Stockholms universitet

Comparison of the Rank-Ordered Logit and Between-Within Regression Models

Marielle Andersson

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2015:7
Matematisk statistik
September 2015

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematiska institutionen

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2015:6**
http://www.math.su.se

Comparison of the Rank-Ordered Logit and Between-Within Regression Models

Marielle Andersson*

September 2015

Abstract

When we have epidemiological data and want to investigate the association between an outcome and an explanatory variable we need to adjust for potential confounders that otherwise can cause a statistically significant association between these factors. In this thesis we compare two regression models that can be applied in order to adjust for confounders when the outcome is continuous by dividing the population into clusters, namely the between-within and the rank-ordered logit model. The between-within model is an extension of the generalized linear model where we include the cluster specific mean of the explanatory variable as an additional covariate. We thereby divide the regression into a within- and a between-effect, where the within-effect is not affected by the confounders shared within a cluster. The rank-ordered logit model assumes that the unknown information has a so called extreme value type I distribution and ranks the outcomes within a cluster. The resulting log likelihood function is equivalent to the likelihood of a stratified Cox proportional hazards model, where all shared confounders within a cluster are matched away. We compare these two models in a simulation study and also apply them on two datasets. The first dataset contains information about blood glucose measurement from the National University Hospital in Singapore and we study the association between the variation in blood glucose level and the mean daily measurement frequency, and we conclude that there is a statistically significant positive association. The second dataset is from the Karolinska mammography project for risk prediction of breast cancer (KAR, 2015), where we analyse risk factors for mammographic density. We focus on post menopausal women and compare the results from the analysis of pairs of unrelated women and pairs of sisters. We find that it is important to adjust for age and body mass index, but not important to adjust for confounding by shared childhood environment and genetic factors. We conclude that whether or not a woman has had hormone replacement therapy or a history of benign breast disease are associated with percent dense volume which is a measure of mammographic density. Also, the age at first birth is associated with mammographic density. We conclude that the between-within model is preferable to the rank-ordered logit model. The advantage of these two models is that we can adjust for unmeasurable confounders. However, both models are biased if the confounders are not completely shared within a cluster and should therefore be applied with caution.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: maan0129@gmail.com. Supervisor: Pieter Trapman.

Acknowledgement

I want to express my gratitude to my supervisor professor Marie Reilly and co-supervisor Nathalie Støer at the Department of Medical Epidemiology and Biostatistics at Karolinska Institutet, thank you for introducing me to this project and for all your advice and guidance throughout the working process of this thesis. I also want to thank my supervisor Pieter Trapman at the Department of Mathematics at Stockholm University, for the feedback, help and support during the writing of this thesis.

Marielle Andersson,
Gothenburg, September 6, 2015

Contents

Chapter 1

Introduction

In the field of epidemiology one is often interested in the association between a certain explanatory variable and a particular outcome. It is common to collect data from observational studies and apply different regression models in order to investigate these relationships. In this thesis we focus on continuous outcomes. If the observations are independent of each other and the explanatory variable we are investigating is associated with the outcome without any influence from other factors, i.e. $X \rightarrow Y$, we can analyse the data with simple regression models. However, in real applications the relationship between the explanatory variable and the outcome is often influenced by several confounding variables. A confounding variable is a factor that is associated with both the outcome and the explanatory variable, which may be, at least partly, the reason for a statistically significant association between the outcome and covariate (Breslow and Day, 1980). This is illustrated in Figure 1.1 where the confounding variable is denoted by C, the outcome by Y and the explanatory variable by X. It is important to adjust for

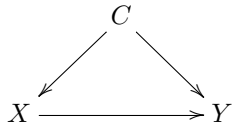


Figure 1.1: The association of the confounding variable, C, on the outcome, Y, and the explanatory variable, X.

these confounders when we study the association between X and Y. There are several ways to adjust for confounders and it can be done either in the design phase of the study or in the analysis phase. In the design phase we can restrict the study population to individuals with certain specific characteristics, thereby obtaining a sample where the investigated association is not influenced by these characteristics (Fletcher and Fletcher, 2005). Another approach is to match the individuals included in the design: for example if the explanatory variable is a binary exposure, we could match the individuals on age and include one exposed and one unexposed individual with the same age in the study. In the analysis phase we can also use stratification, i.e. divide the sample into different strata with respect to some characteristics (Fletcher and Fletcher, 2005).

In the between-within and the rank-ordered logit models we adjust for the confounders by dividing the population into different clusters, where the individuals within a cluster are similar with respect to the confounders we try to adjust for. If the confounders are measurable we use stratification. Since there are potential confounders that are unmeasurable, a useful study design is to use a sample that consists of pairs of siblings, since they share several potential confounders such as maternal factors (Carlin et al., 2005) and family socio-economic status (Begg and Parides, 2003).

There are different methods available to adjust for the confounders that are shared within a cluster. For epidemiological data with continuous outcome a commonly used approach is the between-within model. This is a generalized linear (mixed) model where we include the explanatory variable and the average of the explanatory variable within a cluster. Thus the effect of the explanatory variable is divided into a between- and a within-effect. The within-effect is adjusted for all confounders that are shared within a cluster and is subject specific, while the between-effect is the effect between the clusters (Neuhaus and Kalbfleisch, 1998).

Another method that is used in econometrics literature, but currently is not a common analysis approach for epidemiological data, is the rank-ordered logit model. In this model we assume that we have a linear association between the explanatory variable and the outcome, where the error terms of the model are independent and have a so called extreme value type I distribution. By ranking the outcomes within a cluster in decreasing order the likelihood of the ranked data is equivalent to that of the stratified Cox proportional hazards model. Each cluster corresponds to a stratum and all shared confounders are matched away.

The aim of this thesis is to compare these two regression models in a simulation study with regard to bias, coverage, precision and power. We also apply these models on two different datasets.

The first dataset is collected at National University Hospital in Singapore and contains measurements of blood glucose levels for patients on capillary blood glucose monitoring. We have blood glucose measurements for each individual from one or several monitoring periods, where we define a monitoring period as a period of measurements where two subsequent measurements are no more than two days apart. We are interested in examining the association between the variability in the measurements during the first monitoring period and the mean daily frequency of measuring the blood glucose levels during the same period. In the analysis we divide the patients into eighteen different clusters, with regard to gender, age group and length of first monitoring period.

The second dataset is a subset from the Karma project (KAR, 2015, http://karmastudy.org) and contains information about mammographic density. We use a study population of 2960 women, among whom 879 are pairs of sisters. Mammographic density is known to be associated with breast cancer and we investigate the effect of several reproductive factors on the percent dense volume, when we adjust for confounders. We perform the analysis of these data in three steps. First we analyse the unrelated women, i.e. from the population of 2960 women we exclude one woman from each pair of sisters, so that we assume that the observations are independent. By fitting linear models we estimate the crude and adjusted effects of the different covariates. In the subsequent steps we only consider post menopausal women. In the second step, we perform pairwise analyses with the between-within and the rank-ordered logit models. We apply these models on the sibling data where a pair of siblings corresponds to a cluster. We compare these results with the application of the two models on pairs of randomly chosen unrelated women. Third, we match the population on measurable confounders and divide them into clusters. We apply both the between-within and the rank-ordered logit model on these data.

In Chapter 2 and 3 of this thesis we present the necessary theory: Chapter 2 present the theory behind the between-within model and Chapter 3 examine the rank-ordered logit model. This is followed by Chapter 4 where we explain the programming process, the `R` functions used in this thesis and describe and present the results of the simulation studies. In Chapter 5 we describe the real data analysis, where we apply the between-within and the rank-ordered logit models on data of blood glucose levels and mammographic density, respectively. The discussion and the conclusions are found in Chapter 6.

Chapter 2

The Between-Within Model

The between-within model is mostly used in epidemiological research, when we want to investigate possible exposure-outcome associations and need to adjust for confounders that might not be possible to measure. By dividing the data into different clusters, where the individuals within a cluster share the confounders we can adjust for this unmeasurable confounding. We can use mixed models which have both fixed and random effects, to adjust for the shared effects. By assigning to each cluster its own random effect the similarities in the data within a cluster are modelled as nuisance. We obtain the between-within model by expanding a generalized linear model or a generalized mixed model with the cluster average of the explanatory variable.

A common study design is to study siblings and especially twins, since they share several factors such as genetic and environmental factors. When we study siblings, we have observations that are not independent of each other, since two siblings are more alike than two randomly chosen individuals.

In this chapter we introduce linear mixed models, generalized linear models, generalized linear mixed models and also the between-within models.

Throughout the thesis we denote random variables with capital letters and observed variables with lower case letters.

2.1 Linear Mixed Models

Assume that we want to study the association between an outcome and an explanatory variable in a population which is divided into K different clusters, where a cluster could for example correspond to a family or a pair of siblings. There are different ways to estimate the effect of explanatory variable on outcome. A simple approach is to use the linear regression model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}, \tag{2.1}$$

where the response variable is denoted by Y_{ij} for individual j in cluster i, $i = 1, ..., K$, $j = 1, \ldots, n_i$, X_{ij} is the explanatory variable for this individual, ϵ_{ij} is the error term and n_i denotes the number of individuals in cluster i. In this set up the data is modelled as if the outcomes of all individuals are independent when we condition on the explanatory variable (Carlin et al., 2005). Thus, if a cluster corresponds to a pair of siblings, the model ignores the fact that the siblings might be more alike than two randomly chosen individuals in the sample and by falsely assuming independence one implication is that the standard errors in the fitted model will not be correct (Carlin et al., 2005).

One way to handle this is by using linear mixed models. Linear mixed models are linear models with both fixed and random effects, where a fixed effect is shared by the entire population, while a random

effect is specific for each individual or each cluster of individuals (West et al., 2007). When we have one explanatory variable and want to adjust for a cluster specific effect, we can use a mixed model with random intercept, i.e. we have the model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \gamma_i + \epsilon_{ij}, \qquad i = 1, \ldots, K, \ j = 1, \ldots, n_i, \tag{2.2}$$

where we denote the variance of the random effect, γ_i, by σ^2_γ, where σ^2_ϵ denotes the variance of the error term and thus the within pair correlation is $\sigma^2_\gamma/(\sigma^2_\gamma + \sigma^2_\epsilon)$ (Carlin et al., 2005). In genetic studies the ratio $\sigma^2_\gamma/(\sigma^2_\gamma + \sigma^2_\epsilon)$ is called heritability (Visscher et al., 2008).

We begin this section with an example that illustrates a situation where we need a mixed model with a random intercept.

2.1.1 Example: Simulating siblings

We perform a simulation study where we generate samples of size 1000, consisting of 500 pairs representing siblings. We let the response variable y_{ij} denote the observed blood pressure level of individual j in sibling pair i, $j = 1, 2$ and $i = 1, \ldots, 500$. Let us assume that we want to study the association between blood pressure and age, where x_{ij} denotes the observed age. Since siblings share several genetic and environmental factors, we assign a random effect to the individuals within a pair and this effect is assumed to be normally distributed. We let γ_i denote this effect for pair i. Thus, we have a linear mixed model with a random intercept, a fixed intercept and an explanatory variable, which we centre at the mean age of the sample $\bar{x}$, i.e. $\tilde{x}_{ij} = x_{ij} - \bar{x}$. By centring the age, we model the deviation from the mean age in the sample. The model for the outcome of individual j in pair i is written in the form

$$y_{ij} = \beta_0 + \beta_1 \tilde{x}_{ij} + \gamma_i + \epsilon_{ij}, \qquad i = 1, \ldots, 500, \ j = 1, 2, \tag{2.3}$$

where ϵ_{ij} is the individual error term which is independent and identically distributed for all individuals in the population.

In our simulations each individual's age, x_{ij}, is sampled from a normal distribution with mean value 50 and variance 20 and the sibling effect, γ_i, is sampled from a normal distribution with mean 0 and variance 4. The error terms are simulated from a standard normal distribution. The parameters are assigned the values $\beta_0 = 125$ and $\beta_1 = 0, 1$ and 2 respectively. Thus for a 50-year-old individual, with a genetic effect equal to zero, the expected blood pressure is equal to 125. By varying the parameter β_1, we change the influence of age on blood pressure: when $\beta_1 = 0$ the blood pressure is not associated with age, when $\beta_1 = 1$ ($\beta_1 = 2$) one year increase in age implies a one (two) unit(s) higher blood pressure, when all other factors remain unchanged.

We simulate 1000 samples. For each sample we estimate the parameter values with the `R` function `lmer` in the library `lme4` (see Section 4.1 for function details), using the maximum likelihood estimates.

From the output we save the parameter estimates and the estimated parameters standard error from each simulation. The results of these simulations are presented in Table 2.1. The average bias is the deviation of the estimate from the true value, i.e. $E[\hat{\beta}_1] - \beta_1$, where $E[\hat{\beta}_1] = \frac{1}{1000}\sum_{k=1}^{1000} \hat{\beta}_{1k}$ and k denotes the simulation. In Table 2.1 we see that the estimates are unbiased. For the precision we calculate the standard error of the estimated parameters i.e. $\sqrt{\sum_{k=1}^{1000}(\hat{\beta}_{1k} - E[\hat{\beta}_1])^2}$ which is called the empirical standard error. We compare the empirical standard error to the average standard error, i.e. to the mean of the estimated standard errors on each simulation, to see if the standard error of the parameter which is estimated in the model is accurate. The average standard deviation is equal to 0.0095 for all different parameters while the empirical standard error is between 0.0096 and 0.0102. The type I error is the probability of a false positive, that is, for each sample we calculate the confidence interval of the

	Average Bias	Average SD	Empirical SD	Type I error	Power
$\beta_1 = 0$	-0.0002	0.0095	0.0097	0.06	–
$\beta_1 = 1$	0.0001	0.0095	0.0102	0.059	1
$\beta_1 = 2$	-0.0002	0.0095	0.0096	0.051	1

Table 2.1: Results from the 1000 simulations of siblings which share a genetic effect: the estimated average bias in the samples, i.e. the difference between the estimated parameter and the true parameter, the average standard deviation, i.e. the mean of the estimated standard deviations of the coefficients, the empirical error, i.e. the standard error of the β_1 estimates, the type I error, i.e. the probability of a false positive, and the power, i.e. the probability of correctly rejecting the null hypothesis of the coefficient being zero.

estimated parameter, the type I error is calculated as the number of times the confidence interval does not contain the true parameter value divided by the number of simulations. Thus the type I error should be around 5%, however from the table we see that 5-6% of the estimates are a false positive which is slightly higher than expected. The power is the probability of rejecting the null hypothesis $H_0 : \beta_1 = 0$, when the null hypothesis is false. We calculate the power as the number of times the confidence intervals of the estimated parameter does not contain 0 divided by the number of simulations. In these simulations we see that the probability of this is 1.

