractical as we stated before. A better way to analyze such outcomes is to use logistic quantile regression [4]. In this section, we introduce the logistic quantile regression model. We begin by highlighting some important properties of quantiles and transformations in Section 2.1. In Section 2.2 we describe logistic regression. In Section 2.3 the method used in this thesis, namely, logistic quantile regression is presented. We conclude by explaining bootstrap in Section 2.4.

2.1 Quantiles and transformations

Unlike the mean, quantiles can analyze transformations of the outcome variable. The main reference for this section is Koenker [19].

For a numeric outcome Y , the cumulative distribution function (CDF) can be written as $F(y) = P(Y \leq y)$, and the p th quantile of Y is given by $Q(p) = F^{-1}(p)$. The inverse function of the CDF of F is monotonically increasing and continuous [6, 11, 19].

The useful property quantiles have is the following

$$Q_{h(y)}(p) = h[Q_y(p)], \quad (3)$$

where $h(\cdot)$ is a non-decreasing function. This means that if $g(\cdot)$ is a non-decreasing function, the following equality must hold

$$Q[g(Y)] = g[Q(Y)] \iff Q(Y) = g^{-1}(Q[g(Y)]),$$

thus taking the inverse of $Q[g(Y)]$ will give us $g^{-1}(Q[g(Y)])$, and the inverse of $g[Q(Y)]$ is $Q(Y)$.

The equality in Equation (3) holds, because if $h(\cdot)$ is a monotone function, the following will be implied

$$P(Y \leq y) = P[h(Y) \leq h(y)].$$

2.1.1 Expectation

The advantageous property for quantiles does not hold for transformations of the mean. In other words

$$E[h(Y)] \neq h(E[Y]),$$

although there are exceptions. Since the mean does not have such a useful property as the quantiles, quantile regression might be a better option compared to those regression methods that only take the mean into consideration.

Example 2.1: The function $h(x) = x^2$ is monotonically increasing for $x \in [0, +\infty)$. For simplicity, let therefore x take the following values: 1, 2 and 3. Then $h(x)$ for these values are 1, 4 and 9 respectively. See Table 1 for some descriptive statistics.

Table 1: Mean and median of the x - and the $h(x)$ -values.

				Mean	Median
x:	1	2	3	2	2
h(x):	1	4	9	14/3	4

We can then see that $E[h(x)] \neq h(E[x])$ since

$$E[h(x)] = \frac{1 + 4 + 9}{3} = \frac{14}{3}$$

and

$$h(E[x]) = h(2) = 2^2 = 4.$$

On the other hand

$$Q_{h(x)}(0.5) = 4 = 2^2 = h(2) = h[Q_x(0.5)].$$

2.2 Logistic regression

This part is based on Agresti [1]. Assume that we have an outcome variable with only two different outcomes, for example: success (denoted by 1) and failure (denoted by 0). Let the probability of success be denoted by π , then the odds of a success is given by the following formula

$$\Omega = \frac{\pi}{1 - \pi},$$

which is positive and shows how likely it is that the outcome gives a success compared to a failure. For instance, if $\Omega < 1$, then a failure is more likely to happen than a success. $\Omega = 1$ means that a failure is as likely as a success and finally if $\Omega > 1$, then a failure is not as likely to happen as a success.

Assume that we have n observations and q explained variables, $x_i = (x_{i1}, x_{i2}, \dots, x_{iq})$, where $i = 1, \dots, n$, and let $\pi(x_i)$ denote the probability of success for x_i . Then the logistic regression model can be expressed in the following way

$$\log \left[\frac{\pi(x_i)}{1 - \pi(x_i)} \right] = \text{logit} [\pi(x_i)] = \beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq},$$

and this implies the following equality

$$\frac{\pi(x_i)}{1 - \pi(x_i)} = \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq}) \iff$$

$$\Longleftrightarrow \pi(x_i) = [1 - \pi(x_i)] \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq}) \Longleftrightarrow$$

$$\Longleftrightarrow \pi(x_i) [1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq})] = \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq}) \Longleftrightarrow$$

$$\Longleftrightarrow \pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq})}.$$

2.3 Logistic quantile regression

The fact that the density functions of bounded outcomes can have different shapes is a reason that standard regression models are not appropriate to use. Therefore, quantile regression is preferable, but unfortunately it does not take the bounded interval of the response variables into consideration. In other words, the regression line may not lie between the upper limit and lower limit of the outcome variable. We therefore use a method that does that instead, the logistic quantile regression method, which is a combination of quantile regression and logistic regression [4].

This section is based on Bottai et al. (2010) [4, 29], who found an easy method to analyze bounded response variables. Logistic quantile regression will not only analyze the relationship of the explained variables by looking at the quantiles, but also take the range of bounded outcome values into account.

Assume that we have n observations and q explained variables, $x_i = (x_{i1}, x_{i2}, \dots, x_{iq})$, where $i = 1, 2, \dots, n$, and a bounded continuous response variable y_i , which lies between the known interval (y_{min}, y_{max}) . Note that y_{min} and y_{max} are constants which denote the limits of the outcome variable. Thus, a continuous bounded outcome between 0 and 1 reminds us of a probability. Bottai et al. chose the following logistic link function

$$h(y_i) = \text{logit}(y_i) = \log \left(\frac{y_i - y_{min}}{y_{max} - y_i} \right), \quad (4)$$

which transforms the bounded interval to $[0, 1]$.

The logit function of $Q_{y_i}(p)$ will then be

$$h[Q_{y_i}(p)] = \text{logit}[Q_{y_i}(p)] = \log \left[\frac{Q_{y_i}(p) - y_{min}}{y_{max} - Q_{y_i}(p)} \right],$$

which is equal to Equation (1). The next step is to solve $Q_{y_i}(p)$ from the following equality

$$\log \left[\frac{Q_{y_i}(p) - y_{min}}{y_{max} - Q_{y_i}(p)} \right] = \beta_{p,0} + \beta_{p,1} x_{i1} + \dots + \beta_{p,q} x_{iq}.$$

We begin by taking the exponential on both sides,

$$\frac{Q_{y_i}(p) - y_{min}}{y_{max} - Q_{y_i}(p)} = \exp(\beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq}),$$

multiplying with the denominator of the LHS and collecting the $Q_{y_i}(p)$ -terms gives,

$$Q_{y_i}(p)[1 + \exp(\beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq})] = y_{max} \exp(\beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq}) + y_{min}.$$

Dividing both sides by $[1 + \exp(\beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq})]$ will give the desired function

$$Q_{y_i}(p) = \frac{y_{max} \exp(\beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq}) + y_{min}}{1 + \exp(\beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq})}. \quad (5)$$

To obtain the estimate of the regression coefficients and intercept, $\beta_{p,i}$ for $i = 0, 1, \dots, q$, we should solve

$$(\hat{\beta}_{p,0}, \hat{\beta}_p) = \min_{(\beta_0, \beta) \in \mathbb{R}^{q+1}} \sum_{i=1}^n l_p[h(y_i) - (\beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq})],$$

which can be written as

$$(\hat{\beta}_{p,0}, \hat{\beta}_p) = \min_{(\beta_0, \beta) \in \mathbb{R}^{q+1}} \sum_{i=1}^n l_p[\text{logit}(y_i) - (\beta_{p,0} + \beta_p^T x_i)],$$

where $\beta_p = (\beta_{p,1}, \dots, \beta_{p,q})^T$ [10].

Note that the estimates of the logistic quantile regression can be obtained by using quantile regression where we regress the transformed outcome on x_i [29] (see and compare Equation (2)).

When we obtain the desired estimates, we can take advantage of the favorable property shown in Equation (3) and get inference on $Q_{y_i}(p)$ [10].

2.4 Bootstrap

We use bootstrap to obtain the standard errors, confidence intervals and p -values of the estimates. The reason is that the design matrix bootstrap has been shown to be a better alternative to use for quantile regressions, especially if the data is heteroskedastic (the variance is not constant) [4, 5, 15]. Design matrix bootstrap, also called (x, y) -pairs bootstrap, is an effective method to use when we have independent variables that are not identical [18].

The following parts are based on theory outlined in Andrews and Buchinsky (2000), Buchinsky (1995), Efron and Tibshirani (1994), and elsewhere [2, 5, 12, 14, 18].

Using bootstrap to obtain the standard error:

1. We shall draw k bootstrap samples $(z^{*1}, z^{*2}, \dots, z^{*k})$ with the same size as the original sample. Each observation in a bootstrap sample consist of a pair (x^*, y^*) where $x^* = (x_1^*, \dots, x_q^*)$ and each such pair is drawn with replacement from the n pairs in the sample. Each pair in a bootstrap sample is therefore drawn with equal probability $1/n$.
2. Estimate the regression coefficient for all k new samples, $\hat{\beta}_{p,i}^{*m}$, for $i = 1, \dots, q$ and $m = 1, \dots, k$, where m is the m th bootstrap sample.
3. The next step is to compute the standard error by calculating the standard deviation of the k samples. The reason why we compute the standard error with the standard deviation is because of the unknown variance of the parameter estimates. In other words

$$\widehat{SE}_{boot} = \sqrt{\frac{\sum_{m=1}^k (\hat{\beta}_{p,i}^{*m} - \bar{\beta}^*)^2}{(k-1)}}.$$

Here $\bar{\beta}^*$ denote the mean of the bootstrapped regression coefficients. It can be written as

$$\bar{\beta}^* = \frac{\sum_{m=1}^k (\hat{\beta}_{p,i}^{*m})}{k}$$

Bootstrap for confidence interval with k bootstrap replicates:

The confidence interval can be directly calculated from the non-parametric bootstrap resampling (we will use the xy-pairs bootstrap). The bootstrap confidence interval can be obtained by following steps:

1. Compute the estimates of the regression coefficients.
2. Calculate the bootstrapped standard error.
3. Construct the $100(1 - \alpha)$ -confidence interval of each parameter by using the Wald-interval formula

$$CI = (\hat{\beta}_{p,i} \pm z_{(1-\alpha)/2} \widehat{SE}_{boot}),$$

where $z_{(1-\alpha)/2}$ is the $(1 - \alpha)/2$ quantile of a standard normal distribution. A 95%-confidence interval will therefore be given by

$$(\hat{\beta}_{p,i} \pm 1.96 \cdot \widehat{SE}_{boot}).$$

Obtaining the p -value by using k bootstrap replicates:

The p -value tells us the probability to get at least the same or a more extreme value of a test statistic T than the observed outcome under the null hypothesis, H_0 .

1. We need to compute the T -values for the estimated regression coefficients for all resampled datasets, by using bootstrap.
2. We obtain the p -value by calculating the probability that we get a value of the test statistic from the resamples than for the original dataset. We can define the bootstrapped p -value as

$$\hat{p}_{boot} = \frac{1}{k} \sum_{m=1}^k 1(T_{boot}^{*m} > T),$$

where m denotes the m th bootstrap sample, $m = 1, \dots, k$, so T_{boot}^{*m} is the m th bootstrapped test statistic. T is the test statistic of our original data.

3 Data analysis

The main reference for this section is the article by Orsini et al. [30]. We use the same data as they did except that we have fewer explanatory variables.

3.1 Research question

Lower urinary tract symptoms (LUTS) in men are more common as they get older. The term LUTS describes several different symptoms associated with lower urinary tract problems [3]. One common cause of LUTS is benign prostatic hyperplasia (BPH) [30].

The LUTS can be measured with the International Prostate Symptom Score (I-PSS), where higher scores indicate severe symptoms [30].

Finding what effect lifestyle and other factors such as environment has on health-related problems, is a big part of epidemiological studies [17]. Physical activity is associated with lifestyle and many epidemiological researchers have studied the relationship between physical activity and various kind of diseases [16, 22, 24]. Our research question will be: what is the effect of total physical activity on lower urinary tract symptoms?

3.2 Previous research

A study on long-term physical activity and lower urinary tract symptoms in men showed that total physical activity affects lower urinary tract symptoms significantly. The results showed that inactive men have approximately twice the risk of getting lower urinary tract symptoms than physically active men [30].

Another research study found that the risk for lower urinary tract symptoms was lower for physical active men [9].

Other studies have derived the same results, that decreasing physical activity is associated with higher levels of lower urinary tract symptoms [25, 28].

3.3 Description of data

The data was part of the Cohort of Swedish Men. A questionnaire was sent out between 1997-1998 to all 45 to 79-year-old men that lived in Västmanland and Örebro counties in Sweden. The questions in the survey asked about the presence and severity urination symptoms, physical activity, education and more. Out of the men that handed in the questionnaire, a total of 30,377 were included in the study in the end.

3.3.1 Variables

The data we received had already been adjusted/filtered and consisted of four variables: age, total physical activity (TPA), the Swedish version of International Prostate Symptom Score (I-PSS) and a dummy variable for high I-PSS (HIGH-IPSS), where 1 indicates high I-PSS score and 0 low I-PSS score.

The physical activity score was measured by six questions about physical activity/inactivity for different categories such as occupational activity and household work at different ages. The intensity of these activities is measured in metabolic equivalents (MET), where one MET equals to one kcal/kg/hour. The reported time was multiplied with the intensity to obtain the physical activity score.

I-PSS on the other hand was measured by seven questions concerning the presence of urination symptoms with five alternatives each. One of the questions was asking the participants how many times they go to the toilet each night, with five possible answers: never, once, twice, three times, four times and more than five times. This implies that I-PSS is a bounded variable that lies within the interval 0 and 35. An I-PSS score equal to or is higher than 8 (value 1 of the binary variable "HIGH-IPSS"), means that the person shows moderate to severe lower urinary tract symptoms, while a score of 7 or lower implies the opposite.

See the Table 2 for a briefly summary of the variables.

Table 2: Minimum value, maximum value, mean and median of the variables in the data.

Variable	Min	Max	Mean	Median
IPSS	0	35	5.091	3
TPA	25.5	61.2	41.5	40.7
Age	45	79	59.3	58
HIGH-IPSS	0	1	0.2273	0

3.3.2 Missing data

Dealing with missing data is a very common problem in questionnaire studies. For instance, a total of 100,303 men received the survey, but 51,658 did not give any responses to the questionnaire. Out of those 48,645 men that answer the survey, 92 sent in an empty survey and 15,529 did not answer every question about physical activity and lower urinary tract symptoms. These persons were therefore excluded from the study and additionally, 2,647 were excluded from the study for other reasons.

3.4 Models

We follow the steps described by Orsini and Bottai (2011) [29], our main reference in this section. Another paper by Bottai et al. [4] stated that we can add a small value strictly greater than 0 (ϵ) to prevent getting undefined values such as $\log(0)$ or some numbers divided by 0. We will therefore use the following formula

$$h(y_i) = \text{logit}(y_i) = \log \left(\frac{y_i - y_{\min} + \epsilon}{y_{\max} - y_i + \epsilon} \right), \quad (6)$$

instead of using Equation (4).

From the first reference above,[29], we can see that the authors used $\epsilon = 0.5$. Moreover, since 0.001 and 0.5 are the more commonly used values of ϵ in articles [4, 10, 29], we plotted the distribution of $\text{logit}(\text{I-PSS})$ against total physical activity (TPA) for three different values of ϵ : 0, 0.001 and 0.5 (see Figure 3). From Figure 3, we can see that the plot with $\epsilon = 0.5$ looks more similar to the plot where $\epsilon = 0$ than the plot where $\epsilon = 0.001$ does, so it might therefore be better to choose $\epsilon = 0.5$ rather than $\epsilon = 0.001$.

The four models used in the thesis will be described in Section 3.4.1 and Section 3.4.2, a summary of the models can be found in Table 3.

