NTNU
Norwegian University of
Science and Technology

The use of logistic regression and quantile regression in medical statistics

Solveig Fosdal

Master of Science
Submission date: June 2017
Supervisor: Ingelin Steinsland, IMF
Co-supervisor: Håkon Gjessing, Nasjonalt folkehelseinstitutt

Norwegian University of Science and Technology
Department of Mathematical Sciences

Summary

The main goal of this thesis is to compare and illustrate the use of logistic regression and quantile regression on continuous outcome variables. In medical statistics, logistic regression is frequently applied to continuous outcomes by defining a cut-off value, whereas quantile regression can be applied directly to quantiles of the outcome distribution. The two approaches appear different, but are closely related. An approximate relation between the quantile effect and the log-odds ratio is derived. Practical examples and illustrations are shown through a case study concerning the effect of maternal smoking during pregnancy and mother's age on birth weight, where low birth weight is of special interest. Both maternal smoking during pregnancy and mother's age are found to have a significant effect on birth weight, and the effect of maternal smoking is found to have a slightly larger negative effect on low birth weight than for other quantiles. Trend in birth weight over years is also studied as a part of the case study. Further, the two approaches are tested on simulated data from known probability density functions, pdfs. We consider a population consisting of two groups, where one of the groups is exposed to a factor, and the effect of exposure is of interest. By this we illustrate the quantile effect and the odds ratio for several examples of location, scale and location-scale shift of the normal distribution and the Student t-distribution.

Through this thesis we find that quantile regression often yields an easier interpretation of the estimated effects due to the estimated parameters being on the same measuring scale as the dependent variable of interest. In addition, quantile regression provides easier comparisons of effects in different quantiles of the distribution, where the logistic regression model may easily lead to misinterpretations.

Samandrag

Føremålet med denne oppgåva er å samanlikne og vise bruken av logistisk regresjon og kvantilregresjon på kontinuerlege responsvariablar. I medisinsk statistikk, er logistisk regresjon ofte brukt på kontinuerlege utfall ved å definere ein grenseverdi, mens kvantilregresjon kan bli anvendt direkte på kvantilar i responsfordelinga. Dei to modellane ser ulike ut, men det er ein nær samanheng mellom dei. Ein tilnærma samanheng mellom log-odds ratio og kvantileffekten er utleda. Praktiske eksempel og illustrasjonar er vist gjennom eit eksempel-studie som omhandlar effekt av mors røyking gjennom svangerskapet og mors alder på fødselsvekt, der lav fødselsvekt er spesielt interessant. Både mors røyking og mors alder har ein signifikant effekt på fødselsvekt, og effekten av røyking er litt større ved lav fødselsvekt enn ved andre kvantilar. Endring i fødselsvekt over år er også undersøkt som ei del av eksempel-studiet. Vidare er dei to tilnærmingane brukt til å analysere simulerte data frå kjente sannsynsfordelingar. Vi ser på ein populasjon som består av to grupper der den eine blir påverka av ein faktor, og vi er interessert i effekten av eksponeringa. Ved dette illustrerer vi kvantileffekten og odds ratio for fleire eksempel ved forskyving og spredning av normalfordeling og Student t-fordeling.

Gjennom denne oppgåva ser vi at kvantilregresjon ofte gir enklare fortolking av dei estimerte effektane på grunn av at dei estimerte parametrane er på same måleskala som den avhengige variabelen vi undersøker. I tillegg gir kvantilregresjon enklare samanlikning av effekt i ulike kvantilar, der logistisk regresjon lett kan bli mistolka.

Preface

This Master thesis completes my degree at the Teacher Education program at the Department of Mathematical Sciences at the Norwegian University of Science and Technology (NTNU). The topic of this thesis is comparison and illustration of the use of quantile regression and logistic regression in medical statistics. I want to thank my supervisor Ingelin Steinsland at the Department of Mathematical Sciences, for the guidance and support during the work of this thesis. I also want to thank Håkon K. Gjessing at the Norwegian Institute of Public Health for suggesting this topic, making the data set analyzed in the case studies available to me and for guidance through the work of this thesis.

Solveig Fosdal
Trondheim, Norway
June, 2017

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Several studies has shown that low birth weight is closely related to both foetal and neonatal mortality and morbidity, and in addition it may contribute to a range of poor health outcomes for the babies later in life, like inhibited growth and cognitive development and chronic diseases. The medical definition of low birth weight defined by the World Health Organization (WHO) is a weight at birth less than 2500 g. This cuf-off value is based on epidemiological observations of approximate mortality among infants with low birth weight compared to infants with larger weight. This definition may vary between countries (Wardlaw et al., 2004). A case study conserning birth weight in Norway will be presented in this thesis, and this definition of low birth weight also holds for Norway (Folkehelseinstituttet, 2015).

Due to the possible consequences of low birth weight, there has been a considerable interest in explanatory variables influencing the birth weight. Many of the analyses on birth weight have been carried out using ordinary linear regression models, resulting in estimates of various effects on the conditional mean of birth weight. Nevertheless, it has been recognized that the resulting estimates were not necessarily explanatory for the effect of these factors in the lower tail of the distribution of birth weight. Further, several studies have explored binary response models, like probit models and logistic regression, for the occurrence of low birth weight (Koenker, 2005).

The most common method applied in medical statistics for analysis of binary response is logistic regression (Kirkwood and Sterne, 2013). Logistic regression is used when predicting a dichotomous outcome, and allows us to model how the odds/risk of the dichotomous outcome changes by exposure of an explanatory variable.

An alternative approach is quantile regression, which lets us model how the quantiles of a continuous distribution changes by exposure. The quantile regression model was first introduced in the 1970's by econometricians Koenker and Bassett (1978). This approach has gradually made its way to several other application areas like applications to survival analysis, methods for reference growth charts in medicine (Wei et al., 2006), and applications to environment modelling (Cade and Noon, 2003) (Yu et al., 2003). Still, many remain unaware of it, and the quantile regression model is still not as frequently used as logistic regression in applications in medicine, even when other parts of the distribution than the conditional mean are of interest.

The goal of this thesis is to illustrate the use of both quantile regression and logistic regression, and attempt to enlighten how quantile regression can offer a natural complement to logistic regression, and provide a more complete picture of the effects of the factors of interest. In addition quantile regression may provide several other advantages when it comes to interpretability and understanding of the effects of the factors of interest. In order to illustrate the use of the two approaches, some practical examples and illustrations will be shown where the same data set is analyzed using both regression methods. The response of interest is the birth weight, and especially low birth weight, and the factors of interest will be the smoking habits of the mother and the mother's age. In addition to these examples we will look at the change in birth weight over time, and also how the two methods are affected by censored data. Further some calculations using asymptotic results will be shown to present an approximate relation between the two approaches. The connection between logistic regression and quantile regression will also be shown through analyses using both methods on simulated data sets. The distributions of interest in this part is the normal distribution and the Student-t distribution.

The data set used for the analyses and practical examples will be presented further in Chapter 2, including explanation of the different variables and exploratory analysis of the data. The statistical theory behind the two regression models will be introduced with a special focus on the interpretation of the estimated parameters in Chapter 3. The derivation of an approximate relation between the log-odds ratio and the quantile effect is also presented in Chapter 3. Chapter 4 contains the analyses on the simulated data. The results from the analysis yielding the practical examples is presented in Chapter 5. Further some discussion and conclusion of the findings of the thesis is given in Chapter 6.

Chapter 2

Births in Norway data set and exploratory analyses

The motivating case of this thesis is to study how birth weight is affected by the mother's age and her smoking habits during pregnancy; in particular, it is important to measure how birth weights close to the definition of low birth weight 2500 g are affected. In this chapter we present the data set and do exploratory analyses of the data.

The analyses are performed on an artificially generated data set, which contains the following five variables: birth weight (BW) given in grams, year of birth, whether the mother has been smoking during the pregnancy or not (smoking habits during the pregnancy), the age of the mother when giving birth (in categories), and whether the birth was spontaneous or not.

To generate the data, completely anonymous crude tabular data from the Medical Birth Registry of Norway (MBR) were used as a starting point. The data were then expanded into a full sized data file according to the frequencies in each tabular category. Random noise was then added to the birth weight categories to create a continuous distribution similar to an actual birth weight distribution. The resulting data file thus has a total size and multivariate distributions of its variables that closely resemble those of actual birth registry data; at the same time, none of the "individual" data records in the file refer to any actual patients.

The data cover all births in Norway in the period 1967 to 2009. In total, the data set consists of 2 517 812 births, and five variables. The information about smoking habits was first included in 1998. Therefore, the analyses involving smoking as a factor will be done on a data set containing births in the period 1998 until 2009, on a data set containing 511 447 births. Information about spontaneous births is missing for all births in 2009, so this year is not included in the analyses of spontaneous births.

Data analyses are intended primarily to illustrate the methods and to discuss advantages and disadvantages of presenting results from the two methods; we expect the estimation results presented in this thesis to closely resemble those that would be obtained from real data, while at the same time not needing access to individual data.

A histogram of birth weights in Norway for all births between 1967 and 2009 is shown in Figure 2.1 together with the the corresponding normal-plot. Although the histogram appears Gaussian, it is possible to observe a slightly heavier tail on the left side. The shape of the corresponding normal-plot typically indicates a skewed distribution, and hence supports this observation. In addition we can detect deviation from the normal distribution by the fact that the mean and the median, shown in Table 2.1, are not equal, which would be the case for a normal distribution, and other symmetric distributions. In Figure 2.2 the number of births is plotted for each year.

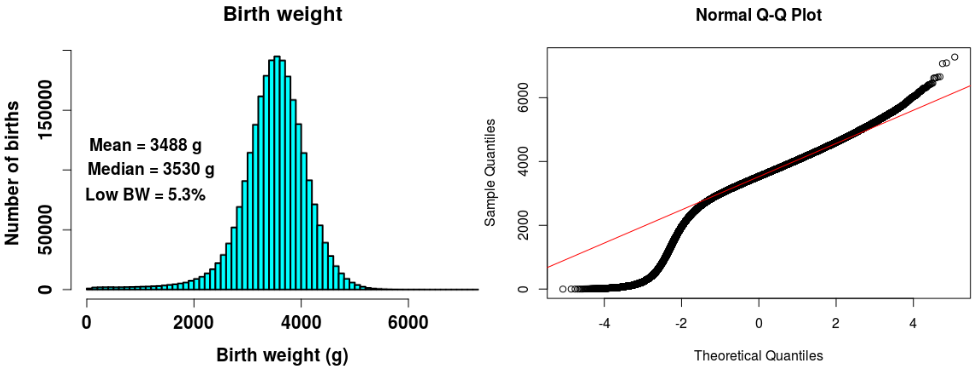

Figure 2.1: Distribution of birth weight.

Table 2.1: The mean, the median, the proportion of birth weights below 2500 grams and the weight corresponding to the 5% quantile.

Mean	3488 g
Median	3530 g
Low BW	5.3%
5% quantile	2468 g

Table 2.2: Description of the smoking categories.

Category	Description
1	the mother did not smoke during the pregnancy
2	the mother smoked sometimes during the pregnancy
3	the mother smoked daily during the pregnancy

The two variables, smoking or not and the age of mother, are categorical variables. The variable containing information about smoking habits is described in Table 2.2. On the left hand side in Figure 2.3 the number of births in each smoking category is plotted for each year. The black curve shows the number of births for the non-smokers, the red

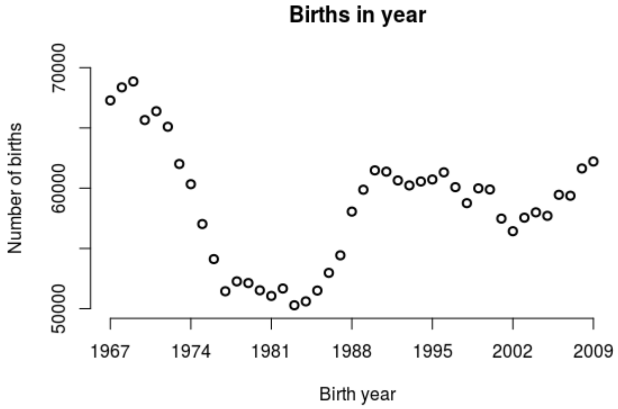

Figure 2.2: Number of births in each year.

curve the smokers, and the blue curve the ones that smoked sometimes. The total number of births in each category is shown in Table 2.3.

Table 2.3: Number of births in each category.

Non-smokers	Smoke sometimes	Daily smokers
446638	6477	59179
87.2%	1.2%	11.6%

In this study we choose to work with only two different categories, 1 denoting non-smokers and 2 denoting smokers. Category 2, smoked sometimes, contains few observations compared to the other two categories, and category 1, non-smokers, is clearly the largest group. So in order to obtain the two categories wanted above, it is reasonable to add the births where the mother smoked sometimes to the category of non-smokers. From the figure on the right hand side in 2.3 we see the number of births in each of the two categories over years. The black curve now denotes both category 1 and 2, and we refer to these as non-smokers and smokers respectively.

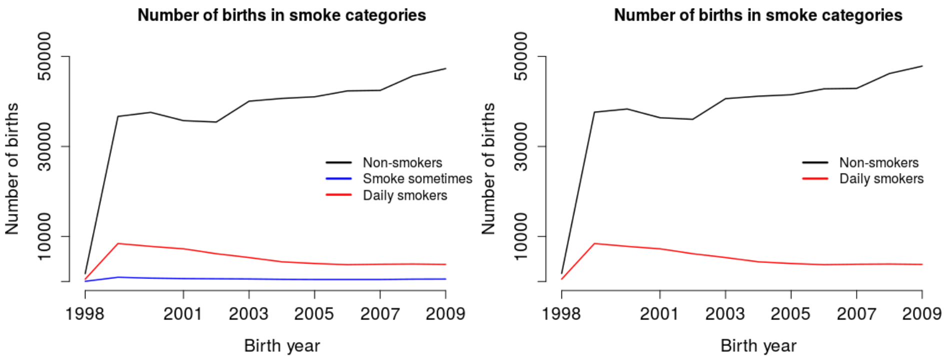

Figure 2.3: Number of births in each category over years.

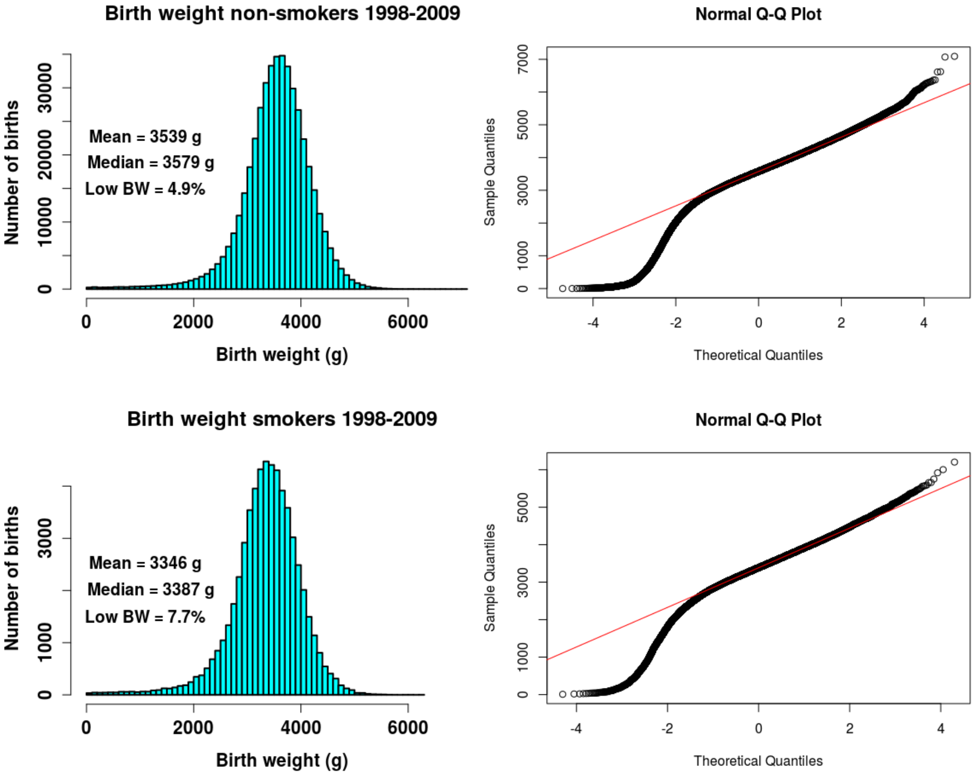

Figure 2.5: A histogram together with the corresponding normal-plot for the non-smokers in the upper panel and the smokers in the lower panel. Both in the period 1998 to 2009.

Figure 2.5 show histograms and corresponding normal-plots for birth weight of non-smokers and smokers separately. The normal QQ-plots indicates deviations from a normal distribution and support the observation of a heavier tail on the left side of both distributions. Notice that the mean and the median are lower for smokers than for non-smokers, and in addition the proportion of children with low birth weight is higher for smokers. These observations are presented in Figure 2.6 in the upper panel. For non-smokers 4.9% of the birth weights has low birth weight, while this proportion increases to 7.7% for smokers. In the lower panel of Figure 2.6 the weight corresponding to the 5% quantile of the two distributions is presented, and we notice how this value decreases from 2508 g for non-smokers to 2293 g for smokers.

In addition to these figures we inspect the boxplot found in Figure 2.7 to obtain more information about the difference between the two groups. The difference in the median is visible from this plot, and we are also led to believe that the distribution of non-smokers has approximately equal variance as the variance for the distribution of smokers.

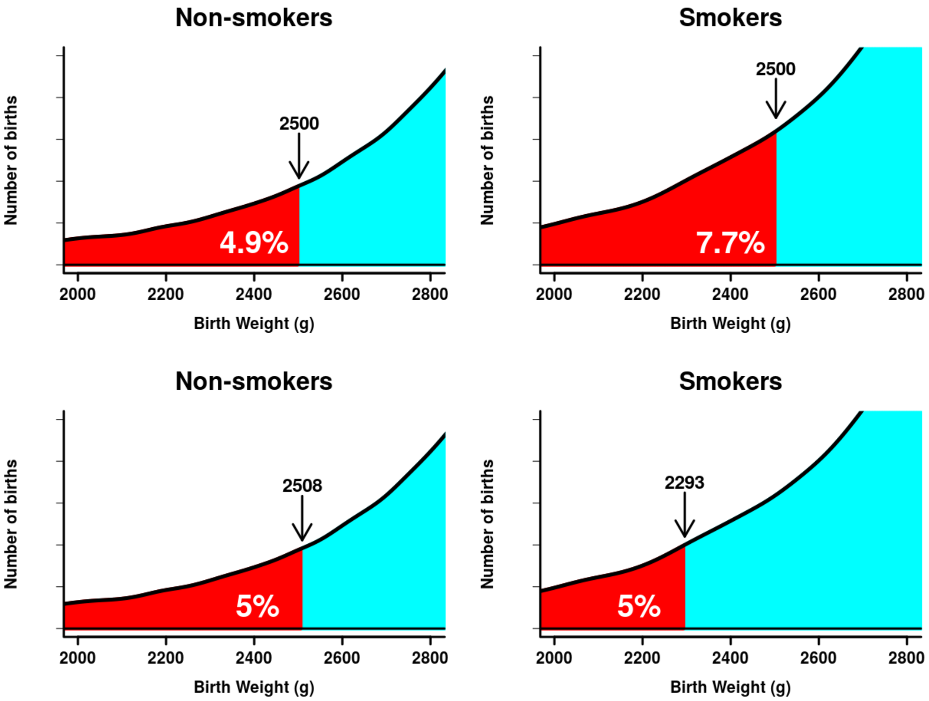

Figure 2.6: The proportion of weights below 2500 g is shown in the upper panel, and the weight corresponding to the 5% quantile in the lower panel. Left panel: non-smokers. Right panel: smokers.