2.1.2 General Description of Linear Mixed Models

In general, we assume that we have a population that is divided into K clusters, where cluster i consist of n_i individuals. A linear mixed model contains both fixed and random effects, where a fixed effect is shared by the entire population while a random effect is specific for every cluster. For example a random intercept model, for data consisting of K clusters of equal size, that is $n_i = n\ \forall i$. A mixed model with a fixed intercept, one explanatory variable and a random intercept models the response for individual j within cluster i is of the form

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \gamma_i + \epsilon_{ij}; \quad i = 1, \ldots, K, \ j = 1, \ldots, n \tag{2.4}$$

(Fahrmeir et al., 2013). In Equation (2.4), the random effect, γ_i, is shared by all individuals in cluster i while the explanatory variable X_{ij} is observed for every individual. This can be written in matrix form, where for each cluster i we have the model

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in} \end{pmatrix} = \begin{pmatrix} 1 & X_{i1} \\ 1 & X_{i2} \\ \vdots & \vdots \\ 1 & X_{in} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \gamma_i + \begin{pmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \vdots \\ \epsilon_{in} \end{pmatrix}, \tag{2.5}$$

i.e. $\boldsymbol{Y}_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \mathbf{1}^T \gamma_i + \boldsymbol{\epsilon}_i$, where $\mathbf{1}$ is a $1 \times n$ vector of ones.

In general, we can have more than a fixed intercept and one fixed explanatory variable and the random effects can also be random slopes. Therefore, the general matrix form of the responses from cluster i, $i = 1, \ldots, K$ with n_i observations, in a linear mixed model with p fixed effect variables and q random effects variables is then

$$\boldsymbol{Y}_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{\gamma}_i + \boldsymbol{\epsilon}_i, \tag{2.6}$$

where the response, $\boldsymbol{Y}_i$ is a vector of dimension $n_i \times 1$, the fixed effects are represented by $\boldsymbol{X}_i^T \boldsymbol{\beta}$, where

$\boldsymbol{X}_i$ is a matrix of dimension $p \times n_i$ and $\boldsymbol{\beta}$ is a column vector of dimension $p \times 1$. The random effects are denoted by $\boldsymbol{Z}_i^T \boldsymbol{\gamma}_i$, where $\boldsymbol{Z}_i$ is a matrix of dimension $q \times n_i$ and $\boldsymbol{\gamma}_i$ is a column vector of dimension $q \times 1$. $\boldsymbol{\epsilon}_i$ is a vector of error terms with dimension $n_i \times 1$, all error terms ϵ_{ij} are assumed to be independent and identically normally distributed (West et al., 2007).

The random effects are assumed to be normally distributed with a covariance matrix $\boldsymbol{S}$, that is $\gamma_i \sim N(0, \boldsymbol{S})$ and since the error terms are independent and identically distributed we have that $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{Q})$, where $\boldsymbol{Q}$ is a diagonal matrix with homogeneous variances (Tutz, 1994, Chapter 7).

The matrix form of the overall mixed model when we have K clusters is then

$$\begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_K \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1^T \\ \boldsymbol{X}_2^T \\ \vdots \\ \boldsymbol{X}_K^T \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \boldsymbol{Z}_1^T & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{Z}_2^T & \dots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \dots & \boldsymbol{Z}_K^T \end{pmatrix} \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \vdots \\ \boldsymbol{\gamma}_K \end{pmatrix} + \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_K \end{pmatrix}, \tag{2.7}$$

i.e. $\boldsymbol{Y} = \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{Z}^T\boldsymbol{\gamma} + \boldsymbol{\epsilon}$. When the random effects are assumed to be normally distributed, then

$$\begin{pmatrix} \boldsymbol{\gamma} \\ \boldsymbol{\epsilon} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{pmatrix} \right) \tag{2.8}$$

where $\boldsymbol{G}$ and $\boldsymbol{R}$ are block diagonal covariance matrices with matrices $\boldsymbol{S}$ and $\boldsymbol{Q}$ on the diagonals, respectively (Fahrmeir et al., 2013). We focus on the linear mixed model with a random intercept.

When the covariance matrices $\boldsymbol{R}$ and $\boldsymbol{G}$, are known, we estimate the parameters in the linear mixed model with a random intercept by maximizing the joint log likelihood of the outcome $\boldsymbol{Y}$ and random intercept $\boldsymbol{\gamma}$, which is equal to

$$l(\boldsymbol{Y}, \boldsymbol{\gamma}) = -0.5 \left(\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{\beta} - \boldsymbol{Z}^T\boldsymbol{\gamma} \right)^T \boldsymbol{R}^{-1} \left(\boldsymbol{Y} - \boldsymbol{X}^T\boldsymbol{\beta} - \boldsymbol{Z}^T\boldsymbol{\gamma} \right) - 0.5\boldsymbol{\gamma}^T\boldsymbol{G}^{-1}\boldsymbol{\gamma}$$

(Fahrmeir et al., 2013). It is however more common that the covariance matrices are unknown and in that situation the linear mixed model with a random intercept can be analysed with maximum likelihood estimation or with restricted maximum likelihood estimation (Fahrmeir et al., 2013, page 372-375). These two approaches estimate the unknown covariance parameters that are necessary in order to estimate the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in Equation (2.6). If we use the notation in Equation (2.8), and let $\boldsymbol{G} = \boldsymbol{G}(\boldsymbol{\vartheta})$ and $\boldsymbol{R} = \boldsymbol{R}(\boldsymbol{\vartheta})$, then the first step is to estimate the parameters $\boldsymbol{\vartheta}$. We introduce the notation $V(\boldsymbol{\vartheta}) = \boldsymbol{Z}^T\boldsymbol{G}(\boldsymbol{\vartheta})\boldsymbol{Z} + \boldsymbol{R}(\boldsymbol{\vartheta})$. In the maximum likelihood method, we look at the log likelihood for the marginal model $\boldsymbol{Y} \sim N(\boldsymbol{X}^T\boldsymbol{\beta}, V(\boldsymbol{\vartheta}))$ and derive the profile log likelihood function, that is we first estimate the $\boldsymbol{\beta}$ parameters while keeping $\boldsymbol{\vartheta}$ constant and then inserting the estimated $\boldsymbol{\beta}$ values into Equation (2.9) below, which is a function of $\boldsymbol{\vartheta}$, to estimate the parameter $\boldsymbol{\vartheta}$. This profile log likelihood function is equal to

$$l_P(\boldsymbol{\vartheta}) = -0.5 \left(\log\left(\mid V(\boldsymbol{\vartheta}) \mid \right) + \left(\boldsymbol{Y} - \boldsymbol{X}^T\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta}) \right)^T V(\boldsymbol{\vartheta})^{-1} \left(\boldsymbol{Y} - \boldsymbol{X}^T\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta}) \right) \right), \tag{2.9}$$

where $\hat{\boldsymbol{\beta}}(\boldsymbol{\vartheta}) = (\boldsymbol{X}V(\boldsymbol{\vartheta})^{-1}\boldsymbol{X}^T)^{-1}\boldsymbol{X}V(\boldsymbol{\vartheta})^{-1}\boldsymbol{Y}^T$ (Fahrmeir et al., 2013). The maximum likelihood method does not adjust the estimates for the fact that we use the same data to estimate the variance parameters and the $\boldsymbol{\beta}$ parameters, therefore for small samples the maximum likelihood estimates of the variances are underestimated (Fitzmaurice et al., 2011). The restricted maximum likelihood method adjust for this and maximizes a slightly different likelihood function, namely the restricted likelihood function which

has an additional term called a penalty term. The restricted likelihood function is equal to

$$l_R(\boldsymbol{\vartheta}) = l_P(\boldsymbol{\vartheta}) - 0.5 \log \left(\mid \boldsymbol{X}^T V(\boldsymbol{\vartheta})^{-1} \boldsymbol{X} \mid \right)$$

(Fahrmeir et al., 2013).

2.2 Generalized Linear Models

For a population that consists of n individuals, a generalized linear model is a function of the expected outcome conditioned on the p explanatory variables. The function is predicted by a linear function,

$$g\left(E\left[Y_j \mid \boldsymbol{X}_j\right]\right) = g\left(\mu_j\right) = \eta_j = \beta_0 + \beta_1 X_{1j} + \beta_2 X_{2j} \ldots + \beta_p X_{pj}, \tag{2.10}$$

where j denotes the individual, $j = 1, ..., n$. Y is a vector of independent observations from an exponential family distribution (Agresti, 2002, Chapter 4). An exponential family with parameter vector θ of dimension k has a probability density (or mass) function which can be written as the product of three different components, one component that is a function of the parameter, one component which is a function of the observations and an exponential function of the product between a function of the observations times the canonical parameter. That is, an exponential family can be written on the form

$$f(\boldsymbol{y} \mid \theta) = a(\theta) b(\boldsymbol{y}) \exp\left(\theta^T \boldsymbol{t}(\boldsymbol{y})\right)$$

where θ is the canonical parameter and $\boldsymbol{t}(\boldsymbol{y})$ is the canonical statistic.

The linear predictor is a function of p explanatory variables and are for individual j equal to a vector of dimension n denoted by η_j, this is called the systematic component and g is the function that links the random and systematic component and g is called a link function. The link function needs to be monotonic and differentiable, thus it has an inverse. Two special link functions are the identity link which is defined as $g(\mu_j) = \mu_j$, another is the canonical link which is the case when $g(\mu_j)$ is equal to the natural parameter, that is $g(\mu_j) = g(\mu_j(\theta_j)) = \theta_j$ (Agresti, 2002, Chapter 4).

2.2.1 Generalized Linear Mixed Models

Generalized linear mixed models are generalized linear models with both fixed and random effects, so that the linear predictor is defined as

$$\eta_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \boldsymbol{Z}_i^T \boldsymbol{\gamma}_i$$

for cluster i, $i = 1, \ldots, K$, where $\boldsymbol{X}_i^T \boldsymbol{\beta}$ is the fixed effect, $\boldsymbol{X}_i$ is a matrix of dimension $p \times n_i$, $\boldsymbol{\beta}$ is a vector of dimension $p \times 1$ and n_i denotes the number of individuals in cluster i. $\boldsymbol{Z}_i^T \boldsymbol{\gamma}_i$ is the random effect, $\boldsymbol{Z}_i$ is a matrix that has dimension $q \times n_i$ also $\boldsymbol{\gamma}$ is a vector of dimension $q \times 1$ (Fahrmeir et al., 2013, Chapter 7).

If we have the generalized linear mixed model

$$g\{E[Y_{ij} \mid X_{ij}]\} = \beta_0 + \gamma_i + \beta_W X_{ij}, \tag{2.11}$$

for cluster i, where γ_i is a random intercept and β_0 is a fixed intercept, Sjölander et al. (2013) present some different analysis approaches for estimating the parameter β_W in Equation (2.11), where the W subscript means the within-effect or the individual specific effect. One approach is to assume that γ_i and $\boldsymbol{X}_i$ are independent and that γ_i has a parametric distribution, which implies that we can use a mixed model, as discussed in Section 2.1 for the identity link situation. Further, Sjölander et al. (2013) discuss

the alternative to maximize the joint likelihood function for β_W and γ_i and estimate the parameters when we assume that γ_i's are fixed parameters. It is also possible to use the conditional likelihood function if g is the canonical link, where for each cluster we condition on the sufficient statistic, $\sum_{j=1}^{n_i} Y_{ij}$, i.e. the conditional likelihood function is the product of $f(\boldsymbol{Y}_i | \sum_{j=1}^{n_i} Y_{ij}, \boldsymbol{X}_i)$ for $i = 1, \ldots, K$. With this approach the random intercepts will be eliminated (Neuhaus and Kalbfleisch, 1998) (Frisell et al., 2012). Another approach is the between-within model, where $\gamma_i = \tilde{\gamma}_i + \beta_B \bar{X}_i$, so the random intercept and the explanatory variable can be associated (Sjölander et al., 2013). We discuss this model in the next section.

2.3 Between-Within Model

If we have the simple linear model

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}, \qquad i = 1, \ldots, K, \ j = 1, \ldots, n_i, \tag{2.12}$$

for an outcome Y_{ij}, an explanatory variable X_{ij} and an error term ϵ_{ij}, we can include the cluster average of the explanatory variable, $\bar{X}_i$, in the model and thus obtain the between-within model. This model divides the linear predictor of a function into a within-effect and a between-effect (Neuhaus and Kalbfleisch, 1998), namely

$$Y_{ij} = \beta_0 + \beta_W \left(X_{ij} - \bar{X}_i\right) + \beta_B \bar{X}_i + \epsilon_{ij}, \tag{2.13}$$

where β_W is the effect within each cluster i which is known as the within-effect while β_B is the effect between the clusters, that is the between-effect and ϵ_{ij} denotes the error term. By including the cluster average in the model we have a variable that captures the cluster effect, thus we adjust for the factors shared within the cluster that have an impact on the explanatory variable. If $\beta_W = 0$ and $\beta_B \neq 0$ then the outcome is associated with a cluster specific effect, while if $\beta_B = 0$ and $\beta_W \neq 0$ the outcome is associated with the deviation of the explanatory variable from the cluster specific average (Carlin et al., 2005).

It is possible to expand this model by including a random intercept for each cluster. So that for individual j in cluster i, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$, we have the model

$$Y_{ij} = \beta_0 + \beta_W \left(X_{ij} - \bar{X}_i\right) + \beta_B \bar{X}_i + \gamma_i + \epsilon_{ij}. \tag{2.14}$$

The advantage of the model in Equation (2.14) compared to Equation (2.13) is that we also adjust for other possible cluster specific effects. The random intercept is the deviation from the fixed intercept β_0 for cluster i. If $\beta_W = \beta_B$, then the cluster average variable is eliminated and the Equations (2.4) and (2.14) are equivalent and the same holds for the fixed intercept models in Equations (2.1) and (2.13) (Fahrmeir et al., 2013).