Table 3: The regression models used in the thesis. The interval variables (Interval1, Interval2, Interval3,...,Interval12) are categorical variables of the variable TPA, where TPA is a continuous variable. Note that Interval1 is our reference variable in model 1 and 2.

Model 1	$h(y_i) = \beta_{p,0} + \beta_{p,1}x_{\text{Interval2}} + \beta_{p,2}x_{\text{Interval3}} + \cdots + \beta_{p,11}x_{\text{Interval12}} + \varepsilon_i$
Model 2	same as model 1 but with the explanatory variable age added
Model 3:	$h(y_i) = \beta_{p,0} + \beta_{p,1}x_{\text{TPA}} + \varepsilon_i$
Model 4:	same as model 3 but with the explanatory variable age added

3.4.1 Categorical explanatory variables

Let us divide the total physical activity, our explanatory variable, into twelve different interval variables (almost equally sized, see Table 4 for some descriptive statistics about these variables). Figure 1 show us boxplots of I-PSS over these variables. The first logistic quantile regression model we will use is

$$h(y_i) = \beta_{p,0} + \beta_{p,1}x_{Interval2} + \beta_{p,2}x_{Interval3} + \dots + \beta_{p,11}x_{Interval12} + \epsilon_i,$$

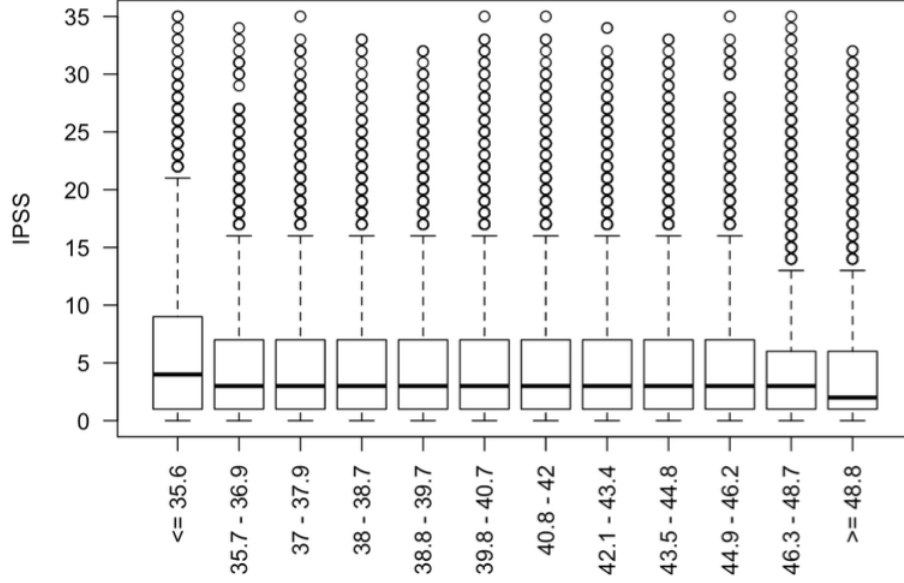
where the lowest interval variable, $x_{Interval1}$, will be our reference variable. We add the variable x_{Age} to the second regression model.

We also use F-test to test the joint hypothesis that none of the interval variables will have an effect on I-PSS.

Table 4: Minimum value, maximum value, mean and median of the interval variables (TPA divided into twelve variables). The lowest interval, Interval 1 (marked with *), is the reference interval.

Variables:	Min	Max	Mean	Median
Interval 1 *	25.5	35.6	34.15121	34.5
Interval 2	35.7	36.9	36.35906	36.4
Interval 3	37	37.9	37.42931	37.5
Interval 4	38	38.7	38.38495	38.4
Interval 5	38.8	39.7	39.28406	39.3
Interval 6	39.8	40.7	40.20631	40.2
Interval 7	40.8	42	41.37865	41.4
Interval 8	42.1	43.4	42.7949	42.8
Interval 9	43.5	44.8	44.17201	44.2
Interval 10	44.9	46.2	45.52942	45.5
Interval 11	46.3	48.7	47.36826	47.3
Interval 12	48.8	61.2	51.68351	51.3

Figure 1: Boxplots of I-PSS over twelve almost equally sized classes of TPA.
Boxplots



3.4.2 Continuous explanatory variables

Instead of having total physical activity (TPA) as categorical predictors, we additionally set up a simple regression model with only TPA as a continuous explanatory variable. The simple regression model can be described with the following equation

$$h(y_i) = \beta_{p,0} + \beta_{p,1}x_{TPA} + \epsilon_i,$$

and by adding the predictor x_{Age} , we get a multiple regression model (which will be our fourth model).

3.4.3 The best fitting epsilon

We use $\epsilon = 0.5$ and it seems to be a good choice, since Orsini and Bottai [29] used it and it showed to give a better fit compared to using $\epsilon = 0.001$ (see Figure 3). Let us investigate if there is another value of the ϵ that gives a better fit and if the results of that ϵ -value differ much.

We follow what Siao et al. [27] did in 2016. They tried twelve different values of ϵ : 10^{-11} , 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.08, 0.1, 0.2, 0.3 and 0.4999. They used two methods with these values, and we also use them when applying the logistic quantile regression models. Then we check the

pseudo- R^2 values for these ϵ -values. The pseudo- R^2 (here denoted R^1) or the goodness of fit for quantile regression can be computed by the following formula according to Koenker and Machado [20]:

$$R^1(p) = 1 - \frac{\hat{V}(p)}{\tilde{V}(p)},$$

where $\hat{\beta}(p)$ is the unrestricted estimated regression coefficient and it is the value obtained from the following minimization problem

$$\hat{V}(p) = \min_{(\beta_{p,0}, \beta) \in \mathbb{R}^{q+1}} \sum_{i=1}^n l_p[h(y_i) - (\beta_{p,0} + \beta_p^T x_i)].$$

The restricted estimated regression coefficient $\tilde{\beta}(p)$ on the other hand is the minimization of

$$\tilde{V}(p) = \min_{(\beta_{p,0}, \beta_{p*}) \in \mathbb{R}^{q+1-t}} \sum_{i=1}^n l_p[h(y_i) - (\beta_{p,0} + \beta_{p*}^T x_i)],$$

where t is the number of effect parameters put to zero for the restricted model. When some regression coefficients in the model are set to zero, we call the model a restricted regression model. To obtain the pseudo R^2 values in the software R, we used the code Koenker [21] wrote. This is one method Siao et al. [27] applied.

The other method that Siao et al. [27] used, is based on the following formula

$$\frac{\sum_{i=1}^n |\hat{Q}_{y_i}(p|\epsilon_1) - \hat{Q}_{y_i}(p|\epsilon_2)|}{n}$$

to compute some numerical values, where ϵ_1 and ϵ_2 are two consecutive ϵ -values and n is the total number of observations. For this method, they tried the following values of ϵ : 10^{-11} , 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.08, 0.1, 0.2, 0.3, 0.4999, 1, 2, 3 and 5. The formula $\hat{Q}_{y_i}(p|\epsilon)$ is obtained by inserting the estimated intercept and regression coefficients, $\hat{\beta}_{p,i}$, into the following equation (we get this formula (Equation (7)) instead of Equation (5) by rewriting Equation (6))

$$Q_{y_i}(p) = \frac{(y_{max} + \epsilon) \exp(\beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq}) - (\epsilon - y_{min})}{1 + \exp(\beta_{p,0} + \beta_{p,1}x_{i1} + \dots + \beta_{p,q}x_{iq})}. \quad (7)$$

In other words

$$\hat{Q}_{y_i}(p|\epsilon) = \frac{(y_{max} + \epsilon) \exp(\hat{\beta}_{p,0} + \hat{\beta}_{p,1}x_{i1} + \dots + \hat{\beta}_{p,q}x_{iq}) - (\epsilon - y_{min})}{1 + \exp(\hat{\beta}_{p,0} + \hat{\beta}_{p,1}x_{i1} + \dots + \hat{\beta}_{p,q}x_{iq})}.$$

Siao et al. [27] then chose a value of ϵ , where the average value of the absolute values of the quantiles between two consecutive ϵ is less than 0.01 for all ϵ greater than or equal to the chosen value (since their outcome was a proportion). The reason is that the \hat{Q}_{y_i} values level off and are about the same above the chosen value. Our data on the other hand is bounded between 0 to 35, so instead of checking if the average of the absolute differences of $\hat{Q}_{y_i}(p|\epsilon)$ is lower than 0.01, we check if the average is smaller than 1.

Lastly, we choose ϵ visually, by plotting the logit of I-PSS against TPA for those quantiles in the previous method [10].

3.5 Software

We use the software R to analyze the data. The following packages are needed: "quantreg" (quantile regression package), "Hmisc" and "car". The quantile regression package is required when estimating logistic quantile regression model. In order to fit the logistic quantile regression model, we use the "rq" (quantile regression) code and we need to insert the logistic transformation of I-PSS as the response variable instead of I-PSS. The second package will be useful when we split the explanatory variable, total physical activity (TPA), into almost equally sized categorical interval variables. We will also use the "car" package to compute F-tests.

3.6 Results

Figure 2: A histogram of I-PSS that shows the 0.05, 0.25, 0.5, 0.75, 0.95 quantiles, the mean and the density function.

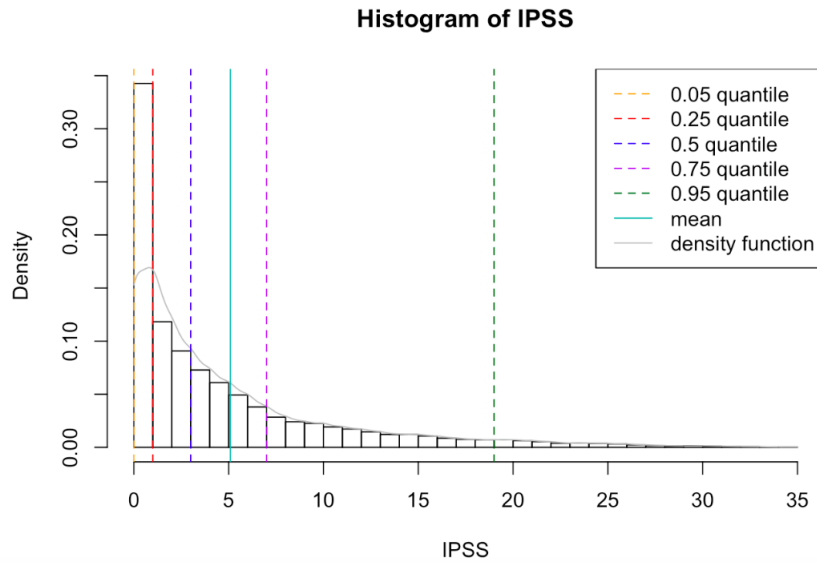
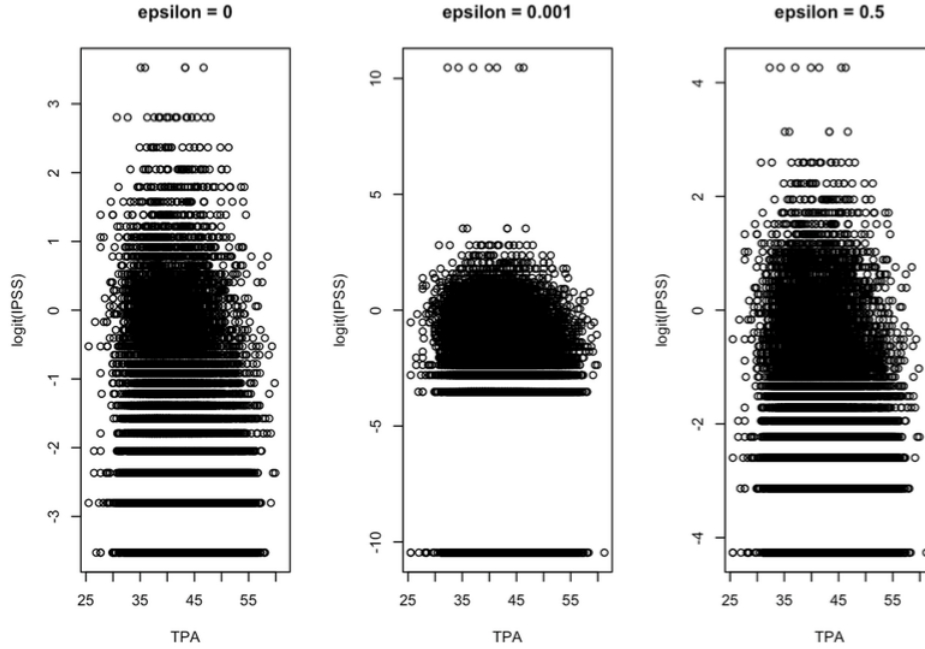


Figure 2 shows that we have a very skewed density function, and we can also see that the mean does not say much about the data, so using standard regression such as OLS would probably not be a good option as we wrote in the introduction part. Instead logistic quantile regression would indeed be a better method to use, thus we can interpret different regressions in different quantiles and get a better view of the data.

We start by showing the results of the regression models where we used $\epsilon = 0.5$. As we stated before, we used the value 0.5 because the plot where $\epsilon = 0.5$ looks more similar to the original sample compared to the plot with $\epsilon = 0.001$ (see Figure 3).

Figure 3: Three different plots that show the distribution of the logistic transformation of I-PSS against TPA with different values on epsilon.



From Figure 1 we can see that higher total physical activity score seems to be associated with lower I-PSS score for our data sample, which is what many other research papers in this field have demonstrated. To see if this is really the case, we look at the regression results of our models.

3.6.1 Regression results of models with interval variables

The result from the first regression model with categorical explanatory variables can be found in Table 5. The table shows that every coefficient is negative and gives a significant result at the 5% significant level, we also see that we can reject the null-hypothesis that the interval variables does not

have an effect on the I-PSS score ($\beta_{0.5,1}, \beta_{0.5,2}, \dots, \beta_{0.5,11} = 0$). Figure 6 in Appendix A1 shows the plots of the estimated regression coefficients over different quantiles.

Table 5: Estimates of the following model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{Interval2} + \beta_{p,2}x_{Interval3} + \dots + \beta_{p,11}x_{Interval12} + \varepsilon_i$. The columns correspond to the quantile (p), variable, estimated coefficient, standard error, t -value, p -value and 95% confidence interval, where 1000 bootstrap samples were drawn to obtain the standard errors, p -values and confidence intervals. The result of the F-test, where we test the hypothesis that none of the intervals do have an effect on I-PSS, can also be seen.

p	Variable	Coeff	SE	t -value	p -value	95% CI
0.5	Intercept	-1.946	0.0252	-77.265	0.000	(-1.995, -1.897)
	Interval 2	-0.283	0.0252	-11.220	0.000	(-0.332, -0.233)
	Interval 3	-0.283	0.0252	-11.220	0.000	(-0.332, -0.233)
	Interval 4	-0.283	0.0308	-9.179	0.000	(-0.343, -0.222)
	Interval 5	-0.283	0.0252	-11.220	0.000	(-0.332, -0.233)
	Interval 6	-0.283	0.0252	-11.220	0.000	(-0.332, -0.233)
	Interval 7	-0.283	0.0252	-11.220	0.000	(-0.332, -0.233)
	Interval 8	-0.283	0.0564	-5.014	0.000	(-0.393, -0.172)
	Interval 9	-0.283	0.0343	-8.235	0.000	(-0.350, -0.215)
	Interval 10	-0.283	0.110	-2.561	0.0105	(-0.499, -0.0663)
	Interval 11	-0.283	0.144	-1.967	0.0492	(-0.564, -0.00103)
	Interval 12	-0.649	0.0252	-25.783	0.000	(-0.699, -0.600)
F-test:						
F-value		18.373	p -value: $<2.2\text{e-}16$			

The result of the second model where the predictor age is added into the first model are reported in Table 6. All the coefficients of the interval variables in Table 6 have a negative value when $p = 0.5$. We can also see from the table that the estimated coefficient of age is positive. To get a better view of the estimated regression coefficients of the variables over different quantiles, see Figure 7 in Appendix A1. The coefficients of the interval variables are decreasing over different quantiles according to the figure.