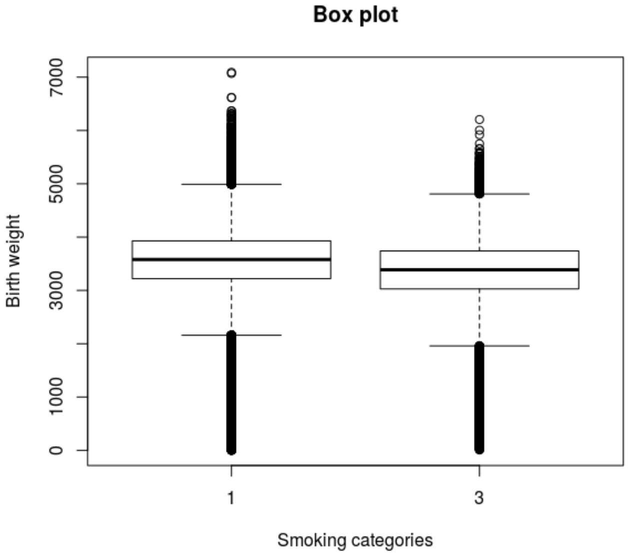

Figure 2.7: Boxplot of birth weight for the smoking categories.

The age variable is grouped in 5 year intervals, giving the 6 categories:

Table 2.4: Description of the age categories.

Category	Age
1	19 years and younger
2	20 - 24
3	25 - 29
4	30 - 34
5	35 - 39
6	40 years and older

The number of births in each category over years is presented in Figure 2.8. By considering the boxplot in Figure 2.9 we are only able to see small differences between the birth weight in each of the age categories.

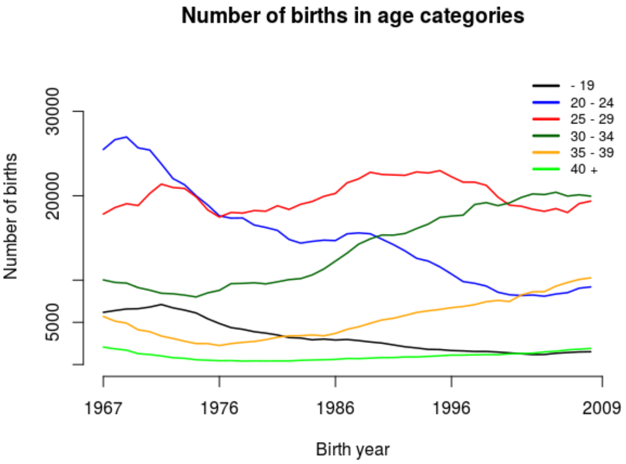

Figure 2.8: Number of births in each age category over years.

Table 2.5: The mean, median, proportion of low birth weight and the weight corresponding to the 5%-quantile for the different age groups.

	≤ 19	$20 - 24$	$25 - 29$	$30 - 34$	$35 - 239$	$40 \geq$
Mean	3365	3447	3508	3530	3505	3462
Median	3418	3486	3545	3577	3566	3533
Low BW	6.9%	5.1%	4.8%	5.2%	6.5%	7.7%
5% quantile	2317	2487	2522	2473	2335	2194

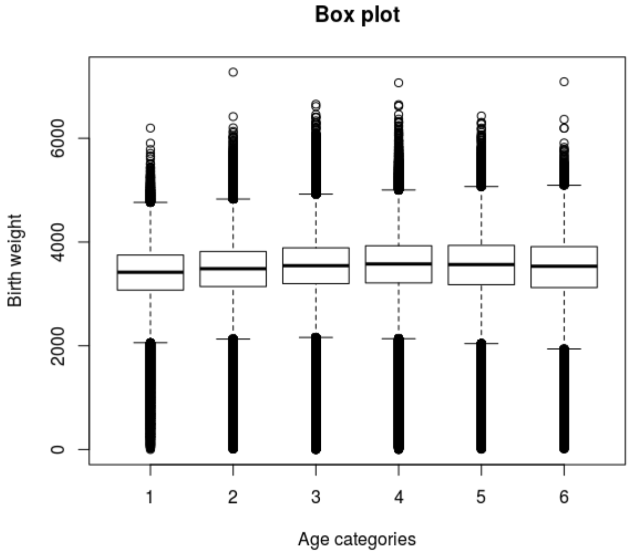

Figure 2.9: Boxplot of birth weight for age categories.

Chapter 3

Statistical methods

In this chapter the statistical theory of logistic regression and quantile regression is introduced with a special focus on the interpretation of the estimated parameters.

3.1 Logistic Regression

The logistic regression model is one out of several possible generalized linear models (GLM). To introduce this model we start by presenting the theoretical framework of generalized linear models using notation as presented by Rodriguez (2013). The standard linear model will be introduced at first, and further how this can be generalized in two steps to obtain the logistic regression model.

3.1.1 Generalized linear model theory

Let $y_1, ..., y_n$ denote n independent observations of a response that are defined to be realizations of the random variable Y_i. We assume that Y_i has a normal distribution with mean μ_i and variance σ^2, that is assumed to be equal for all of the n observations,

$$Y_i \sim N(\mu_i, \sigma^2). \tag{3.1}$$

The expected value, μ_i, can be expressed as

$$\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}, \tag{3.2}$$

where we assume that the expected value is a linear function of the r predictors taking the values $\boldsymbol{x_i^T} = (x_{i1}, ..., x_{ir})$ for the i-th observation and $\boldsymbol{\beta}$ is a vector of the unknown parameters, also called regression coefficients, that needs to be estimated. The response can now be expressed as

$$\begin{aligned} y_i =& \mu_i + \epsilon_i \\ y_i =& \boldsymbol{x_i}^T \boldsymbol{\beta} + \epsilon_i, \end{aligned} \tag{3.3}$$

where, ϵ_i is a Gaussian error, $\epsilon_i \sim N(0, \sigma^2)$. This is the standard linear model that the generalized linear models are based on, and this generalization requires two steps. First the observations need to come from a distribution in the exponential family and be expressed as

$$f(y_i) = \exp\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\}, \tag{3.4}$$

where θ_i and ϕ_i are parameters, and the functions $a(\phi_i)$, $b(\theta_i)$ and $c(y_i, \phi)$ are known. The exponential family includes distributions such as the normal, binomial, Poisson, gamma and exponential distributions.

Further, a transformation of the mean is introduced in order to obtain a model that is not directly of the mean. This is a one-to-one, continuous differentiable transformation given by the function $g(\mu_i)$, called the $link\ function$. It is assumed that this transformed mean follows a linear model, and the linear predictor is introduced,

$$\eta_i = g(\mu_i) \tag{3.5}$$

$$\eta_i = x_i^T \beta. \tag{3.6}$$

Since this transformation is one-to-one it can be inverted to obtain the expected value. For the standard linear model the linear predictor is simply

$$\eta_i = \mu_i, \tag{3.7}$$

known as the identity. For the logistic regression model this function is called the $logit$ transformation, that we will come back to after defining the stochastic structure of the data.

3.1.2 The model and logit link

The logistic regression model is a model for dichotomous data, where the response takes one of two possible outcomes, either “success” or “failure” for a given event.

We consider a binary response y_i defined as,

$$y_i = \begin{cases} 1 & \text{success} \\ 0 & \text{failure,} \end{cases} \tag{3.8}$$

being realizations of a random variable Y_i that takes the values one and zero with probabilities p_i and $1 - p_i$ respectively. Then the distribution of Y_i is the Bernoulli distribution on the form,

$$P(Y_i = y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}. \tag{3.9}$$

This distribution is a special case of the binomial distribution of size one, equivalent to considering individual data. Another possibility is to work with grouped data where, it is assumed that the data can be divided into groups such that all individuals in a group have identical values of all predictors. Then n_i denotes the number of observations and y_i denotes the number of "successes" in group i. The random variable Y_i can then take the values $0, 1, ..., n_i$, and when the y_i's are independent have the same probability p_i for "success", then

$$Y_i \sim Binomial(n_i, p_i). \tag{3.10}$$

$$P(Y_i = y_i) = \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i} \tag{3.11}$$

Since the observations come from a binomial distribution, it is known that a change in the probability p would affect both the expected value and the variance. Therefore, constant variance is not an assumption for the logistic regression model.

The binomial distribution belongs to the exponential family, so the general theory for generalized linear models holds. The logistic regression model is a generalized linear model with binomial response and link $logit$ that we will now define. Based on the standard linear regression model, we may first suggest a model assuming a linear function for the probability,

$$p_i = \boldsymbol{x_i^T \beta}.$$

By doing this, the probability can take any real value, and restrictions on the explanatory variables and the estimated parameter is needed to ensure that no probability takes values less than 0 or larger than 1. To avoid these restrictions we use the logit transformation, a transformation of the probability through the odds to the logit. The odds is defined as,

$$\text{odds} = \frac{p_i}{1-p_i}, \tag{3.12}$$

the ratio of the probability to its compliment. Let the logit of the probability be defined as,

$$\text{logit}(p_i) = \eta_i, \tag{3.13}$$

The logit is connected to the probability through the odds, where the logit is the logarithm of the odds, defined as the linear predictor in Equation (3.6), so that

$$\eta_i = \text{logit}(p_i) = \ln(\text{odds}) = \ln(\frac{p_i}{1-p_i}) = \boldsymbol{x_i^T \beta}. \tag{3.14}$$

The logit can take any real value, and by exponentiating the logit the odds is obtained. Since the logit transformation is one-to-one, and the inverse transformation makes it possible to go from logit to probabilities, and this leads us to the next section on how to interpret the parameters in the logistic regressoin model. Figure 3.1 shows the logit and the probabilities plotted together.

3.1.3 Interpretation of the parameters - comparing two groups

The regression coefficients $\boldsymbol{\beta}$ in the linear model defined in Equation (3.14) can be interpreted in the same way as for the standard linear model, remembering that the left-hand

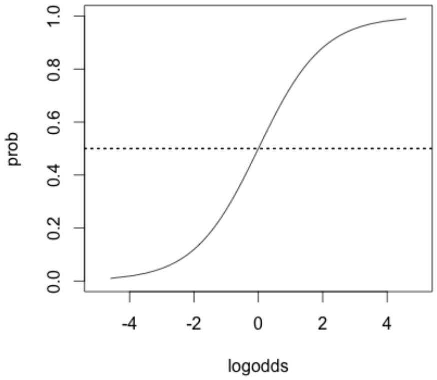

Figure 3.1: Logarithm of the odds vs the probability.

side is the logit, not the mean. β_j would now represent a change in the $logit$ of the probability when changing the $j'th$ predictor with one unit, while the other predictors, if any, are held constant.

By exponentiating Equation (3.14) we obtain the expression for the odds for the $i'th$ unit

$$\begin{aligned}\frac{p_i}{1-p_i} &= \exp\{\boldsymbol{x_i^T\beta}\}\\ \text{odds} &= \exp\{\boldsymbol{x_i^T\beta}\},\end{aligned}\tag{3.15}$$

that may be more familiar and easier to interpret than the logit scale. If the probability of an event is 50%, then the odds are one-to-one, called even, and the logit would be zero. A probability below 50% would take a negative value for the logit and a probability above 50% would take a positive value. This can be seen in Figure 3.1.

From this model it can also be seen how the odds of a dichotomous response changes when being exposed to a factor, allowing us to compare the two groups where one is exposed to the factor and the other is not. Equation (3.15) represents a multiplicative model for the odds. Recall that β_j represents the change in the logit when changing the $j'th$ predictor by one unit. When exponentiating β_j we obtain the $odds\,ratio$, that is more familiar to interpret than the effect of the logit. The odds ratio now gives information about the effect on the odds.

For the effect on the odds, it is also possible to differ between effects, the gross effect and the net effect. The gross effect of the parameters the β-coefficient in a simple model that only contains one covariate. The net effect is the β-coefficient in a model with several covariates, and it is then the effect on the response of changing the covariate one unit and keeping the other constant. In this study we will mostly work with simple models, and obtain the gross effect.

When looking at the effect of an explanatory variable, it is worth mentioning that although we are able to go from logit to probabilities by the inverse of the logit function, this may not be very helpful when interpreting the effect of a factor. There is no simple way to express this effect on the probability since an effect that appears constant in the logit scale would transform to varying effects in the probability scale depending on both β_j and the probability. This can also be seen in Figure 3.1. From the logit link we are able to go back to probabilities by

$$p_i = \frac{\exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{x}_i^T \boldsymbol{\beta}\}}. \tag{3.16}$$

This leads to several ways of interpreting the results of logistic regression. Odds ratio (OR) and relative risk (RR) are the two most widely used in epidemiology (Schmidt and Kohlmann, 2008). It is discussed what is the best scale for presenting results. When interpreting rare events that occurs in less than 10% of the cases, RR is considered to yield an acceptable approximation of the OR (Schmidt and Kohlmann, 2008). Another measure scale is the risk difference, and we will consider these three scales of measuring in Chapter 5.

3.1.4 Estimation of parameters

The parameters in a logistic regression model are estimated by maximum likelihood estimation. The observations are independent, so the joint density is used to find the likelihood function.

$$L(\beta_i, y_i) = \prod_{i=1}^{n} f_i(y_i; p), \tag{3.17}$$

further finding the logarithm of this function, the log-likelihood function

$$\text{ln}L(\beta_i, y_i) = \sum_{i=1}^{n} \text{ln}f_i(y_i; p). \tag{3.18}$$

For the n independent binomial observations, the log-likelihood function takes the form,

$$\text{ln}L(\beta_i, y_i) = \sum_{i=1}^{n} y_i \text{ln}(p) + (n - y_i)\text{ln}(1 - p) \tag{3.19}$$

The estimates for β is found by using the connection between the probability p and the covariates x_i and β through the logit. Then the parameter β is then estimated by optimizing the likelihood function or the log-likelihood function, these are equivalent, and choosing the β than makes the data that are observed as likely as possible. This can be expressed as:

$$\text{ln}L(\hat{\boldsymbol{\beta}}, \boldsymbol{y}) \geq \text{ln}L(\boldsymbol{\beta}, \boldsymbol{y}) \tag{3.20}$$

for all $\boldsymbol{\beta}$.

One way of finding this maximum likelihood estimator is to set what is called the score vector equal to zero. The score vector is the first derivative of the log-likelihood function,

$$\boldsymbol{u}(\boldsymbol{\beta}) = \frac{\partial \text{ln}L(\boldsymbol{\beta}, \boldsymbol{y})}{\partial \boldsymbol{\beta}}, \tag{3.21}$$

then

$$\boldsymbol{u}(\hat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{3.22}$$

From this the Hessian matrix can be found by,

$$\boldsymbol{H}(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{u}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}. \tag{3.23}$$

For many problems the maximum likelihood estimation requires iterative procedures. One way of doing this is to expand the score function using a first order Taylor series and further using the Hessian matrix to obtain the first order approximation,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 - \boldsymbol{H}^{-1}(\boldsymbol{\beta}_0)\boldsymbol{u}(\boldsymbol{\beta}_0) \tag{3.24}$$

From this the Newton-Raphson technique can be used by, given a trial value, the equation above (3.24) is used to obtain an improved estimate and repeating the procedure until the difference between estimates are sufficiently close to zero. Another suggestion for a procedure is known as Fisher scoring which gives an imporved estimate. In this procedure the Hessian matrix is replaced by the information matrix, that is its expected value, and the improved estimate is then given by,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 - \boldsymbol{I}^{-1}(\boldsymbol{\beta}_0)\boldsymbol{u}(\boldsymbol{\beta}_0) \tag{3.25}$$

The information matrix is defined as,

$$\text{Var}[\boldsymbol{u}(\boldsymbol{\beta})] = E[\boldsymbol{u}(\boldsymbol{\theta})\boldsymbol{u}'(\boldsymbol{\theta})] = \boldsymbol{I}(\boldsymbol{\beta}). \tag{3.26}$$

Or under mild regularity conditions,

$$\text{Var}[\boldsymbol{u}(\boldsymbol{\beta})] = \boldsymbol{I}(\boldsymbol{\beta}) = -E[\frac{\partial^2 \text{ln} L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}}]. \tag{3.27}$$

3.1.5 Hypotheses testing and confidence intervals

To test the significance of a parameter, β_j, look at the hypothesis,

$$H_0 : \beta_j = 0. \tag{3.28}$$

The asymptotic results for the MLE is that under the certain regularity conditions, when $n \to \infty$ the estimated parameter ($\hat{\beta}_j$) are normally distributed. From this the Wald test follows. The Wald-statistic, or the z-statistic, given by,

$$z = \frac{\hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta_j})}}, \tag{3.29}$$

is the critical value that is used at a chosen α-level of significance to decide if β_j is significant.

This z-statistic can also be used to construct a $(1-\alpha)\%$-confidence interval. Using the equation above (3.29), the two-sided $(1-\alpha)\%$-confidence interval is given by,

$$\hat{\beta}_j = \pm z_{1-\alpha/2}\sqrt{\text{Var}(\hat{\beta}_j)}, \tag{3.30}$$

where $\sqrt{\text{Var}(\hat{\beta}_j)}$ is the standard error. An estimate of this variance can be given by the the inverse of the expected information matrix,

$$\hat{\text{var}}(\hat{\boldsymbol{\beta}}) = \text{I}^{-1}(\hat{\boldsymbol{\beta}}). \tag{3.31}$$

The observed information matrix from Equation (3.27) is also possible to use for this estimation (Rodriguez, 2013).

3.2 Quantile Regression

The key idea of quantile regression was first introduced by Koenker and Bassett (1978). In this section we introduce the statistical theory of the quantile regression model mainly using theory and notation as presented by Koenker (2005) and Yu et al. (2003). For the choice of methods of inference we refer to Kocherginsky et al. (2005) and He and Hu (2002).

3.2.1 Definition of the quantile of a random variable

To define the term quantile we start by looking at a well known example of a quantile, the median, known as the 50% quantile. The sample median can be defined as the middle value of a set of ordered data (or the halfway between the two middle values). This means that the sample median would split the data into two parts, and these two parts would contain equally many observations. Let the variable Y be defined on a population and let m denote the population median, that usually can be estimated by the sample median. For Y a continuous random variable, the median m is the value that solves

$$F(m) = \frac{1}{2}, \tag{3.32}$$

where the belonging cumulative distribution function of Y is $F(y) = P(Y \leq y)$. Since the median splits the data into two equal parts, we have

$$P(Y \leq m) = P(Y \geq m) = \frac{1}{2}. \tag{3.33}$$

Figure 3.2 show the 50%-quantile value, found on the horizantal axis, on the cumulative distribution function and the probability density function.