Equation (2.14) can be rewritten as

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_W \left(X_{ij} - \bar{X}_i\right) + \beta_B \bar{X}_i + \gamma_i + \epsilon_{ij} \\ &= \beta_0 + \beta_W X_{ij} - \beta_W \bar{X}_i + \beta_B \bar{X}_i + \gamma_i + \epsilon_{ij} \\ &= \beta_0 + \beta_W X_{ij} + \left(\beta_B - \beta_W\right) \bar{X}_i + \gamma_i + \epsilon_{ij} \\ &= \beta_0 + \beta_W X_{ij} + \beta_B^* \bar{X}_i + \gamma_i + \epsilon_{ij}, \end{aligned} \tag{2.15}$$

which is preferable to Equation (2.14) since it is easier to interpret and contains the same information (Begg and Parides, 2003). Similarly, we can rewrite Equation (2.13) as a function of the explanatory

variables X_{ij} and $\bar{X}_i$ instead, i.e.

$$Y_{ij} = \beta_0 + \beta_W X_{ij} + \beta_B^* \bar{X}_i + \epsilon_{ij}. \tag{2.16}$$

The coefficient of the between-effect, β_B^*, is equal to $\beta_B - \beta_W$.

The β_W parameter of Equation (2.15) and Equation (2.14) is interpreted as the expected change in the outcome when X_{ij} increases by one unit and all other variables remain unchanged. That means, if we look at two individuals who belong to the same cluster but with one unit difference in the explanatory variable, then their difference in outcome is equal to β_W. The interpretation of the between-effect differs in the two models. For Equation (2.15) β_B^* is the expected change in the outcome when the mean value of the explanatory variable in cluster i increase by one unit and all other variables remain unchanged. That is, the difference in outcome between two individuals with the same explanatory variable but that belong to two different clusters, with one unit difference in cluster average and equal random effect. For Equation (2.14) on the other hand, β_B is the expected change in the outcome when the mean of the pair increases by one unit while the deviation from the mean of the pair remains unchanged. That is, β_B is the difference in outcome between two individuals that belong to two different clusters, with one units difference in cluster average, but equal random effect and equal deviation between the explanatory variable and cluster average.

In the subsequent sections of this thesis we refer to the representation in Equation (2.15) when we discuss the between-within model unless otherwise indicated. If $\beta_B^* = 0$ that is equivalent to $\beta_B = \beta_W$ and we do not have a cluster specific effect on the outcome due to shared confounders. If $\beta_W = 0$ then $\beta_B^* = \beta_B$ and we only have cluster specific effects associated with the outcome. If $\beta_B = 0$ then $\beta_B^* = -\beta_W$ and we have an effect of the deviation from the mean exposure within a cluster and a possible random effect.

The relationship between the outcome and the explanatory variable does not have to be an identity link. The general form of the between-within model is an extension of the generalized linear model

$$g\left(E\left[Y_{ij}|\bar{X}_i, X_{ij}\right]\right) = \beta_0 + \beta_W X_{ij} + \beta_B \bar{X}_i,$$

or similarly an extension of the generalized linear mixed model.

2.3.1 Between-Within Models with Confounders

If we are in the situation where there is a confounding variable present and the value of this confounder is equal for the members of the cluster, e.g. for the two siblings in a sibling pair study, we can model the outcome as a linear mixed model with a random intercept which is usually an unobservable quantity. In this section we assume that we are studying pairs of siblings, but it can be generalized to bigger clusters. We assume that we have K different clusters and since the value of the confounder is equal for two siblings we have that $C_{i1} = C_{i2} = C_i$, $i = 1, \ldots, K$. For individual j in cluster i we have a linear mixed model

$$Y_{ij} = \beta_0 + C_i + \beta_1 X_{ij} + \epsilon_{ij}, \ i = 1, \ldots, K, \ j = 1, 2, \tag{2.17}$$

where ϵ_{ij} is an error term assumed to be independent of C_i and X_{ij} (Neuhaus and McCulloch, 2006). If the relationship between exposure and confounder is linear, we have

$$X_{ij} = C_i + \eta_{ij}, \tag{2.18}$$

where η_{ij} is a random error term. By inserting Equation (2.18) into Equation (2.17), the linear model can be rewritten accordingly

$$\begin{aligned} E\left[Y_{ij}|C_i, X_{ij}\right] &= \beta_0 + C_i + \beta_1 X_{ij} \\ &= \beta_0 + C_i + \beta_1\left(X_{ij} - \bar{X}_i\right) + \beta_1 \bar{X}_i \\ &= \beta_0 + \underbrace{C_i + \beta_1\left(C_i + \bar{\eta}_i\right)}_{\gamma_i} + \beta_1\left(X_{ij} - \bar{X}_i\right) \\ &= \beta_0 + \gamma_i + \beta_1\left(X_{ij} - \bar{X}_i\right) \end{aligned} \tag{2.19}$$

where the last term is independent of γ_i and can thus be analysed as generalized linear mixed model with identity link and the same holds for generalized linear models with other link functions (Neuhaus and McCulloch, 2006). Equation (2.19) can also be rewritten by replacing the term C_i with $\tilde{\gamma}_i + \beta_B \bar{X}_i$ where we assume that the new random intercept is independent of the explanatory variables (Sjölander et al., 2013).

The confounding variable is not always entirely shared. Frisell et al. (2012) have shown that if the true causal model is

$$\begin{aligned} Y_{ij} &= \beta_1 X_{ij} + \beta_2 C_{ij} + \epsilon_{ij} \\ X_{ij} &= \beta_3 C_{ij} + \eta_{ij}, \end{aligned} \tag{2.20}$$

then the estimated within-effect β_W in (2.16) might be more biased than the estimated effect β_1 from the crude linear model in Equation (2.12) if the confounding are not perfectly shared by the siblings, note that we do not include a random intercept.

When the response variable and the explanatory variable are continuous the within-effect can be calculated exactly in Equation (2.20) and is equal to

$$\hat{\beta}_W = \beta_1 + \frac{\beta_2 \beta_3 \mathrm{Var}(C)}{\beta_3^2 \mathrm{Var}(C) + \mathrm{Var}(\epsilon_X)\frac{1-\rho_X}{1-\rho_C}}, \tag{2.21}$$

where β_1 is the true coefficient of the explanatory variable, i.e. the β_1 in Equation (2.20) (Frisell et al., 2012). ρ_X denotes the correlation of the exposure within a cluster, i.e. $\rho_X = \mathrm{Cor}(X_{i1}, X_{i2}) = \frac{\mathrm{Cov}(X_{i1}, X_{i2})}{\sqrt{\mathrm{Var}(X_{i1})}\sqrt{\mathrm{Var}(X_{i2})}}$. Similarly, ρ_C denotes the correlation of the confounder within a cluster, i.e. $\rho_X = \mathrm{Cor}(C_{i1}, C_{i2})$. If the confounding variable is completely shared within the pairs $\hat{\beta}_W = \beta_1$ (Frisell et al., 2012). If ρ_X are decreasing to zero and ρ_C is not equal to one, then the bias of the within estimate will increase.

2.3.2 Example: Simulate a Confounding Effect

We simulate from a model similar to the one introduced in Frisell et al. (2012) but with continuous outcome and sample size of 1000 from the model (2.17) and (2.18). We have a sample that consist of $K = 500$ pairs of siblings, within each pair the confounding variable is perfectly shared and dichotomous,

furthermore the exposure is also correlated. We make the following assumptions for pair i

$$
\begin{aligned}
C_i^{(n)} &\sim N\left(\begin{pmatrix}50\\50\end{pmatrix}, \begin{pmatrix}\sigma_C^2 & \rho_C \times \sigma_C^2\\ \rho_C \times \sigma_C^2 & \sigma_C^2\end{pmatrix}\right)\\
C_{ij} &= \mathbb{I}\{C_{ij}^{(n)} \geq 50\}\\
\boldsymbol{\eta}_i &= \begin{pmatrix}\eta_{i1}\\ \eta_{i2}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_X^2 & \rho_X \times \sigma_X^2\\ \rho_X \times \sigma_X^2 & \sigma_X^2\end{pmatrix}\right)\\
Y_{ij} &= \beta_0 + \beta_1 X_{ij} + \alpha C_{ij} + \epsilon_{ij} = 120 + 0.8X_{ij} + C_{ij} + \epsilon_{ij}\\
X_{ij} &= 0.5C_{ij} + \eta_{ij},
\end{aligned}
$$

where ρ_X is the correlation for the exposure within each pair. We let $\rho_X = 0.3,\ 0.5,\ 0.6$ and 0.8, ρ_C is the correlation of the confounding variable within a sibling pair and is equal to 1. $\mathbb{I}$ denotes the indicator function which is equal to one if the value of the confounder for individual j in pair i is larger than or equal to 50 and otherwise the function is equal to zero. The variances are assumed to be 20 and 2 for σ_C^2 and σ_X^2, respectively. The data is analysed by a simple linear regression, model (2.1), and by fitting different between-within models.

We simulate 500 samples of sample size 1000 from this model. The average bias of the estimated coefficient β_1 is presented in Table 2.2. We can see that the estimated average coefficient for the within-effect, β_1, is equal for all models which are divided into a between- and a within-effect. So it does not matter in terms of bias in this simulation if we use the model in Equation (2.14), (2.15) or the model in Equation (2.16), neither does it matter if we include a random intercept in the model or not. We note that the average bias of the estimated within-effect does not seem to be affected by ρ_X when the confounder is perfectly shared within a pair, i.e. $\rho_C = 1$. Furthermore, the within-effect is less biased then the estimated β_1 in the model $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}$.

Model	Function	ρ_X=0.3	ρ_X=0.5	ρ_X=0.6	ρ_X=0.8
$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}$	lm	0.06	0.06	0.06	0.07
$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 \bar{X}_i + \epsilon_{ij}$	lm	−0.01	0	0	0.02
$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \beta_2 \bar{X}_i + \gamma_i + \epsilon_{ij}$	lmer	−0.01	0	0	0.02
$Y_{ij} = \beta_0 + \beta_1 (X_{ij} - \bar{X}_i) + \beta_2 \bar{X}_i + \gamma_i + \epsilon_{ij}$	lmer	−0.01	0	0	0.02

Table 2.2: The bias of the estimated β_1 values, when the true parameter value is 0.8.

In the simulations presented in Section 4.2 we study the behaviour of the between-within model when the confounder is not perfectly shared within a pair.

Chapter 3

The Rank-Ordered Logit Model

3.1 Introduction to the Rank-Ordered Logit Model

Discrete choice models are useful when analysing data that are collected from surveys, where the participants are choosing between several different choices and are either asked to choose the most preferable alternative or to rank all of the alternatives. One of the most common models is the logit model, which is used when the individuals are faced with several different alternatives and choose one of them. In this model one models an individuals utility as a linear function of observable explanatory variables and some unknown information which is assumed to follow an extreme value type I distribution. From this model the rank-ordered logit model is developed, in which we use the information about the ranking of all alternatives. The resulting log likelihood function is equivalent to the stratified Cox proportional hazards model. Therefore this chapter begins with a description of the properties of the extreme value type I distribution, then we continue by describing survival models, present the theory for the rank-ordered logit model and describe how it can be applied to data with continuous outcome.

3.1.1 Extreme Value Type I Distribution

The extreme value type I distribution (also called Gumbel distribution) is continuous and defined for all real numbers, $x \in \mathbb{R}$ (Forbes et al., 2010, Chapter 19). For a sample of independent and identically distributed continuous random variables the distribution models the limiting maximum value when the sample size tends to infinity (Alves and Neves, 2011). The distribution function is defined as

$$F_X(x) = \exp\left(-\exp\left(-\frac{x-a}{b}\right)\right) \tag{3.1}$$

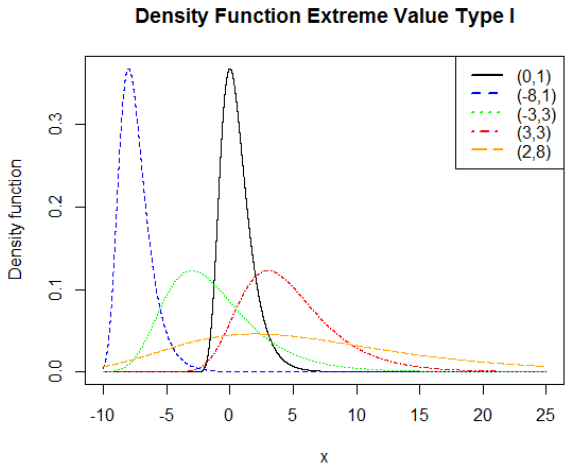
where a is the location parameter, $a \in \mathbb{R}$, and b is the scale parameter, which by definition is larger than 0, thus the density function is

$$f_X(x) = \frac{1}{b}\exp\left(-\frac{x-a}{b}\right)\exp\left(-\exp\left(-\frac{x-a}{b}\right)\right), \tag{3.2}$$

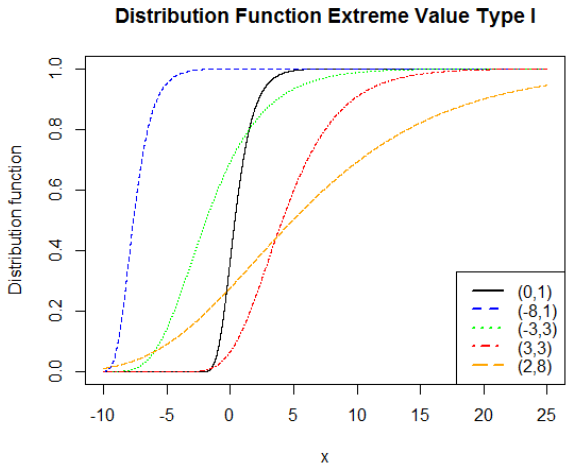
with mean value equal to $a + b\gamma$, where γ is the Euler constant and approximately equal to 0.577 and the variance is equal to $\frac{\pi^2 b^2}{6} \approx 1.645 \cdot b^2$ (Forbes et al., 2010). The special case when $a = 0$ and $b = 1$ we refer to as a standard extreme value type I distribution in this thesis.

The density function and distribution function is shown for different values of the location and scale parameters in Figure 3.1a and 3.1b, respectively. In these figures we can see that the distribution is right skewed. Skewness is defined as $E[(\frac{X-\mu}{\sigma})^3]$, where μ is the mean value of the random variable X and σ

is the standard deviation of X. The skewness of the extreme value type I distribution is equal to 1.14 (Forbes et al., 2010).



(a) The density function of an extreme value type I distribution with parameters (a,b).



(b) The distribution function of an extreme value type I distribution with parameters (a,b).

Figure 3.1: Illustrations of the extreme value type I distribution.