Table 6: Estimates for the model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{Interval2} + \beta_{p,2}x_{Interval3} + \dots + \beta_{p,11}x_{Interval12} + \beta_{p,12}x_{Age} + \varepsilon_i$. The quantile (p), variables, estimated coefficients, standard errors, t -values, p -values and 95% confidence intervals are shown, where we used 1000 bootstrap replicates to obtain the standard errors, p -values and confidence intervals. The F-test, where we test the joint hypothesis ($\beta_{0.5,1}, \beta_{0.5,2}, \dots, \beta_{0.5,11} = 0$), are also shown.

p	Variable	Coeff	SE	t -value	p -value	95% CI
0.5	Intercept	-4.276	0.0864	-49.498	0.000	(-4.445, -4.107)
	Interval 2	-0.0766	0.0478	-1.602	0.109	(-0.170, 0.0171)
	Interval 3	-0.130	0.0449	-2.894	0.00381	(-0.218, -0.0419)
	Interval 4	-0.153	0.0490	-3.119	0.00182	(-0.249, -0.0568)
	Interval 5	-0.115	0.0508	-2.257	0.0240	(-0.214, -0.0151)
	Interval 6	-0.244	0.0476	-5.131	0.000	(-0.338, -0.151)
	Interval 7	-0.229	0.0446	-5.141	0.000	(-0.317, -0.142)
	Interval 8	-0.306	0.0497	-6.144	0.000	(-0.403, -0.208)
	Interval 9	-0.306	0.0501	-6.103	0.000	(-0.404, -0.207)
	Interval 10	-0.344	0.0523	-6.577	0.000	(-0.446, -0.241)
	Interval 11	-0.420	0.0541	-7.764	0.000	(-0.526, -0.314)
	Interval 12	-0.435	0.0588	-7.404	0.000	(-0.551, -0.320)
	Age	0.0382	0.00131	29.258	0.000	(0.0356, 0.0408)
F-test:						
	F-value	18.515	p -value:	<2.2e-16		

3.6.2 Regression results of models with continuous variables

The estimates of the simple regression model where TPA was considered as a continuous variable instead of categorical predictors can be found in Table 7 and the results of the multiple regression model (where the explanatory variable age is added to the simple regression model) are summarized in Table 8. Figure 8 and Figure 9 in Appendix A1 show some plots of the estimated coefficients over different quantiles (p) for these two logistic quantile regression models.

Table 7: Estimates for the model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{TPA} + \varepsilon_i$. We used 1000 bootstrap replicates to get the standard errors, p -values and confidence intervals.

p	Variable	Coeff	SE	t -value	p -value	95% CI
0.25	Intercept	-3.135	0.000	-9.621e+15	0.000	(-3.135 , -3.135)
	TPA	0.000	0.000	-3.808	1.4e-04	(-3.55e-17, -1.14e-17)
0.5	Intercept	-1.314	0.417	-3.148	0.00165	(-2.132, -0.496)
	TPA	-0.0229	0.0105	-2.194	0.0282	(-0.0434, -0.00245)
0.75	Intercept	-0.264	0.101	-2.608	0.00910	(-0.463, -0.0657)
	TPA	-0.0264	0.00236	-11.181	0.000	(-0.0310, -0.0218)
0.95	Intercept	1.014	0.171	5.943	0.000	(0.679, 1.348)
	TPA	-0.0222	0.00429	-5.174	0.000	(-0.0306, -0.0138)

Table 8: Estimates for the following model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{TPA} + \beta_{p,2}x_{Age} + \varepsilon_i$, where $p = 0.25, 0.5, 0.75$ and 0.95 . 1000 bootstrap replicates were used to calculate the standard errors, p -values and 95% confidence intervals.

p	Variable	Coeff	SE	t -value	p -value	95% CI
0.25	Intercept	-3.849	0.124	-31.111	0.000	(-4.091, -3.606)
	TPA	-0.0291	0.00386	-7.535	0.000	(-0.0367, -0.0215)
	Age	0.0325	0.00263	12.350	0.000	(0.0273, 0.0376)
0.5	Intercept	-3.375	0.103	-32.841	0.000	(-3.577, -3.174)
	TPA	-0.0268	0.00216	-12.456	0.000	(-0.0311, -0.0226)
	Age	0.0379	0.00107	35.331	0.000	(0.0358, 0.0400)
0.75	Intercept	-2.871	0.107	-26.908	0.000	(-3.080, -2.662)
	TPA	-0.0233	0.00189	-12.320	0.000	(-0.0271, -0.0196)
	Age	0.0417	0.00111	37.567	0.000	(0.0395, 0.0439)
0.95	Intercept	-1.773	0.186	-9.556	0.000	(-2.137, -1.409)
	TPA	-0.0179	0.00332	-5.386	0.000	(-0.0244, -0.0114)
	Age	0.0422	0.00181	23.284	0.000	(0.0387, 0.0458)

3.6.3 Pseudo- R^2 and average of the absolute differences of $\hat{Q}_{y_i}(p|\epsilon)$

Let us now summarize the result from the pseudo- R^2 values in Table 9. It shows that $\epsilon = 0.4999$ gives the highest value for all four regression models.

The results of the average values of $|\hat{Q}_{y_i}(p|\epsilon_1) - \hat{Q}_{y_i}(p|\epsilon_2)|$, (where the variable age was excluded from the model), can be found in Table 10. Table 11 in Appendix A1 gives the results of the average values when the variable age was added into the model. The tables show that every value of the average of the absolute differences of $\hat{Q}_{y_i}(p|\epsilon)$ are strictly less than 1. According to the method in [27], one should therefore choose $\epsilon = 10^{-11}$, although Table 10 indicates that ϵ has a small impact on the result. The regression results when ϵ is set to be 10^{-11} can be seen in Appendix A2 (page 40).

Table 9: Summary of the pseudo- R^2 values for the four different regression models (Model 1 indicates the first categorical model and Model 2 the second categorical model, while Model 3 stands for the first continuous regression model and Model 4 for the second continuous regression model) with $p = 0.5$, the median. The values in blue depict the highest pseudo- R^2 values.

ϵ	Model 1	Model 2	Model 3	Model 4
10^{-11}	0.00078428	0.0079329	8.9896e-05	0.0079198
0.0001	0.0016112	0.016299	0.00018468	0.016272
0.0005	0.0018006	0.018216	0.00020639	0.018185
0.001	0.0018965	0.019186	0.00021738	0.019154
0.005	0.0021627	0.021882	0.00024786	0.021846
0.01	0.0023002	0.023277	0.00026358	0.023239
0.05	0.0026834	0.027188	0.00030717	0.027143
0.08	0.0028116	0.028512	0.00032159	0.028465
0.1	0.0028742	0.029164	0.00032859	0.029116
0.2	0.0030708	0.031243	0.00035018	0.031192
0.3	0.0031809	0.032446	0.00036179	0.032392
0.4999	0.0033022	0.033837	0.00037372	0.033779

Table 10: The average values of $|\hat{Q}_{y_i}(p|\epsilon_1) - \hat{Q}_{y_i}(p|\epsilon_2)|$. ϵ_1 and ϵ_2 are two consecutive ϵ -values. The first row shows the results of the average value when $\epsilon_1 = 10^{-11}$ and $\epsilon_2 = 0.0001$. The regression model we used was $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{TPA} + \varepsilon_i$. The last row gives NA (not available values) thus we do not have ϵ_2 -value when $\epsilon_1 = 5$.

ϵ	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 0.95$
10^{-11}	6.2106e-16	1.6512e-06	7.4814e-07	5.0602e-08
0.0001	3.2288e-16	6.6032e-06	2.9923e-06	2.0240e-07
0.0005	2.8092e-16	8.2508e-06	3.7400e-06	2.5298e-07
0.001	8.0213e-16	6.5875e-05	2.9902e-05	2.0230e-06
0.005	3.8783e-16	8.2019e-05	3.7334e-05	2.5268e-06
0.01	2.5547e-16	0.00064342	0.00029691	2.0136e-05
0.05	2.2848e-16	0.00046819	0.00022065	1.5010e-05
0.08	2.6786e-16	0.00030554	0.00014614	9.9638e-06
0.1	6.5048e-17	0.0015798	0.00071945	4.9306e-05
0.2	1.1318e-16	0.0045108	0.00070117	4.8466e-05
0.3	1.1793e-15	0.0011085	0.0013494	9.4442e-05
0.4999	1.2652e-15	0.0048082	0.0023189	0.00022284
1	1.5259e-15	0.0062339	0.0051577	0.00039469
2	7.2892e-16	0.0038600	0.0042437	0.00033776
3	1.2625e-15	0.0045061	0.0040694	0.00054429
5	NA	NA	NA	NA

Plotting the logit of I-PSS against TPA with the ϵ -values seen in Table 10 (see Figure 10 in Appendix A1) shows that ϵ between 0.4999 and 2 gives a plot resembling the original data. We therefore also perform regression analysis with $\epsilon = 1$ and the regression results with this ϵ -value can be found in Appendix A3 (page 46).

3.7 Interpretation

In this section, we interpret the results we summarized in the tables of the result section. We also briefly interpret the results in Appendix A2 and A3.

3.7.1 First model

Let us interpret the results from the first model (see Table 5) first. The estimated coefficients in that table show that a lower TPA score seems to lead to a higher chance of getting higher I-PSS score. We see that if a person has a TPA score between 35.7-36.9 (Interval 2), he will probably get a logit transformed I-PSS score with approximately 0.283 lower value than a person with a TPA score lower than 35.7 (Interval 1). We also see from Table 5 that a person with a TPA score that lies within Interval 3 to Interval 11 will more likely get 0.283 lower value of the transformed I-PSS compared to a

person with TPA score that lies between the values in Interval 1. Similarly, we find that a person with a TPA score within Interval 12 will more likely get 0.649 lower value of the transformed I-PSS score than a person with TPA score in Interval 1.

We can also see that every coefficient gives a significant result at the 5% confidence level (thus the p -values are strictly less than 0.05), so the hypothesis that the difference of the coefficients of Interval 1 and each interval variable in the table is equal to zero can be rejected at the 5% level. The result from the F-test tells us that the joint hypothesis can be rejected at an extremely small percentage level. This implies that the interval variables seem to have an effect on the I-PSS score, so total physical activity may affect LUTS.

3.7.2 Second model

The results from the second model (see Table 6),

$$h(y_i) = \beta_{p,0} + \beta_{p,1}x_{Interval2} + \beta_{p,2}x_{Interval3} + \dots + \beta_{p,11}x_{Interval12} + \beta_{p,12}x_{Age} + \varepsilon_i,$$

(where $p = 0.5$) also show that higher TPA scores are more likely associated with lower I-PSS scores. The estimated coefficients of each interval variable are negative, which implies that the I-PSS score is lower for an individual with a TPA score greater than 35.6 compared to a person with TPA score within Interval 1. The estimated coefficient of Interval 2 is -0.0766 , so a person with a TPA score within Interval 2 will probably get 0.0766 smaller logit transformed I-PSS score than a person with TPA score in Interval 1. The lowest coefficient of the interval predictors is -0.435 and it is the estimated coefficient of Interval 12. This implies that a person with a TPA score within the highest interval will more likely have a 0.435 lower value of the logit I-PSS compared to an individual with a TPA score lower than or equal to 35.6 (Interval 1). We can also see from Table 6 that the estimated coefficient of the variable age is 0.0382, which means that raising the age by 1 year will lead to an increase of the logit of I-PSS by 0.0382 units. In other words, the older a person is, the more likely it is to get a higher I-PSS score.

Almost all p -values are smaller than 0.05, with Interval 2 as an exception, which means that the variables in the table give a significant result at the 5% level. This implies that we cannot reject the hypothesis that the difference of the estimated coefficient of Interval 1 and Interval 2 is zero. The results from the F-test tell us that the joint hypothesis that every coefficient is all equal to zero can be rejected.

The overall results seem to show that total physical activity and age do have some sort of effect on the I-PSS score. A weaker effect of total physical activity on lower urinary tract symptoms can be seen when the explanatory variable age is added to the regression model for TPA score in Interval 12 and activity score lower than Interval 8. On the other hand, the effect when

TPA score is between Interval 8 and 11 gives lower values when the age variable is included in the model.

3.7.3 Third model

Table 7 shows that the estimated coefficient of the variable TPA is approximately 0.000, -0.0229 , -0.0264 , -0.0222 for p equal to 0.25, 0.5, 0.75 and 0.95 respectively of the third regression model. This means that increasing the TPA by one unit, the logit of I-PSS will increase by a small value, approximately zero, when $p = 0.25$ and the logit of I-PSS will decrease by 0.0229, 0.0264 and 0.0222 units in the model where $p = 0.5, 0.75$ and 0.95 respectively. TPA seems to not have an effect on I-PSS for the logistic quantile regression model with lower quantiles, but starting somewhere between $p = 0.25$ and $p = 0.5$ it seems to have a negative effect on LUTS. From Figure 8 we see that the estimated coefficient of TPA starts to decrease and becomes negative around $p = 0.3$ and 0.4, and the coefficient seems to stay between -0.03 and -0.02 for the model where the quantile is greater than the median. The estimated coefficient of the intercept is negative for $p = 0.25, 0.5, 0.75$, while it gives a positive value when $p = 0.95$ (Figure 8 show that the coefficient gets positive around $p = 0.8$).

Every coefficient from Table 7 are significant at the 5% level, thus the p -values are all strictly less than 0.05. This implies that the null hypothesis that TPA does not have an effect on I-PSS can be rejected, in other words TPA seems to affect I-PSS.

3.7.4 Fourth model

The results of the multiple regression model (see Table 8) tells us that the logit of I-PSS score will decrease by 0.0291 when the TPA score increases by one unit where $p = 0.25$, and the logit of I-PSS will decrease by 0.0268, 0.0233 and 0.0179 when we increase the TPA level by one unit where $p = 0.5, 0.75$ and 0.95 respectively. We can also see a positive trend between age and the logit of I-PSS, so older men will more likely have moderate or severe lower urinary tract symptoms (higher I-PSS score). It is shown in Table 8 that the estimated coefficient of age when $p = 0.25$ is 0.0325, which implies that increasing the age by one year will lead to an increase in the logit of I-PSS by 0.0325 units.

The p -values of all estimated coefficients are strictly smaller than 0.05, which gives significant results. It is therefore very reasonable to assume that both continuous explanatory variables TPA and age do have some effects on the LUTS.

We also see that the effect TPA has on I-PSS is stronger when age is excluded from the regression model when $p = 0.75$ or 0.95 and weaker when $p = 0.25$ and 0.5.

3.7.5 The model with the best fitted epsilon

The ϵ -value that gives the highest pseudo- R^2 value:

From Table 9 we see that the ϵ -value with the highest pseudo- R^2 for each model is $\epsilon = 0.4999$, which is approximately 0.5, so the results should not differ much compared to our analyses with $\epsilon = 0.5$.