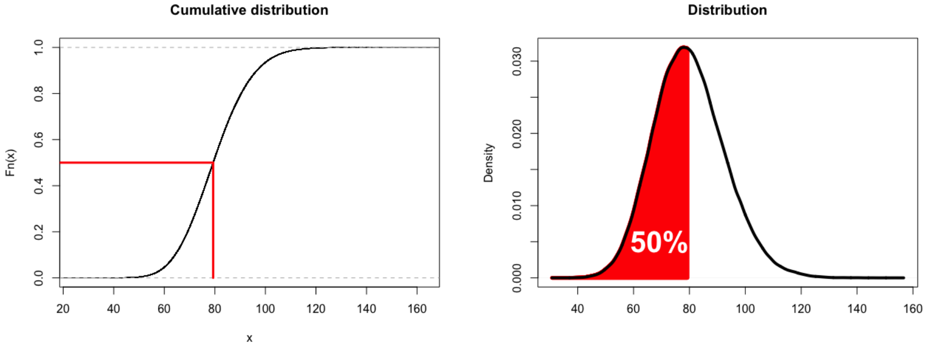

Figure 3.2: The density and the cumulative distribution function.

Another example is to split the ordered data into four parts, with the proportions of one quarter, a half and three quarters. We would then get the 25% quantile, also known as the lower quartile, the median and the 75% quantile, also known as the upper quartile of the population. As before, for the continuous case, $F(y) = \frac{1}{4}$ and $F(y) = \frac{3}{4}$ for the lower and upper quartile respectively. In general we define τ to take values so that $0 < \tau < 1$. Then, for the continuous case, the $100\tau\%$ quantile is the value of y that solves $F(y) = \tau$. Note also that the $100\tau\%$ quantile is equivalent to the $100\tau'th$ percentile (Yu et al., 2003).

3.2.2 Linear quantile regression models

We recall the standard linear model presented in Equation (3.3), where we fit a model to the conditional mean of relationship between the explanatory variables and the response. An alternative, that is known to be more robust, is to fit a model to the median instead of the mean. The basic idea is then to estimate $\boldsymbol{\beta}$ by minimizing the absolute value of the error,

$$\min_{\beta} \sum_{n=1}^{n} \mid y_i - x_i^T \beta \mid . \tag{3.34}$$

This is the basis for the median regression model, and this can be extended to other quantiles in addition to the median, and we obtain the conditional quantile regression model. The relationship between the $100\tau\%$ quantile of the response and the explanatory variables $\boldsymbol{x}_i$ is given by

$$Q_{Y_i}(\tau|x) = \boldsymbol{x_i}^T \boldsymbol{\beta}(\tau) \tag{3.35}$$

So by the quantile regression model, it will be possible to look at the relationship between the explanatory variables and the response in different quantiles of the distribution, not only the mean. Figure 3.3 How the estimates of $\boldsymbol{\beta}$ is estimated will be explained in section 3.2.4, and in the next section the interpretation of the quantile regression model will be explained.

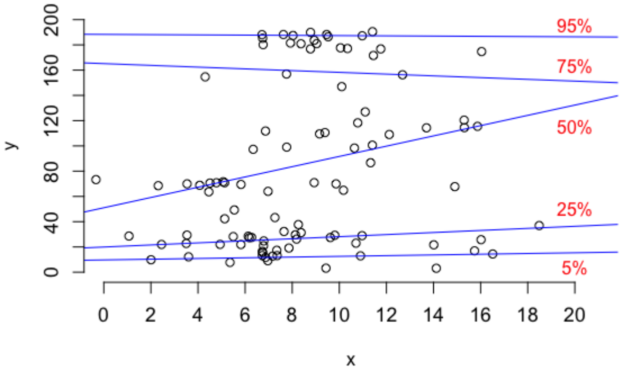

Figure 3.3: Toy example with one covariate shows the relationship between the explanatory variable and the response in different quantiles of the distribution. The red percentages denote the quantiles.

3.2.3 Interpretation of the quantile effect-interpretation of the regression coefficients

Quantile regression models the relationship between the $\tau'th$ sample quantile of the response and the different explanatory variables in the model. The model usually contain an intercept, that is the sample quantile with no effect from any of the explanatory variables present. We usually interpret the effect of a explanatory variable by looking at what happens in the response when we change this with one unit. The effect would typically be the change in the response that is necessary to keep the response at the same quantile as before the change in the explanatory variable was made.

To explain this further Koenker (2005) presents the simplest regression model, the two-sample treatment-control model, as introduced by Lehmann and Doksum (1974) (Koenker, 2005). By this model we assume that the response of an untreated observation would be x, and that the treatment then would add an amount $\Delta(x)$ to the response. The random variable X is distributed according to F, and the random variable $X + \Delta(X)$ is distributed according to a G. We can now define $\Delta(x)$ as the "horizontal distance" between F and G at x, giving

$$F(x) = G(x + \Delta(x)). \tag{3.36}$$

Now, $\Delta(x)$ is uniquely defined and we express it as,

$$\Delta(x) = G^{-1}(F(x)) - x. \tag{3.37}$$

We now recall the definition of the quantile, and use $\tau = F(x)$ to change variables and obtain the quantile treatment effect,

$$\delta(\tau) = \Delta(x) = \Delta(F^{-1}(\tau)) = G^{-1}(\tau) - F^{-1}(\tau). \tag{3.38}$$

This quantile treatment effect can be estimated by

$$\hat{\delta}(\tau) = \hat{G}_n^{-1}(\tau) - \hat{F}_m^{-1}(\tau), \tag{3.39}$$

where we let $\hat{G}_n^{-1}$ and $\hat{F}_m^{-1}$ denote the empirical distribution functions of the treatment and control observations, with n and m observations, respectively. Figure 3.4 illustrates an example of this setting. We now recall the quantile regression model, and formulate the model for this problem,

$$Q_{Y_i}(\tau|x_i) = \beta_0(\tau) + \beta_1(\tau)x_i. \tag{3.40}$$

In this binary problem the x_i denotes the treatment indicator, where $x_i = 1$ indicates treatment and $x_i = 0$ indicates the control. The estimates for the parameters in this model will then be $\hat{\beta}_0(\tau) = \hat{F}_m^{-1}$ and $\hat{\beta}_1(\tau) = \hat{\delta}(\tau)$. β_0 is the intercept in a model, where none of the covariates is present. The parameter β_1 denotes the quantile treatment effect, and is the slope of the model. We can interpret more complicated models in this way as well. Often, to find the effect of a covariate we might look at the derivative of the model with respect to the parameter of interest. For additive models we would obtain the effect of the covariate of interest. If we should have a model with interaction effects as well, we need to be aware of the additional interaction effect.

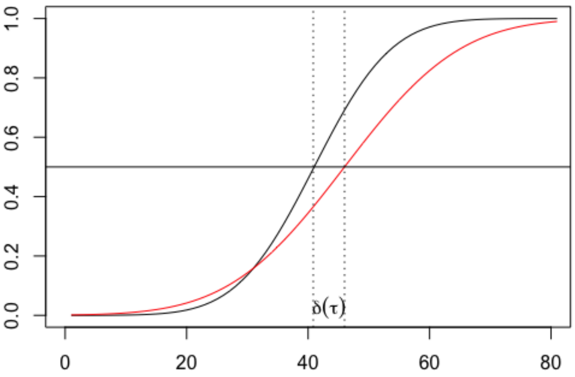

Figure 3.4: Illustration of an example of empirical cumulative distribution functions F and G. The quantile effect is shown by $\delta(\tau)$.

3.2.4 Estimation of the parameters

The estimation of the parameters in the quantile regression model is computationally more demanding than for both the standard linear model and the logistic model. The basic idea, elaborated by Koenker and Basset (1978), is to minimize the sum of absolute errors. This can be done by minimizing,

$$\min_{\beta} \sum_{n=1}^{n} \rho_{\tau} \mid y_i - x_i^T \boldsymbol{\beta} \mid, \tag{3.41}$$

where ρ_τ is dependent on the quantile of interest, and often known as the check function. An associated loss function can be written as $|u|$ where $u = y_i - \boldsymbol{x_i}^T\boldsymbol{\beta}$, where y_i is observation number i, $i = 1, ...n$, and $\boldsymbol{\beta}$ contains the parameters to be estimated ,and we extend this to a more convenient loss function, the check function

$$\rho_{\tau}(u) = \tau u I_{[0,\infty)}(u) - (1-\tau)uI_{(-\infty,0)}(u), \tag{3.42}$$

where

$$I_A = \begin{cases} 1, & u \in A \\ 0, & \text{otherwise} \end{cases} \tag{3.43}$$

(Yu et al., 2003). Since this check function is not differentiable in the origin, there is no explicit solution for the parameters in this model. To solve the minimization problem, Koenker (2005, 1978) show that minimizing the loss actually leads to a simple optimization problem that can be solved by linear programming. We formulate the quantile regression problem (3.41) as a linear program on the form,

$$\min_{\beta,u,v} \{\tau 1_n^T u + (1-\tau)1_n^T v | X\beta + u - v = y\}, \tag{3.44}$$

where X denotes the usual $n \times p$ regression design matrix. The variables u and v are introduced as artificial variables $u_i, v_i : 1, ..., n$ representing positive and negative parts of the residual vector $y - X\beta$. The solutions, $\hat{\beta}(\tau)$, that we call regression quantiles, follow the properties of solutions of linear program, known as basic solutions. There are several options for what algorithm to use to find these solutions. In the `quantreg package` in R we find several available methods, many implemented by Koenker (2015). For large samples, like the data set in our case study, the method used for optimization is the Frisch-Newton interior point method.

3.2.5 Inference for regression quantiles

Several approaches to statistical inference for quantile regression applications exists, and we can classify these into three categories: direct estimation of the variance-covariance matrix, rank-score method and resampling methods; bootstrapping, pairwise or residual. Several authors, including Koenker (2005), have considered the different methods, and monte carlo simulations have been performed to compare the different methods. Based on these comparisons and recommendations we choose, in this study, to apply a variant of the bootstrap estimates called Markov chain marginal bootstrap, MCMB, proposed by He and Hu (2002). For large problems the two common methods, bootstrapping pairwise and bootstrapping residuals can be very time consuming as they require repeated calculation of regression quantile estimates. This method is especially attractive for large problems, with np between $10\,000$ and $2\,000\,000$, and has shown robustness against certain deviations from homoscedasticity (He and Hu, 2002). He and Hu (2002) show that the least absolute deviation regression estimator is mcm bootstrappable, and this approach has been adapted for quantile regression by Kocherginsky et al. (2005).

From the bootstrapping algorithm yielding K bootstrap samples, a sequence of $\beta^{(1)}, ..., \beta^{(K)}$ is returned, and under the assumption of iid error models

$$y_i = x_i^T \beta(\tau) + \epsilon_i \tag{3.45}$$

He and Hu (2002) have shown that the sample variance of $\beta^{(k)}$ consistently approximates the variance-covariance matrix. The sequence is a Markov chain, and some modifications of the basic algorithm has been made to eliminate the correlation of the estimated sequence, obtaining the MCMB-A method. This is done by standardizing the design matrix X by $\tilde{X} = (X^T X)^{-1/2} X$, and transform back at the end by $\beta^{(k)}(X^T X)^{-1/2}\tilde{\beta}^{(k)}$, leading back to the original parameter space. This method has been implemented by Kocherginsky et al. (2005) to the `quantreg` package, called `mcmb`.

When using bootstrapping to estimate the variance-covariance matrix between 50 and 200 bootstrap replications are reccommended to obtain a decent estimate (Kocherginsky et al., 2005). To construct a confidence interval based on the percentiles of the bootstrap estimates, more replications are needed. This way of constructing confidence intervals is often preferred, but on the other hand, for large data sets where we would not afford a much larger number of bootstrap replicates, an SD-based confidence interval is generally adequate (Kocherginsky et al., 2005). For our case study we therefore use this SD-based confidence interval on the form

$$\hat{\beta}(\tau) \pm z_{\alpha/2}\mathrm{SD}(\hat{\beta}(\tau)). \tag{3.46}$$

In addition to testing significance and creating confidence interval, it is possible to carry out a test to find out whether the effect is constant over all the quantiles. This can be done by the anova-function in the `quantreg` package that perform a joint test of equality of slopes in the models for different quantiles of interest (Koenker, 2015).

3.3 Derivation of a relation between odds ratio (OR) and quantile effect

In this section an approximate relation between the odds ratio, or more precisely the logarithm of the odds ratio, and the quantile effect will be derived.

We let $\hat{p}_1$ and $\hat{p}_2$ denote the estimated proportion of observations with value less than the critical values $\hat{x}_1$ and $\hat{x}_2$, respectively. Let $F(x)$ denote the cumulative distribution function with corresponding density $f(x)$. Then $\hat{p}_1 = F(\hat{x}_1)$ and $\hat{p}_2 = F(\hat{x}_2)$, and $\hat{x}_2 = F^{-1}(\hat{p}_2)$ and $\hat{x}_1 = F^{-1}(\hat{p}_1)$. By definition,

$$\text{ln (OR)} = \ln\left(\frac{\hat{p}_2/(1-\hat{p}_2)}{\hat{p}_1/(1-\hat{p}_1)}\right) = \ln\left(\frac{F(\hat{x}_2)/(1-F(\hat{x}_2)}{F(\hat{x}_1)/(1-F(\hat{x}_1))}\right). \tag{3.47}$$

Define

$$g(x) = \ln\left(\frac{F(x)}{1-F(x)}\right). \tag{3.48}$$

Then the log-odds ratio in Equation (3.47) can now be expressed as

$$\text{ln(OR)} = g(\hat{x}_2) - g(\hat{x}_1), \tag{3.49}$$

where

$$g(\hat{x}_1) = \ln\left(\frac{F(\hat{x}_1)}{1-F(\hat{x}_1)}\right) \text{ and } g(\hat{x}_2) = \ln\left(\frac{F(\hat{x}_2)}{1-F(\hat{x}_2)}\right). \tag{3.50}$$

It is in general adequate to use only the first order Taylor expansion; the first derivative, and we ignore the remainder (Casella and Berger, 2002). By first order Taylor expansion of $g(\hat{x}_2)$ around $\hat{x}_1$,

$$g(\hat{x}_2) \approx g(\hat{x}_1) - \frac{d}{dx}g(\hat{x}_1)(\hat{x}_2 - \hat{x}_1) \tag{3.51}$$

where

$$\begin{aligned} \frac{d}{dx}g(x) =& \frac{d}{dx}\left(\ln(F(x)) - \ln(1-F(x))\right) \\ =& \left(\frac{1}{F(x)} - \frac{1}{1-F(x)}\right)f(x) \\ =& \frac{f(x)}{F(x)(1-F(x))}. \end{aligned} \tag{3.52}$$

The approximate relation between the odds ratio and the quantile effect is thus,

$$\begin{aligned} g(\hat{x}_2) \approx& g(\hat{x}_1) - \frac{d}{dx}g(\hat{x}_1)(\hat{x}_2 - \hat{x}_1) \\ g(\hat{x}_2) - g(\hat{x}_1) \approx& -\frac{f(\hat{x}_1)}{F(\hat{x}_1)(1-F(\hat{x}_1))}(\hat{x}_2 - \hat{x}_1) \\ \text{ln (OR)} \approx& -\frac{f(\hat{x}_1)}{F(\hat{x}_1)(1-F(\hat{x}_1))}(\hat{x}_2 - \hat{x}_1), \end{aligned} \tag{3.53}$$

Chapter 4

Synthetic case studies

The aim of this chapter is to explore analyses using quantile regression and logistic regression for known pdfs, both analytically and by simulations. The quantile effect and the odds ratio for location, scale and location-scale shifts of the normal distribution and the Student t-distribution will be demonstrated, and further see how the two approaches are related. To do this quantile regression and logistic regression are tested for simulated data sets.

The simulated data sets has the same number of observations as the data set being analyzed for effect of smoking on birth weight in Chapter 5, i.e $511\,447$ observations. These observations are categorized in two groups, where the observations of Group 2 are being exposed to a factor that the observations of Group 1 are not. In Group 2 there are $59\,070$ of the observations, yielding $452\,377$ observations in Group 1. This is similar to the number of smokers and the number of non-smokers, respectively.

Further some illustrations on how a gradual increase in location, scale and location-scale shifts affects the the results will be shown for both the normal distributed cases and the t-distributed cases. This shows a trend in how the effect of the factor changes when the difference between the two groups gradually increases. The distribution of Group 1 is constant and we change the distribution of Group 2 by changing $\Delta\mu$ and/or $\Delta\sigma$.

4.1 Normally distributed cases

We let the observations of Group 1 be normally distributed with mean μ_1 and standard deviation σ_1, and the observations of Group 2 be normally distributed with mean $\mu_1 + \Delta\mu$ and standard deviation $\sigma_1 + \Delta\sigma$. Let X be an observation from our data set, then,

$$\begin{aligned} X|\text{Group 1} &\sim N(\mu_1,\, \sigma_1^2) \\ X|\text{Group 2} &\sim N(\mu_1 + \Delta\mu,\, (\sigma_1 + \Delta\sigma)^2). \end{aligned} \tag{4.1}$$

When analyzing the data sets using quantile regression we let τ be the quantile of interest. When the logistic regression model is used, we are interested in the odds of the event that a given observation is smaller than a chosen cut-off value, denoted by c. We analyze the data for five different values for τ, and the values for $c^{(\tau)}$ is set to be equal to the intercept value obtained by quantile regression, that is the estimated quantile value for Group 1. Let x_i be defined as

$$x_i = \begin{cases} 1 & \text{if the observation comes from Group 2} \\ 0 & \text{if the observation comes from Group 1.} \end{cases} \tag{4.2}$$

Then the response in the τ-th quantile is modelled by Equation (3.35), yielding

$$Q_{Y_i}(\tau|x_i) = \beta_0(\tau) + \beta_1(\tau)x_i \tag{4.3}$$

and by logistic regression the probability for an event below $c^{(\tau)}$ is modelled through the logit link by Equation (3.14), yielding

$$\text{logit}(p_i^{(\tau)}) = \beta_0^{(\tau)} + \beta_1^{(\tau)}x_i \tag{4.4}$$

For the synthetic case studies we will look at three different cases that will illustrate the effect on quantiles and odds ratio of location, scale and location-scale shifts of the normal distribution. The three cases are presented in Table 4.1. By letting the observations be normally distributed it is, in addition to analyses on the data set, possible to calculate the theoretical values for the effect of the factor in the data set. The derivation of these theoretical values is the subject of Section 4.1.2 and Section 4.1.3, after presenting some properties of the normal distribution.

Table 4.1: An overview of the three different simulated cases.

	μ_1	$\Delta\mu$	σ_1	$\Delta\sigma$
Case 1	3539	-193	635	0
Case 2	3539	0	635	100
Case 3	3539	-193	635	100

4.1.1 Properties of the normal distribution

Let Y be a random variable from a normal distribution with mean μ and variance σ^2,

$$Y \sim N(\mu, \sigma^2).$$

Then Y has the probability density function (pdf),

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma}\text{exp}\left(-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}\right)$$

Considering the random variable $Z \sim N(0,1)$ we obtain the standard normal curve on the form,

$$f(z) = \frac{1}{\sqrt{2\pi}}\text{exp}\left(-\frac{z^2}{2}\right),$$

and the cumulative distribution function is given by

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\Big(-\frac{t^2}{2}\Big) dt. \tag{4.5}$$

The standard normal curve can be used to find this probability, by finding the area under the curve.