Example: Extreme Value Type I or Normally Distributed Errors

When we have independent observations assumed to follow a linear model and know the distribution of the error terms, we can estimate the parameters in the model. In this example we examine the robustness of two analysis approaches, namely if we assume that the error terms are normally distributed and extreme value type I distributed, respectively. We do this by analysing data where the true underlying distribution of the errors is standard extreme value type I distributed, i.e. mean 0.577 and variance $\pi^2/6$, or normally distributed with mean 0 and variance $\pi^2/6$, thus the variance of the error terms are the same in the two models to facilitate comparison. We assume that we have a population of individuals where a linear model describes the relationship between the response variable, Y_j, (blood pressure) for individual j and the explanatory variables 'age', X_{1j} and 'gender', X_{2j}. The variable 'age' is continuous and centred, i.e. $\tilde{X}_{1j} = X_{1j} - \bar{X}$, where $\bar{X}$ denotes the mean age in the population, the variable 'gender' is binary,

$$X_{2j} = \begin{cases} 1, & \text{if individual } j \text{ male,} \\ 0, & \text{if individual } j \text{ female.} \end{cases}$$

We generate the responses from two linear models, that differ only in terms of error distribution. Model 1 is

$$Y_j^{M1} = \beta_0 + \beta_1 \tilde{X}_{1j} + \beta_2 X_{2j} + \epsilon_j, \quad j = 1, \ldots, 1000, \tag{3.3}$$

where ϵ_j is a normally distributed error term with mean 0 and variance $\pi^2/6$. Model 2 is defined as

$$Y_j^{M2} = \beta_0 + \beta_1 \tilde{X}_{1j} + \beta_2 X_{2j} + \eta_j, \quad j = 1, \ldots, 1000, \tag{3.4}$$

where η_j is a standard extreme value type I distributed error term. The parameters in these two models

are

$$\beta_0 = 125,\ \beta_2 = 0.5,$$
$$\beta_1 = 0, 1, 2 \text{ respectively.}$$

We generate X_1 and X_2 as

$$X_{1j} \sim N(50, 20),$$
$$X_{2j} \sim Be(0.5),$$

where Be denotes the Bernoulli distribution. We simulate 1000 chains of Model 1 and Model 2, respectively and for each chain (i.e. each sample of 1000) we fit the two models and compare the results of analysis when we assume that the error terms are standard extreme value type I distributed and when they are assumed to be normally distributed, both when the true underlying model is Equation (3.3) and when it is Equation (3.4). Therefore, we generate 1000 samples of Model 1 and Model 2, respectively, where each sample is of size 1000. To each sample we fit two different regression models in order to estimate the parameter β_1. To fit the linear model which assumes a normal distribution we use the function `lm` in `R`. To find the maximum likelihood estimates of the log likelihood function of a linear model with error terms that have a standard extreme value distribution, namely

$$l(\boldsymbol{\beta} \mid x_{2j}, x_{1j}, y_j) = \log\left(L\left(\boldsymbol{\beta} \mid x_{2j}, x_{1j}, y_j\right)\right) = \log\left(\prod_{j=1}^{n} f_\eta\left(y_j - \beta_0 - \beta_1 x_{1j} - \beta_2 x_{2j}\right)\right)$$

we make the temporary substitution $z_j = y_j - \beta_0 - \beta_1 x_{1j} - \beta_2 x_{2j}$,

$$\begin{aligned}
&= \log\left(\prod_{j=1}^{n} f_\eta\left(z_j\right)\right) \\
&= \sum_{j=1}^{n}\left(\log\left(\frac{1}{1}\exp\left(-\frac{z_j - 0}{1}\right)\exp\left(-\exp\left(-\frac{z_j - 0}{1}\right)\right)\right)\right) \\
&= \sum_{j=1}^{n}\left(\log\left(\exp\left(-z_j\right)\exp\left(-\exp\left(-z_j\right)\right)\right)\right) \\
&= \sum_{j=1}^{n}\left(-z_j + \left(-\exp\left(-z_j\right)\right)\right) = -\sum_{j=1}^{n}\left(z_j + \left(\exp\left(-z_j\right)\right)\right)
\end{aligned}$$

and by inserting $z_j = y_j - \beta_0 - \beta_1 x_{1j} - \beta_2 x_{2j}$ we obtain the log likelihood function

$$= -\sum_{j=1}^{n}\left(y_j - \beta_0 - \beta_1 x_{1j} - \beta_2 x_{2j} + \exp\left(-\left(y_j - \beta_0 - \beta_1 x_{1j} - \beta_2 x_{2j}\right)\right)\right)$$

which we find the maximum likelihood estimates of with the `R` function `optim` described in Section 4.1.

The focus of the comparison of the two different analysis approaches is the parameter β_1, the impact of age in Equation (3.3) and (3.4). We look at the bias, precision, type I error and power. We defined these in the example in Section 2.1.1.

The result of these simulations are presented in Table 3.1. The two different error distributions have the same variance but different means, 0 and 0.577 respectively, but this difference does not influence the parameter estimates of β_1, only the intercept. With regard to the bias of the two models they perform equally well, but we also note from the empirical standard deviation and the type I errors that the normal

True parameter	True Distribution	Assumed Distribution	Average Bias	Empirical SD	Average SD	Type I error	Power
$\beta_1 = 0$	Normal	Normal	0	0.009	0.009	0.056	–
		EVT1	0	0.013	0.007	0.287	–
	EVT1	Normal	0	0.009	0.009	0.051	–
		EVT1	−0.001	0.007	0.007	0.048	–
$\beta_1 = 1$	Normal	Normal	−0.001	0.009	0.009	0.057	1
		EVT1	−0.001	0.014	0.007	0.34	1
	EVT1	Normal	0	0.009	0.009	0.055	1
		EVT1	−0.001	0.007	0.007	0.053	1
$\beta_1 = 2$	Normal	Normal	0	0.009	0.009	0.045	1
		EVT1	−0.001	0.014	0.007	0.323	1
	EVT1	Normal	0	0.009	0.009	0.05	1
		EVT1	0	0.007	0.007	0.055	1

Table 3.1: Estimates of the coefficient of interest, β_1, in Model 1 where $\epsilon \sim N(0, \pi^2/6)$ and in Model 2 where $\eta \sim EVT1(0, 1)$, where EVT1 denotes a standard extreme value type I distribution. The models used to generate tha data are given in column 2 ('True distribution') and the assumed distribution in the analysis is given in column 3.

distribution approach is more robust than the extreme value type I approach. When the true underlying distribution is normal but the method of analysis is the extreme value type I approach, the standard error is underestimated by approximately 0.006, as can be seen by comparing with the average standard error. A consequence of this is that the type I error rate is above 5% when the true underlying error distribution is normal but we perform the analysis as if the errors were extreme value type I distributed. When $\beta_1 = 2$ the type I error is equal to 0.323, which implies that the 95 % confidence interval of the estimate only contains the true parameter value 67.7 % of the simulations. We can also see from the table that the power is equal to 1 in all models indicating that we never fail to detect a significant effect of the explanatory variable.

3.1.2 Survival Models

Cox Proportional Hazards Regression Model

In survival analysis one models the time to certain events, for example time of death, described by the random variable T (Lachin, 2011, Chapter 9). At the end of the study all events have not necessarily occurred, that is we might have some censored data (Fox, 2002). If we for example study mortality some individuals may be alive at the end of the study. If T is continuous the probability that an event will occur before time t, $t > 0$, is described by the distribution function $F(t) = P(T \leq t)$ and corresponding density function $f(t)$. Thus, the probability that an event has not yet occurred at time t is defined by the survival function, $S(t)$, which is equal to

$$S(t) = P(T > t) = 1 - F(t).$$

The hazard function is defined as the probability of an event to occur instantaneous by time t, given that it has not yet occurred before t (Lachin, 2011, Chapter 9), i.e.

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t},$$

(Klein and Moeschberger, 1997), which for continuous time is equal to

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{3.5}$$

By rewriting Equation (3.5) we have $f(t) = \lambda(t)S(t)$,

For a set of explanatory variables $\boldsymbol{x}$, $\boldsymbol{x} = (x_1, \ldots, x_p)$, the proportional hazards model was defined by Cox (1972) as

$$\lambda(t; \boldsymbol{x}) = \exp\left(\boldsymbol{x}^T \boldsymbol{\beta}\right) \lambda_0(t), \tag{3.6}$$

where $\lambda_0(t)$ is an arbitrary function called the baseline hazard. That is, the logarithm of the hazard is assumed to be approximately linear, i.e.

$$\log(\lambda(t; \boldsymbol{x})) = \log(\lambda_0(t)) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

from this we can see that the ratio of hazards for two independent observations are independent of the time, t (Fox, 2002).

The Partial Likelihood Function

In this section we follow the heuristic derivation of the partial likelihood function in Lachin (2011). We let n denote the population size and the d observed event times denoted by $t_{(1)}, \ldots, t_{(d)}$, where we assume that none of these events are tied. For individual j the vector of explanatory variables is denoted $\boldsymbol{x}_{(j)}$. We want to calculate the conditional probability that an event occurs at time $t_{(j)}$ for individual j, who has not yet experienced an event by then, given that an event actually occurs for some individual in the population time $t_{(j)}$, i.e. $P(t_{(j)}$ event time for individual $j \mid$ an event is observed time $t_{(j)})$. We have that

$$\begin{aligned}
&P(t_{(j)} \text{ event time for individual } j \mid \text{ an event is observed time } t_{(j)}) \\
&= \frac{P(t_{(j)} \text{ event time for individual } j, \text{ an event is observed time } t_{(j)})}{P(\text{ an event is observed time } t_{(j)})} \\
&= \frac{P(t_{(j)} \text{ event time for individual } j, \mid j \text{ at risk time } t_{(j)})}{P(\text{an event occurs instantaneously at time } t_{(j)} \mid j \text{ at risk time } t_{(j)})}
\end{aligned}$$

(Therneau, 2015). We have from Equation (3.6) that the probability in the numerator is equal to

$$\lambda(t_{(j)}; \boldsymbol{x}_{(j)}) = \exp\left(\boldsymbol{x}_{(j)}^T \boldsymbol{\beta}\right) \lambda_0(t_{(j)}).$$

By summing over all individuals in the population who have not yet experienced the event, the total conditional probability of an event time $t_{(j)}$ is

$$\sum_{l:t_l \geq t_{(j)}} \exp\left(\boldsymbol{x}_l^T \boldsymbol{\beta}\right) \lambda_0(t_{(j)}).$$

From this we can obtain the likelihood function by multiplying the conditional distribution over all event times that have been observed (Lachin, 2011, Chapter 9),

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{d} \frac{\exp\left(\boldsymbol{x}_j^T \boldsymbol{\beta}\right)}{\sum_{l:t_l \geq t_{(j)}} \exp\left(\boldsymbol{x}_l^T \boldsymbol{\beta}\right)}, \tag{3.7}$$

where the time-dependent term $\lambda_0(t)$ is eliminated.

Stratified Cox Proportional Hazards Regression Model

If the population can be sorted into K different clusters, that is we have K different strata, then each stratum may have a specific baseline hazard function

$$\lambda_i(t;\boldsymbol{x}) = \lambda_i(t)\exp\left(\boldsymbol{x}^T\boldsymbol{\beta}\right)$$

where $i = 1,\ldots,K$. By similar reasoning as in the previous subsection we obtain the likelihood function

$$L_S(\boldsymbol{\beta}) = \prod_{i=1}^{K}\prod_{j=1}^{n_i}\left[\frac{\exp\left(\boldsymbol{x}_{j(i)}^T\boldsymbol{\beta}\right)}{\sum_{\iota:t_{\iota(i)}\geq t_{j(i)}}\exp\left(\boldsymbol{x}_{\iota(i)}^T\boldsymbol{\beta}\right)}\right]^{\delta_{j(i)}} \tag{3.8}$$

where n_i denotes the number of individuals in stratum i, $j(i)$ denotes individual j in stratum i and $\delta_{j(i)}$ is an indicator function which is equal to one if individual j in stratum i experiences an event and equal to 0 if not (Lachin, 2011).

For the case where we have K strata and p explanatory variables, an one unit change in x_{ih} where $i = 1,\ldots,K$, $h = 1,\ldots,p$ correspond to a multiplicative change of $\exp(\beta_h)$ on the hazard, so that if $\beta_h > 0$ the risk of the event increase when the covariate x_{ih} increases and if $\beta_h < 0$ the risk of the event decrease when the covariate increase when all other covariates are given. The parameters in the partial likelihood function are interpreted in the same manner.

3.2 Logit Models

Discrete choice models are widely used in the econometrics literature where one is often interested to forecast the demand for a new product. By collecting data from surveys, where the participating individuals are asked to choose the most preferred alternative or to rank all of the alternatives, it is possible to analyse the consumers choice. The logit model is a discrete choice models where the most preferable alternative is modelled. This model is derived with the underlying assumption that the errors have a standard extreme value type I distribution. Since ranked data contains more information than data of the most preferable alternative, the rank-ordered logit model, which is originating from the logit model, has been developed. In this section we introduce the rank-ordered logit model, for describing how individual alternatives in a dataset are ranked. We begin in Section 3.2.1 by introducing the discrete choice models and in Section 3.2.2 introduce the logit model which is described by McFadden (1974). We then follow the derivations by Beggs et al. (1981) in order to introduce in Section 3.2.3 the rank-ordered logit model and its application to the analysis of continuous data.

3.2.1 Discrete Choice Model

In a population where the individuals are faced with several different choices, it is of interest to find a model that describes the choices made. In the economics literature it is common to use the discrete choice model in order to study buying behaviour when choosing between several alternative products (Train, 2003). Participants in a survey are asked to order the items according to their preferences where a high ranking implies a high utility, i.e. the most prefered item. The response variable in these models is the individuals utility, which is not observed.

There are some underlying assumptions of the discrete choice model: when an individual chooses one alternative this means that none of the other alternatives are chosen, i.e. they are mutually exclusive,

the sample space includes all possible choices and is finite (Train, 2003, page 15). It is assumed that all individuals make the decision that maximizes their utility, but all factors that affect an individual's utility cannot be observed. When there are L possible alternatives the utility of alternative l for individual j is defined as

$$U_{jl} = V_{jl} + \epsilon_{jl}, \tag{3.9}$$

where V_{jl} is a linear function of observed factors and ϵ_{jl} is an error term corresponding to random unobserved factors (Train, 2003). The observed part of the utility function for individual j may be a function of both the attributes $\boldsymbol{x}_{jl}$ of the possible choices and the attributes of the individual s_j, $V_{jl} = V(\boldsymbol{x}_{jl}, s_j)$ (Beggs et al., 1981). Several different discrete choice models have been developed: logit, probit, generalized extreme value and mixed logit, but we focus on the logit model in this thesis.