The ϵ -value obtained by looking at the average of the absolute differences of $\hat{Q}_{y_i}(p|\epsilon)$:

The results from the average values of $|\hat{Q}_{y_i}(p|\epsilon_1) - \hat{Q}_{y_i}(p|\epsilon_2)|$ (shown in Table 10 and 11) are all smaller than 1 ($\epsilon \geq 10^{-11}$ gives average values less than 1), which means that the values of $\hat{Q}(p|\epsilon)$ are nearly the same no matter of the choice of ϵ . It should therefore be appropriate to choose $\epsilon = 0.5$ as we did (in Model 1, 2, 3 and 4).

Siao et al. [27] chose the smallest ϵ_1 -value that satisfy

$$\frac{\sum_{i=1}^n (|\hat{Q}_{y_i}(p|\epsilon_1) - \hat{Q}_{y_i}(p|\epsilon_2)|)}{n} < 0.01,$$

for each ϵ_1 -value equals to and is greater than that ϵ_1 -value (e.g. if the values of $1/n \sum_{i=1}^n (|\hat{Q}_{y_i}(p|\epsilon_1) - \hat{Q}_{y_i}(p|\epsilon_2)|) < 0.01$ for each $\epsilon_1 \geq 0.001$ and for each quantiles, then we should choose $\epsilon = 0.001$).

They used 0.01 because the outcome variable was a proportion between 0 and 1. Our response variable is bounded between 0-35, which is not a proportion, we therefore check if the average values are smaller than 1 instead of 0.01. Let us interpret the results of the regression models when ϵ is set to 10^{-11} ($\epsilon \geq 10^{-11}$ gives values smaller than 1). The results of the simple categorical regression model can be found in Table 14 in the Appendix A2. We can see that higher TPA scores are associated with lower I-PSS scores from the results. A male individual with TPA score within Interval 2, 3, ..., 11 are assumed to have approximately 0.319 lower logit I-PSS score than a man with a TPA score strictly lower than 35.7 (Interval 1). The table also shows that the logit I-PSS score will be almost 0.756 lower for a man with TPA score in Interval 12 compared to a male with a TPA level within Interval 1. The coefficients show significance on the 5% confidence level (with exception for Interval 11) and the F-test concludes that the joint null hypothesis of no effect of TPA score can be rejected.

The results from the model where the variable age is included (Table 13 in Appendix) also seem to show that higher TPA lead to lower I-PSS values.

The continuous regression models (see Table 14 and Table 15) show that increasing TPA lead to a decreasing I-PSS score.

The ϵ -value obtained visually:

Figure 10 in Appendix plots the logit of the I-PSS distribution against the desired explanatory variable TPA. We see that ϵ -values around 0.4999 and 2

give plots more similar to the original data and this is the reason why $\epsilon = 1$ was used as we stated before.

The results of each regression model where we take $\epsilon = 1$ are summarized in Table 16, 17, 18 and 19. These show similar results as when ϵ was set to 0.5 and 10^{-11} , in other words a negative relationship between higher TPA score and I-PSS can be seen. The results do not really differ much for the three different ϵ values we used.

4 Discussion

In this thesis we have used logistic quantile regression in order to investigate the effect total physical activity has on lower urinary tract symptoms. When fitting the regression models to data, the results show (regardless if ϵ was set to 10^{-11} , 0.5 or 1) a negative trend between total physical activity and International Prostate Symptom Score, I-PSS, which means that physical inactivity seems to be related to a higher I-PSS score (more severe lower urinary tract symptoms). It does not seem to matter which values of ϵ we choose, as we stated in the interpretation section, the models fits for different ϵ give similar results and the average of the absolute differences of $\hat{Q}_{y_i}(p|\epsilon)$ gave nearly the same values regardless of which ϵ we used.

4.1 Limitations

Orsini et al. [30] discussed some possible limitations, and one of them was that the result might differ if those who did not answer to the questionnaires would have answered. For example, those men who did not respond to the survey might exercise more or less compared to those we observed, which can lead to over- and underestimation of the total physical activity score. They also discussed that the mean of MET is not an optimal measurement for total physical activity, because the males in the survey are not assumed to perform different physical activities.

Another limitation might be difficulties of measuring the exact total physical activity. Especially when one of the questions in the questionnaire asked the men about the usual number of hours spent on physical activities when they were 30 years old, but it can be hard to remember that. Orsini et al. [30] briefly discussed that these reported physical activity values might therefore not be valid.

Our data consists of three variables, one of which is the response variable. There are probably many other explanatory variables that might affect the I-PSS. Some of these may have a confounding effect on the relation between physical activity and I-PSS, for instance.

4.2 Possible future work

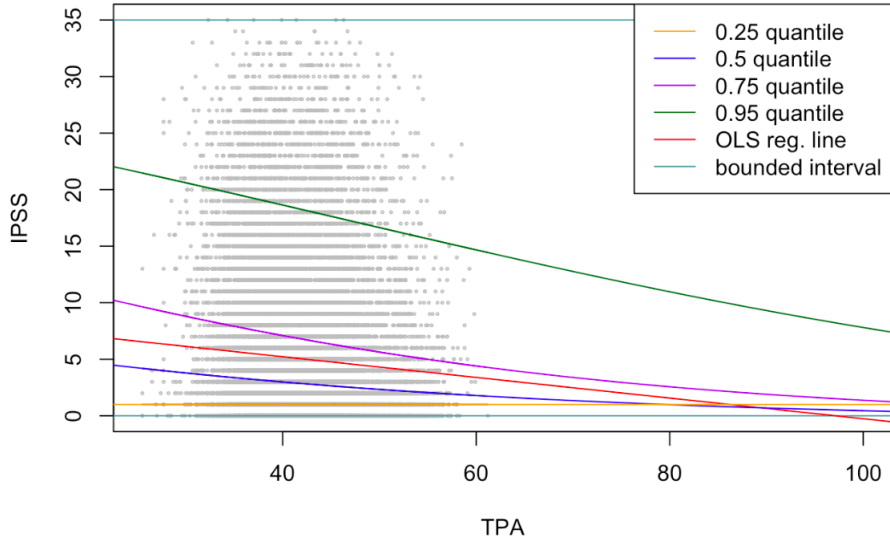
Future studies should take physical activity into consideration, because total physical activity is shown to have some sort of negative effect on lower urinary tract symptoms in our study and several other researches have also seen a negative trend between physical activity and I-PSS [9, 25, 28, 30]. Thus, since it seems like higher physical activity is associated with having milder symptoms, we should encourage both younger and older people to exercise and be more physically active.

4.3 Personal Considerations

We have a quite large sample, which leads to more accurate results. Another strength is that we used logistic quantile regression in order to see how the data behaves for different quantiles (see Figure 4). Thirdly, choosing $\epsilon = 0.5$ was a good choice since two out of three methods showed that the value 0.5 gave better fits and more accurate results. The results of the method when we calculated the average of the absolute differences of $\hat{Q}_{y_i}(p|\epsilon)$ showed that the $\hat{Q}_{y_i}(p|\epsilon)$ -values was nearly the same for all the ϵ -values we tried, so this implies that choosing $\epsilon = 0.5$ will probably not change the regression results much.

Figure 4: The simple OLS regression line (red) does not take the bounded interval into account, thus when the TPA score is between 90 and 100, the I-PSS score is assumed to be lower than 0. The predicted transformed values of the logistic quantile regression lines ($p = 0.25, 0.5, 0.75, 0.95$) on the other hand, approaches the lower limit ($y_{min}=0$) when TPA increases (they take the bounded interval into account).

Predicted regression lines



As we stated in Section 4.1 Limitations, I-PSS might depend on some other variables that we do not have data for, like genetic causes, biological causes, different environmental conditions, different health conditions and more.

We computed the F-statistics for OLS models, which may lead to biased results. Another possibility is measurement errors in the observations due to that some individuals were not honest when answering some question in the survey.

5 References

- [1] Agresti, A. (2002): *Categorical Data Analysis*(Second Edition). Hoboken, New Jersey: Wiley-Interscience.
- [2] Andrews., D.W.K., & Buchinsky, M. (2000): A Three-Step Method for Choosing the Number of Bootstrap Repetitions. *Econometrica*, Volume 68, No. 1, 23-51.
- [3] Andrology Australia (2014): Lower Urinary Tract Symptoms (LUTS) in men. Available at: <https://andrologyaustralia.org/your-health/lower-urinary-tract-symptoms-luts-in-men/> [Accessed 12 Apr. 2018].
- [4] Bottai, M., Cai, B., & McKeown, R.E. (2010): Logistic quantile regression for bounded outcomes. *Statistics in Medicine*, Volume 29, Issue 2, 309-317. DOI: 10.1002/sim.3781.
- [5] Buchinsky, M. (1995): Estimating the asymptotic covariance matrix for quantile regression models A Monte Carlo study. *Journal of Econometrics*, Volume 68, Issue 2, 303-338. DOI: 10.1016/0304-4076(94)01652-G.
- [6] Cade, B.S., & Noon, B.R. (2003): A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, Volume 1, Issue 8, 412-420. DOI: 10.1890/1540-9295(2003)001[0412:AGITQR]2.0.CO;2.
- [7] Chernick M.R. (2008): *Bootstrap Methods: A Guide for Practitioners and Researchers*(Second Edition). Hoboken, New Jersey: Wiley-Interscience.
- [8] Chernick, M.R., & Friis, R.H. (2003): *Introductory Biostatistics for the Health Sciences: Modern Applications Including Bootstrap*. Hoboken, New Jersey: Wiley-Interscience. ISBN 0-471-41137-X.
- [9] Choo M. S., Han J.H., Shin T.Y., Ko K., Lee W.K., Cho S.T., Lee S.K., & Lee S.H. (2015): Alcohol, Smoking, Physical Activity, Protein, and Lower Urinary Tract Symptoms: Prospective Longitudinal Cohort. *International Neuourology Journal*, Volume 19, Issue 3, 197-206. DOI: 10.5213/inj.2015.19.3.197.
- [10] Columbu, S., & Bottai, M. (2016): Logistic Quantile Regression to Model Cognitive Impairment in Sardinian Cancer Patients. In *Topics on Methodological and Applied Statistical Inference, Studies in Theoretical and Applied Statistics*, Di Battista, T., Moreno E. & Racugno, W. eds. Springer, Cham, Switzerland, 65-73. DOI: 10.1007/978-3-319-44093-4_7.

- [11] Davino, C., Furno, M., & Vistocco, D. (2013): *Quantile Regression : Theory and Applications*. Wiley. ISBN 9781119975281.
- [12] Efron, B. & Tibshirani, R.J. (1994): *An introduction to the bootstrap*. New York: Chapman & Hall.
- [13] Ernstgård, L., & Bottai, M. (2011): Visual analogue scales: how can we interpret them in experimental studies of irritation in the eyes, nose, throat and airways? *Journal of Applied Toxicology*, Volume 32, Issue 10, 777-782. DOI: 10.1002/jat.1681.
- [14] Held, L. & Sabanés Bové, D. (2014): *Applied Statistical Inference: Likelihood and Bayes*. Berlin, Heidelberg: Springer Berlin Heidelberg.
- [15] Holbrook, J.R., Cuffe, S.P., Cai, B., Visser, S.N., Forthofer, M.S., Bottai, M., Ortaglia, A., & McKeown, R.E. (2016): Persistence of Parent-Reported ADHD Symptoms From Childhood Through Adolescence in a Community Sample. *Journal of Attention Disorders*, Volume 20, Issue 1, 11–20. DOI: 10.1177/1087054714539997.
- [16] Hu, F.B., Stampfer, M.J., Colditz, G.A., Ascherio, A., Rexrode K.M., Willett. W.C., & Manson J.E. (2000): Physical Activity and Risk of Stroke in Women. *JAMA*. 2000;283(22):2961–2967. DOI: 10.1001/jama.283.22.2961.
- [17] Karolinska Institutet (2018): Epidemiology and Public Health Sciences. Available at: <https://ki.se/en/research/epidemiology-and-public-health-sciences> [Accessed 13 Apr. 2018]
- [18] Koenker, R. (2005): *Quantile Regression* (Econometric Society Monographs). Cambridge: Cambridge University Press. DOI: 10.1017/CBO9780511754098.
- [19] Koenker, R. (2000): *Quantile regression*. Department of Economics, University of Illinois, Urbana-Champaign, Champaign. Version: 10 Nov. 2000. Available at: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.394.3210&rep=rep1&type=pdf>.
- [20] Koenker R., & Machado J.A.F. (1999): Goodness of Fit and Related Inference Processes for Quantile Regression. *Journal of the American Statistical Association*, Volume 94, No. 448, 1296-1310. DOI: 10.2307/2669943.
- [21] Koenker. R. (2006): Pseudo R for Quant Reg. Available at: <https://stat.ethz.ch/pipermail/r-help/2006-August/110386.html> [Accessed 9 Apr. 2018].

- [22] Laurin, D., Verreault, R., Lindsay, J., MacPherson, K., & Rockwood, K. (2001): Physical Activity and Risk of Cognitive Impairment and Dementia in Elderly Persons. *Arch Neurol.* 2001;58(3):498–504. DOI: 10.1001/archneur.58.3.498.
- [23] Lesafre, E., Rizopoulos, D., & Tsonaka, R. (2007): The logistic transform for bounded outcome scores. *Biostatistics*, Volume 8, Issue 1, 72–85. DOI: 10.1093/biostatistics/kxj034.
- [24] Lynch, B.M., Neilson, H.K., & Friedenreich C.M. (2010): Physical Activity and Breast Cancer Prevention. In: Courneya K., Friedenreich C. (eds) *Physical Activity and Cancer*. Recent Results in Cancer Research, Volume 186, 13–42. Springer, Berlin, Heidelberg.
- [25] Parsons J.K., & Kashefi, C. (2008): Physical activity, benign prostatic hyperplasia, and lower urinary tract symptoms. *European Urology*, Volume 53, Issue 6, 1228–1235. DOI: 10.1016/j.eururo.2008.02.019.
- [26] Salottolo, K., Stewart Levy, A., & Slone, D.S (2014): The Effect of Age on Glasgow Coma Scale Score in Patients With Traumatic Brain Injury. *JAMA Surg.* 2014;149(7):727–734. DOI: 10.1001/jamasurg.2014.13.
- [27] Siao, J.S., Hwang, R.C., & Chu C.K. (2016): Predicting recovery rates using logistic quantile regression with bounded outcomes. *Quantitative Finance*, 16:5, 777–792. DOI: 10.1080/14697688.2015.1059952.
- [28] Smith, D.P., Weber, M.F., Soga, K., Korda, R.J., Tikellis, G., Patel, M.I., Clements, M.S., Dwyer, T., Latz, I.K., & Banks, E. (2014): Relationship between Lifestyle and Health Factors and Severe Lower Urinary Tract Symptoms (LUTS) in 106,435 Middle-Aged and Older Australian Men: Population-Based Study. *PLoS ONE*, Volume 9, Issue 10, e109278. DOI: 10.1371/journal.pone.0109278.
- [29] Orsini, N., & Bottai, M. (2011): Logistic quantile regression in Stata. *The Stata Journal*, Volume 11, Number 3, 327–344.
- [30] Orsini, N., RashidKhani, B., Andersson, S-O., Karlberg, L., Johansson, J-E. & Wolk, A. (2006): Long-Term Physical Activity and Lower Urinary Tract Symptoms in Men. *The Journal of Urology*, Volume 176, Issue 6, 2546–2550. DOI: 10.1016/j.juro.2006.07.030.

Appendixes

A1 Plots and a table

Plots

Figure 5: The tilted absolute value function (or loss function) can be seen in the following plots. The second plot illustrate the tilted absolute value function for four different quantiles ($p = 0.25, 0.5, 0.75, 0.95$).