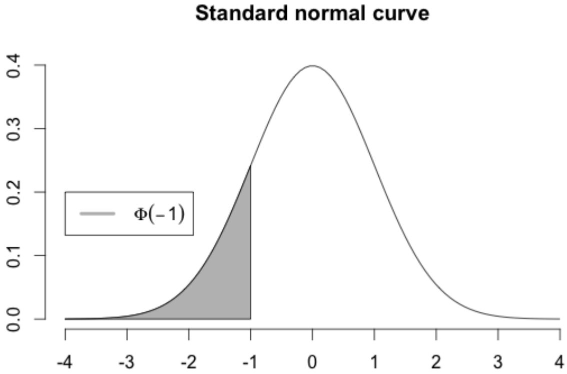

Figure 4.1: Standard normal curve

To find $\Phi(z)$ we use a $normal\,table$ or statistical software, that is based on numerical integration in Equation (4.5). If $\Phi(z)$ is known z can be found by that same methods. This probability can be found for the variable Y as well by doing a $Z\,transformation$ where,

$$Z = \frac{Y - \mu}{\sigma}. \tag{4.6}$$

Then

$$P(Y \leq y) = P(\frac{Y - \mu}{\sigma} \leq \frac{y - \mu}{\sigma}) = P(Z \leq \frac{y - \mu}{\sigma}) = \Phi(\frac{y - \mu}{\sigma})$$

(Larsen and Marx, 2012).

4.1.2 Theoretical quantile effect

We let τ denote the quantile of interest, and to calculate the theoretical values for the response associated with τ, we use the properties of the normal distribution. We have

$$P(X < k_1|\text{Group 1}) = \tau \text{ and } P(X < k_2|\text{Group 2}) = \tau,$$

where k_1 and k_2 denotes the response associated with τ for Group 1 and Group 2 respectively. Further, we use the $Z\,transformation$ from Equation (4.6) and obtain

$$Z_1 = \frac{X - \mu_1}{\sigma_1} \text{ and } Z_2 = \frac{X - (\mu_1 + \Delta\mu)}{(\sigma_1 + \Delta\sigma)}$$

Recall Equation (4.5) and let

$$\Phi(z_\tau) = P(Z \leq z_\tau) = \tau.$$

We are now able to find expressions for k_1 and k_2 associated with τ. We have

$$z_\tau = \frac{k_1 - \mu_1}{\sigma_1} \text{ and } z_\tau = \frac{k_2 - (\mu_1 + \Delta\mu)}{(\sigma_1 + \Delta\sigma)} \tag{4.7}$$

$$k_1 = z_\tau \sigma_1 + \mu_1 \quad \text{and} \quad k_2 = z_\tau(\sigma_1 + \Delta\sigma) + \mu_1 + \Delta\mu$$

The difference $k_2 - k_1$ would now be the change in the response on the τth quantile caused by the factor exposed to Group 2. This change denotes the effect of the factor on the $\tau - th$ quantile.

$$\begin{aligned}
\text{effect} = k_2 - k_1 &= (z_\tau(\sigma_1 + \Delta\sigma) + \mu_1 + \Delta\mu) - (z_\tau\sigma_1 + \mu_1) \\
&= z_\tau(\sigma_1 + \Delta\sigma) + \mu_1 + \Delta\mu - z_\tau\sigma_1 - \mu_1 \\
\text{effect} &= z_\tau\Big((\sigma_1 + \Delta\sigma) - \sigma_1\Big) + \Delta\mu \qquad (4.8) \\
\text{effect} &= z_\tau(\sigma_1 + \Delta\sigma - \sigma_1) + \Delta\mu \\
\text{effect} &= z_\tau\Delta\sigma + \Delta\mu
\end{aligned}$$

4.1.3 Theoretical odds ratio

When applying the logistic regression model we are interested in the odds for the event that an observation X is smaller than a given value c. The odds ratio, OR, is the effect on the odds for the given event when the observation comes from Group 2.

Let $odds_1$ denote the odds for the event given Group 1, and $odds_2$ denote the odds for the event given Group 2. The odds ratio is then,

$$\text{OR} = \frac{\text{Odds}_2}{\text{Odds}_1} = \frac{P(X \leq c|\text{Group2})/(1 - P(X \leq c|\text{Group2}))}{P(X \leq c|\text{Group1})/(1 - P(X \leq c|\text{Group1}))}. \tag{4.9}$$

Standardizing using Equation (4.6) yields

$$\begin{aligned}
P(X \leq c|\text{Group1}) &= P(Z \leq \frac{c - \mu}{\sigma_1}) \\
P(X \leq c|\text{Group2}) &= P(Z \leq \frac{c - \mu_1 - \Delta\mu}{(\sigma_1 + \Delta\sigma)}).
\end{aligned} \tag{4.10}$$

By combining Equations (4.9), (4.10) and (4.5) we obtain the formula for the odds ratio,

$$\text{OR} = \frac{\Phi(\frac{c - \mu_1 - \Delta\mu}{(\sigma_1 + \Delta\sigma)})/(1 - \Phi(\frac{c - \mu_1 - \Delta\mu}{(\sigma_1 + \Delta\sigma)}))}{\Phi(\frac{c - \mu_1}{\sigma_1})/(1 - \Phi(\frac{c - \mu_1}{\sigma_1}))}, \tag{4.11}$$

that would be the effect of the factor in Group 2 on the odds for the event of interest.

4.1.4 Case 1 - results

For the first set of analyses on data simulated from a normal distribution we let the two groups be distributed as presented in Equation (4.1) by the values found for ”Case 1” in Table 4.1. The pdf’s and and the cdf’s of the two groups are shown in Figure 4.2. Since $\Delta\sigma = 0$ the distributions have the same shape, and because of the change in the mean value, the distribution of Group 2 is shifted to the left with a distance to Group 1 equal to $\Delta\mu = -193$.

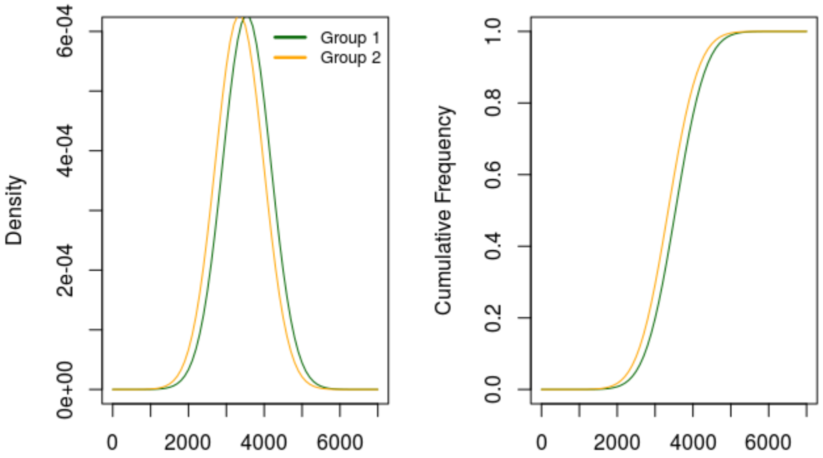

Figure 4.2: Left panel: pdf’s of the two groups. Right panel: cdf’s of the two groups. Green curve: Group 1. Yellow curve: Group 2.

We analyze the data for five different values for τ, and the values for c corresponding to the τ-th quantile is found in Table 4.2.

Table 4.2: The quantiles of interest, τ, and corresponding values for c of interest.

τ	0.05	0.25	0.50	0.75	0.95
$c^{(\tau)}$	2493	3109	3540	3968	4583

The results obtained by quantile regression for this simulated data set are shown in Figure 4.3. The effect appears as a straight, horizontal line approximately at -200 on the vertical axis. This implies that the effect of the factor is estimated to be equal for all quantiles of the distribution. Recalling Section 3.2.3 and Equation (3.39) the effect on the quantiles can be expressed as the horizontal distance between the curves of the empirical cumulative distribution function. By construction, this distance is equal over the entire distribution, and equal to the difference in the mean between the two groups, $\Delta\mu = -193$.

This is reflected in the estimated quantile effect presented by the curve in Figure 4.3, and as expected. Since Group 2 is shifted to the left of Group 1, the quantile effect is negative.

The results obtained by logistic regression for the same simulated data are found in Figure 4.4. The curve denoting the odds ratio does not give the impression of a constant effect on the logit scale. The odds ratio is larger than 1 for all cut-off values, implying that the odds is increasing by exposure of the factor of interest, and hence more likely with an observation below $c^{(\tau)}$ with exposure. It is not as easy to see exactly how a change in location affects the results from logistic regression, that we will also come back to later.

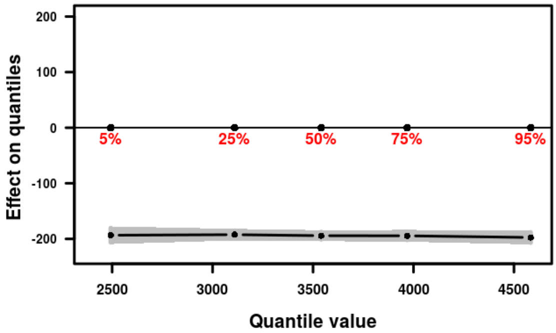

Figure 4.3: Results from quantile regression. The black curve denotes the quantile effect shown on the y-axis, plotted against the values for the τ-th quantile when the factor of interest is not present, that is the quantiles of Group 1, on the x-axis. The τ's of interest are presented by the red percentages. A 95%-confidence interval is indicated in gray.

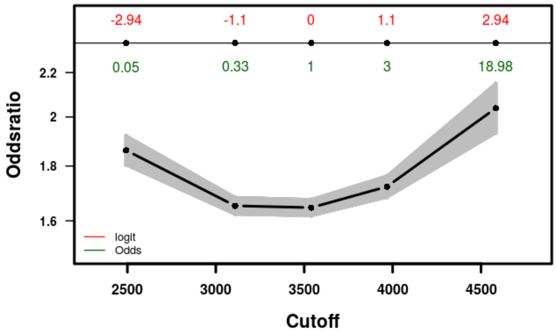

Figure 4.4: Results from logistic regression. The red numbers denote the logit$(p_i^{(\tau)})$ for an observation being below the cut-off value $c^{(\tau)}$ found on the x-axis. The green numbers denote the corresponding odds. The black curve denotes the odds ratio, plotted on a log-scale, i.e it is the values on logit scale that are plotted, but the labels on the y-axis show actual odds ratio values. A 95%-confidence interval is indicated in gray.

Further we let the location shift to the left gradually increase, and let $\Delta\mu$ take the values presented in Table 4.3. The location shifts are illustrated in Figure 4.5 showing the pdf and the cdf of the different changes. Note that the black curve now shows the distribution of Group 1. The results from both quantile regression and logistic regression are presented in Figure 4.6.

Table 4.3: The different changes in location, with corresponding colours for Figure 4.5.

black	$\Delta\mu = 0$	$\Delta\sigma = 0$
red	$\Delta\mu = -100$	$\Delta\sigma = 0$
green	$\Delta\mu = -200$	$\Delta\sigma = 0$
blue	$\Delta\mu = -300$	$\Delta\sigma = 0$

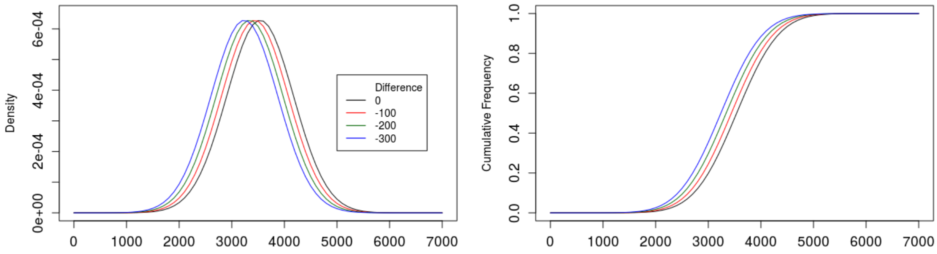

Figure 4.5: Illustrates location shift to the left. Left panel: pdf's. Right panel: cdf's. Note that the black curve represents the distribution of Group 1.

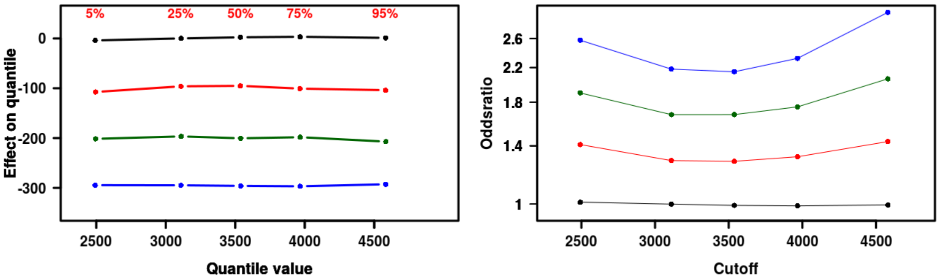

Figure 4.6: Left panel: the results from quantile regression. Right panel: the results from logistic regression, plotted on log-scale. Note that the black curve now illustrates how the effect is presented if the two groups were equally distributed. The colours are presented in Table 4.3.

From these results we notice that for the results obtained by quantile regression the curves presenting the quantile effect still appear as horizontal lines, and thus constant in all quantiles. As the distribution of Group 2 gradually shifts more to the left, the effect gradually increases, and each curve lies approximately at the value of $\Delta\mu$. For the results obtained by logistic regression the curves presenting the odds ratio gradually take larger values. When the two groups are equally distributed the odds ratio is naturally approximately constant at 1. In addition to the increase in the odds ratio by location shift, we also notice that the curves change shapes, and it might seem as they get gradually more asymmetric around the mean as well. When the quantile effect is approximately 0 the odds ratio is approximately 1. This can also be seen in Figure 4.7 where the odds ratio is plotted against the quantile effect. The approximate relation derived in Section 3.3 is also reflected in this figure, and we notice the approximate linear relation between the quantile effect and the log-odds ratio.

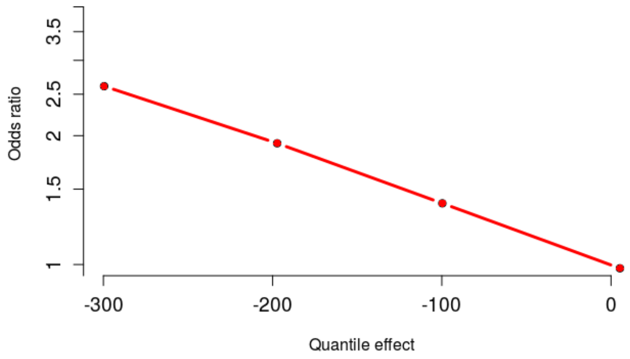

Figure 4.7: The odds ratio (log-scale) plotted against the quantile effect. Calculated for the 5%-quantile.

4.1.5 Case 2 - results

For the second set of analyses on data simulated from a normal distribution, we consider how the analyses are affected by a difference in the standard deviation, the scale parameter, between the two groups. We let the two groups be distributed as in Equation (4.1) by the values found in Table 4.1 by ”Case 2”. An overview of the quantile of interest and the corresponding cut-off value for $c^{(\tau)}$ is found in Table 4.4.

Table 4.4: The quantiles of interest, τ, and corresponding values for c of interest.

τ	0.05	0.25	0.50	0.75	0.95
$c^{(\tau)}$	2492	3107	3537	3967	4583

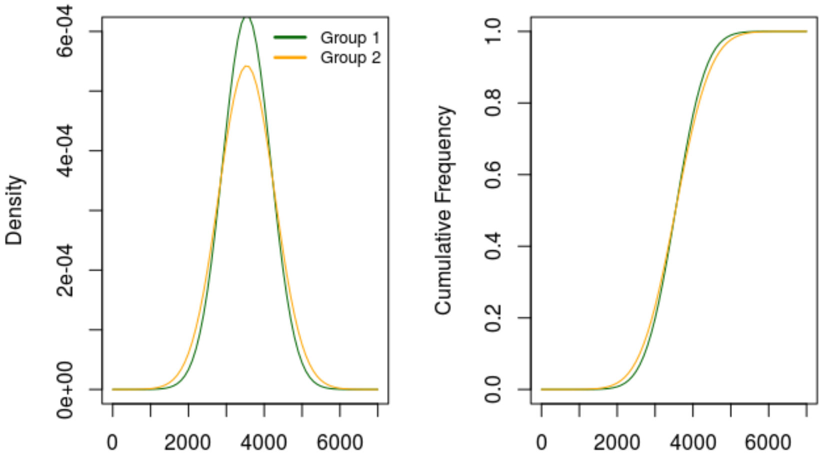

Figure 4.8: Left panel: pdf’s of the two groups. Right panel: cdf’s of the two groups. Green curve: Group 1. Yellow curve: Group 2.

The pdf’s and cdf’s of the two groups are shown in Figure 4.8. By construction, the two groups have the same mean value, and as we move further towards the tails of the distribution then the distance between them increases.

The results from quantile regression are found in Figure 4.9. The black curve, that denotes the effect of exposure of Group 2, now takes on a different shape than in the previous case, where the mean was different and the standard deviation was constant. Instead of being a straight horizontal line, we now observe a curve with constant slope. The effect is both positive and negative, and zero exactly at the 50%-quantile. For the normal distribution the mean is equal to the median, i.e. at the 50%-quantile. Keeping in mind that the mean of the two distributions now is equal, this is as expected. Further, the effect is positive as we move towards the upper tail, and negative as we move towards

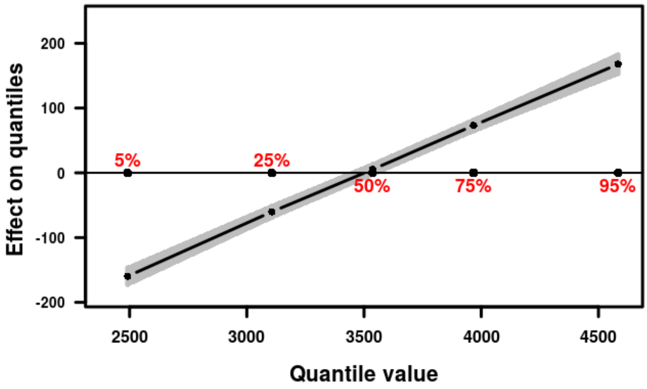

Figure 4.9: Results from quantile regression. The black curve denotes the quantile effect shown on the y-axis, plotted against the values for the τ-th quantile when the factor of interest is not present, that is the quantiles of Group 1, on the x-axis. The τ's of interest are presented by the red percentages. A 95%-confidence interval is indicated in gray.

the lower tail. Recalling Section 3.2.3 and Equation (3.39) the effect on the quantiles can be expressed as the horizontal distance between the curves of the empirical cumulative distribution function. The distribution of the observations in Group 2 is more spread on both tails in opposite directions. This explains why the effect is positive in the upper tail and negative is the lower tail, and in addition, the absolute value of the effect would be equal. This is consistent with the black curve in Figure 4.9.

The results from logistic regression are found in Figure 4.10. The black curve denoting the odds ratio takes values larger than 1 and smaller than 1. It reaches the value 1 when the cut-off value is equal to the mean value. Since the mean is equal for Group 1 and Group 2, the factor that Group 2 is exposed to would not give an effect in this point. An odds ratio of 1 when $c =$ mean is then consistent with the effect being 0 at the 50%-quantile. As before we do not get the impression that the absolute value of the effect is equal in the upper and lower tail.

Further we let $\Delta\sigma$ gradually increase, and obtain more spread in each direction of the distribution. The values of $\Delta\sigma$ is presented in Table 4.5 with corresponding color for the figures. The curves presenting the results from quantile regression in Figure 4.11 are all increasing and cross 0 in the mean, as expected. Further we observe the same as before, the effect is as positive in the upper tail, as it is negative in the lower tail. The results from logistic regression show that all curves are decreasing and cross 1 at the mean value. This is as expected. The curves are not linear, and by the odds ratio curves it is difficult to draw any conclusions on how the effect varies over the distribution. The relation between the two approaches is also for this case presented in Figure 4.13. The same trend is observed for this case; as the quantile effect approaches 0 the odds ratio approaches 1, and there is an approximate linear relation.