3.2.2 The Logit Model

Assume a population of size N, where every individual j has a vector of attributes s_j and is presented with a choice set of L alternatives. We obtain the logit model by assuming that the error terms are independent and have a standard extreme value type I distributed (Train, 2003, Chapter 3). We begin by studying one individual and for clarity ignore the subscript indicating which individual we are studying. The probability that an individual's utility from choice k is less than z is

$$P(U_k \leq z) = P(\epsilon_k \leq z - V_k).$$

Thus, the probability that the utility from alternative l is larger than the utility of alternative k is

$$P(U_k \leq U_l) = P(\epsilon_k \leq U_l - V_k) = \exp\left(-\exp\left(-\left(V_l + \epsilon_l - V_k\right)\right)\right)$$

where $k,\ l = 1, \ldots, L$ and $l \neq k$. We begin with the case where we choose between two alternatives: l and k. The choice maker chooses the alternative with greatest utility, we are therefore interested in the probability that the utility from alternative l is larger than from alternative k, which is

$$\begin{aligned}
P(U_l > U_k,\ l \neq k) &= \int_{-\infty}^{\infty} P(U_l > U_k, l \neq k \mid \epsilon_l) f_\epsilon(\epsilon_l) d\epsilon_l = \int_{-\infty}^{\infty} e^{-e^{-(V_l+\epsilon_l-V_k)}} e^{-e^{-\epsilon_l}} e^{-\epsilon_l} d\epsilon_l \\
&= \int_{-\infty}^{\infty} e^{-e^{-\epsilon_l}(e^{-(V_l-V_k)}+1)} e^{-\epsilon_l} d\epsilon_l \\
&= \begin{bmatrix} Substitution & s = e^{-\epsilon_l} \\ & ds = -e^{-\epsilon_l} d\epsilon_l \\ & \infty < s < 0 \end{bmatrix} \\
&= -\int_{\infty}^{0} e^{-s\left(e^{-(V_l-V_k)}+1\right)} ds = \int_{0}^{\infty} e^{-s\left(e^{-(V_l-V_k)}+1\right)} ds \\
&= \left[-\frac{e^{-s\left(e^{-(V_l-V_k)}+1\right)}}{e^{-(V_l-V_k)}+1} \right]_0^{\infty} = \frac{1}{e^{-(V_l-V_k)}+1} \\
&= \frac{e^{V_l}}{e^{V_l}+e^{V_k}}
\end{aligned} \tag{3.10}$$

(Train, 2003). We can extend the results from Equation (3.10) to find the probability that the utility of alternative l is larger than the utility of all other alternatives for the individual. We use the fact

the ϵ_l's are independent and identically distributed extreme value type I which implies that the U_l's are independent (Cramer, 2003), thus

$$\begin{aligned}
P(U_l > U_k, \forall k \neq l) &= \int_{-\infty}^{\infty} \prod_{m \neq l} e^{-e^{-(U_l - V_m)}} e^{-\epsilon_l} e^{-e^{-\epsilon_l}} d\epsilon_l \\
&= \int_{-\infty}^{\infty} \prod_{m=1}^{L} e^{-e^{-(V_l + \epsilon_l - V_m)}} e^{-\epsilon_l} d\epsilon_l = \int_{-\infty}^{\infty} \exp\left(-e^{-\epsilon_l} \sum_{m=1}^{L} e^{-(V_l - V_m)}\right) e^{-\epsilon_l} dU_l \\
&= \int_{0}^{\infty} \exp\left(-s \sum_{m=1}^{L} e^{-(V_l - V_m)}\right) ds \\
&= \left[-\frac{e^{-s \sum_{m=1}^{L} e^{-(V_l - V_m)}}}{\sum_{m=1}^{L} e^{-(V_l - V_m)}} \right]_0^{\infty} \\
&= \frac{1}{\sum_{m=1}^{L} e^{-(V_l - V_m)}} = \frac{e^{V_l}}{\sum_{m=1}^{L} e^{V_m}} = P_l. \qquad (3.11)
\end{aligned}$$

This is the logit model, for the most preferred alternative (Train, 2003), which has the property of independence of irrelevant alternative (IIA), that is the ratio of the probability to choose alternative l and the probability to choose alternative k, is independent of the probabilities of all other possible alternatives. From Equation (3.11), $P_l/P_k = \exp(V_l)/\exp(V_k)$, that is the sum in the denominator cancels out and what is left is not affected by the other possible choices (Train, 2003).

A common example when the IIA property can cause problems is given by McFadden (1974) where one assumes that there are only two transportation methods possible, by car or by bus. Assume one third of the population travels by bus and two thirds travel by car, then the ratio of these two probabilities is equal to two, i.e. $\frac{\text{P(Car)}}{\text{P(Original bus)}} = \frac{2/3}{1/3} = 2$. If we assume that another bus is introduced. The ratio of the probability to travel by car divided by the probability to travel by the original bus remains unchanged and equal to two by the IIA property. However, if the two buses are similar and the travellers have equal probability to travel by both buses. This implies that the fraction between the probabilities these two alternatives is equal to one, $\frac{\text{P(Original bus)}}{\text{P(New bus)}} = 1$. Thus $\text{P(Car)} = 2 \cdot \text{P(Original bus)} = 2 \cdot \text{P(New bus)}$ and therefore the fraction of the population who travel by car decreases to a half, which is probably not the case, since the fraction who takes the car should not be affected by the introduction of a new bus. Therefore, it is not appropriate to use the model if the different alternatives cannot be considered independent for the individuals (McFadden, 1974).

3.2.3 The Rank-Ordered Logit Model

In the rank-ordered logit model we take into account that we know the ordinal ordering of the choice set. The IIA property is equivalent to the fact that when we condition on the order of all choices that are available to the decision maker, the distribution of utility from the most preferred choice is independent of the way in which the other choices are ordered (Beggs et al., 1981). We show by following the results given by Beggs et al. (1981) that the conditional distribution of the utility from alternative l being less than or equal to z given that the individual obtains highest utility from that alternative, is independent of how the other choices are ordered.

In these derivations the error terms are assumed to be standard extreme value type I distributed and we use the induced distribution for the utility of alternative l, namely

$$H(U_l) = \exp(-e^{-(U_l - V_l)})$$

(Beggs et al., 1981). We begin by calculating that the probability that the utility from alternative l is less then or equal to z conditioned on the utility from alternative l being larger than the utility from choice k and use the result from Equation (3.10), namely

$$\begin{aligned}
P(U_l \leq z \mid U_l > U_k, l \neq k) &= \frac{P(U_l \leq z, U_l > U_k, l \neq k)}{P(U_l > U_k, l \neq k)} = \frac{P(z \geq U_l > U_k, l \neq k)}{P(U_l > U_k, l \neq k)} \\
&= \frac{\int_{-\infty}^{z}\int_{-\infty}^{U_l} e^{-(U_l - V_l)} e^{-e^{-(U_l - V_l)}} e^{-(U_k - V_k)} e^{-e^{-(U_k - V_k)}} dU_k dU_l}{\frac{e^{V_l}}{e^{V_l}+e^{V_k}}} \\
&= \frac{e^{V_l}+e^{V_k}}{e^{V_l}} \int_{-\infty}^{z} e^{-(U_l - V_l)} e^{-e^{-U_l}\left(e^{V_l}+e^{V_k}\right)} dU_l = \begin{bmatrix} e^{-U_l} = z \\ -e^{-U_l} dU_l = dz \end{bmatrix} \\
&= -\frac{e^{V_l}+e^{V_k}}{e^{V_l}} e^{V_l} \int_{\infty}^{e^{-U_l}} e^{-z\left(e^{V_l}+e^{V_k}\right)} dz \\
&= \left(e^{V_l}+e^{V_k}\right) \int_{e^{-U_l}}^{\infty} e^{-z\left(e^{V_l}+e^{V_k}\right)} dz \\
&= \left(e^{V_l}+e^{V_k}\right) \left[-\frac{e^{-z\left(e^{V_l}+e^{V_k}\right)}}{\left(e^{V_l}+e^{V_k}\right)}\right]_{e^{-U_l}}^{\infty} = e^{-e^{-U_l}\left(e^{V_l}+e^{V_k}\right)} \\
&= e^{-e^{-\left(U_l - \log\left(e^{V_l}+e^{V_k}\right)\right)}}
\end{aligned} \tag{3.12}$$

and when $U_l > \max_{k \neq l, k \in \jmath} U_k$ where $\jmath$ denotes the choice set presented to the individual, we have

$$\begin{aligned}
P(U_l \leq z \mid U_l > U_k, \forall l \neq k) &= \frac{P(U_l \leq z, U_l > U_k, \forall l \neq k)}{P(U_l > U_k, \forall l \neq k)} \\
&= \frac{e^{V_l}\int_{-\infty}^{z} \prod_{m=1}^{L} e^{-e^{-(U_l - V_m)}} e^{-U_l} dU_l}{\frac{e^{V_l}}{\sum_{m=1}^{L} e^{V_m}}} \\
&= \left(\sum_{m=1}^{L} e^{V_m}\right) \int_{-\infty}^{U_l} e^{-e^{-z}\sum_m e^{V_m}} e^{-z} dz \\
&= e^{-e^{-\left(U_l - \log\left(\sum_{m \in \jmath} e^{V_m}\right)\right)}},
\end{aligned} \tag{3.13}$$

which is the distribution function for an extreme value type I distribution with location parameter $\log\left(\sum_{m \in \jmath} e^{V_m}\right)$ and scale parameter 1. From the results derived in this section we can now express the conditional probability that given the ranking of the L alternatives presented to the individual, the probability that the largest utility in that subset is at most z is

$$\begin{aligned}
P(U_1 \leq z \mid U_1 > U_2 > \ldots > U_L) &= \frac{P(U_1 \leq z, U_1 > U_2 > \ldots > U_L)}{P(U_1 > U_2 > \ldots > U_L)} \\
&= \frac{P(z \geq U_1 > U_2 > \ldots > U_L)}{P(U_1 > U_2 > \ldots > U_L)}
\end{aligned} \tag{3.14}$$

where the probability in the numerator is equal to

$$
\begin{aligned}
&P(z \geq U_1 > U_2 > \ldots > U_L) \\
&= \int_{-\infty}^{z} \int_{-\infty}^{U_1} \int_{-\infty}^{U_2} \cdots \int_{-\infty}^{U_{(L-1)}} \prod_{m=1}^{L} \exp\left(-e^{-(U_m - V_m)}\right) e^{-(U_m - V_m)} dU_L \cdots dU_1 \\
&= \int_{-\infty}^{z} \int_{-\infty}^{U_1} \int_{-\infty}^{U_2} \cdots \int_{-\infty}^{U_{(L-2)}} \underbrace{\prod_{m=1}^{L-2} \exp\left(e^{-(U_m - V_m)}\right) e^{-(U_m - V_m)}}_{A} \\
&\cdot \exp\left(-e^{-U_{(L-1)}} \left(e^{V_{(L-1)}} + e^{V_L}\right)\right) e^{-(U_{(L-1)} - V_{(L-1)})} dU_{(L-1)} \cdots dU_1 = \left[\begin{array}{c} Substitute \\ s = e^{-U_{(L-1)}} \end{array}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= -e^{V_{(L-1)}} \int_{-\infty}^{z} \int_{-\infty}^{U_1} \int_{-\infty}^{U_2} \cdots \int_{e^{-U_{(L-2)}}}^{\infty} A \cdot \exp\left(-s\left(e^{V_{(L-1)}} + e^{V_L}\right)\right) dU_{(L-1)} \cdots dU_1 \\
&= \frac{e^{V_{(L-1)}}}{e^{V_{(L-1)}} + e^{V_L}} \int_{-\infty}^{z} \int_{-\infty}^{U_1} \int_{-\infty}^{U_2} \cdots \int_{-\infty}^{U_{(L-3)}} A e^{-e^{-U_{(L-2)}}\left(e^{V_{(L-1)}} + e^{V_L}\right)} dU_{(L-2)} \cdots dU_1 \\
&= \ldots = \prod_{m=2}^{L-1} \left(\frac{e^{V_m}}{\sum_{l=m}^{L} e^{V_l}}\right) \int_{-\infty}^{z} e^{-e^{-U_1}\left(\sum_{l=2}^{L} e^{V_l}\right)} \exp\left(-e^{-(U_1 - V_1)}\right) e^{-(U_1 - V_1)} dU_1 \\
&= -\prod_{m=2}^{L-1} \left(\frac{e^{V_m}}{\sum_{l=m}^{L} e^{V_l}}\right) \int_{\infty}^{e^{-U_1}} e^{-z\left(\sum_{l=1}^{L} e^{V_l}\right)} e^{V_1} dz \\
&= \prod_{m=1}^{L} \left(\frac{e^{V_m}}{\sum_{l=m}^{L} e^{V_l}}\right) \exp\left(-e^{-\left(U_1 - \log\left(\sum_{l=1}^{L} e^{V_l}\right)\right)}\right). \qquad (3.15)
\end{aligned}
$$

For the denominator, we have

$$
\begin{aligned}
&P(U_1 > U_2 > \ldots > U_L) \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{U_1} \int_{-\infty}^{U_2} \cdots \int_{-\infty}^{U_{(L-1)}} \prod_{m=1}^{L} \exp\left(-e^{-(U_m - V_m)}\right) e^{-(U_m - V_m)} dU_L \cdots dU_1 \\
&= \ldots = \prod_{m=2}^{L-1} \left(\frac{e^{V_m}}{\sum_{l=m}^{L} e^{V_l}}\right) \int_{-\infty}^{\infty} e^{-e^{-U_1}\left(\sum_{l=2}^{L} e^{V_l}\right)} \exp\left(-e^{-(U_1 - V_1)}\right) e^{-(U_1 - V_1)} dU_1 \\
&= \prod_{m=2}^{L-1} \left(\frac{e^{V_m}}{\sum_{l=m}^{L} e^{V_l}}\right) \int_{-\infty}^{\infty} e^{-e^{-U_1}\left(\sum_{l=1}^{L} e^{V_l}\right)} e^{-(U_1 - V_1)} dU_1 = \left[\begin{array}{c} Substitute \\ s = e^{-U_1} \end{array}\right] \\
&= \prod_{m=2}^{L-1} \left(\frac{e^{V_m}}{\sum_{l=m}^{L} e^{V_l}}\right) e^{V_1} \int_{0}^{\infty} e^{-s\left(\sum_{l=1}^{L} e^{V_l}\right)} ds = \prod_{m=2}^{L-1} \left(\frac{e^{V_m}}{\sum_{l=m}^{L} e^{V_l}}\right) e^{V_1} \left[-\frac{e^{-s\left(\sum_{l=1}^{L} e^{V_l}\right)}}{\sum_{l=1}^{L} e^{V_l}}\right]_0^{\infty} \\
&= \prod_{m=1}^{L} \left(\frac{e^{V_m}}{\sum_{l=m}^{L} e^{V_l}}\right). \qquad (3.16)
\end{aligned}
$$