The tilted absolute value function for quantile regression

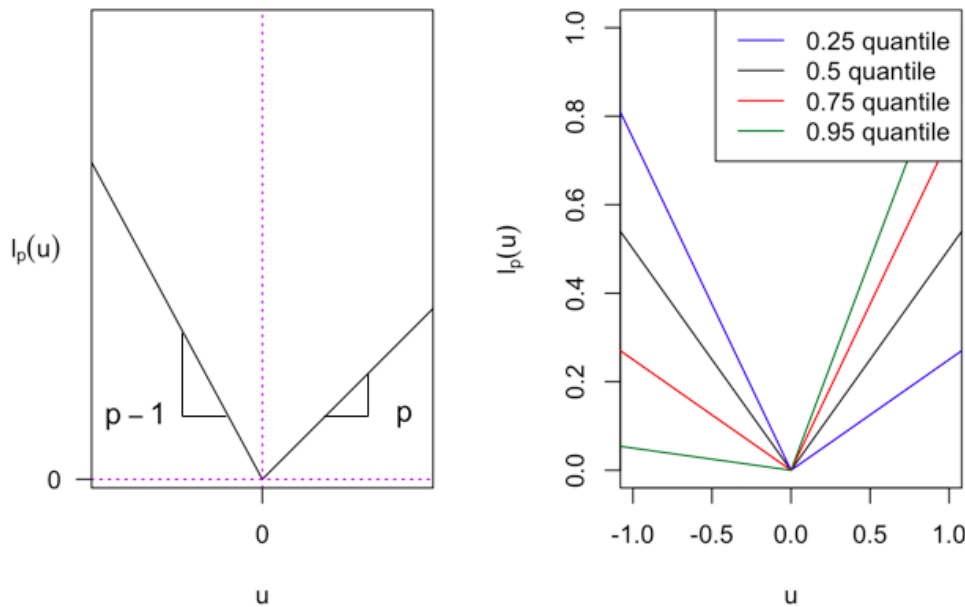


Figure 6: Plots of the coefficient behavior of each variable in our first model over different quantiles (p). The black dots in the plots denote the estimated coefficient of the specific variable (see the title of each plot) for the following values of p : $i/20$ for $i = 1, 2, \dots, 19$. We can also see a grey colored shade around the approximate regression line (black dash-dotted curve) which show the 95% confidence intervals of the coefficients. The red straight line shows the corresponding OLS estimate of the regression coefficient and the red dashed lines around the OLS estimate show the 95% confidence interval.

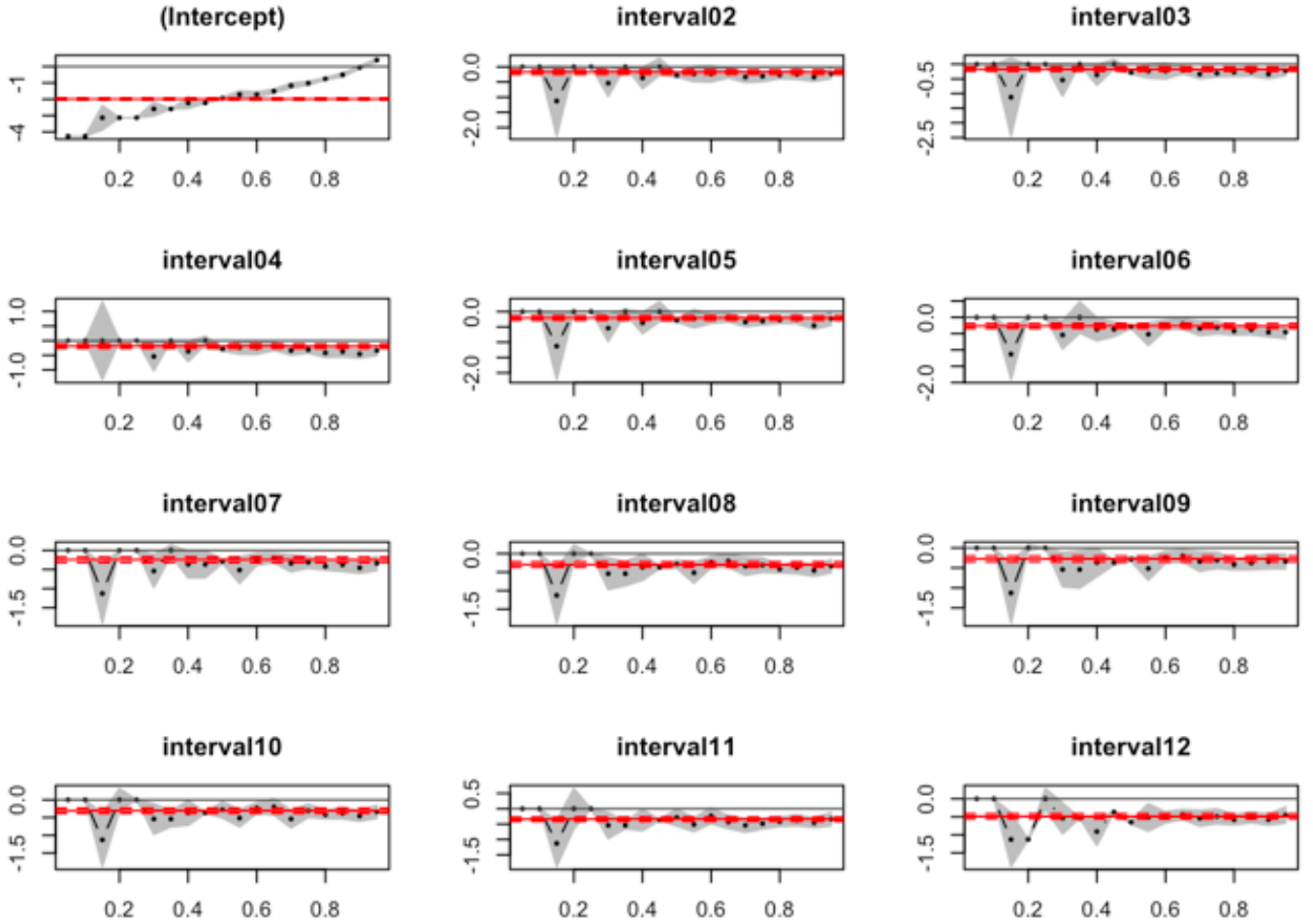


Figure 7: The regression coefficients of the second model, where the variable age is included in the categorical regression model, are shown. We have the quantile p on the x-axis and the black dots show the estimated coefficient of the variable for $p = 0.05, 0.1, 0.15, \dots, 0.95$. The black lines between the dots show the approximate regression line and the grey shades are the 95% confidence intervals of the estimated regression coefficients. The red dashed lines show the 95% confidence interval of the estimated OLS regression coefficients when OLS method is used instead of logistic quantile regression. The red solid lines show us the estimated OLS regression coefficients of the variables.

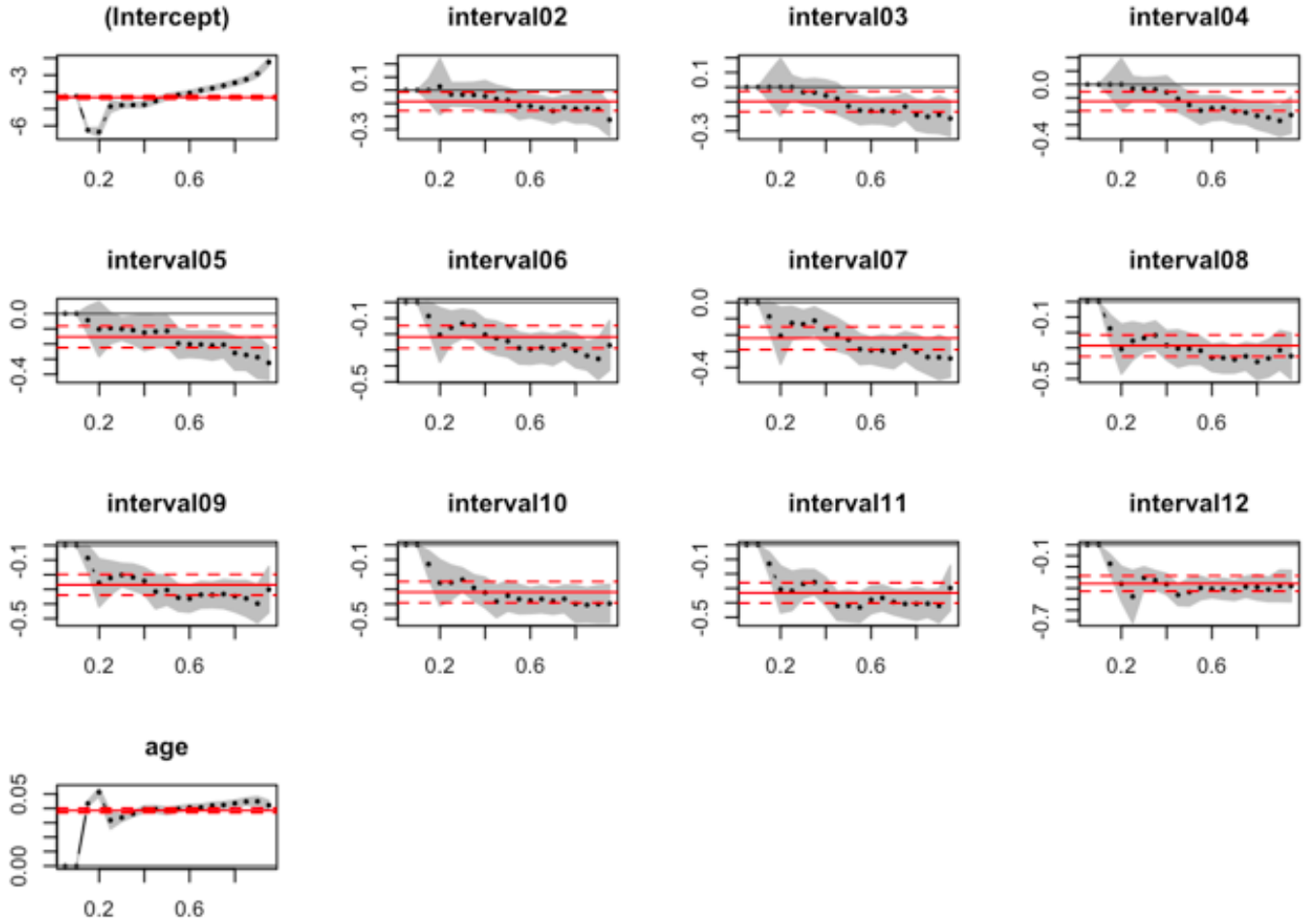


Figure 8: Plots of the estimated coefficients of the intercept and the TPA variable of the third model ($h(y_i) = \beta_{p,0} + \beta_{p,1}x_{TPA} + \varepsilon_i$). The black dots in the plots show the estimated coefficient of a certain variable (intercept in one plot and effect of TPA in the other) for the following quantiles (p): 0.05, 0.1, 0.15,..., 0.95. An approximate regression line can be shown with the black lines between the dots. The grey shades around the regression line show the 95% confidence intervals of the regression coefficients. The red lines illustrate the estimated OLS regression coefficients and their 95% confidence intervals.

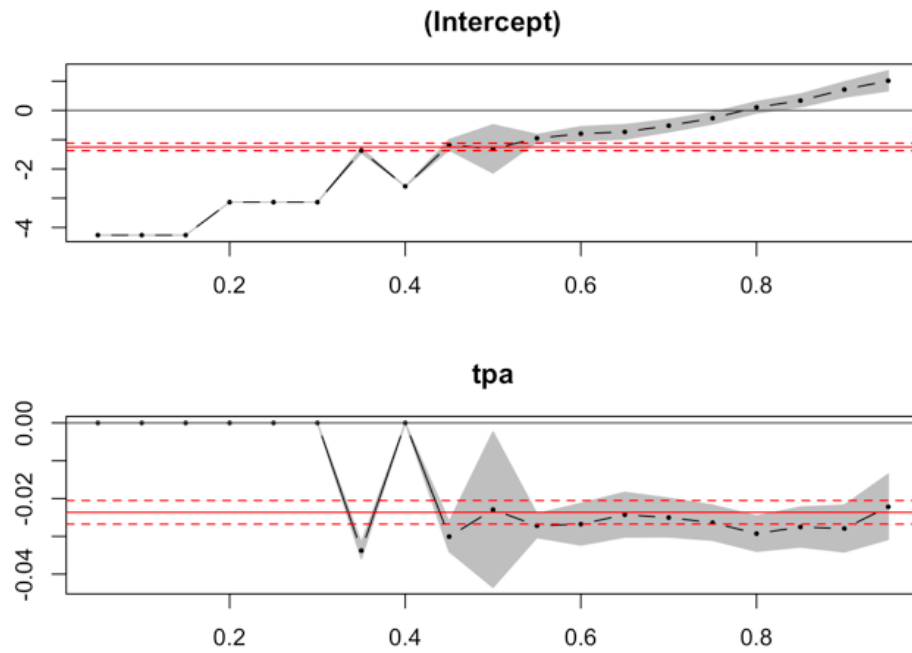


Figure 9: Plots of the estimated coefficients of the variables: intercept, TPA and age over different quantiles, for the fourth model. The grey parts show the 95% confidence intervals of the estimated coefficients and the red lines illustrate the estimated OLS regression coefficients and the 95% confidence intervals of these coefficients.