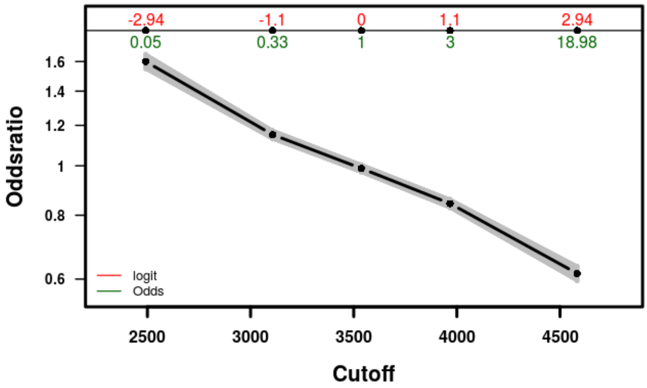

Figure 4.10: Results from logistic regression. The red numbers denote the logit($p_i^{(\tau)}$) for an observation being below the cut-off value c found on the x-axis. The green numbers denote the corresponding odds. The black curve denotes the odds ratio, plotted on a log-scale. A 95%-confidence interval is indicated in gray.

Table 4.5: The different changes in scale, with corresponding colours for Figure 4.11.

black	$\Delta\mu = 0$	$\Delta\sigma = 0$
red	$\Delta\mu = 0$	$\Delta\sigma = 50$
green	$\Delta\mu = 0$	$\Delta\sigma = 100$
blue	$\Delta\mu = 0$	$\Delta\sigma = 150$

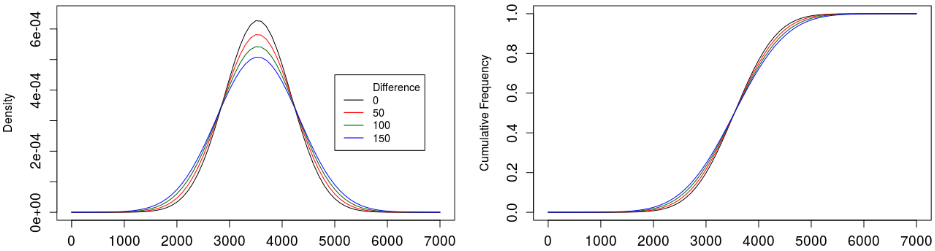

Figure 4.11: Illustrates the difference in the standard deviation. Left panel: pdf's. Right panel: cdf's. Note that the black curve represents the distribution of Group 1.

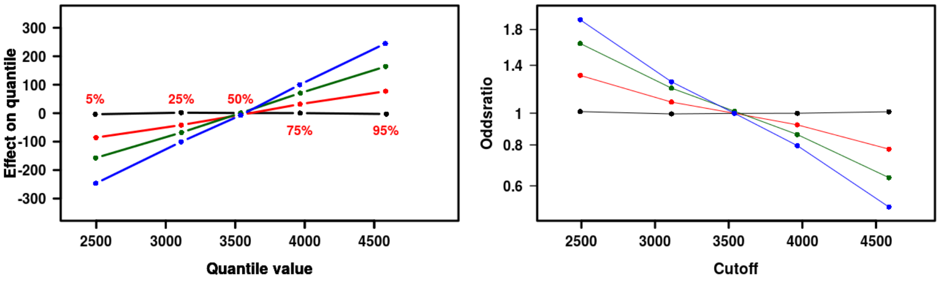

Figure 4.12: Left panel: the results from quantile regression. Right panel: the results from logistic regression, plotted on log-scale. Note that the black curve now illustrates how the effect is presented if the two groups were equally distributed. The colours are presented in Table 4.5.

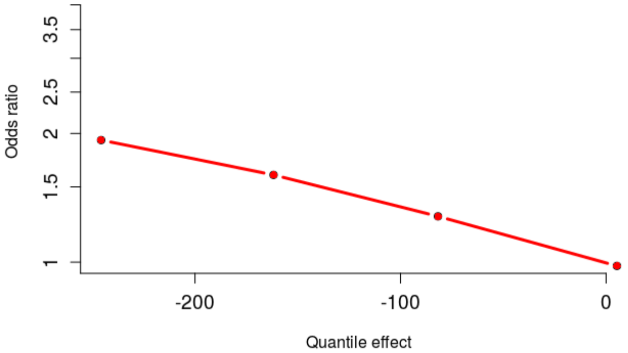

Figure 4.13: The odds ratio (log-scale) plotted against the quantile effect. Calculated at the 5%-quantile.

4.1.6 Case 3 - results

For the third set of analyses on data simulated from a normal distribution we consider how the analysis is affected when both the mean and the standard deviation is different between the two groups, location-scale shift. We let the two groups be distributed as in Equation (4.1) by the values found in Table 4.1 denote ”Case 3”. An overview of the quantiles of interest and the corresponding cut-off values $c^{(\tau)}$ are found in Table 4.6.

Table 4.6: The quantiles of interest, τ, and corresponding values for c of interest.

τ	0.05	0.25	0.50	0.75	0.95
$c^{(\tau)}$	2495	3110	3540	3968	4584

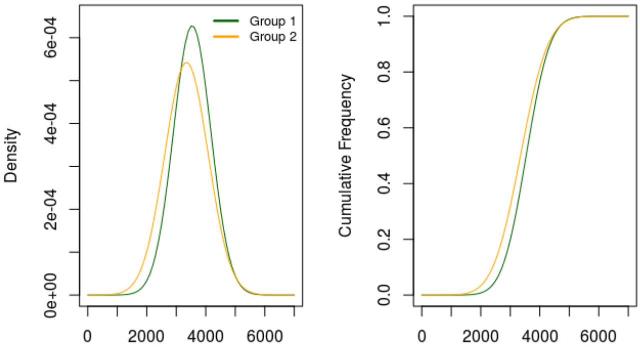

Figure 4.14: Left panel: pdf’s of the two groups. Right panel: cdf’s of the two groups. Green curve: Group 1. Yellow curve: Group 2.

The pdf’s and cdf’s of the two groups are shown in Figure 4.14. A clear trend where the distance between the distributions gets larger in the lower tail, and coinciding in the upper tail is observed.

The result from quantile regression are found in Figure 4.15. The shape of the black curve that denotes the effect on the quantile looks similar to the shape of the effect curve in case 2, but this curve shows negative effect on all quantiles. In the 95%-quantile the effect is close to zero. This is consistent with what we would expect when examining the distributions, keeping in mind that they are approximately coinciding in the upper tail. We also notice clearly how this linear curve is consistent with the theoretical expression derived in Section 4.1.2.

The results from logistic regression is shown in Figure 4.16. The odds ratio is larger than 1 for all values of $c^{(\tau)}$. It is largest in the lower tail of the distribution and decreases until it approaches 1 when we consider the upper tail. This is also natural considering the shape of the cdf.

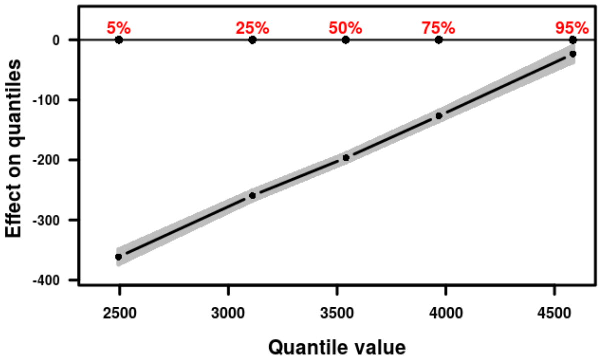

Figure 4.15: Results from quantile regression. The black curve denotes the quantile effect shown on the y-axis, plotted against the values for the τ-th quantile when the factor of interest is not present, that is the quantiles of Group 1, on the x-axis. The τ's of interest is presented by the red percentages. A 95%-confidence interval is indicated in gray.

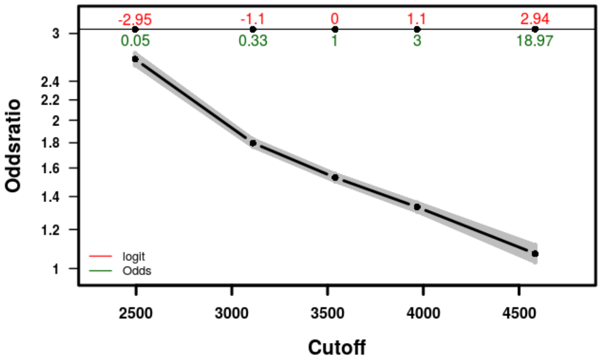

Figure 4.16: Results from logistic regression. The red numbers denote the logit$(p_i^{(\tau)})$ for an observation being below the cut-off value c found on the x-axis. The green numbers denote the corresponding odds. The black curve denotes the odds ratio, plotted on a log-scale. A 95%-confidence interval is indicated in gray.

Further we let both $\Delta\mu$ and $\Delta\sigma$ gradually change obtaining a gradually clearer location-scale shift. In Figure 4.17 we observe a gradually larger distance between the groups in the lower tail, and in the upper tail, all of the distribution curves are coinciding. Values of the change and corresponding colors is found in Table 4.7.

Table 4.7: The different location-scale shits, with corresponding colours for Figure 4.17.

black	$\Delta\mu = 0$	$\Delta\sigma = 0$
red	$\Delta\mu = -100$	$\Delta\sigma = 50$
green	$\Delta\mu = -200$	$\Delta\sigma = 100$
blue	$\Delta\mu = -300$	$\Delta\sigma = 150$

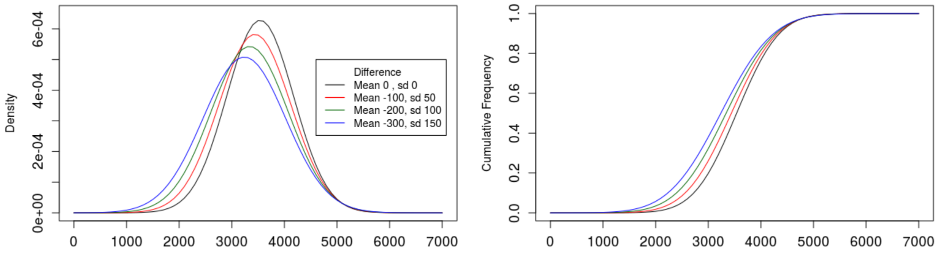

Figure 4.17: Illustration of the difference in the scale and location. Left panel: pdf's. Right panel: cdf's. Note that the black curve represents the distribution of Group 1.

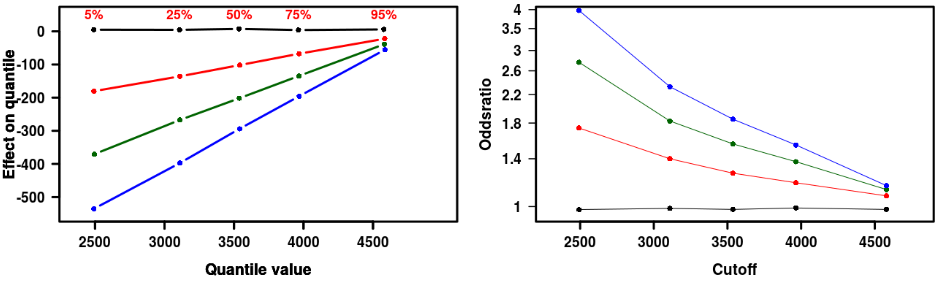

Figure 4.18: Left panel: the results from quantile regression. Right panel: the results from logistic regression, plotted on log-scale. Note that the black curve now illustrates how the effect is presented if the two groups were equally distributed. The colours are presented in Table 4.7.

The results are presented in Figure 4.18. The curves denoting the quantile effect is still linear, but the negative effect gradually increases, and the slope gradually increases as the groups are more shifted. In the upper tail all curves are close to 0 for the effect and approximately equal. For logistic regression the odds ratio gradually increases and in addition the shape of the curves gradually gets more different as the two groups are more shifted. As expected the odds ratio approaches 1 in the upper tail. In Figure 4.19 the odds ratio is plotted against the quantile effect, and we also here observe an approximately linear relation, and as the quantile effect approaches 0 the corresponding odds ratio approaches 1.

From the three cases discussed above we get an impression of how the results from quantile regression and logistic regression relate to each other. To summarize, an odds ratio of 1 obtained from the logistic regression model is equivalent to an effect of 0 on the quantile of interest. When the effect on quantiles approaches 0, the odds ratio approaches 1. We have also seen that the curves presenting the quantile effect yield results as expected when considering the corresponding cdf's, and the results are easy to interpret as they are on the same scale. In addition we can easily detect where the effect is found to be largest, either advantageous or disadvantageous. The results from logistic regression are not as easy to connect to the location, scale and location-scale shifts of the distribution. We are able to find out it the effect of exposure yields increasing or decreasing odds. When examining how the effect varies on different quantiles of the distribution, the odds ratio alone does not answer this without taking the original probability into account. This has not been discussed in this chapter, but it will be considered in Chapter 5.

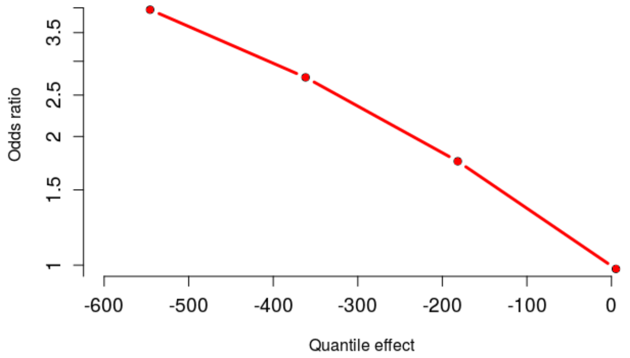

Figure 4.19: The odds ratio (log-scale) plotted against the quantile effect. Calculated at the 5%-quantile.

4.2 Theoretical values and simulation study

In this section we present a simulation study that has been performed to evaluate the performance of the confidence intervals obtained for the estimated parameters. The simulation study is performed for one of the above cases, and we have chosen the case where both the mean and the standard deviation differ between the two groups, with $\Delta\mu = -300$ and $\Delta\sigma = 150$. Both quantile regression and logistic regression have been applied through 1000 simulations from this distribution. Further we consider the coverage of the confidence intervals. The theoretical values needed for this part are calculated using the Equation (4.8) and Equation (4.11) derived in Sections 4.1.2 and 4.1.3. In Figure 4.20 the theoretical values for the quantile effect and the theoretical odds ratio are shown for all the cases in Table 4.7. The blue curve denotes the case of interest in the simulation study, but we choose to present the curves for all the cases as it is of interest to see that these curves appear similar as the curves in Figure 4.18.

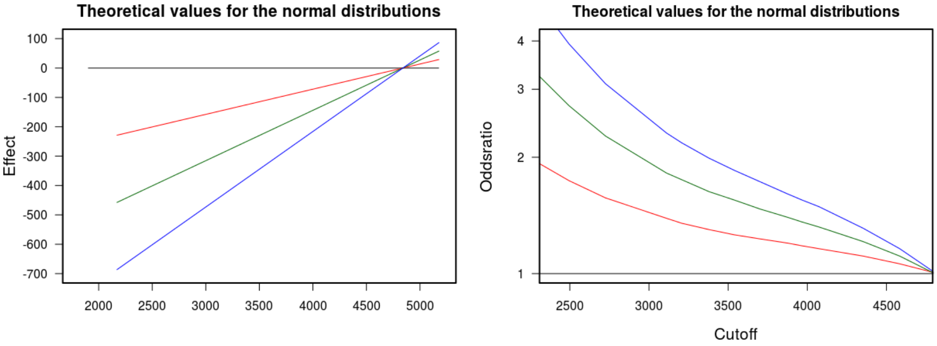

Figure 4.20: Left panel: the theoretical quantile effect. Right panel: the theoretical odds ratio.

Results from the simulations are shown in Table 4.8 and in Table 4.9. The mean difference between the estimated odds ratio and the theoretical odds ratio is found to be 0.00482, and for the quantile effect the mean difference is found to be -0.1064, thus we may consider both estimators to be unbiased for this sample size. Further we consider the coverage of the confidence intervals, and find that the theoretical value is covered by the 95%-confidence interval for 94.5% of the estimates by logistic regression. By quantile regression the theoretical values for the quantile effect is covered by the 95%-confidence interval in 95.14% of the estimates. From these findings there is no reason to believe that either of the confidence intervals is too optimistic or too narrow.

Table 4.8: Result for logistic regression obtained by 1000 simulations.

Average of estimated OR	Theoretical value
3.924	3.910
2.307	2.300
1.843	1.840
1.550	1.550
1.161	1.160

Table 4.9: Result for quantile regression obtained by 1000 simulations.

Average of estimated effect	Theoretical value
-547.071	-546.728
-401.425	-401.173
-300.057	-300.000
-198.899	-198.827
-53.080	-53.272

4.3 Student t-distribution cases

To explore analyses using quantile regression and logistic regression on distributions with heavier tails, similar analyses have been performed on simulated data from a Student t-distribution with 3 degrees of freedom, although not as extensively. We adapt the three cases of changes in the location and scale, presented in Table 4.3, Table 4.5 and Table 4.7. Let

$$T \sim t(3) \tag{4.12}$$

Then

$$\begin{aligned} X|\text{Group 1} &= \mu_1 + \sigma_1 T \\ X|\text{Group 2} &= (\mu_1 - \Delta\mu) + (\sigma_1 - \Delta\sigma)T, \end{aligned} \tag{4.13}$$

denotes the distributions of the two groups from the Student t-distribution.

The models obtained by quantile regression and logistic regression are defined in Equation (4.3) and Equation (4.4). Results from the analyses considering location, scale and location-scale shifts are presented in Figure 4.21. From comparing these results to the corresponding results for the normally distributed cases (Figure 4.6, 4.12, 4.18) we find many of the same trends when we consider when the quantile effects are positive or negative, and the odds ratios are smaller than or larger than 1. The curves denoting the quantile effects behaves similar as the curves obtained by the normal distributed cases. The results from logistic regression yields odds ratio curves that behaves quite differently and now has a clearly different shape than for the cases of normal distributed data. In Figure 4.24 the odds ratio is plotted against the quantile effect for the three cases. The approximate linear relation is still observed, and a quantile effect of approximately 0 yields an odds ratio of approximately 1. From this we understand that the interpretation of the odds ratio depend both on the original probability, and the shape of the distribution, and if this is not taken into account one may draw the wrong conclusion about the effect of the exposure. It does not make sense to compare odds ratios alone in different parts of the distribution and neither to compare odds ratios when the distributions have different shapes. From this we also see that by knowing that there is a pure location, scale or location-scale shift between the two groups, it is possible to know how to expect the quantile effects to behave, whereas this is not the case for logistic regression since the odds ratios are dependent on the shape of the distribution. This will also be considered further in Chapter 5 where we consider a case study concerning the effect of smoking and mother's age on birth weight.