By inserting the probabilities in Equation (3.15) and (3.16) into Equation (3.14), we get the following result

$$
P(U_1 \leq z \mid U_1 > U_2 > \ldots > U_L) = \exp\left(-e^{-\left(U_1 - \log\left(\sum_{l=1}^{L} e^{V_l}\right)\right)}\right) \qquad (3.17)
$$

which is equivalent to the probability presented in Equation (3.13) i.e. $P(U_1 \leq z \mid U_1 > U_2 > \ldots > U_L) = P(U_1 \leq z \mid U_l > U_k, l \neq k)$. Thus we can conclude that this conditional distribution for the utility of the choice with the highest ranking is independent of the order of the other choices (Beggs et al., 1981). If we

let V_l be a linear function such that $V_l = \boldsymbol{x}_l^T\boldsymbol{\beta} = \beta_1 x_{l1} + \ldots + \beta_p x_{lp}$, then Equation (3.16) is equal to the likelihood of the Cox proportional hazard function in Equation (3.7) when we do not have any censored events. If we assume that we observe the ranking $R = (r_1, r_2, \ldots, r_{L_i})$ for a decision maker j who faces L_j different choices, then the probability of the observed ranking is equal to

$$P(U_{r_1} > U_{r_2} > \ldots > U_{r_{L_j}}) = \prod_{l=1}^{L_j}\left[\frac{e^{\boldsymbol{x}_{r_l}^T\boldsymbol{\beta}}}{\sum_{m=l}^{L_j} e^{\boldsymbol{x}_{r_m}^T\boldsymbol{\beta}}}\right] \tag{3.18}$$

and the likelihood for a sample of individuals is obtained by taking the product over all individuals so that for N individuals the log likelihood is

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{N}\sum_{l=1}^{L_j}\left(\boldsymbol{x}_{jr_l}^T\boldsymbol{\beta} - \log\left(\sum_{m=l}^{L_j}\exp\left(\boldsymbol{x}_{ir_m}^T\boldsymbol{\beta}\right)\right)\right), \tag{3.19}$$

where j denotes the individual and l the alternative. As stated above, the linear predictor V_{jl} may be a function of both the attributes of the alternative and attributes of the individual j, that is $V_{jl} = V(\boldsymbol{x}_{jl}, \boldsymbol{s}_j)$, but in Equation (3.18) we see that the individual's attributes, which only depend on the individual and not the choice, are eliminated in the likelihood function. Since the likelihood function is concave (Beggs et al., 1981), a unique maximum likelihood estimate exists.

As we showed above, the rank-ordered logit model possesses the property of independence of irrelevant alternatives.

3.3 Application of the Rank-Ordered Logit Model

Even though the rank-ordered logit model allows for the coefficients in the linear predictor to be different for different individuals we only consider the simplified situation where we assume that the coefficients are identical. We apply the rank-ordered logit model to data with a continuous outcome variable and we divide the data into several different clusters, where individuals within a cluster are matched on potential confounders. If we use sibling studies each cluster consists of one pair of siblings. A cluster corresponds to a choice set in the discrete choice models, thus, instead of modelling the ranked alternatives for an individual, we model the ranked outcomes within a cluster. The consequence from dividing the individuals with similar confounding profile into clusters is that in this set up the confounders correspond to the individuals attributes in the discrete choice notation. Thus, the confounders are matched away in the log likelihood function and we only need to include the explanatory variables of interest in the regression. This implies that we can use these models when we have unmeasurable confounders, as long as we know which individuals have a similar confounding profile. When we rank the outcomes higher ranking correspond to the best alternatives in the rank-ordered logit model we rank the outcomes so that they are positive and in decreasing order.

Since we have obtained a log likelihood that is equivalent to a Cox stratified proportional hazards model, we can use the function `coxph` in `R` to fit the rank-ordered logit model (see Section 4.1 for further details).

In Equation (3.11) we saw that when we compare two outcomes, the probability that the outcome of alternative l where larger than the outcome of alternative k is equal to $P_l = \exp(V_l)/(\exp(V_l) + \exp(V_k))$. The covariates that are equal in the linear predictors cancel out. The model can then be considered a conditional logistic regression if the underlying assumptions of the rank-ordered logit model are fulfilled. The coefficients can therefore be interpreted as log odds ratios.

3.3.1 Scaled Parameters and the Delta Method

When we fit a rank-ordered logit model, we assume the error terms are standard extreme value type I distributed. This implies that we force the variance to be equal to $\pi^2/6$, even though this might not be true and therefore we normalize the parameters when we apply the rank-ordered logit model. The parameters of the utility function can be scaled without affecting the ordering of the utilities, that is for a model

$$U_{jl} = V_{jl} + \epsilon_{jl} \tag{3.20}$$

with $\text{Var}(\epsilon_{jl}) = b^2 \cdot \pi^2/6$ we can divide the parameters by b so that the variance is equal to $\pi^2/6$. Thus we transform the model in Equation (3.20) to the model

$$U'_{jl} = V_{jl}/b + \epsilon'_{jl},$$

where the ϵ'_{jl} is standard extreme value type I distributed (Train, 2003). When fitting the rank-ordered logit model to data we actually estimate the scaled parameters, i.e. if $V_{jl} = \boldsymbol{x}_{jl}^T\boldsymbol{\beta}$ in Equation (3.20) we want to estimate the parameters $\boldsymbol{\beta}$ but estimated is the parameter $\boldsymbol{\beta}/b$, so in order to find the true effect of the covariate we need to multiply the estimate by b.

In the rank-ordered logit model we have a linear predictor, V_{jl}, where the covariates are both explanatory variables and confounders, and the shared confounders are eliminated as explained following Equation (3.18) above. If the confounders are measurable and their influence on the outcome can be described linearly, we can obtain the estimates from the rank-ordered logit model using the linear regression model with extreme value type I distributed error terms. An advantage from fitting a linear model with extreme value type I distributed error terms, is that by looking at the residuals of the linear model we can determine if it is reasonable to assume that the error terms actually have an extreme value type I distribution. However, the confounders that are eliminated and matched away in the rank-ordered logit model must be included in the linear model. If the models are comparable, these two approaches should yield similar estimates, apart from the fact that the rank-ordered logit model estimates are scaled. Therefore, we need to estimate both the coefficients, $\boldsymbol{\beta}$, and the scale parameter, b, in the linear regression model. If the scale parameter is equal to 1 we can interpret the estimates of the rank-ordered logit model in the same manner as in the linear model with extreme value type I distributed errors. Since both the coefficients and the scale parameter are estimated, we need to take this into consideration when determining the standard error of the estimate $\boldsymbol{\beta}/b$. We use the delta method (Cox, 2005), to estimate a parameter θ we use large sample theory and assume that the estimator is asymptotically normally distributed, with mean θ and variance $\sigma^2(\theta)$. We let g be a function which is differentiable at the true parameter value θ and for values close to θ, and the first order derivative of g is nonzero at the true value of θ. If $\hat{\theta}_n$ denotes the estimated parameter from a sample of size n, then from the delta method we have

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow{d} N\left(0, \sigma^2(\theta)\left(g'(\theta)\right)^2\right).$$

If we have two estimated parameters, $\boldsymbol{\theta} = (\theta_1, \theta_2)$, we assume that our estimator, $\hat{\boldsymbol{\theta}}_n$, is asymptotically bivariate normally distributed, with mean $\boldsymbol{\theta}$ and covariance matrix $\Sigma(\boldsymbol{\theta})/n$, where n is our sample size (Agresti, 2002, page 589). Just as in the case of a single parameter, we assume that the derivative of $g(\boldsymbol{\theta})$ is nonzero at the true values of $\boldsymbol{\theta}$. The multivariate delta method is then

$$\sqrt{n}\left(g(\hat{\boldsymbol{\theta}}_n) - g(\boldsymbol{\theta})\right) \xrightarrow{d} N\left(0, g'(\boldsymbol{\theta})^T\Sigma(\boldsymbol{\theta})g'(\boldsymbol{\theta})\right).$$

If $g(\boldsymbol{\theta}) = g(\theta_1, \theta_2) = \theta_1/\theta_2$, where θ_1 denotes the estimated coefficient and θ_2 denotes the scale parameter, the derivatives of $g(\boldsymbol{\theta})$ are $\partial g(\boldsymbol{\theta})/\partial\theta_1 = 1/\theta_2$ and $\partial g(\boldsymbol{\theta})/\partial\theta_2 = -\theta_1/\theta_2^2$, the mean of $g(\boldsymbol{\theta})$ is equal to θ_1/θ_2 and the variance is

$$\text{Var}\left(g\left(\boldsymbol{\theta}\right)\right) = \begin{pmatrix} \frac{\partial g(\boldsymbol{\theta})}{\partial\theta_1} & \frac{\partial g(\boldsymbol{\theta})}{\partial\theta_2} \end{pmatrix} \begin{pmatrix} \text{Var}(\theta_1) & \text{Cov}(\theta_1,\theta_2) \\ \text{Cov}(\theta_1,\theta_2) & \text{Var}(\theta_2) \end{pmatrix} \begin{pmatrix} \frac{\partial g(\boldsymbol{\theta})}{\partial\theta_1} \\ \frac{\partial g(\boldsymbol{\theta})}{\partial\theta_2} \end{pmatrix},$$

which is estimated by plugging in our estimates $\hat{\theta}_1$ and $\hat{\theta}_2$.

3.3.2 Example: Impact of Scaling the coefficients

We continue examining the simulations from the linear model in Equation (3.3) and Equation (3.4), where the outcome was a function of two explanatory variables, age and gender. In Equation (3.3) the error terms were normally distributed with mean 0 and variance $\pi^2/6$, while in model (3.4) the error terms were standard extreme value type I distributed. In Section 3.1.1 the models were analysed both by assuming the true underlying distribution of the error terms were normal and by assuming it was standard extreme value type I distributed. However, in real data applications we do not know the scale parameter of the underlying distribution and thus need to estimate it together with the coefficients. As indicated in Section 3.3.1 we need to scale the coefficients of a linear model with extreme value type I distributed errors in order to compare them with the estimates of the rank-ordered logit model. We therefore estimate the parameters in the log likelihood function

$$l(\boldsymbol{\beta}, b, a) = -\sum_{j=1}^{N}\left(\log(b) + \frac{1}{b}\left(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta} - a\right) + \exp\left(-\frac{1}{b}\left(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta} - a\right)\right)\right) \tag{3.21}$$

with the function `optim` in `R`. In Equation (3.21) j denotes the individual, N the number of individuals in the sample, b the scale parameter and a the location parameter. The estimates of the coefficient of the age variable, β_1, and the scale parameter for the simulated data in Section 3.1.1 are presented in Table 3.2. We also present the results from the analysis of the same models with a variance of the error terms equal to $4\pi^2/6$, i.e. $b = 2$. We compare the estimates from the linear model with extreme value type I

True error distribution	True (β_1, b)	Linear model $\hat{\beta}_1$	Linear model $\hat{b}$	Linear model $\hat{\beta}_1/b$	Rank-ordered logit $\hat{\beta}_1^{RO}$
EVT1	(0,1)	−0.001	0.998	−0.001	0
	(1,1)	0.999	0.998	1.001	1.001
	(2,1)	1.999	0.999	2.002	2
Normal	(0,1)	0	1.274	0	0
	(1,1)	0.999	1.276	0.783	0.791
	(2,1)	1.999	1.277	1.565	1.601
EVT1	(0,2)	0	1.996	0	0
	(1,2)	1	1.996	0.501	0.5
	(2,2)	2	1.997	1.001	1
Normal	(0,2)	0	2.548	0	0
	(1,2)	0.999	2.551	0.392	0.387
	(2,2)	2	2.555	0.783	0.79

Table 3.2: Results from data generated from Equation (3.3) and (3.4). The true error distribution is given in the first column, in the fitted linear model we assume that the error distribution is extreme value type I. The estimates of the coefficient for the age variable, β_1, and the scale parameter, b, are given in the third and fourth column of the table. In the fifth column we find the corresponding scaled estimate and in the last column the estimate from the rank-ordered logit model is presented.

distributed errors with the estimates of the rank-ordered logit model, when we in the rank-ordered logit model divide the population into two clusters with respect to gender.

We see in Table 3.2 that the scaled estimates from the linear model is equal to those from the rank-ordered logit model. When the true underlying distribution is extreme value type I, the scale parameter is correctly estimated and by scaling the parameters we obtain a model with variance $\pi^2/6$. On the other hand, when the true underlying distribution is normal, the estimate of the scale parameter is significantly larger than the true scale parameter. But, the rank-ordered logit model is reacting the same way, thus we can use a linear model with extreme value type I distributed errors to estimate the size of the scale parameter in this setting when we fully adjust for the random effect in the linear predictor and have measurable confounders.

Chapter 4

Computational Methods

In this thesis we have used the software `R` for all programming. In this chapter we describe the libraries and functions we have used. Also, we present the simulation studies we have performed in order to compare the between-within and the rank-ordered logit regression models.

4.1 R Programming

All simulations and data analyses in this thesis are performed in `R`, in this section we present the libraries and functions that we use.

4.1.1 Library 'stats'

In the library `stats` the function `optim` is defined. This function finds the values of the parameters which maximizes the likelihood function. There are different estimate methods available. In this thesis we use the `L-BFGS-B` approach, and in some analyses the `BFGS` approach, which are iterative methods to estimate the parameters that optimizes the log likelihood function (https://stat.ethz.ch/R-manual/R-patched/library/stats/html/optim.html).

4.1.2 Library 'lmtest'

When two models are nested, it is of interest to compare the two models to determine which one of the two that is preferable to the other. One test for this is the likelihood ratio test. In this test we compare the likelihood of the models when the maximum values of the parameters are used. The simple model is the true model under the null hypothesis and we denote the parameters in this models by the vector $\boldsymbol{\beta}_0$ and we denote the likelihood function of this model by $l_0(\hat{\boldsymbol{\beta}}_0) = l_0$, where $\hat{\boldsymbol{\beta}}_0$ denotes the maximum likelihood estimates. The model with some additional covariates is the true model under the alternative hypothesis, this model have the parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1)$, the likelihood of this model we denote by $l_1(\hat{\boldsymbol{\beta}}) = l_1$, where $\hat{\boldsymbol{\beta}}$ denotes the maximum likelihood estimates. We then test the null hypothesis $\boldsymbol{\beta}_1 = \mathbf{0}$ against the alternative $\boldsymbol{\beta}_1 \neq \mathbf{0}$ with the statistic $-2\log(l_0/l_1)$ which has a $\chi^2(k)$ distribution, where k is the degrees of freedom and is equal to the dimension of $\boldsymbol{\beta}_1$ (Agresti, 2002, Page 11). The package `lmtest` contains the function `lrtest` which performs the likelihood ratio test on two linear models (Horthorn et al., 2015).