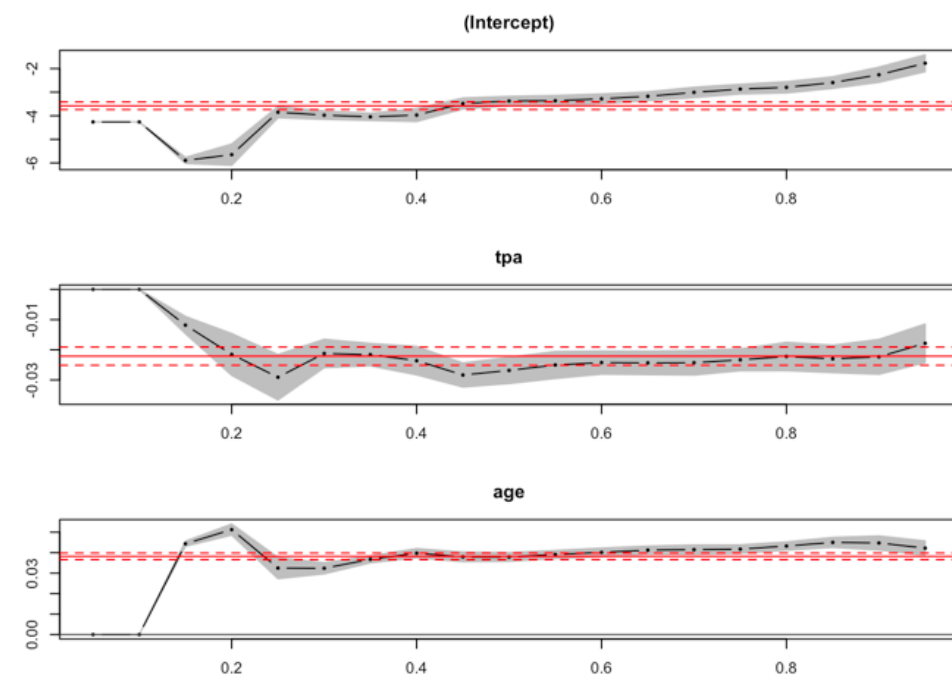
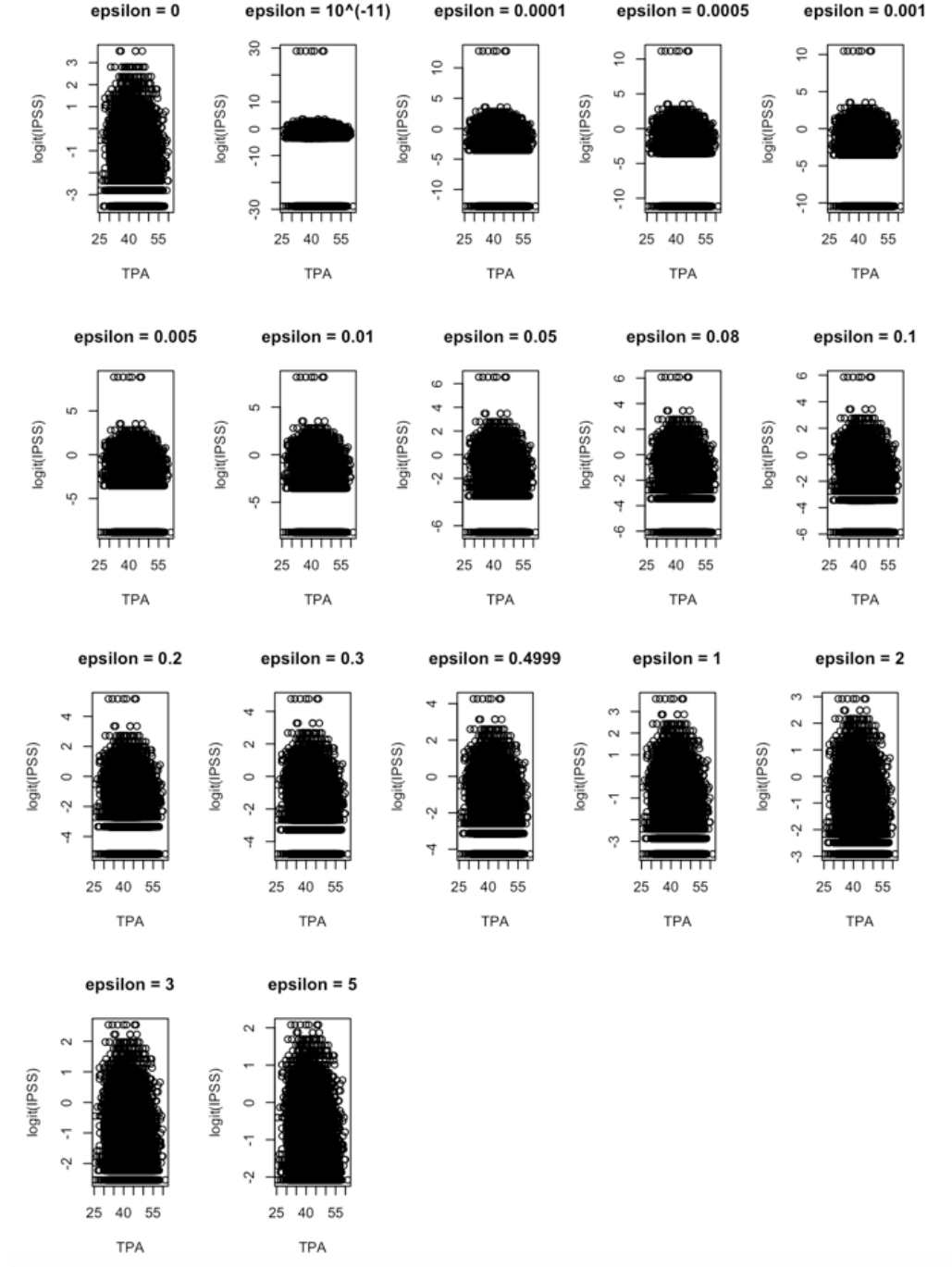


Figure 10: Different plots over the $\text{logit}(\text{IPSS})$ distribution against TPA for all the ϵ values we tested to find the best fit. The first plot with $\epsilon = 0$ corresponds to the original data of the $\text{logit}(\text{IPSS})$ distribution as function of TPA.



The average values of the absolute differences of $\hat{Q}_{y_i}(p|\epsilon)$

Table 11: The average values of $|\hat{Q}_{y_i}(p|\epsilon_1) - \hat{Q}_{y_i}(p|\epsilon_2)|$, where ϵ_1 and ϵ_2 are two consecutive ϵ values. The first row shows the results of the average value when $\epsilon_1 = 10^{-11}$ and $\epsilon_2 = 0.0001$. The regression model we used was $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{TPA} + \beta_{p,2}x_{Age} + \varepsilon_i$. Note that NA stands for non available value in the last row. This is due to not having ϵ_2 -value when $\epsilon_1 = 5$.

ϵ	$p = 0.25$	$p = 0.5$	$p = 0.75$	$p = 0.95$
10^{-11}	8.3369e-06	6.9578e-06	6.4128e-06	6.2909e-07
0.0001	3.3337e-05	2.7827e-05	2.5649e-05	2.5162e-06
0.0005	4.1649e-05	3.4774e-05	3.2055e-05	3.1450e-06
0.001	0.00033229	0.00027781	0.00025622	2.5150e-05
0.005	0.00041313	0.00034632	0.00031973	3.1411e-05
0.01	0.0032183	0.0025196	0.0018752	0.00025023
0.05	0.0023171	0.0019020	0.0025947	0.00018645
0.08	0.0015011	0.0017690	0.0014016	0.00012373
0.1	0.0070264	0.0065230	0.0041050	0.00061180
0.2	0.0063196	0.0060435	0.0042855	0.00060062
0.3	0.010912	0.011026	0.010026	0.0013419
0.4999	0.020546	0.021724	0.018067	0.0028356
1	0.019136	0.031081	0.029358	0.0047892
2	0.011339	0.019005	0.023112	0.0040606
3	0.010865	0.025957	0.033136	0.013225
5	NA	NA	NA	NA

A2 Results when $\epsilon = 10^{-11}$

Table 12: Estimates when $\epsilon = 10^{-11}$ for the model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{Interval2} + \beta_{p,2}x_{Interval3} + \dots + \beta_{p,11}x_{Interval12} + \varepsilon_i$. The regression variables, estimated coefficients, standard errors, t -values, p -values and 95% confidence intervals when $p = 0.5$ are shown. 1000 bootstrap replicates were used to get the standard errors, p -values and confidence intervals. We can also see the result of the F-test, where we test the joint hypothesis ($\beta_{0.5,1} = \beta_{0.5,2} = \dots = \beta_{0.5,11} = 0$).

p	Variable	Coeff	SE	t -value	p -value	95% CI
0.5	Intercept	-2.048	0.0285	-71.923	0.000	(-2.103, -1.992)
	Interval 2	-0.319	0.0285	-11.220	0.000	(-0.375, -0.264)
	Interval 3	-0.319	0.0285	-11.220	0.000	(-0.375, -0.264)
	Interval 4	-0.319	0.0348	-9.179	0.000	(-0.388, -0.251)
	Interval 5	-0.319	0.0285	-11.220	0.000	(-0.375, -0.264)
	Interval 6	-0.319	0.0285	-11.220	0.000	(-0.375, -0.264)
	Interval 7	-0.319	0.0285	-11.220	0.000	(-0.375, -0.264)
	Interval 8	-0.319	0.0664	-4.814	0.000	(-0.449, -0.189)
	Interval 9	-0.319	0.0397	-8.040	0.000	(-0.397, -0.242)
	Interval 10	-0.319	0.131	-2.439	0.0147	(-0.576, -0.0627)
	Interval 11	-0.319	0.171	-1.873	0.0611	(-0.654, 0.0148)
	Interval 12	-0.756	0.0285	-26.542	0.000	(-0.811, -0.700)
F-test:						
F-value		11.354	p -value:		<2.2e-16	

See Figure 11 on the next page for a better view of the estimated coefficients.

Figure 11: Plots of the estimated coefficients of the intercept and the interval variables (Interval 2, 3,..., 12) when $\epsilon = 10^{-11}$. The black dots in the plots denote the estimated coefficient of the specific variable for the following quantiles (p): $i/20$ where $i = 1, 2, \dots, 19$. The grey color depicts the 95% confidence intervals of the coefficients, the red lines show the estimated OLS regression coefficients and the red dashed lines correspond to 95% confidence intervals of the estimated OLS coefficients.

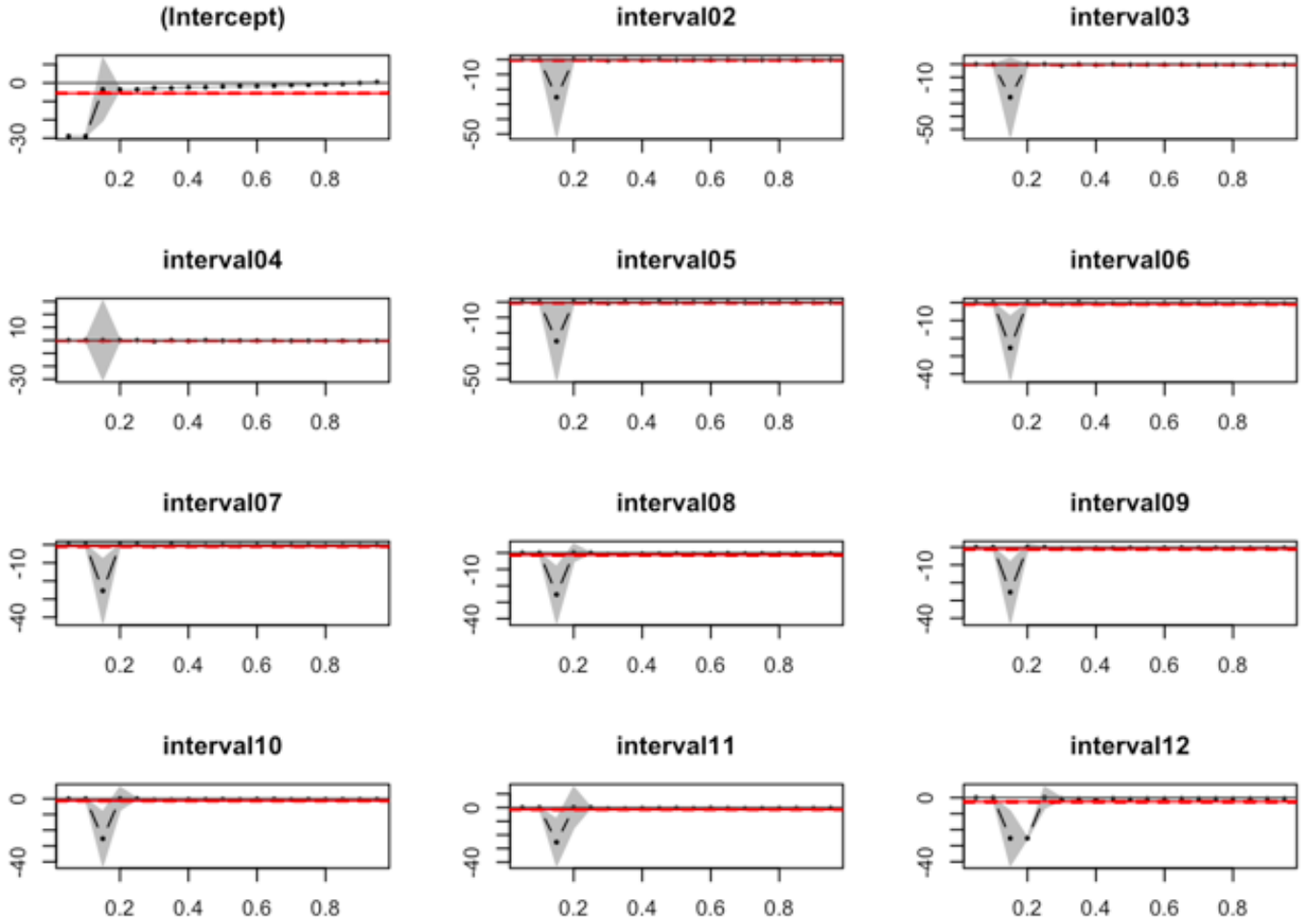


Table 13: Results of fitting the regression model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{Interval2} + \beta_{p,2}x_{Interval3} + \dots + \beta_{p,11}x_{Interval12} + \beta_{p,12}x_{Age} + \varepsilon_i$, where $\epsilon = 10^{-11}$. The table lists the variables, estimated coefficients, standard errors, t -values, p -values and 95% confidence intervals where $p = 0.5$, where 1000 bootstrap replicates were used to obtain the standard errors, p -values and confidence intervals. The result of the F-test, where we test the joint hypothesis ($\beta_{0.5,1} = \beta_{0.5,2} = \dots = \beta_{0.5,11} = 0$) is also shown.

p	Variable	Coeff	SE	t -value	p -value	95% CI
0.5	Intercept	-4.742	0.0867	-54.687	0.000	(-4.912, -4.572)
	Interval 2	-0.0880	0.0569	-1.545	0.122	(-0.200, 0.0236)
	Interval 3	-0.131	0.0526	-2.511	0.0121	(-0.235, -0.0289)
	Interval 4	-0.172	0.0559	-3.082	0.00206	(-0.282, -0.0627)
	Interval 5	-0.128	0.0581	-2.208	0.0272	(-0.242, -0.0144)
	Interval 6	-0.264	0.0541	-4.876	0.000	(-0.370, -0.158)
	Interval 7	-0.260	0.0525	-4.955	0.000	(-0.363, -0.157)
	Interval 8	-0.348	0.0568	-6.130	0.000	(-0.460, -0.237)
	Interval 9	-0.348	0.0584	-5.965	0.000	(-0.463, -0.234)
	Interval 10	-0.392	0.0610	-6.428	0.000	(-0.512, -0.273)
	Interval 11	-0.480	0.0619	-7.756	0.000	(-0.602, -0.359)
	Interval 12	-0.484	0.0693	-6.980	0.000	(-0.620, -0.348)
	Age	0.0440	0.00124	35.600	0.000	(0.0416, 0.0464)
F-test:						
	F-value	11.736	p -value:	<2.2e-16		

We can see Figure 12 on the next page to get a better picture of the estimated coefficients over different quantiles.

Figure 12: Plots of the estimated coefficients of the intercept, the interval variables and the variable age when $\epsilon = 10^{-11}$. The black dots in the plots show the estimated coefficient of the specific variable for the following quantiles: $i/20$ where $i = 1, 2, \dots, 19$. The grey color depicts the 95% confidence intervals of the coefficients, the red lines show the estimated OLS regression coefficients and the red dashed lines correspond to the 95% confidence intervals of the estimated OLS coefficients.