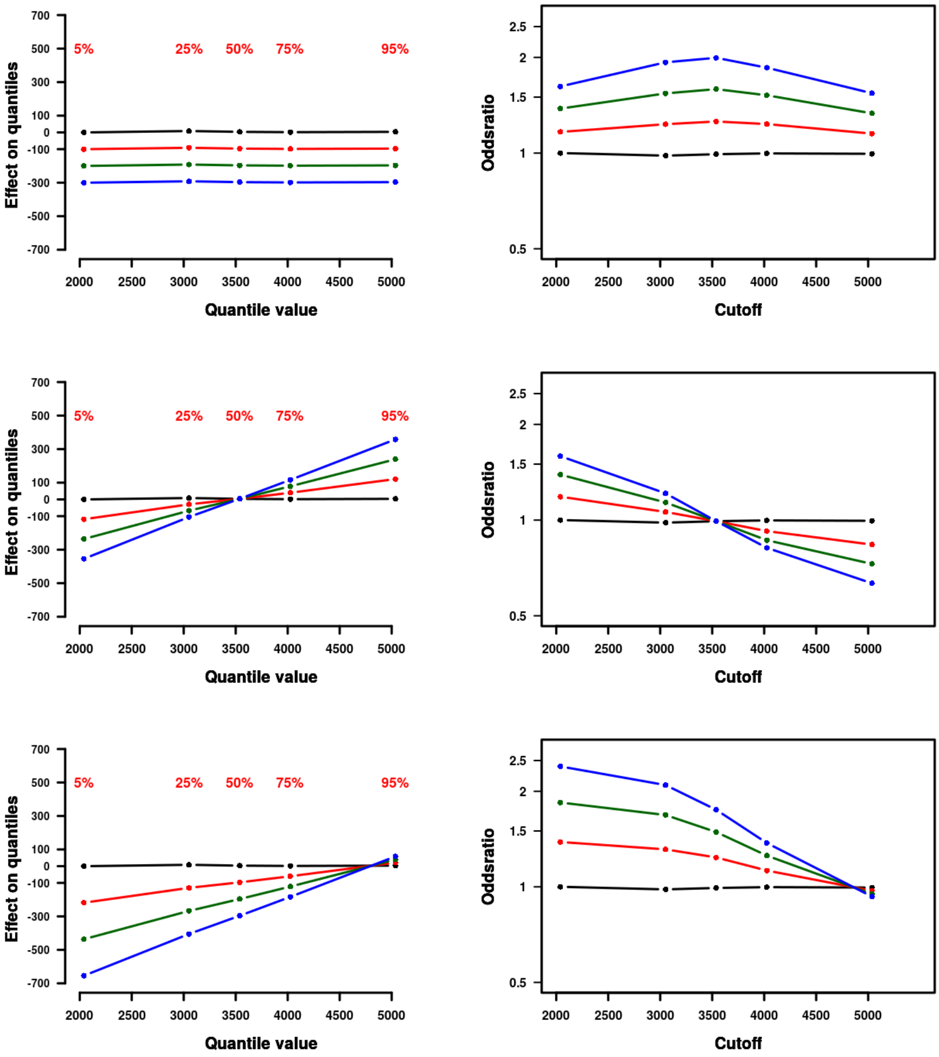

Figure 4.21: Left panel: the results from quantile regression. Right panel: the results from logistic regression, plotted on log-scale. Note that the black curve now illustrates how the effect is presented if the two groups were equally distributed. Upper panel: location shift. Middle panel: scale shift. Lower panel: location-scale shift.

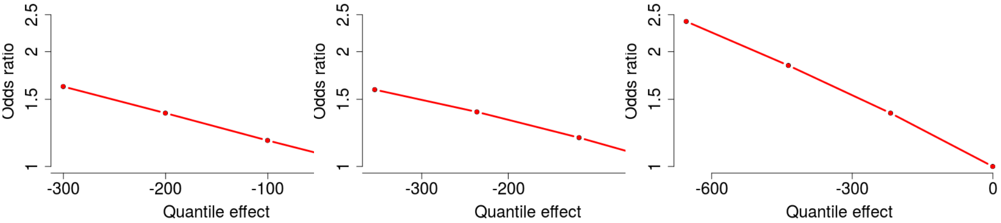

Figure 4.24: Quantile effect vs OR, the 5%-quantile, all cases.

Chapter 5

Case study of birth weight

In this chapter the birth weight data set introduced in Chapter 2 is analyzed using both logistic regression and quantile regression, as introduced in Chapter 3. We are especially interested in the effect of smoking and mother's age on low birth weight, but we will also model the effect of these factors for other quantiles. For the analysis of the effect of mother's age on birth weight we will also use the spontaneous birth variable as a selection variable, doing the same analysis on a data set containing only spontaneous births.

The quantiles of interest are presented together with corresponding cut-off values used when fitting the logistic regression models. These cut-off values are set to be equal to the intercept values from quantile regression, yielding the corresponding quantile value when the factor of interest is not present. This means that the cut-off value will change between data sets, and the values of interest will be presented in tables before each case of analysis. For each case we notice that the value corresponding to the 5%-quantile is close to the definition of low birth weight, 2500g.

The results from both logistic regression and quantile regression modelling the effect of smoking on birth weight are presented in Section 5.1. The results from modelling the effect of the mother's age on birth weight, also including only spontaneous births, are presented in Section 5.2.

5.1 Birth weight and smoking

5.1.1 Model and results from logistic regression

To obtain the results using logistic regression we fit in total five models to the data set, one for each of the cut-off values presented in Table 5.1.

Table 5.1: Quantiles of interest and corresponding cut-off values.

τ	0.05	0.25	0.5	0.75	0.95
cut-off$^{(\tau)}$	2508g	3221g	3579	3929g	4458g

Let the binary response be $y_i^{(\tau)}$, so that,

$$y_i^{(\tau)} = \begin{cases} 1 & \text{if the i-th observation has BW below cut-off}^{(\tau)} \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

The $y_i^{(\tau)}$ is now realizations of a Bernoulli random variable $Y_i^{(\tau)}$ that takes the value 1 (BW below cut-off$^{(\tau)}$) with probability $p_i^{(\tau)}$ and 0 (BW above cut-off$^{(\tau)}$) with probability $(1-p_i^{(\tau)})$. By logistic regression we model the probability of BW below cut-off$^{(\tau)}$ through the logit link,

$$\text{logit}(p_i^{(\tau)}) = \hat{\beta}_0^{(\tau)} + \hat{\beta}_1^{(\tau)} x_i \tag{5.2}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated regression coefficients, and x_i is a factor, taking the values

$$x_i = \begin{cases} 1 & \text{if the mother is smoking} \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

First, we consider the special case of interest, the odds and the effect of smoking on the odds for low birth weight, and further discuss the results for all cut-off values.

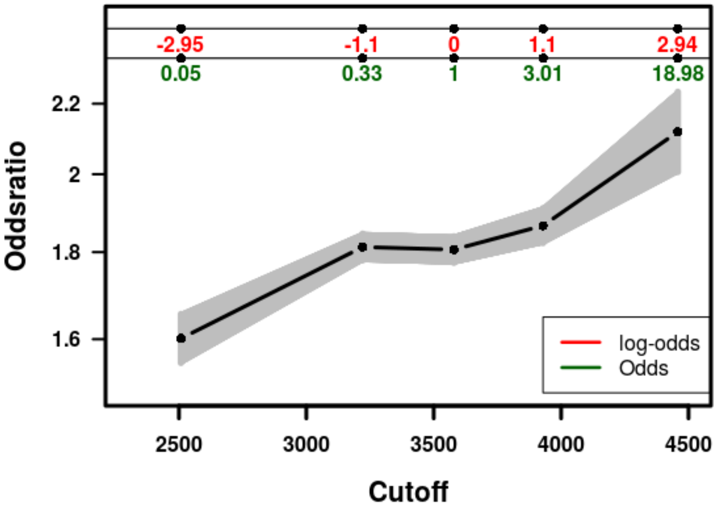

Figure 5.1: Illustration of the odds and the odds ratio in the given birth weights.

Results from fitting the regression models to the data are presented in Figure 5.1. The red numbers denotes $\text{logit}(p_i^{(\tau)}) = \hat{\beta}_0^{(\tau)}$ and the green numbers, obtained by Equation (3.15), denotes the odds for birth weight below the cut-off value found on the x-axis, for a non-smoking mother. The black curve shows the odds ratio obtained for each of the cut-off values. The odds ratio, as described in Section 3.1.3, is in this case the effect of smoking on the odds for birth weight below the cut-off value. The curve is plotted on a log-scale, meaning that $\hat{\beta}_1^{(\tau)}$ is plotted, but we choose to express the y-axis in terms of the odds ratio, for easier interpretation. The 95%-confidence interval, obtained by Equation (3.30), is indicated in gray. The results are also presented in Table 5.2. By the hypothesis test presented in Section 3.1.5 the z-value and p-value in the table are obtained. These values imply that smoking has a significant effect at all chosen cut-off values.

We are especially interested in modelling the odds of a child being born with low birth weight, and if maternal smoking during pregnancy has any effect on the odds for low birth weight. The cut-off value of 2508 g is approximately equal to the definition of low birth weight, and we will use this to draw conclusions about the odds for low birth weight obtained by these models. To calculate the odds and the odds ratio we use Equation (3.15). The odds for low birth weight is found to be 0.05 for a non-smoking mother, whereas the odds for low birth weight when the mother is smoking is found to be 0.083. To consider the effect of smoking during pregnancy on the odds, we now turn to the odds ratio, which is 1.6. This value is larger than 1, implying that the odds for low birth weight increases by smoking, which we also notice by the higher value for the odds, 0.083, when the mother is smoking.

From knowing that the odds ratio is larger than 1, and thus a higher value for the odds is obtained, we conclude that it is more likely that a child is born with low birth weight if the mother is smoking. This can also be found by using Equation (3.16) to go from the logit scale to probabilities. The effect on logit scale alone cannot be transformed to an effect on probability scale, so the original probability needs to be taken into account. The logit for the underlying probability for low birth weight for a non-smoking mother is found to be -2.9457 that translates to a probability of 0.05. When the mother is smoking the probability of low birth weight is found to be 0.078, yielding an effect on probability scale of 2.8 percentage points, that is known as the risk difference (RD). In addition to these interpretations for the effect of smoking we recall that, when modelling the lower part of the distribution it is possible to express this result in terms of relative risk (RR). We can then express the odds ratio (OR) of 1.6 as 60% higher odds for low birth weight for smokers.

In addition to considering the effect of smoking on the odds for low birth weight, we are also interested in the effect of smoking for other values of the cut-off. The results from the regression analyses show that the odds for birth weight below the given cut-off value, for a non-smoking mother, increase as this cut-off value increases, which of course is natural. The odds ratio is larger than 1 for all cases, implying that the odds of birth weight below the cut-off value, in general, increases when the mother is smoking during the pregnancy. At first glance, the curve showing the odds ratio may give the impression that the effect of smoking is larger in the upper part of the distribution at the cut-off value of 4458 g, as the value for the odds ratio of 2.12 is the largest. However, by recalling from Section 3.1.3 that a constant effect in the logit scale will translate to varying effects on the probability scale, we understand that we are not able to compare effects by considering the odds ratios without taking the original probability into account. For the case of low birth weight we considered the cut-off value of 2508 g and found the effect of smoking to be 2.8 percentage points in terms of probabilities. Equivalent calculations considering the cut-off value at 4458 g yields an effect of smoking of 2.5 percentage points. This implies that the effect of smoking is in fact found to be larger at 2508 g.

(Notice that the only estimate that appears not to be significant from considering the p-values is the intercept value for cut-off 3579. This is reasonable since the value for the logit in this model is close to zero implying even odds, and it is about as likely with a birth weight above the 3579 as below 3579).

Table 5.2: Results from logistic regression, the logit values. The 95%-confidence interval is shown i parantheses below each estimate of $\hat{\beta}_1$.

cut-off		Value	Std. Error	z value	Pr(> |z|)
2508	$\hat{\beta}_0$	-2.9457	0.0068	-431.6	< 2e-16
	$\hat{\beta}_1$	0.4710 (0.438, 0.504)	0.0168	28.0	< 2e-16
3221	$\hat{\beta}_0$	-1.0981	0.0034	-319.85	< 2e-16
	$\hat{\beta}_1$	0.5946 (0.577, 0.613)	0.0092	64.92	< 2e-16
3579	$\hat{\beta}_0$	-0.0027	0.0030	-0.898	0.369
	$\hat{\beta}_1$	0.5912 (0.573, 0.609)	0.0091	65.051	< 2e-16
3929	$\hat{\beta}_0$	1.1008	0.0034	320.42	< 2e-16
	$\hat{\beta}_1$	0.6236 (0.600, 0.647)	0.0120	52.03	< 2e-16
4458	$\hat{\beta}_0$	2.9433	0.0068	431.7	< 2e-16
	$\hat{\beta}_1$	0.7504 (0.696, 0.804)	0.0276	27.2	< 2e-16

5.1.2 Model and results from quantile regression

We now turn to quantile regression, and the results obtained from fitting these models. Let τ take the values presented in Table 5.1 and x_i be defined by Equation (5.3). By quantile regression we model the birth weight depending on τ on the form of equation (3.40),

$$Q_{Y_i}(\tau|x_i) = \hat{\beta}_0(\tau) + \hat{\beta}_1(\tau)x_i. \qquad (5.4)$$

From this model we obtain an estimate of the birth weight for the τ-th quantile when the mother is not smoking, given by $\hat{\beta}_0(\tau)$. When the the mother is smoking, the estimate for the birth weight is $\hat{\beta}_0(\tau) + \hat{\beta}_1(\tau)$. The birth weight corresponding to the τ-th quantile changes, and $\hat{\beta}_1(\tau)$ denotes an estimate of this change, and what we consider as the effect of smoking on the birth weight, or the quantile effect.

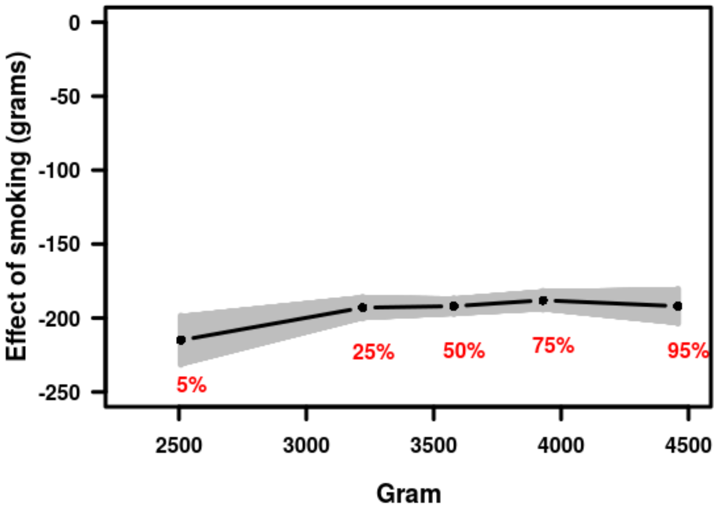

Figure 5.2: Results obtained from quantile regression.

Results from the quantile regression analysis are presented in Figure 5.2. The black curve represents the effect of smoking, $\hat{\beta}_1(\tau)$, plotted against the birth weight corresponding on the τ-th quantile for a non-smoking mother, $\hat{\beta}_0(\tau)$. The red percentages denote the quantiles of interest, τ, and the 95%-confidence interval for the estimates is indicated in gray. The effect of smoking on birth weight is found to be negative in all the quantiles of interest. This implies that the birth weight, in general, is less when the mother has been smoking during the pregnancy, than for a non-smoking mother.

The estimated parameters are also presented in Table 5.3, together with the standard errors and SD-based 95%-confidence interval, as presented in Section 3.2.5. From these results we conclude that smoking during pregnancy has a significant effect on birth weight for all quantiles.

In addition to considering whether the effect is significant, it is also interesting to know if the effect is constant over all quantiles of interest. From the curve in Figure 5.2 we observe an approximately constant negative effect around -190 grams for the 25%, 50%, 75% and 95% quantiles, and a slightly larger negative effect in the 5%-quantile. By testing for equal slopes in each of the models, by using the anova-function, we find a p-value of 0.052. From this result we may conclude that at a 0.1-significance level we reject the hypothesis that the effect is constant over all quantiles, which supports the observations on the curve, that the negative effect of smoking is slightly larger for the 5%-quantile.

A special case of interest is the 5% quantile since the corresponding weight for the non-smokers is approximately equal to the definition of low birth weight, 2500g. For the 5%-quantile the estimated effect is -215 g; this is a slightly larger negative effect on the birth weight than for the other quantiles. These results imply that the babies that are already small and with low birth weight, get even more affected by smoking during pregnancy.

Table 5.3: Results from quantile regression.

τ		Value	Std. Error	t value	Pr(>|t|)
0.05	$\hat{\beta}_0$	2508.000 (2501.7, 2514.3)	3.237	774.764	0.000
	$\hat{\beta}_1$	-215.000 (-232.3, -197.8)	8.842	-24.317	0.000
0.25	$\hat{\beta}_0$	3221.000 (3218.8, 3223.2)	1.129	2853.690	0.000
	$\hat{\beta}_1$	-193.000 (-199.5, -186.5)	3.319	-58.152	0.000
0.5	$\hat{\beta}_0$	3579.000 (3577.0, 3581.0)	1.027	3484.822	0.000
	$\hat{\beta}_1$	-192.000 (-197.1, -186.9)	2.588	-74.201	0.000
0.75	$\hat{\beta}_0$	3929.000 (3926.8, 3931.2)	1.140	3446.287	0.000
	$\hat{\beta}_1$	-188.000 (-193.7, -182.3)	2.920	-64.383	0.000
0.95	$\hat{\beta}_0$	4458.000 (4454.8, 4461.2)	1.616	2758.620	0.000
	$\hat{\beta}_1$	-192.000 (-203.5, -180.5)	5.861	-32.757	0.000

The fact that the value for the birth weight at the 5%-quantile for smoking mothers is 215 g less than for non-smokers would then lead to a higher proportion of births defined as low birth weight for the smokers, and this leads us to the next section where we discuss the interpretation of the quantile regression model and the logistic regression model and how the results are related.

5.1.3 Interpretation of the models, with focus on low birth weight

In Section 5.1.1 the special case of interest was the effect of smoking on the odds for low birth weight. We obtained an estimate for the odds ratio that was larger than 1, implying that low birth weight is more likely for a smoking mother than for a non-smoking mother. This is consistent with the results from the previous section, 5.1.2, where a negative effect on the birth weight at the 5%-quantile was found. Even if these effects are on different scales and appear very different, they are closely related. A factor that has a negative effect on the quantile value, would lead to more observations falling below the original cut-off value, and consequently lead to a higher proportion of observations below the cut-off, yielding higher odds, and an odds ratio larger than 1. The fact that the results from quantile regression are in gram scale, the same scale as the weight, it is by simple interpretation possible to draw conclusions on how much the birth weight is directly affected.

The results from fitting logistic regression models to the data presented a higher odds ratio for the cut-off value 4458 g than for 2508 g, but the effect of smoking was still found to be larger for 2508 g as cut-off. When fitting the quantile regression models we found that the effect could be said to have the largest negative value at the 5%-quantile, hence these results are consistent. When it comes to examining how the effects of a factor may vary between different parts of the distribution, quantile regression yields an easier interpretation. As the scale of the effect is directly on the birth weight, it gives a basis for such comparisons, whereas logistic regression may be misunderstood, if not carefully interpreted.