4.1.3 Library 'evd'

In order to simulate and calculate the quantiles of the extreme value type I distribution, we use the functions in the package `evd` (Stephenson, 2015).

4.1.4 Library 'lme4'

The `lme4` package in `R` fits linear mixed models with the function `lmer` (Bates, 2010). Into the `lmer` function we insert the formula that describes the model with the response variable, fixed effects and random effects (Bates et al., 2015) and choose to analyse the data with the restricted maximum likelihood method discussed in Section 2.1.2 by adding the command `REML=TRUE`. If we want to test if the fixed effects are significantly different from zero we need the package `lmerTest`.

4.1.5 Library 'lmerTest'

The package `lmerTest` contains an extended version of the `lmer` function in the library `lme4`, which also calculates the p-value of F-tests for the significance of the fixed-effects estimates (Kuznetsova et al., 2015). These p-values are estimated by Satterthwaite's approximations, we do not cover the theory of these calculations in this thesis, for more information see Kuznetsova et al. (2015).

4.1.6 Library 'survival'

The package `survival` in `R` contain the function `coxph` which fits the Cox proportional hazards model, discussed in Section 3.1.2, by inserting the formula, with a survival function as response vector and the explanatory variables together with a vector which tells which strata the different observations correspond to (Therneau, 2015). The function return the estimates that maximizes the partial likelihood function in Equation (3.8) using the Newton-Raphson method (Therneau and Grambsch, 2000), that is by first choosing an initial guess of the parameter vector $\boldsymbol{\beta}^{(0)}$ and then update this guess stepwise by inserting it into the function

$$\boldsymbol{\beta}^{(k+1)} = \boldsymbol{\beta}^{(k)} - \left(\boldsymbol{H}^{(k)}\right)^{-1} \boldsymbol{u}^{(k)}; \quad k = 0, 1, 2, \ldots$$

where $\boldsymbol{H}^{(k)}$ is the Hessian matrix and $\boldsymbol{u}^{(k)}$ is the score vector of the likelihood function for $\boldsymbol{\beta}$. k denotes the k:th step of the iteration which stops when the contribution to likelihood between two subsequent steps are small enough (Agresti, 2002). If there are any ties among the time of events in the data, the function `coxph` by default use the Efron approximation to handle these (Therneau, 2015), where the likelihood contribution of the ties are weigthed. The approximation is described in Therneau and Grambsch (2000).

In the output of the function `coxph` we also find the result from a Wald test of the null hypothesis that the coefficient is equal to zero. That is, for the coefficient β, the Wald test statistic of the null hypothesis $H_0 : \beta = 0$ is $z = \frac{\hat{\beta}}{\sqrt{\text{Var}(\hat{\beta})}}$ and $z^2 \sim \chi^2(1)$ (Agresti, 2002, Page 10).

4.2 Simulation Study

In order to compare the rank-ordered logit and the between-within regression models simulation studies have been performed where we study the association between a certain outcome and a particular explanatory variable. In these simulations the association is also influenced by a random effect or a random confounder.

We simulate samples of 500 pairs, so that a natural clustering method is to use the pairs which here illustrates pairs of siblings. In our first scenario we have a linear model with an explanatory variable and a random genetic effect that is completely shared within a pair of siblings. Furthermore, we want to examine the models ability to estimate the effect from the explanatory variable in more realistic settings. Therefore, we investigate the models in the setting where the genetic effect is a confounder that is completely shared within a pair of siblings and also when the confounder is not one hundred percent shared within a pair.

For each scenario we apply different analysis approaches. In Section 4.2.1 we assume that the effect that we are trying to adjust for is unmeasurable. We fit the unadjusted linear regression model with the explanatory variable as the only covariate and the random intercept model which adjust for the shared cluster specific factors that are not included in the model by treating them as nuisance. We also use matched designs, in this case we use sibling pairs, where we divide the population into clusters where a cluster consists of a pair of siblings and then apply the rank-ordered logit and the between-within models.

In Section 4.2.2 we assume that the random effects are measurable. When all covariates are measurable we can adjust for the covariates we are not interested in by including them in the linear model and thus obtain an adjusted linear model. We can also apply the between-within and the rank-ordered logit model, where we divide the individuals into different clusters with respect to their random effects, i.e. instead of treating each pair as a cluster we create larger clusters where the individuals within a cluster have similar random effects.

We simulated 1000 samples, each sample consisting of 1000 individuals.

4.2.1 Unmeasurable Random Effects

A Shared Genetic Effect

The purpose of the simulation study is to compare the performance of the between-within and the rank-ordered logit model with regard to bias, precision, coverage and power. Coverage is defined as the percentage of times the estimated confidence intervals include the true parameter value. Thus, with a 95% significance level the coverage should be 95%. We look at both the mean of the estimated standard errors in the fitted models, i.e. the average standard deviation, and the standard error of the estimated parameters in the fitted models, i.e. the empirical standard deviation, in order to determine the precision of the models.

In this section we describe a model where the outcome and explanatory variable are continuous. We assume that we have samples of siblings. Since siblings are similar in genetic and maternal aspects it is not appropriate to analyse it as an independent sample. We begin with a simple setting where we generate a genetic effect that also influences the outcome. This effect is continuous and completely shared by the pair of siblings. That is, we generate data from the model

$$Y_{ij} = \beta_0 + \beta_1 \times X_{ij} + \beta_2 \times C_i + \epsilon_{ij}, \tag{4.1}$$

where Y_{ij} denotes the outcome for individual j in sibling pair i, $j = 1, 2$ and $i = 1, \ldots, 500$. The explanatory variable is denoted by X_{ij}, the random effect that is specified for each pair is denoted by C_i while the error term is denoted by ϵ_{ij}. The error distribution is standard extreme value type I and normal with mean 0 and variance $\pi^2/6$, respectively. The parameters are equal to

$$\begin{aligned} \beta_0 &= 0.2, \\ \beta_1 &= 0,\ 0.25,\ 0.5, \\ \beta_2 &= 1, \end{aligned}$$

i.e. we vary the coefficient β_1 in order to investigate the importance of the magnitude of the explanatory variable. The random effects are simulated from a multivariate normal distribution with mean 2, variance 1 and a within pair correlation 1, i.e.

$$\boldsymbol{C} = \begin{pmatrix} C_{i1} \\ C_{i2} \end{pmatrix} \sim mN\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \right). \tag{4.2}$$

The explanatory variable, X_{ij}, is assumed to have a normal distribution with mean 0 and variance 1.

We estimate the parameter β_1 with different models, namely

- the unadjusted model with the explanatory variable as the only covariate, i.e. $y_{ij} = \beta_0 + \beta_1 x_{ij} + \epsilon_{ij}$,
- a mixed model with a random intercept, i.e. $y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma_i + \epsilon_{ij}$,
- the rank-ordered logit model, where each pair of siblings form one stratum, thus we do not need to include the confounder in the regression model,
- the between-within model, where each pair correspond to a matched set i, and we include a random intercept, γ_i, in the model, i.e. $y_{ij} = \beta_1 x_{ij} + \beta_B \bar{x}_i + \gamma_i + \epsilon_{ij}$.

A Shared Confounder

We expand the model above by making the explanatory variable a function of the random effect, i.e. the random effect becomes a confounder. The model of the outcome is still equal to Equation (4.1) and the random effects are simulated from a multivariate normal distribution with mean 2, variance 1 and correlation 1. But now the explanatory variable is defined as

$$X_{ij} = \alpha_0 + \alpha_1 \times C_i + \eta_{ij} = 0.4 + C_i + \eta_{ij}, \tag{4.3}$$

where η_{ij} is an error term. The error terms are independent and identically distributed. We simulate the error terms from a normal distribution with mean 0 and variance 0.5. We apply the same analysis approaches as in the simulations described above for the model with a shared genetic effect.

A Correlated Confounder

In real life the random effect may not be entirely shared by two siblings. We therefore expand the model with a confounder further by letting the random effect be positively correlated within a pair. That is, for each pair i, we induce a positive correlation between C_{i1} and C_{i2} so that $\text{cor}(C_{i1}, C_{i2}) = \rho$, where ρ denotes the correlation and is equal to 0.7 and 0.9, respectively. The confounders for the individuals within a pair, are sampled from the multivariate normal distribution, with a mean equal to 2 and a covariance matrix where the variance is equal to 1 and the covariance is equal to $1 \times \rho$, thus

$$\boldsymbol{C} = \begin{pmatrix} C_{i1} \\ C_{i2} \end{pmatrix} \sim mN\left(\begin{pmatrix} 2 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right). \tag{4.4}$$

We generate data from

$$Y_{ij} = \beta_0 + \beta_1 \times X_{ij} + \beta_2 \times C_{ij} + \epsilon_{ij}, \tag{4.5}$$

and

$$X_{ij} = 0.4 + C_{ij} + \eta_{ij}. \tag{4.6}$$

Where the parameters and the distributions of the error terms are equal to those in the previous simulation descriptions. We apply the same analysis approaches as in the simulations described above for the model with a shared genetic effect.

4.2.2 Measurable Random Effects

If we instead want to adjust for confounders which we believe are similar within a cluster of individuals that share measurable covariates, we can apply the adjusted linear model or a matched stratum design. Where the former is a linear model with the explanatory variable of interest and the random effect or confounder as covariates, in the latter we divide the population into several different clusters with respect to some measurable covariates and we apply the rank-ordered logit and the between-within models to these designs.

We analyse the data generated from the models in Section 4.2.1 with the analysis approaches that are applicable when the potential cluster specific effects are measurable. That is, we have a sample with 500 pairs of siblings who share a genetic effect, share a confounder or have a correlated confounder within pair and estimate the parameter β_1 with

- the adjusted model with exposure and the random effect or confounder as explanatory variables, i.e. $y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 c_{ij} + \epsilon_{ij}$,
- the stratified rank-ordered logit model where we divide the population into 10 different strata based on their random effect or confounding variable,
- the stratified between-within model where we divide the population into 10 different matched sets based on their random effect or confounding variable and also include a random intercept.

In these stratified approaches we create clusters where the individuals within a cluster are similar with respect to the potential confounders and for a pair of siblings the two individuals may belong to two different clusters.

4.3 Simulation Result

The results from the simulation studies in Chapter 4.2 are presented in this section. We first look at the simulations where we assume unmeasurable effects and then continue with the simulations where we assume the random effects to be measurable.

4.3.1 Unmeasurable Random Effects

A Shared Genetic Effect

We have applied four different methods on the samples. In the tables 'Unadj' denotes the unadjusted linear regression model, with one explanatory variable and 'Mixed Model' denotes the mixed model. 'RO' and 'BW' denotes pairwise comparisons of the rank-ordered logit model and the between-within model, respectively.

The results from the simulations where the random effect is shared within a pair are presented in Table 4.1, from these tables we can conclude that when the errors have a standard extreme value type I distribution none of the models are biased. When we compare the results with the normally distributed error terms we see that the estimates of the rank-ordered logit model are slightly biased. The unadjusted and the mixed models have smaller standard error than the rank-ordered logit and the between-within models and have therefore better precision. We know that the observations are correlated and expect the

		$\epsilon \sim$EVT1(0,1)					$\epsilon \sim$ N$(0, \pi^2/6)$				
True β_1	Model	Avg Bias	Avg SD	Emp SD	Cov	Power	Avg Bias	Avg SD	Emp SD	Cov	Power
0	Unadj	−0.002	0.052	0.052	0.948	–	−0.001	0.051	0.051	0.955	–
	Mixed Model	−0.003	0.048	0.048	0.953	–	−0.001	0.048	0.048	0.944	–
	RO	−0.002	0.063	0.066	0.944	–	0	0.063	0.065	0.949	–
	BW	−0.004	0.057	0.058	0.947	–	−0.001	0.057	0.058	0.943	–
0.25	Unadj	0.002	0.051	0.049	0.962	0.999	−0.001	0.051	0.052	0.945	0.997
	Mixed Model	0.002	0.048	0.045	0.96	1	−0.002	0.048	0.048	0.942	0.999
	RO	0.002	0.066	0.067	0.938	0.964	−0.028	0.066	0.066	0.918	0.929
	BW	0.002	0.057	0.057	0.949	0.987	−0.002	0.057	0.059	0.941	0.988
0.5	Unadj	0.001	0.051	0.05	0.948	1	0.001	0.051	0.052	0.941	1
	Mixed Model	0	0.048	0.046	0.951	1	0.001	0.048	0.049	0.935	1
	RO	0.004	0.074	0.073	0.952	1	−0.05	0.072	0.074	0.877	1
	BW	−0.001	0.057	0.055	0.958	1	0	0.057	0.06	0.94	1

Table 4.1: Results from the simulations of the model with a shared genetic effect described in Section 4.2.1. The error terms in the fitted model have a standard extreme value type I distribution in the first column and a normal distribution with mean 0 and variance $\pi^2/6$ in the second.

standard error of the models which does not take this into consideration to underestimate the variation, but we have a large sample and the standard deviation is not affected by the correlation. The standard deviation is approximately the same in the mixed and unadjusted models and the average and empirical estimates are almost identical. The coverage is approximately 0.95 for all models with standard extreme value type I distributed errors in Table 4.1, while in the models with normally distributed errors we see that the rank-ordered logit model is not robust enough to deal with the normal errors. A consequence is that it has somewhat lower coverage and a higher difficulty to detect a significant effect for $\beta_1 = 0.25$, i.e. lower power.

A Shared Confounder

In Section 4.2.1 we described a model where we have a confounder that is completely shared within a pair of siblings. The results from the simulations of this model is presented in Table 4.2.

		$\epsilon \sim$EVT1(0,1)					$\epsilon \sim$ N$(0, \pi^2/6)$				
True β_1	Model	Avg Bias	Avg SD	Emp SD	Cov	Power	Avg Bias	Avg SD	Emp SD	Cov	Power
0	Unadj	0.666	0.036	0.038	0	–	0.666	0.036	0.039	0	–
	Mixed Model	0.651	0.037	0.04	0	–	0.651	0.037	0.041	0	–
	RO	0.001	0.09	0.092	0.945	–	−0.001	0.09	0.091	0.951	–
	BW	−0.002	0.081	0.084	0.952	–	0	0.081	0.079	0.96	–
0.25	Unadj	0.667	0.036	0.036	0	1	0.665	0.036	0.037	0	1
	Mixed Model	0.651	0.037	0.038	0	1	0.651	0.037	0.038	0	1
	RO	0.002	0.092	0.092	0.951	0.79	−0.032	0.091	0.097	0.918	0.654
	BW	0.003	0.081	0.08	0.948	0.87	−0.007	0.081	0.081	0.954	0.857
0.5	Unadj	0.666	0.036	0.038	0	1	0.666	0.036	0.038	0	1
	Mixed Model	0.651	0.037	0.039	0	1	0.651	0.037	0.04	0	1
	RO	0.005	0.098	0.098	0.955	1	−0.052	0.096	0.096	0.914	0.999
	BW	−0.002	0.081	0.082	0.944	1	0.002	0.081	0.082	0.948	1

Table 4.2: Results from the simulations where we a confounder that is completely shared within a pair of siblings. In the first column the error terms have a standard extreme value type I distribution, while they have a normal distribution with mean 0 and variance $\pi^2/6$ in the second column.