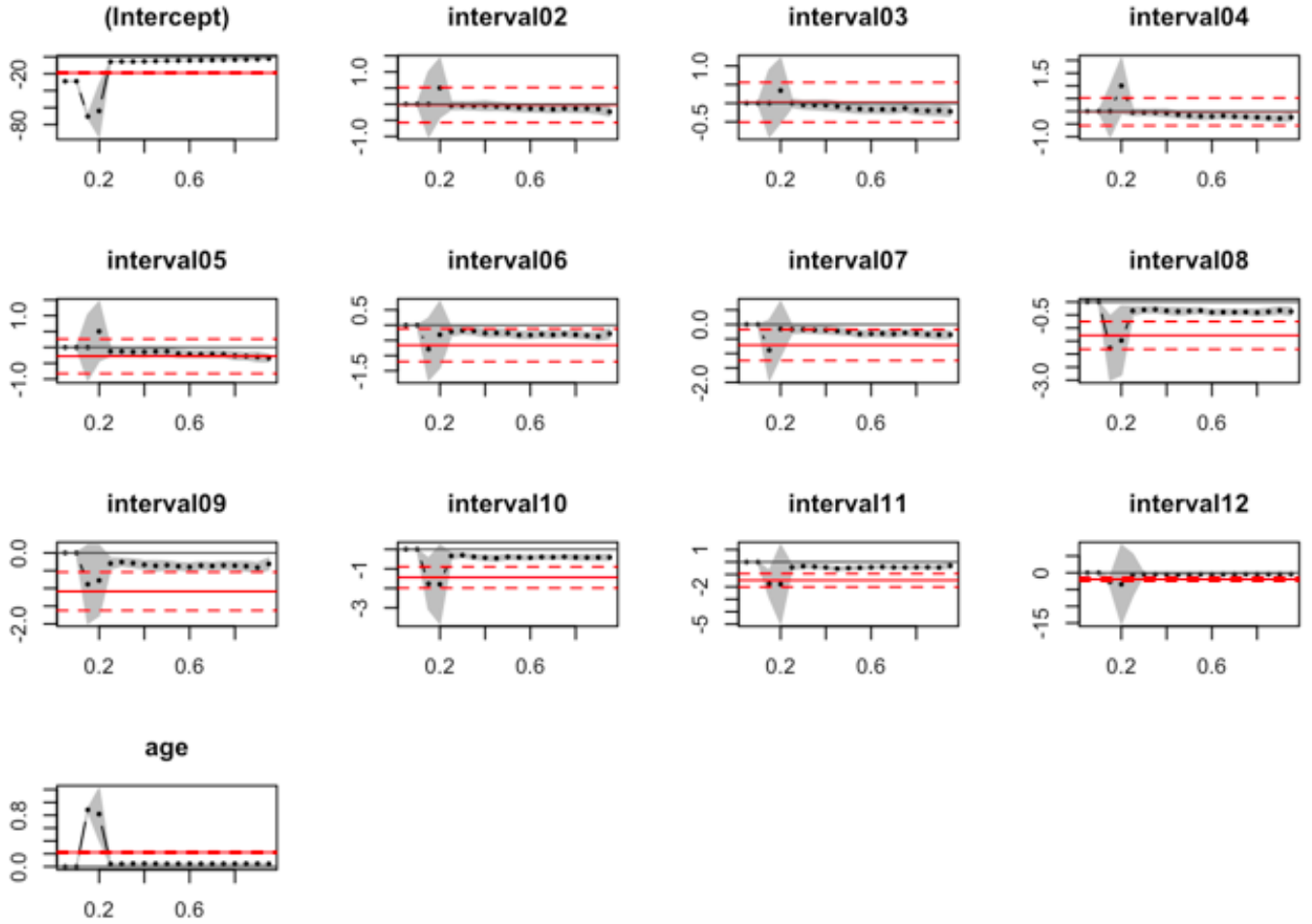


Table 14: Results of fitting the model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{TPA} + \varepsilon_i$, where $\epsilon = 10^{-11}$. The variables, estimated coefficients, bootstrapped standard errors, t -values, p -values and 95% confidence intervals of four different quantiles (0.25, 0.5, 0.75, 0.95) are shown. 1000 bootstrap replicates were used to obtain the standard errors, p -values and confidence intervals.

p	Variable	Coeff	SE	t -value	p -value	95% CI
0.25	Intercept	-3.526	0.000	-2.611e+15	0.000	(-3.526, -3.526)
	TPA	0.000	0.000	-1.286	0.199	(-1.051e-16, 2.183e-17)
0.5	Intercept	-1.295	0.496	-2.611	0.00903	(-2.268, -0.323)
	TPA	-0.0269	0.0124	-2.169	0.0301	(-0.0513, -0.00259)
0.75	Intercept	-0.255	0.107	-2.377	0.0174	(-0.466, -0.0449)
	TPA	-0.0279	0.00251	-11.105	0.000	(-0.0328, -0.0229)
0.95	Intercept	1.043	0.176	5.942	0.000	(0.699, 1.387)
	TPA	-0.0228	0.00441	-5.174	0.000	(-0.0315, -0.0142)

See Figure 13 below for a more complete picture of the estimated coefficients over different quantiles.

Figure 13: The estimated coefficients of the variables intercept and TPA over different quantiles when $\epsilon = 10^{-11}$ (the black dotted values are the estimated regression coefficients, while the black lines between the dots show an approximation of the regression coefficients in between). The grey color in the plots shows the 95% confidence intervals of the coefficients. The red lines on the other hand show the estimated OLS coefficients and their 95% confidence intervals.

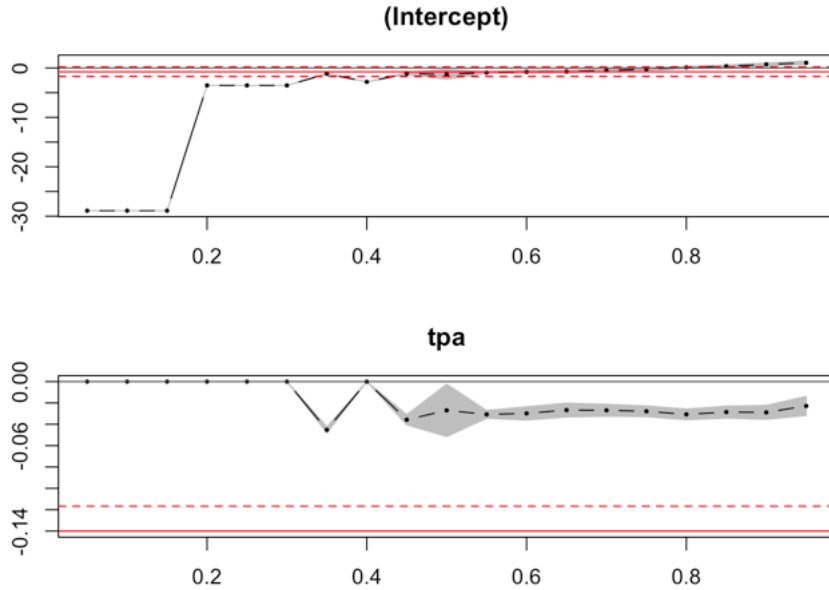
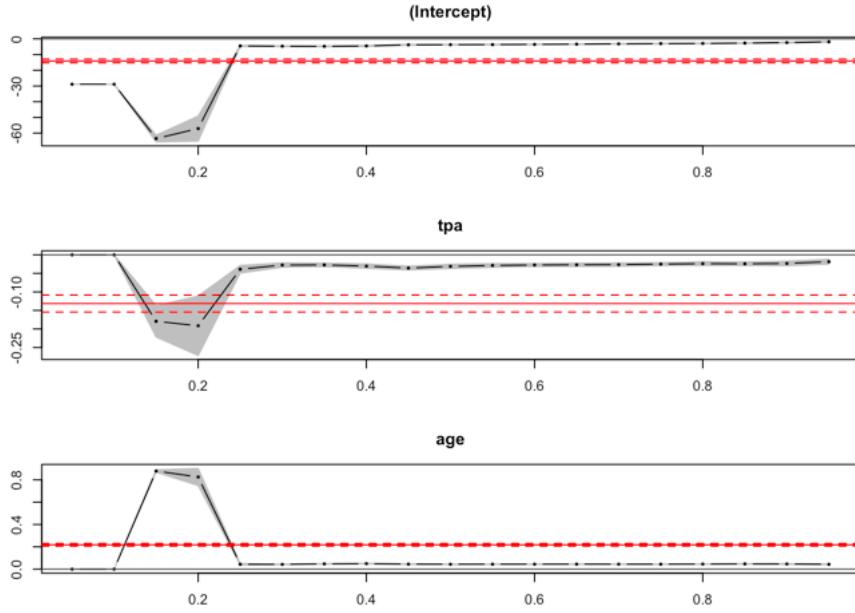


Table 15: Results of fitting the model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{TPA} + \beta_{p,2}x_{Age} + \varepsilon_i$. We used $\epsilon = 10^{-11}$. 1000 bootstrap replications were used to get the standard errors, p -values and 95% confidence intervals.

p	Variable	Coeff	SE	t -value	p -value	95% CI
0.25	Intercept	-4.481	0.165	-27.139	0.000	(-4.805, -4.157)
	TPA	-0.0389	0.00520	-7.487	0.000	(-0.0491, -0.0287)
	Age	0.0435	0.00350	12.408	0.000	(0.0366, 0.0504)
0.5	Intercept	-3.695	0.122	-30.384	0.000	(-3.933, -3.457)
	TPA	-0.0312	0.00277	-11.265	0.000	(-0.0366, -0.0258)
	Age	0.0438	0.00127	34.471	0.000	(0.0413, 0.0462)
0.75	Intercept	-3.017	0.115	-26.187	0.000	(-3.243, -2.792)
	TPA	-0.025	0.00205	-12.182	0.000	(-0.0291, -0.0210)
	Age	0.0444	0.00118	37.616	0.000	(0.0421, 0.0467)
0.95	Intercept	-1.824	0.191	-9.542	0.000	(-2.198, -1.449)
	TPA	-0.0185	0.00343	-5.391	0.000	(-0.0252, -0.0118)
	Age	0.0435	0.00187	23.228	0.000	(0.0398, 0.0472)

The following figure (Figure 14) gives a more detailed picture of the estimated coefficients.

Figure 14: The estimated coefficients of each variable over different quantiles when $\epsilon = 10^{-11}$. The grey color in the plots shows the 95% confidence intervals of the coefficients and the red lines on the other hand illustrate the estimated OLS coefficients and their 95% confidence intervals.



A3 Results when $\epsilon = 1$

Table 16: Results of fitting the model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{Interval2} + \beta_{p,2}x_{Interval3} + \dots + \beta_{p,11}x_{Interval12} + \varepsilon_i$ with $\epsilon = 1$. 1000 bootstrap replicates were used to compute the standard errors, p -values and confidence intervals. We can also see the result of the F-test, where the joint hypothesis ($\beta_{0.5,1} = \beta_{0.5,2} = \dots = \beta_{0.5,11} = 0$) was tested.

p	Variable	Coeff	SE	t -value	p -value	95% CI
0.5	Intercept	-1.856	0.0226	-82.024	0.000	(-1.901, -1.812)
	Interval 2	-0.254	0.0226	-11.220	0.000	(-0.298, -0.210)
	Interval 3	-0.254	0.0226	-11.220	0.000	(-0.298, -0.210)
	Interval 4	-0.254	0.0277	-9.179	0.000	(-0.308, -0.200)
	Interval 5	-0.254	0.0226	-11.220	0.000	(-0.298, -0.210)
	Interval 6	-0.254	0.0226	-11.220	0.000	(-0.298, -0.210)
	Interval 7	-0.254	0.0226	-11.220	0.000	(-0.298, -0.210)
	Interval 8	-0.254	0.0492	-5.164	0.000	(-0.350, -0.158)
	Interval 9	-0.254	0.0303	-8.374	0.000	(-0.313, -0.194)
	Interval 10	-0.254	0.0957	-2.654	0.00796	(-0.441, -0.0664)
	Interval 11	-0.254	0.125	-2.039	0.0415	(-0.498, -0.00982)
	Interval 12	-0.571	0.0226	-25.251	0.000	(-0.616, -0.527)
F-test:						
F-value		18.067	p -value:	<2.2e-16		

Figure 15 on the next page gives a more detailed picture of the estimated coefficients.

Figure 15: The estimated coefficients of each variable over different quantiles when $\epsilon = 1$. The grey color in the plots illustrates the 95% confidence intervals of the coefficients and the red lines on the other hand illustrate the estimated OLS coefficients and their 95% confidence intervals.

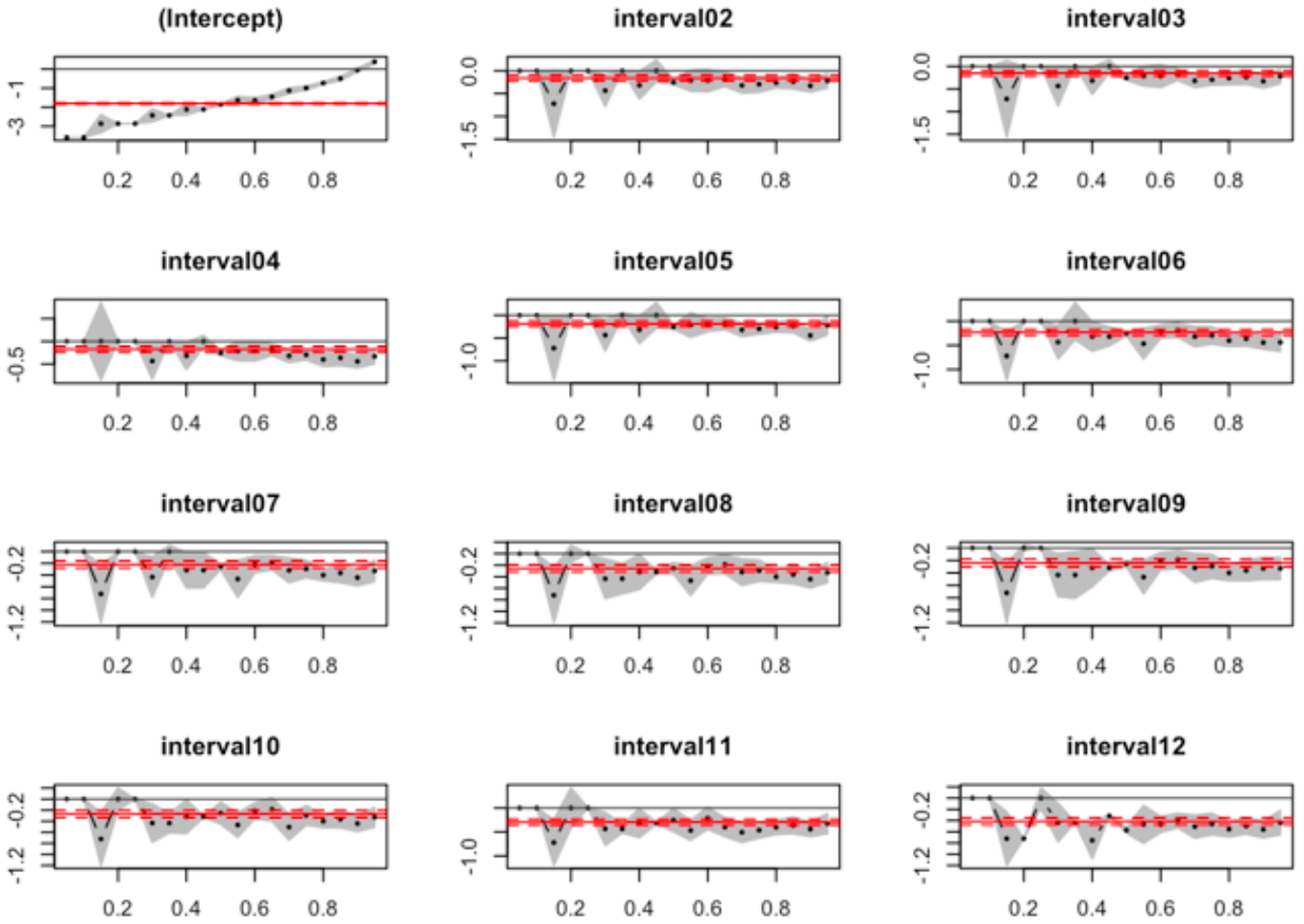


Table 17: A summary of fitting the regression model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{Interval2} + \beta_{p,2}x_{Interval3} + \dots + \beta_{p,11}x_{Interval12} + \beta_{p,12}x_{Age} + \varepsilon_i$, with $\epsilon = 1$. The table lists the variables, estimated coefficients, standard errors, t -values, p -values and 95% confidence intervals where $p = 0.5, 1000$ replicates were used to get the standard errors, p -values and confidence intervals. The result of the F-test, where we test the joint hypothesis ($\beta_{0.5,1} = \beta_{0.5,2} = \dots = \beta_{0.5,11} = 0$) is also shown.

p	Variable	Coeff	SE	t -value	p -value	95% CI
0.5	Intercept	-3.907	0.0756	-51.653	0.000	(-4.055, -3.759)
	Interval 2	-0.0672	0.0426	-1.578	0.115	(-0.151, 0.0163)
	Interval 3	-0.103	0.0389	-2.653	0.00798	(-0.179, -0.0270)
	Interval 4	-0.134	0.0413	-3.255	0.00113	(-0.215, -0.0535)
	Interval 5	-0.101	0.0435	-2.317	0.0205	(-0.186, -0.0156)
	Interval 6	-0.220	0.0414	-5.320	0.000	(-0.301, -0.139)
	Interval 7	-0.202	0.0396	-5.0973	0.000	(-0.279, -0.124)
	Interval 8	-0.269	0.0438	-6.145	0.000	(-0.355, -0.183)
	Interval 9	-0.269	0.0432	-6.220	0.000	(-0.354, -0.184)
	Interval 10	-0.302	0.0469	-6.454	0.000	(-0.394, -0.211)
	Interval 11	-0.370	0.0469	-7.879	0.000	(-0.462, -0.278)
	Interval 12	-0.388	0.0522	-7.435	0.000	(-0.491, -0.286)
	Age	0.0336	0.00115	29.217	0.000	(0.0314, 0.0359)
F-test:						
	F-value	18.08	p -value:	<2.2e-16		

See Figure 15 on the next page for a more complete picture of the estimated coefficients over different quantiles.