5.2 Birth weight and mother's age

Similar analyses as presented in the previous section for the effect of smoking on birth weight has been carried out for the effect of mother's age on birth weight. Due to the changes over time of both spontaneous births and mothers in different age groups, these analyses is performed for both all births and only spontaneous births separately. We use age category 3: 25-29 years as reference category, as the number of births in this age group is the most stable over time, as can be seen in Figure 2.8. Recall Table 2.4 for description of the age categories. Analyses using both logistic regression and quantile regression have been carried out with all births included, and further equivalent analyses on a data set only including the spontaneous births. The cut-off values used for the logistic regression models are presented with the corresponding quantile values relative to age group 3 for all births in Table 5.4, and for spontaneous births in Table 5.5. The data set containing all briths appart from year 2009 consist of 2 451 197 births, and the data set only including spontaneous birth consist of 2 034 889 births.

Table 5.4: Quantiles of interest and corresponding cut-off values for all births.

τ	0.05	0.25	0.5	0.75	0.95
cut-off$^{(\tau)}$	2522g	3197g	3545	3888g	4403g

Table 5.5: Quantiles of interest and corresponding cut-off values for spontaneous births.

τ	0.05	0.25	0.5	0.75	0.95
cut-off$^{(\tau)}$	2615g	3214g	3550	3882g	4384g

The results from logistic regression are presented in the next section, 5.2.1, and in Section 5.2.2 we present the results obtained by quantile regression.

5.2.1 Model and results from logistic regression

We define an index variable, $Y_i^{(\tau)}$ as in Section 5.1.1, by Equation (5.1). The probability of a birth weight below each cuf-off value is modelled through the logit link,

$$\text{logit}(p_i^{(\tau)}) = \boldsymbol{x_i}^T \hat{\boldsymbol{\beta}}^{(\boldsymbol{\tau})}, \tag{5.5}$$

where $\hat{\boldsymbol{\beta}}^{(\boldsymbol{\tau})}$ is a column vector of the estimated regression coefficients and $\boldsymbol{x_i}^T$ a row vector of zeros and ones, depending on what age group the $i'th$ observation belongs to. The first element in this vector is always 1, denoting the reference group, age group 3.

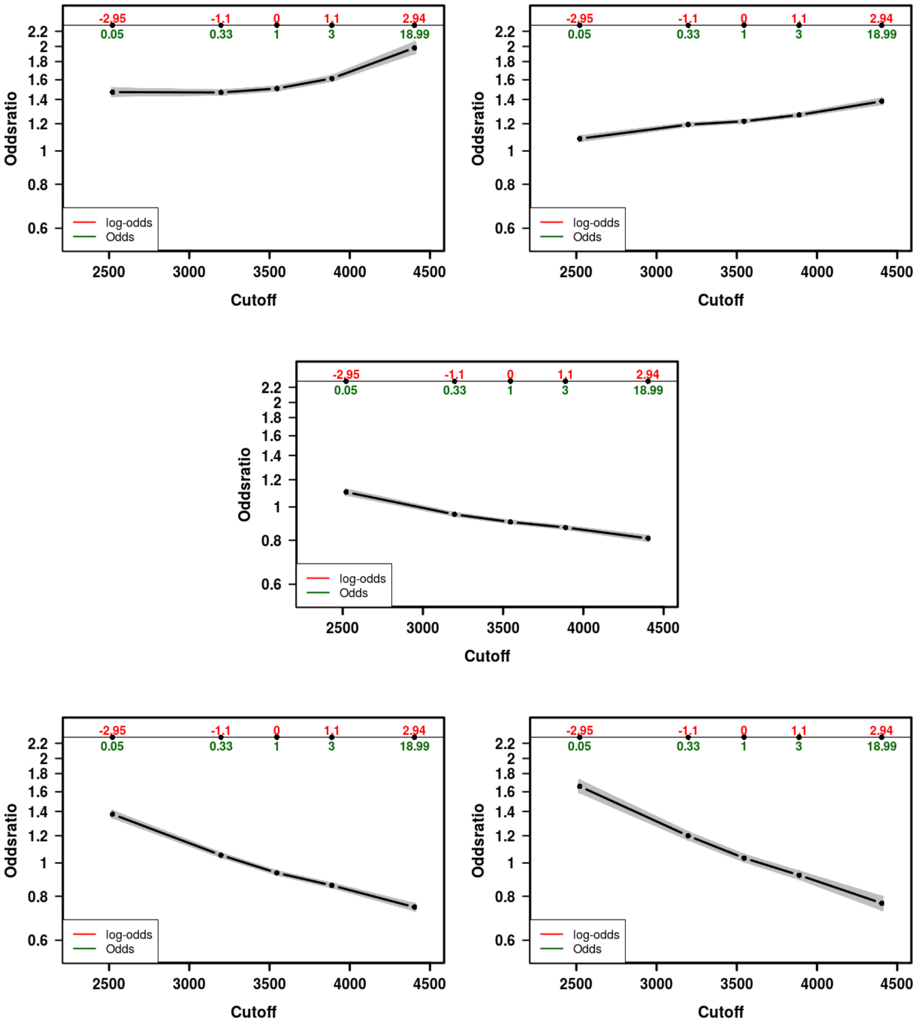

Figure 5.3: Results from logistic regression all births.

The first set of analyses of the effect of mother's age on birth weight fit logistic regression models, one for each cut-off value, to the data set containing all births. By the hypothesis test presented in Section 3.1.5 we find that the effect of each of age group on the birth weight, relative to category 3, is significant. This conclusion holds for all five values for the cut-off. Figure 5.3 presents the effect of mother's age on the odds of birth weight below the given cut-off value found on the x-axis. The odds ratio obtained for each age group relative to the reference group is plotted against the cut-off value, and the 95%-confidence interval is indicated in gray. In the upper panel we find the odds ratio curve for age group 1 and 2 relative to group 3. Both curves lie over 1 for all the cut-off values implying that the odds of being below the given cut-off value is increasing when the mother is younger than 25 years.

In the lower panel we find the results for the effect of age group 4, 5 and 6. For these three older age groups, a small cut-off value gives an odds ratio larger than 1 implying higher odds for birth weight below the cut-off value, while a large value for the cut-off yields an odds ratio smaller than 1, and hence lower odds for birth weight below the cut-off values. From these results we also notice that the curve presenting the odds ratio for age group 4 is the curve that is found closest to 1, and hence the age group where the effect is the smallest relative to the reference group. We do not attempt to examine in which of the cut-off values the effect of each age group is the largest as this will be easier to interpret by quantile regression.

A special case of interest is the effect of mother's age on the odds for low birth weight. The cut-off value at 2522 g is close to the definition of low birth weight, and thus it is natural to use this cut-off value to consider the effect of mother's age on low birth weight. The odds ratio is found to be larger than 1 at all ages, implying that for both younger and older mothers the odds for low birth weight increases. Since we are now interested in the effects at one particular cut-off value, it makes sense to compare the odds ratios for the different age groups as the original probability will be equal for all. The odds ratio is found to be largest for mothers at 40 years or older, taking the value 1.65, and second largest for the youngest mothers at 19 years or younger, taking the value 1.48. By expressing the results as relative risks we find that the risk of low birth weight is 65% higher for a mother at age 40 or older than for a mother at age $25 - 29$, and 48% higher for a mother at 19 years or younger.

In the second set of analyses we fit the logistic regression models to the data set only including spontaneous births. The curves denoting the odds ratio for each age group relative to the reference group is shown in Figure 5.4. In the upper panel we find the odds ratio curve for age group 1 and 2 relative to group 3, the results for the effect of age group 4, 5 and 6 is found in the lower panel. We observe many of the same trends from these results as we did from the analyses using all births. For all age groups we observe slightly larger values for the odds ratios than we did in the previous analyses on all births. By construction, since each cut-off value corresponds to a given quantile, the original probability for weight below each of the cut-off values, are equal to the original probability for the corresponding cut-off values in the previous case. This makes it possible to compare the results from the regression analyses on all births to these results for spontaneous births.

For the special case of interest, the odds for low birth weight, we also observe the same trend as before, all value for the odds ratio is larger than one, yielding an increase in the odds by mother's age.

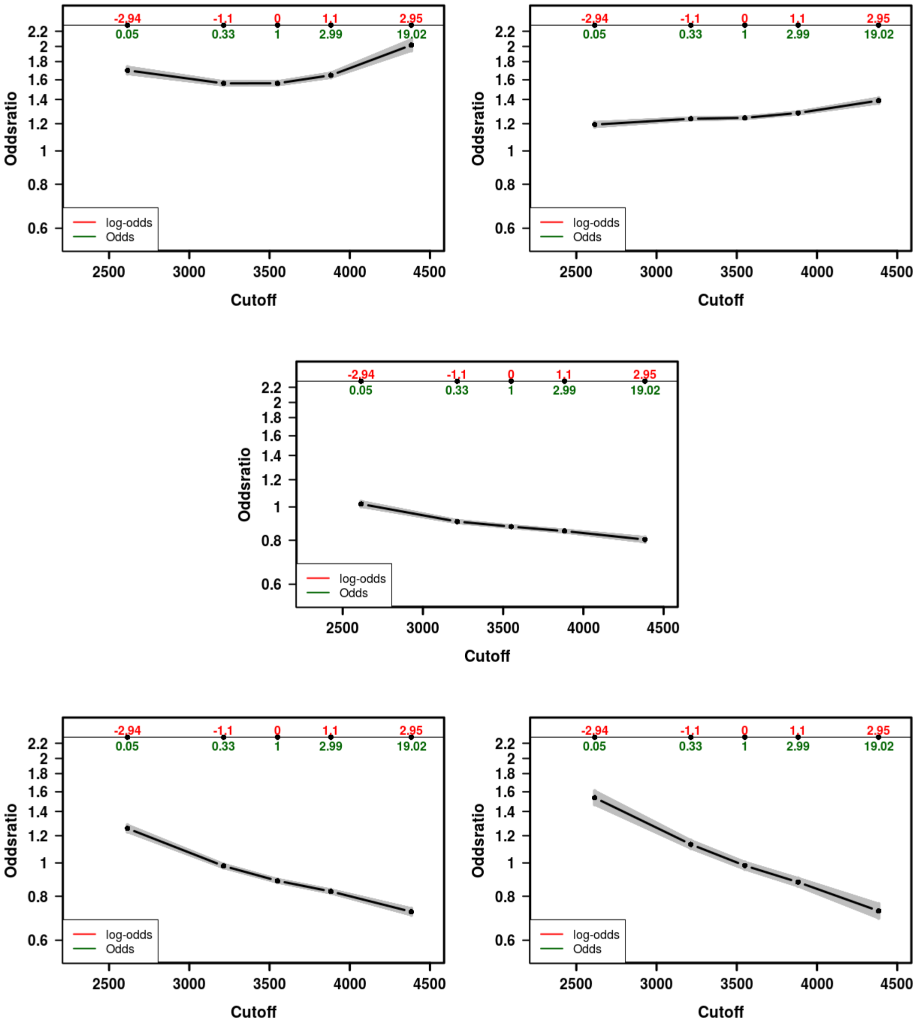

Figure 5.4: Results from logistic regression spontaneous births.

5.2.2 Model and results from quantile regression

In this section we present the model and results from fitting quantile regression models to both the data set containing all births and to the data set only including spontaneous births. The quantiles, τ, of interest was presented in Table 5.5, and we model the birth weight dependent on τ by,

$$Q_{Y_i}(\tau|\boldsymbol{x_i}^T) = \boldsymbol{x_i}^T\hat{\boldsymbol{\beta}}(\tau), \tag{5.6}$$

where $\hat{\boldsymbol{\beta}}(\tau)$ is a column vector of the estimated regression coefficients and ${\boldsymbol{x_i}}^T$ a row vector of zeros and ones, depending on what age group the $i'th$ observation belongs to. The first element in this vector is always 1, denoting the reference group, that is age group 3.
The results from the regression analysis on all births are presented in Figure 5.5. The curves showing the quantile effects for age group 1 and 2 are plotted in the upper panel. Both curves are consistently below 0 for all quantiles, implying that for all the quantiles the mothers younger than 25 years have smaller babies relative to age group 3. This is also consistent with the odds ratio larger than 1 found in the previous section. It is also clear that the negative effect is larger, in all quantiles, for the mothers at age 19 year or younger than for the mothers at age $20-24$ years. The negative effect for the youngest mothers is always larger than 100 g, whereas for age group 2 the effect is found to be less than -100 g in all quantiles.
In the lower panel we find the results for age groups 4, 5 and 6. All three curves are decreasing and crosses zero at some point, implying that the age effect can be both positive and negative depending on the quantile. In general, for these older age groups, we observe that the smallest babies, with weight around the 5%-quantile, get even smaller with mother's age, and the larger babies, with weight around the 95%-quantile, get even larger. This trend is most clear by considering the effect of age group 6, 40 years or older, where the effect on the 5%-quantile is -328 g and the effect on the 95%-quantile is $+74$ g. By the regression output the effect of all age groups is found to be significant effect all quantiles. In addition we find varying effects depending on the quantile.
Considering the special case of interest, the effect of mother's age on low birth weight, we find that mothers belonging to any of the other age group have smaller babies in the 5%-quantile than the mother in the reference group, $25-29$ years. The negative effect on the birth weight in the 5%-quantile is largest for age group 6, for mothers at age 40 or older, and found to be -328 g. The second largest effect of -204 g, is found for the youngest group, mothers who are 19 years or younger.

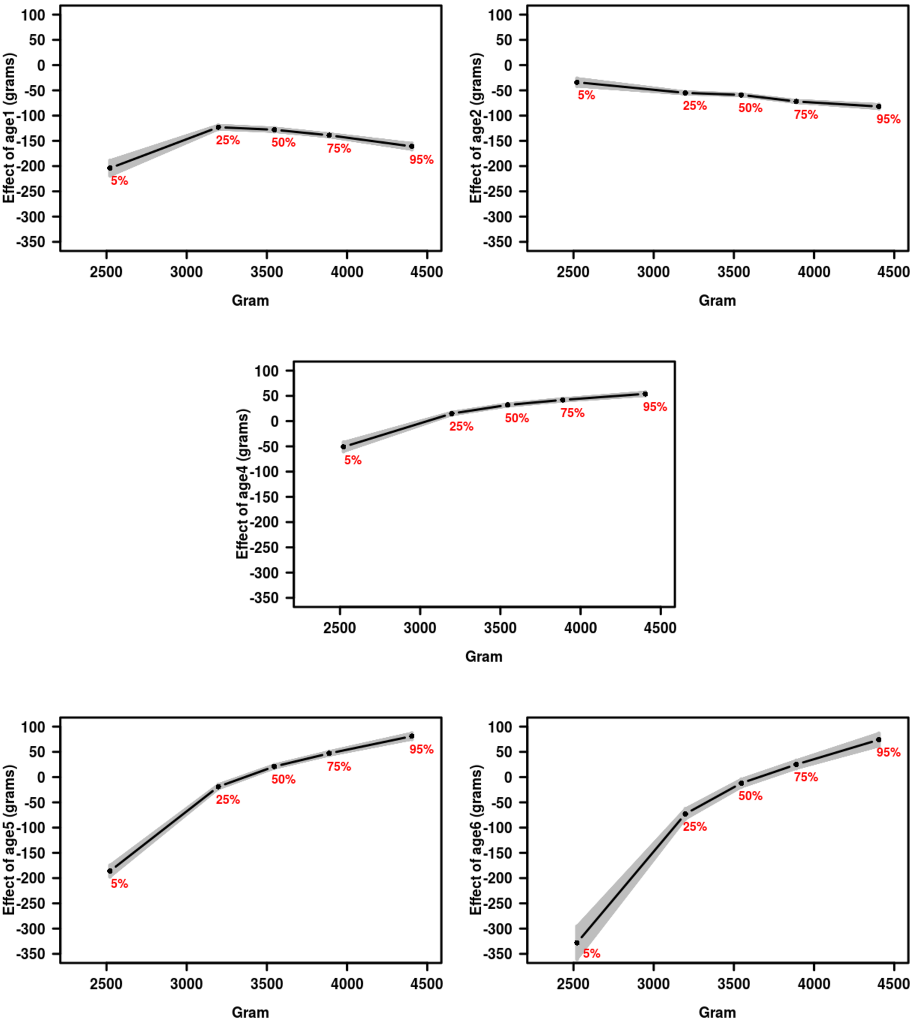

Figure 5.5: Results from quantile regression all births.

Equivalent analyses has been carried out for spontaneous births. These results are shown in Figure 5.6, with the effect of age group 1 and 2 shown in the upper panel, and age group 4, 5 and 6 is shown in the lower panel. We observe the same overall trend as observed for all births, a negative effect on all quantiles for age group 1 and 2, and an increasing curve taking both positive and negative effects in the older age groups, depending on the quantile. The difference between the effects obtained from all births and only spontaneous births is most visible in the lower quantiles. Also in this case all age groups has a significant effect for all quantiles and various effects depending on quantiles.

For age group 1 and 2 we notice that the negative effect for the spontaneous births are not very different from the effect for all births, except in the 5%-quantile where a larger negative effect is observed for the spontaneous births. The opposite trend is observed for age group 4, 5 and 6, where we now can observe a decrease in the negative effect, in the lower quantiles.

When considering the special case of interst, the effect on low birth weight, we notice that the weight corresponding to the 5%-quantile takes a larger value for the spontaneous births than in the previous case when we considered all births. This value of 2615 g is not as close to the definition of low birth weight, but we can still draw some conclusions by considering effects on the 5%-quantile. First of all, this larger value for the birth weight in the 5%-quantile for mothers between $25 - 29$ years implies that less babies are born with low birth weight when the birth is spontaneous. In fact, less than 5% of the births can be considered low birth weight. As mentioned above the opposite trend is observed in the 5%-quantile for the mothers at 19 years or younger and the mothers at 40 years or older. For spontaneous births the negative effect increases from -205 (all births) to -258 for the youngest mothers, and decrease from -328 to -226 for the mothers at 40 or older. This may imply that several of the births that was considered low birth weight may not have been spontaneous. For the youngest mothers this may imply that inducing labor is not especially common for the smallest babies.

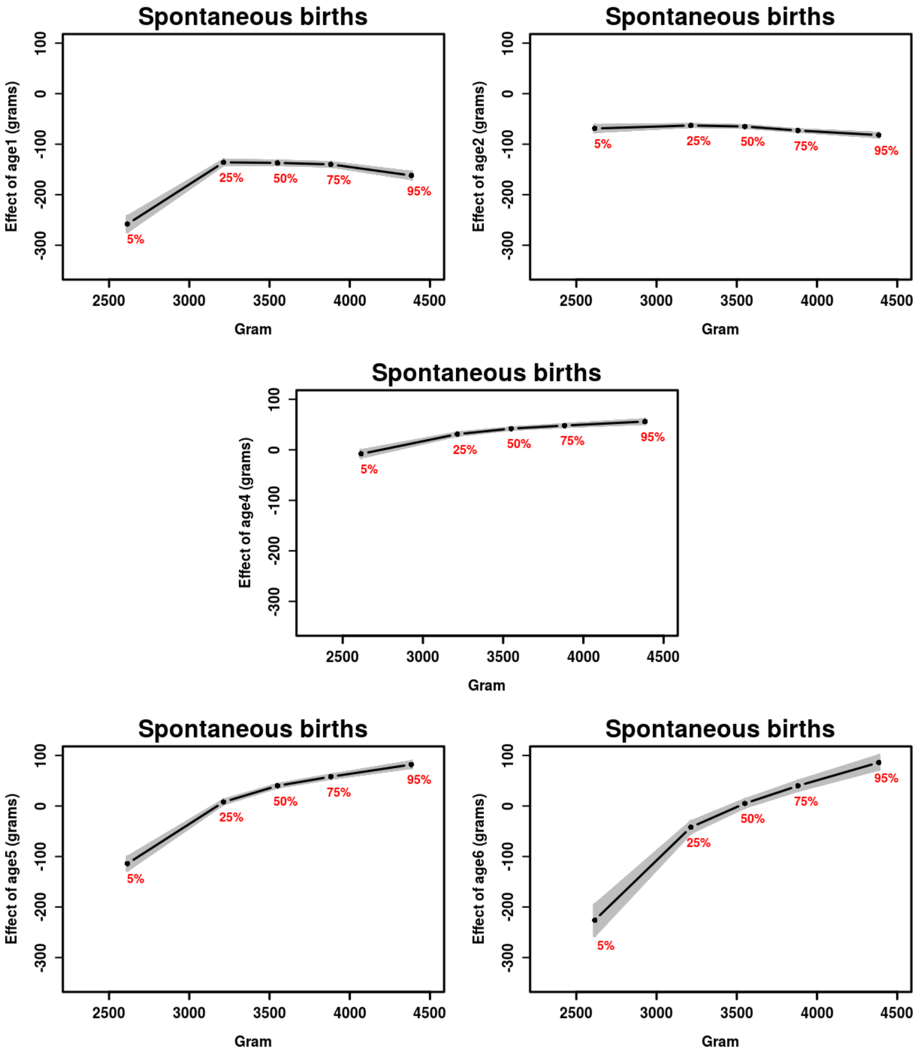

Figure 5.6: Results from quantile regression spontaneous.

5.3 Trend in birth weight

In addition to model the birth weight using smoking habits and mother's age as predictors, another case of interest is to see how the birth weight changes over years. In this section we will discuss how both logistic regression and quantile regression can be applied to illustrate a trend between years. We use the first year available in our data set, 1967, as a reference year, and further the changes that appears in the birth weight, relative to this year, will be illustrated. Several figures will be presented where the odds ratio, mean and quantile values will be considered. The results will be presented for both the data set including all births, and the data set where only spontaneous births are included.

The first example is shown in Figure 5.7. The black curve, with corresponding axis on the left hand side, indicates the gain in the mean birth weight in each year relative to the reference year, 1967. From 1967 until 1975 we observe a gain in mean birth weight close to zero which implies that there are only small changes in the mean birth weight in these years relative to 1967. From 1975 until 1981 we observe an increase in mean birth weight, with a gain around 50 g. Further, the largest gain is found in 1999 where we observe a gain of 75 g.

The red curve, with corresponding axis on the right hand side, indicates the odds ratio (OR) for low birth weight in each year relative to 1967. In 1967 the OR for low birth weight is, naturally, equal to 1. Further we observe a decrease in the OR until the smallest value, an OR of around 0.85,is reached in 1979. As this value is smaller than 1, the odds of low birth weight was higher in 1967 than in 1975, and fewer babies had low birth weight in 1979. Through the following years the OR increases, but remains close to 1 from around 1991 until 2009.

When comparing these two curves, we notice that it is not clear how to expect the OR curve to look like based on the gain in mean birth weight curve. Through the first years, from 1967 until 1991, it seems like a gain in the mean birth weight can be reflected as a decrease in the OR for low birth weight. On the other hand, when we observed an increase in the years from 1991 until 2005 the OR still takes values around 1, implying that the odds for low birth weight does not change much relative to 1967 even in there is an observed gain in the mean birth weight.

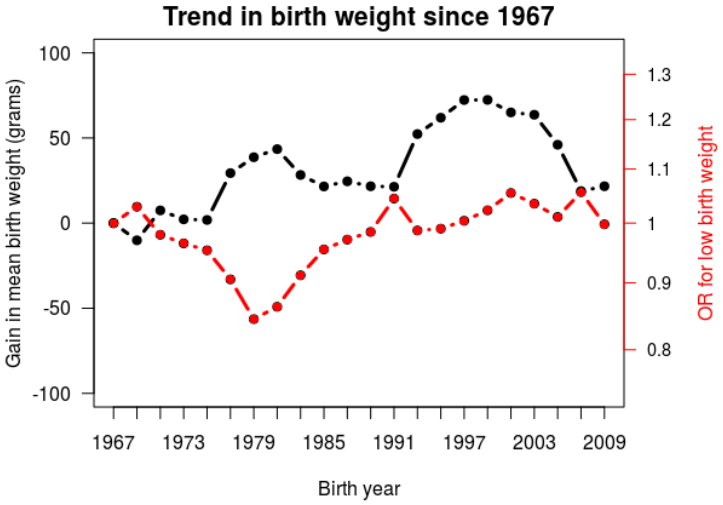

Figure 5.7: Illustration of the gain in mean birth weight and the OR for low birth weight, for all births. x-axis: years. Axis to the left: gain in mean birth weight in gram scale. Axis to the right: the OR, plotted on a log-scale.

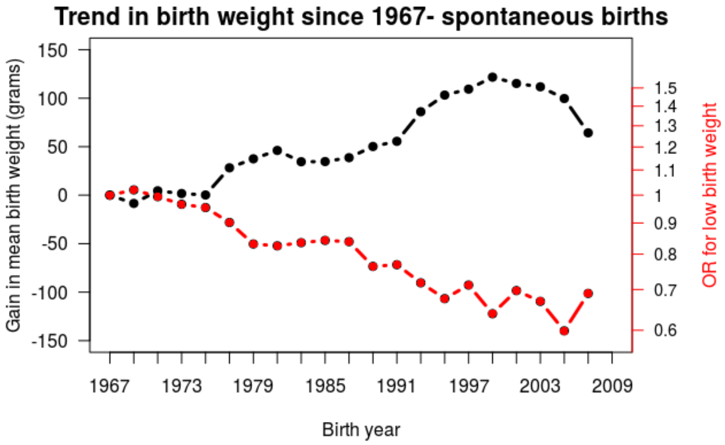

Figure 5.8: Illustration of the gain in mean birth weight and the OR for low birth weight, for spontaneous births. x-axis: years. Axis to the left: gain in mean birth weight in gram scale. Axis to the right: the OR, plotted on a log-scale.

A different illustration of change in birth weight over years is shown in Figure 5.9. The figure shows the gain in the birth weight corresponding to three different quantiles for each year relative to 1967. The black curve denotes the 5% quantile, the red curve the 50% quantile and the green curve the 95% quantile. From 1979 the gain in the birth weight corresponding to the 95%-quantile keeps increasing and reaches its maximum value in 1999, whereas the birth weight corresponding to the 5%-quantile keeps decreasing over the same years. This implies that the smallest babies are getting smaller and the heavier babies are getting larger.

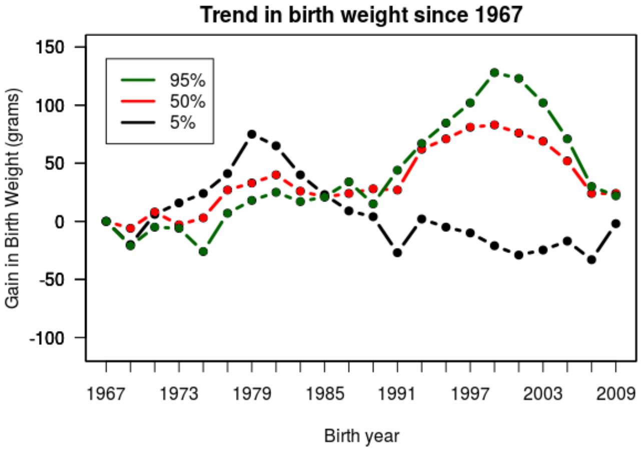

Figure 5.9: Illustration of the gain in birth weight at the corresponding quantiles, for all births.

Figure 5.10 shows a similar plot of how the odds for birth weight below given cut-off values changes over years. These cut-off values corresponds to the quantiles 5%, 50% and 95%, and are shown in Table 5.6. The colours of the curves is presented in the figure, these are the same as in the previous figure.

Table 5.6: Quantiles of interest and corresponding cut-off values, all births.

τ	0.05	0.5	0.95
cut-off	2468g	3530g	4399

Table 5.7: Quantiles of interest and corresponding cut-off values, spontaneous births.

τ	0.05	0.5	0.95
cut-off	2565g	3535g	4378

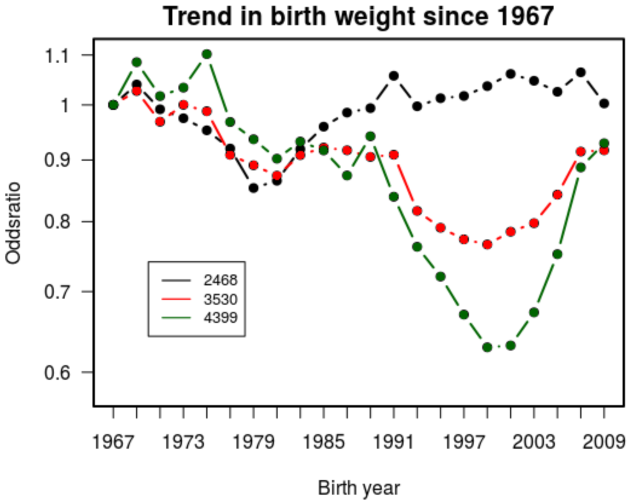

Figure 5.10: Illustration of the odds ratio of being below the birth weight corresponding to the quantiles 5%, 50% and 95%. All births included.

The two figures 5.9 and 5.10 show how the curves reflect each other. As the black curve increase in figure 5.9 it decrease through the same period in 5.10 and opposite. For the green curve we see this trend clearly. As the odds ratio decreases (relative to 1967), the gain in the weight that corresponds to the 95% quantile increases (relative to 1967). This basically means that as the weight that corresponds to the 95% quantile increases, more of the birth weights will be found above the original weight that corresponds to the 95% quantile in 1967, and therefore, the odds for being below this weight will be smaller than what it was in 1967, hence an odds ratio smaller than 1.

In Figure 5.11 and Figure 5.12, similar results has been carried out for the data set including only spontaneous births to see how the two methods integrates with the censored

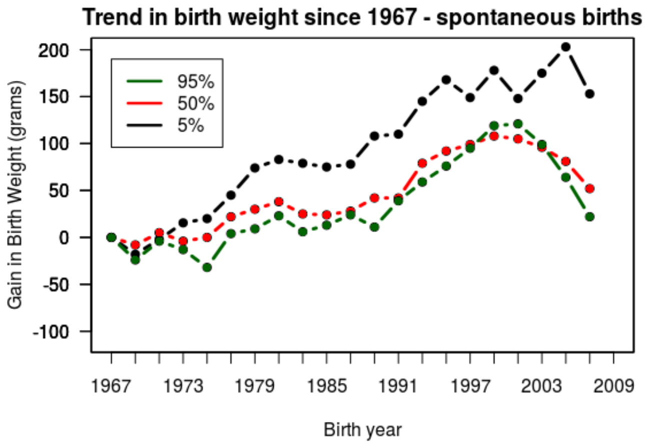

Figure 5.11: Illustration of the gain in birth weight at the corresponding quantiles, for spontaneous births.

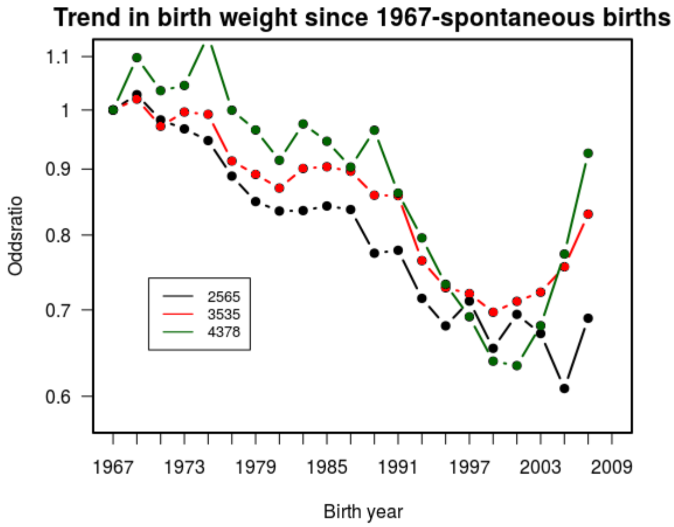

Figure 5.12: Illustration of the odds ratio of being below the birth weight corresponding to the quantiles 5%, 50% and 95%. Only spontaneous births included.

data. From these results a steady increase in the birth weight corresponding to all three quantiles is observed through the years, relative to 1967. The curve denoting the 5%-quantile has the largest gain through times, that may indicate that many of the births at low birth weight may not have been spontaneous, and that inducing labor may have become more common for these cases over years. The green and the red curve are not as different from the green and the red curve showing the trend for all births. The trend observed in

Figure 5.11 for the birth weight in the quantiles is also reflected in the odds ratio for birth weight below the given cut-off value in Figure 5.12. The green and the red curve do not change as much when only considering spontaneous births, but the black curve shows a decrease in the odds ratio for a birth weight below 2565 g for spontaneous births.

Chapter 6

Discussion and conclusion

Through this thesis we have studied the use of both logistic regression and quantile regression. When other parts of the distribution than central location are of interest, the logistic regression model for binary outcome is frequently used for continuous outcome variables. The logistic regression model then allows us to model any part of the distribution by choosing the cut-off value of interest. From the analyses on known distributions in Chapter 4 we studied how location, scale and location-scale shifts affected the regression analyses, and we found that this was reflected differently depending on the shape of the distribution for the results from logistic regression. Hence the odds ratio is affected by both the original probability and the shape of the distribution. For the logistic regression model an important issue is on what scale the results should be presented. This was discussed further in Chapter 5, concerning the case study on birth weight where the difference between OR, RR and RD was discussed. If one is not aware of what scale the result are presented in, it may lead to misinterpretation. We saw that by only studying the odds ratio we were not able to correctly compare the estimated effect of smoking and mother's age over the entire distribution.

In addition to being known to be robust and handle heteroscedasticity, quantile regression yields interpretation on the same scale as the distribution being modelled, like standard linear regression. This simplifies the interpretation, and provides easier comparison of effects over the entire distribution directly. In Chapter 4 it was shown that the results from quantile regression appeared similar for the two different distribution by location, scale and location-scale shifts.

An approximate relation between the two approaches was derived in Section 3.3, and we find that they are closely related even if they appear different. This was especially shown in Chapter 5 when we considered the trend in birth weight over years. When comparing the curves of the change in the given quantiles with the corresponding OR we found that they expressed the same trend, however quantile regression gave results that are easier to interpret. Due to this we see in what way quantile regression can yield a useful contribution to the analyses. From this we suggest that quantile regression should definitely be applied more often for continuous outcome variables, that one usually solve by defining a cut-off value and applying logistic regression. However, it need to be mentioned that to be able to use quantile regression, the outcome need to be continuous. For outcomes that are truly dichotomous it is not an option to use quantile regression.

Bibliography

Cade, B. S., Noon, B. R., 2003. A gentle introduction to quantile regression for ecologists. Front Ecol Environ 1.

Casella, G., Berger, R. L., 2002. Statistical Ínference. Brooks/Cole, Cenage Learning.

Folkehelseinstituttet, 2015. Fødselsvekt i norge.
URL https://www.fhi.no/hn/statistikk/statistikk3/fodselsvekt-i-norge-faktaark/

He, X., Hu, F., 2002. Markov chain marginal bootstrap. Journal of the American Statistical Association 97:459.

Kirkwood, B. R., Sterne, J. A., 2013. Essential Medical Statistics, 2nd Edition. Blackwell Science Ltd.

Kocherginsky, M., He, X., Mu, Y., 2005. Practical confidence intervals for regression quantiles. Journal of Computational and Graphic Statistics 14:1.

Koenker, R., 2005. Quantile Regression. Cambridge University Press.

Koenker, R., 2015. Quantile regression in r: A vignette.
URL http://cran.r-project.org

Koenker, R., Bassett, G., 1978. Regression quantiles. Econometrica 46 (1), 33–50.

Larsen, R. J., Marx, M. L., 2012. An introduction to mathematical statistics and its applications. Pearson Education.

Rodriguez, G., 2013. Lecture notes in generalized linear models.
URL http://data.princeton.edu/wws509

Schmidt, C. O., Kohlmann, T., 2008. When to use odds ratio or the relative risk? Int J Public Health 53, 165–167.

Wardlaw, T., Organization, W. H., UNICEF., 2004. Low Birthweight: Country, Regional and Global Estimates. UNICEF, Editorial and Publications Section.

Wei, Y., Pere, A., Koenker, R., He, X., 2006. Quantile regression methods for reference growth charts. Statistics in Medicine 25.

Yu, K., Lu, Z., Stander, J., 2003. Quantile regression: applications and current research areas. The Statistician 52.

Appendix

The analyses are done in R-studio. Some of the code is found in this Appendix.

```
load("~/Documents/Master /d.Sol.RData")
library(MASS)
library(quantreg)

#A vector of birth weight
vectorBw<-(d.Sol$bw)
# Check observations with negative value
neg <- subset(d.Sol, bw<0, select=c(c.year, bw,
    smoke.end, m.age.c6, spontaneous))
#set negative birth weight to NA
#(will exclude this observation)
vectorBw[vectorBw ==-1] <-NA
d.Sol$bw <- vectorBw

 #Quantile regression in R:

 #Assign smoke 1 and 0
.bw <- d.Sol$bw
.smoke <- d.Sol$smoke.end
.smoke[.smoke %in% 1:2] <- 0
.smoke[.smoke %in% 3] <- 1
.tmpd <- data.frame(bw = .bw, smoke = .smoke)
# Remove NA
.tmpd <- na.omit(.tmpd)

## QUANTILE REGRESSION
# quantiles of interest
tau <- c(0.05, 0.25, 0.5, 0.75, 0.95)
.res.rq <- rq(bw ~ smoke, tau = tau,
            data = .tmpd, method = "fn")
#Regression output using the mcmb-method
.sum <- summary(.res.rq, se="boot", bsmethod = "mcmb", R=200)
```

```
# Logistic regression in R:
cutOff <- tab.int$Value  #set the cut-off values equal
    # to the intercept from quantile regression

for (i in seq(along=cutOff) ){
.tmpd$cutbw <- (.tmpd$bw < cutOff[i]) + 0
.res.glmCut <- glm(cutbw ˜ smoke, family = binomial(link = "logit"),
                data = .tmpd)
# Regression output
sumCut <- summary(.res.glmCut)

###################################

# Example from the synthetic case study -
# generating normally distributed data
# location-scale shift
    n1 <- 452377
    n2 <- 59070
    mean1 <- 3539
    mean2 <- 3539 -193
    sd1 <- 635
    sd2 <- 635 +100
    NsumNoSmoke <- n1
    NsumSmoke <- n2
set.seed()
NormDataNS <- data.frame(BirthW = rnorm(n=n1, mean=mean1, sd=sd1),
  smokE =rep(0, NsumNoSmoke) ,class="Non-smokers")
NormDataS <- data.frame(BirthW = rnorm(n=n2, mean=mean2, sd=sd2),
  smokE=rep(1, NsumSmoke), class="Smokers")

NormData <- rbind(NormDataNS, NormDataS)
#NormData
NormDataVec <- (NormData$BirthW)
```