Figure 16: The estimated coefficients of each variable over different quantiles when $\epsilon = 1$. The grey color in the plots shows the 95% confidence intervals of the coefficients. The red lines on the other hand show the estimated OLS coefficients and their 95% confidence intervals.

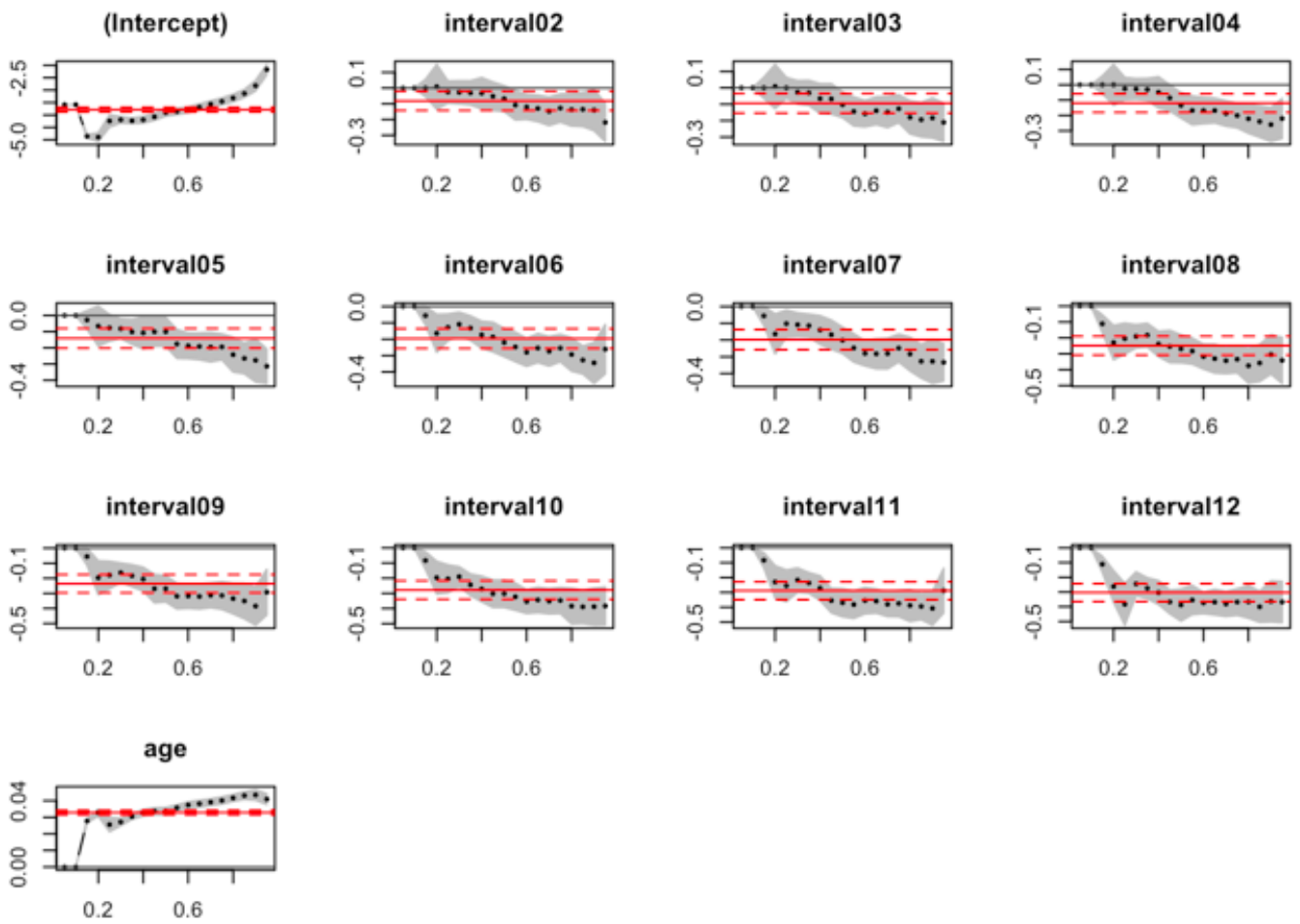


Table 18: Summary of fitting the model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{TPA} + \varepsilon_i$, where $\epsilon = 1$. The variables, estimated coefficients, bootstrapped standard errors, t -values, p -values and 95% confidence intervals of four different quantiles (0.25, 0.5, 0.75, 0.95) are shown in the table. We used 1000 bootstrap replicates to obtain the standard errors, p -values and confidence intervals.

p	Variable	Coeff	SE	t -value	p -value	95% CI
0.25	Intercept	-2.862	0.000	-8.874e+15	0.000	(-2.862, -2.862)
	TPA	0.000	0.000	-3.169	1.530e-03	(-2.937e-17, -6.924e-18)
0.5	Intercept	-1.318	0.362	-3.646	0.00027	(-2.027, -0.610)
	TPA	-0.0199	0.00905	-2.193	0.0283	(-0.0376, -0.00211)
0.75	Intercept	-0.273	0.0958	-2.853	0.00434	(-0.461, -0.0855)
	TPA	-0.0250	0.00223	-11.228	0.000	(-0.0294, -0.0206)
0.95	Intercept	0.986	0.166	5.944	0.000	(0.661, 1.311)
	TPA	-0.0216	0.00417	-5.175	0.000	(-0.0297, -0.0134)

Figure 17 below plots the estimated coefficients over different quantiles.

Figure 17: Illustration of the estimated regression coefficients of the intercept in the first plot and the estimated regression coefficients of the explanatory variable TPA in the second plot for different quantile values. Here we use $\epsilon = 1$. The grey color depicts the 95% confidence intervals of the coefficients, the red lines show the estimated OLS regression coefficients and the red dashed lines correspond to 95% confidence intervals of the estimated OLS coefficients.

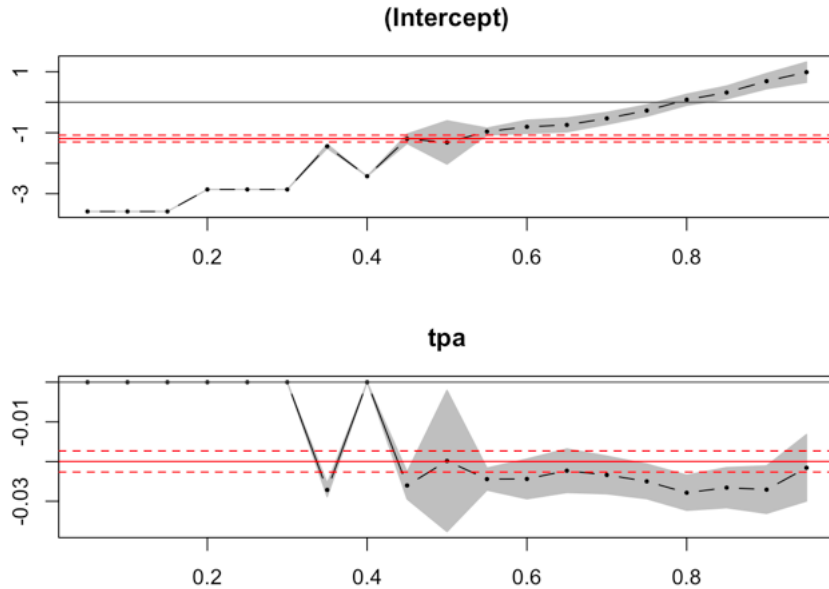
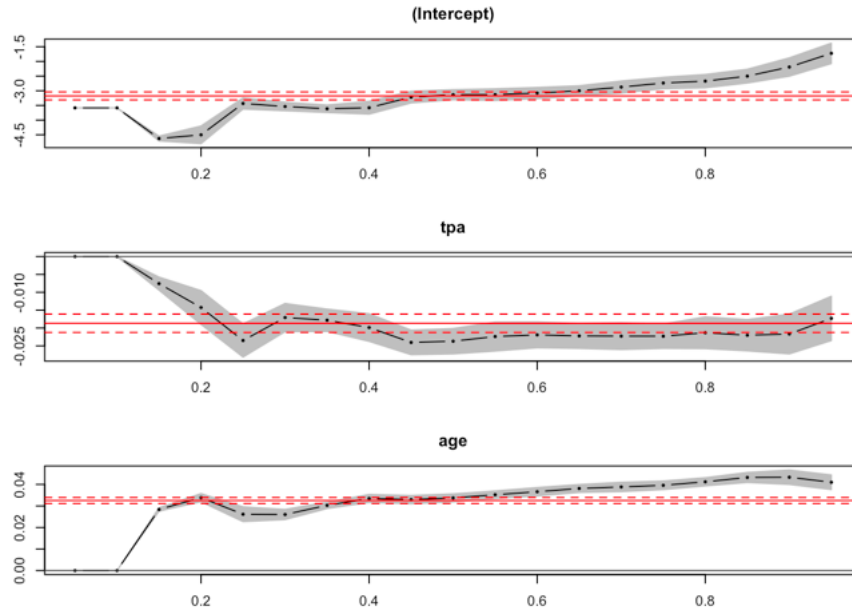


Table 19: Results for the model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{TPA} + \beta_{p,2}x_{Age} + \varepsilon_i$ with $\epsilon = 1$. The standard errors, p -values and 95% confidence intervals were obtained by 1000 bootstrap replicates.

p	Variable	Coeff	SE	t -value	p -value	95% CI
0.25	Intercept	-3.433	0.0978	-35.094	0.000	(-3.625, -3.241)
	TPA	-0.0235	0.00237	-9.919	0.000	(-0.0281, -0.0189)
	Age	0.0262	0.00176	14.851	0.000	(0.0227, 0.0296)
0.5	Intercept	-3.134	0.0866	-36.213	0.000	(-3.304, -2.965)
	TPA	-0.0237	0.00172	-13.766	0.000	(-0.0271, -0.0203)
	Age	0.0338	0.00096	35.182	0.000	(0.0319, 0.0356)
0.75	Intercept	-2.736	0.101	-27.088	0.000	(-2.934, -2.538)
	TPA	-0.0223	0.00182	-12.245	0.000	(-0.0259, -0.0187)
	Age	0.0396	0.00105	37.882	0.000	(0.0375, 0.0416)
0.95	Intercept	-1.725	0.184	-9.374	0.000	(-2.086, -1.364)
	TPA	-0.0173	0.00315	-5.487	0.000	(-0.0235, -0.0111)
	Age	0.0410	0.00180	22.771	0.000	(0.0375, 0.0445)

Figure 18 below gives a more complete picture of the estimated coefficients over different quantiles.

Figure 18: Plots of the estimated coefficients of the intercept, TPA and age over different quantiles when $\epsilon = 1$. The grey color in the plots shows the 95% confidence intervals of the coefficients. The red lines on the other hand illustrate the estimated OLS coefficients and their 95% confidence intervals.



Abbreviations

Table 20: Some abbreviations used in the thesis

I-PSS (sometimes IPSS in figures and tables)	International Prostate Symptom Score
LUTS	Lower urinary tract symptoms
TPA (sometimes tpa in figures)	Total physical activity

List of Figures

1	Boxplots of I-PSS	15
2	A histogram of I-PSS	17
3	Distribution of the logistic transformation of I-PSS when $\epsilon =$ 0, 0.001 and 0.5	18
4	Predicted regression lines	29
5	The tilted absolute value function for quantile regression . . .	33
6	The regression coefficients of the first model when $\epsilon = 0.5$. .	34
7	The regression coefficients of the second model when $\epsilon = 0.5$.	35
8	The regression coefficients of the third model when $\epsilon = 0.5$. .	36
9	The regression coefficients of the fourth model when $\epsilon = 0.5$.	37
10	Distribution of the logistic transformation of I-PSS against TPA	38
11	The regression coefficients of the first model when $\epsilon = 10^{-11}$.	41
12	The regression coefficients of the second model when $\epsilon = 10^{-11}$	43
13	The regression coefficients of the third model when $\epsilon = 10^{-11}$	44
14	The regression coefficients of the fourth model when $\epsilon = 10^{-11}$	45
15	The regression coefficients of the first model when $\epsilon = 1$. . .	47
16	The regression coefficients of the second model when $\epsilon = 1$. .	49
17	The regression coefficients of the third model when $\epsilon = 1$. .	50
18	The regression coefficients of the fourth model when $\epsilon = 1$. .	51

List of Tables

1	Example of transformation	6
2	Descriptive statistics of the data	12
3	The regression models	13
4	Descriptive statistics of the interval variables	14
5	Estimates for the first model when $\epsilon = 0.5$	19
6	Estimates for the second model when $\epsilon = 0.5$	20
7	Estimates for the third model when $\epsilon = 0.5$	21
8	Estimates for the fourth model when $\epsilon = 0.5$	21
9	Pseudo- R^2 values	22
10	The average of the absolute differences of $\hat{Q}_{y_i}(p \epsilon)$ for the model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{TPA} + \varepsilon_i$	23
11	The average of the absolute differences of $\hat{Q}_{y_i}(p \epsilon)$ for the model $h(y_i) = \beta_{p,0} + \beta_{p,1}x_{TPA} + \beta_{p,2}x_{Age} + \varepsilon_i$	39
12	Estimates for the first model when $\epsilon = 10^{-11}$	40
13	Estimates for the second model when $\epsilon = 10^{-11}$	42
14	Estimates for the third model when $\epsilon = 10^{-11}$	44
15	Estimates for the fourth model when $\epsilon = 10^{-11}$	45
16	Estimates for the first model when $\epsilon = 1$	46
17	Estimates for the second model when $\epsilon = 1$	48
18	Estimates for the third model when $\epsilon = 1$	50
19	Estimates for the fourth model when $\epsilon = 1$	51
20	Abbreviations	52