Neither the unadjusted model nor the mixed model can cope with this situation, the estimates are positively biased. The mixed model cannot figure out that we also have a confounding effect affecting the outcome and gives all weight to the exposure estimate. This implies that the models do not cover the true

value. This is no surprise, since we saw in Equation (2.19) that if we want to estimate the parameter β_1 with a mixed model, we should model the deviation from the mean of the explanatory variable within the cluster and not the explanatory variable. The between-within model is unbiased, has a good coverage and a power slightly below 1, no matter which error distribution we are studying. Thus, this model is robust and behaves very well when we have an unmeasurable, completely shared confounder. The estimates of the rank-ordered logit model are somewhat biased when the error distribution is normal. The bias is negative which indicate that the estimated coefficients are underestimated. It also has a smaller power than the between-within model in both situations and is less precise.

In the two simulations we have studied so far, the samples illustrates 500 pairs of monozygotic twins. Carlin et al. (2005) studied the situation when we have a sample of twins and recommended the use of the between-within model. From our simulations we can agree with their conclusion and also recommend the use of the rank-ordered logit model if the error terms have a standard extreme value type I distribution.

A Correlated Confounder

In this section we look at a more realistic situation, we have a confounder in our model, but this confounder is not completely shared by the siblings. We start by assuming that the confounder has a correlation equal to 0.9 within a pair of siblings, the results from this model are presented in Table 4.3. We then continue by lowering the correlation even further, to 0.7 within the pair. If we have siblings and not monozygotic twins this is a more realistic situation. The result of this simulations are presented in Table 4.4.

		$\epsilon \sim$EVT1(0,1)					$\epsilon \sim N(0, \pi^2/6)$				
True β_1	Model	Avg Bias	Avg SD	Emp SD	Cov	Power	Avg Bias	Avg SD	Emp SD	Cov	Power
0	Unadj	0.666	0.036	0.037	0	–	0.666	0.036	0.038	0	–
	Mixed Model	0.655	0.037	0.039	0	–	0.653	0.037	0.04	0	–
	RO	0.165	0.083	0.084	0.491	–	0.143	0.083	0.083	0.591	–
	BW	0.172	0.076	0.078	0.381	–	0.167	0.076	0.076	0.396	–
0.25	Unadj	0.667	0.036	0.037	0	1	0.667	0.036	0.037	0	1
	Mixed Model	0.655	0.037	0.039	0	1	0.655	0.037	0.039	0	1
	RO	0.155	0.088	0.088	0.575	0.999	0.108	0.086	0.087	0.776	0.989
	BW	0.167	0.076	0.076	0.414	1	0.164	0.076	0.077	0.424	0.999
0.5	Unadj	0.667	0.036	0.038	0	1	0.666	0.036	0.037	0	1
	Mixed Model	0.655	0.037	0.039	0	1	0.655	0.037	0.038	0	1
	RO	0.14	0.095	0.094	0.71	1	0.084	0.093	0.096	0.853	1
	BW	0.162	0.076	0.075	0.424	1	0.168	0.076	0.08	0.408	1

Table 4.3: Results from the simulations where we a correlated confounder in the model. The correlation of this confounder within pair is equal to 0.9. The error terms have a standard extreme value type I distribution and a normal distribution with mean 0 and variance $\pi^2/6$, respectively.

From Table 4.3 and 4.4 we see that the unadjusted model is performing really badly. This is not surprising, since the confounder is ignored. The mixed model generates biased estimates. The between-within model is biased, it is more biased when the correlation is 0.7 than when it is 0.9. This agrees with the theoretical within-effect presented in Equation (2.21). When the correlation is equal to 0.7, the coverage is zero as a consequence of an overestimated coefficient. The rank-ordered logit model has the best performance among the models presented in the table. However, its estimates are still biased and the coverage is above 0.5 when the correlation is 0.9 but close to zero when the correlation is 0.7. Thus it is not a good analysis method when the confounder is not fully shared.

		ϵ ∼EVT1(0,1)					$\epsilon \sim N(0, \pi^2/6)$				
True β_1	Model	Avg Bias	Avg SD	Emp SD	Cov	Power	Avg Bias	Avg SD	Emp SD	Cov	Power
0	Unadj	0.666	0.036	0.036	0	–	0.667	0.036	0.036	0	–
	Mixed Model	0.658	0.037	0.036	0	–	0.66	0.037	0.037	0	–
	RO	0.344	0.076	0.075	0	–	0.317	0.075	0.073	0.004	–
	BW	0.372	0.068	0.068	0	–	0.375	0.068	0.067	0.001	–
0.25	Unadj	0.666	0.036	0.037	0	1	0.666	0.036	0.037	0	1
	Mixed Model	0.659	0.037	0.038	0	1	0.659	0.037	0.038	0	1
	RO	0.332	0.084	0.084	0.012	1	0.288	0.082	0.083	0.041	1
	BW	0.375	0.068	0.066	0	1	0.379	0.068	0.065	0	1
0.5	Unadj	0.665	0.036	0.037	0	1	0.666	0.036	0.037	0	1
	Mixed Model	0.657	0.037	0.037	0	1	0.659	0.037	0.037	0	1
	RO	0.313	0.095	0.093	0.058	1	0.262	0.092	0.088	0.158	1
	BW	0.368	0.068	0.065	0	1	0.376	0.068	0.065	0	1

Table 4.4: Results from the simulations where we a correlated confounder in the model. The correlation of this confounder within pair is equal to 0.7. The error terms have a standard extreme value type I distribution and a normal distribution with mean 0 and variance $\pi^2/6$, respectively.

4.3.2 Measurable Random Effect

A Genetic Effect

In Section 4.2.1 we assumed that the random effects were unmeasurable. If we instead assume that the random effects actually were measurable then we can use either an adjusted linear model where we include the measured random effect in the linear regression model, or the stratified rank-ordered logit and the stratified between-within models. In the stratified rank-ordered logit and between-within models we divide the population into 10 different clusters with respect to the random effects so that individuals with similar random effects belong to the same cluster. Thus, with these models the individuals within a cluster do not necessarily have a completely shared random effect.

		ϵ ∼EVT1(0,1)					$\epsilon \sim N(0, \pi^2/6)$				
True β_1	Model	Avg Bias	Avg SD	Emp SD	Cov	Power	Avg Bias	Avg SD	Emp SD	Cov	Power
0	Adj	−0.002	0.041	0.042	0.935	–	−0.001	0.041	0.04	0.953	–
	RO Match	0	0.032	0.033	0.942	–	0	0.033	0.032	0.956	–
	BW Match	−0.002	0.051	0.043	0.977	–	−0.001	0.051	0.041	0.988	–
0.25	Adj	0.001	0.041	0.039	0.949	1	−0.001	0.041	0.041	0.938	1
	RO Match	−0.007	0.033	0.032	0.954	1	−0.072	0.033	0.034	0.41	1
	BW Match	0.001	0.051	0.04	0.986	1	−0.001	0.051	0.042	0.982	1
0.5	Adj	0.001	0.041	0.039	0.953	1	0.002	0.041	0.041	0.94	1
	RO Match	−0.013	0.036	0.036	0.927	1	−0.138	0.034	0.037	0.024	1
	BW Match	0	0.051	0.04	0.984	1	0.002	0.051	0.042	0.979	1

Table 4.5: Results from the simulations of the model with a genetic effect that is shared within a pair of siblings, described in Section 4.2.1. The error terms have a standard extreme value type I distribution and a normal distribution with mean 0 and variance $\pi^2/6$, respectively. The results described assumes that the random effect is measurable and individuals within a cluster are matched on the genetic effect in the 'RO Match' and 'BW Match' models.

In Table 4.5 we present the results from the simulations where data are generated from Equation (4.1), where we have a genetic effect that is shared within a pair of siblings (and not within a cluster). We can conclude that the adjusted model works really well. Both stratified approaches works well when the errors are extreme value type I distributed. The between-within model performs well with both underlying error distributions in terms of bias, but with a bit too high coverage. The rank-ordered logit model is biased when the errors are incorrect with regard to the model assumptions. It performs worse

than in the paired analysis where the random effect was completely shared within a cluster and the coverage decrease for increasing effect size. When $\beta_1 = 0.5$ the model covers the true parameter in only 2.4 % of the simulations.

A Shared or Correlated Confounder

For the measurable confounders we present the simulations from two different set ups, when $\rho = 1$ and when $\rho = 0.9$ within a pair. We do not present the results of the models when the within-pair correlation is $\rho = 0.7$ since they are approximately the same as from the model with $\rho = 0.9$. That is, we generate data from Equation (4.1), where the confounders are generated from Equation (4.3) and the explanatory variables are generated from Equation (4.2). We apply the models described in Section 4.2.2 and the results of these simulations are presented in Table 4.6. Here, the confounder is completely shared within a pair, but not necessarily within a cluster in the rank-ordered logit and between-within models. Furthermore, for $\rho = 0.9$ we generate data from Equation (4.5), Equation (4.4) and Equation (4.6) and apply the models described in Section 4.2.2. The results are presented in Table 4.7.

		$\epsilon \sim$EVT1(0,1)					$\epsilon \sim N(0, \pi^2/6)$				
True β_1	Model	Avg Bias	Avg SD	Emp SD	Cov	Power	Avg Bias	Avg SD	Emp SD	Cov	Power
0	Adj	0	0.057	0.058	0.945	–	−0.002	0.058	0.057	0.956	–
	RO Match	0.069	0.044	0.043	0.678	–	0.052	0.044	0.044	0.792	–
	BW Match	0.075	0.056	0.057	0.74	–	0.075	0.056	0.056	0.731	–
0.25	Adj	0	0.058	0.058	0.943	0.988	−0.003	0.057	0.058	0.942	0.989
	RO Match	0.059	0.045	0.045	0.754	1	−0.021	0.045	0.046	0.915	1
	BW Match	0.076	0.056	0.057	0.729	0.999	0.074	0.056	0.057	0.737	1
0.5	Adj	−0.001	0.057	0.06	0.937	1	0.001	0.058	0.056	0.963	1
	RO Match	0.052	0.047	0.049	0.817	1	−0.088	0.046	0.048	0.503	1
	BW Match	0.074	0.056	0.059	0.728	1	0.077	0.056	0.056	0.708	1

Table 4.6: Results from the simulations of the model with a confounder which is completely shared within a pair of siblings. The individuals in the population are matched on their confounders and divided into 10 different strata in the 'RO match' and 'BW match' models. The error terms have a standard extreme value type I distribution and a normal distribution with mean 0 and variance $\pi^2/6$, respectively. The results described assumes that the random effect is measurable.

		$\epsilon \sim$EVT1(0,1)					$\epsilon \sim N(0, \pi^2/6)$				
True β_1	Model	Avg Bias	Avg SD	Emp SD	Cov	Power	Avg Bias	Avg SD	Emp SD	Cov	Power
0	Adj	0.002	0.057	0.059	0.947	–	−0.003	0.058	0.059	0.944	–
	RO Match	0.07	0.044	0.045	0.634	–	0.051	0.044	0.045	0.775	–
	BW Match	0.078	0.056	0.059	0.698	–	0.074	0.056	0.058	0.729	–
0.25	Adj	−0.003	0.057	0.056	0.955	0.989	−0.001	0.058	0.055	0.955	0.996
	RO Match	0.058	0.045	0.045	0.755	1	−0.02	0.045	0.045	0.922	0.999
	BW Match	0.073	0.056	0.056	0.732	1	0.076	0.056	0.056	0.737	1
0.5	Adj	0	0.057	0.056	0.964	1	0.001	0.057	0.057	0.951	1
	RO Match	0.053	0.047	0.048	0.8	1	−0.086	0.046	0.048	0.522	1
	BW Match	0.075	0.056	0.056	0.736	1	0.077	0.056	0.055	0.721	1

Table 4.7: Results from the simulations where we a correlated confounder within a pair of siblings, $\rho = 0.9$. The error terms have a standard extreme value type I distribution and normal distribution with mean 0 and variance $\pi^2/6$, respectively. The results described assumes that the random effect is measurable.

From the simulation results which are presented in Table 4.6 and Table 4.7 we see that none of the models are affected by the degree of correlation of the confounder within a pair. This is not surprising since we use a stratified approach where two individuals who are generated as a pair does not necessarily

belong to the same stratum/cluster. We can see that with regard to bias the rank-ordered logit model performs better than the between-within model when the underlying error distribution is the extreme value type I distribution, even though the difference between the two models are small in that setting. Also the coverage is slightly higher for the rank-ordered logit model. For the normally distributed error terms the rank-ordered logit model is less biased when the magnitude of the explanatory variable on the outcome is low, while when the impact of the explanatory variable is 0.5 the between-within model has a better performance, both with regard to coverage and bias since the estimates are slightly less biased. The estimates of the rank-ordered logit model are more precise than the estimates of the between-within model. As expected the adjusted model that includes all covariates works fine in both set ups.

Thus, when the confounder is not completely shared within a pair, the estimates in these matched designs generates less biased estimates than the analysis approaches presented in Section 4.3.1. Thus, if we have a measurable confounder or another measurable factor that is strongly correlated with the confounder we want to adjust for, we can conclude from our simulations that the estimates of the between-within and the rank-ordered logit models are less biased when we match the individuals into larger clusters than when we use the paired design unless the confounder is completely shared within a pair.

Chapter 5

Real Data Analysis

In this thesis we apply the rank-ordered logit and the between-within regression models on two different datasets, which are presented in this chapter. The first dataset contain collected blood samples from patients that have been admitted at National University Hospital in Singapore and we study the association between the variation in measurement and the daily reading frequency. The other dataset contains mammographic density measurements from the Karolinska mammography project for risk prediction of breast cancer (KAR, 2015) and we study the impact of different risk factors on mammographic density, which is known to be associated with breast cancer.

5.1 Blood Glucose Study

The dataset is provided by the National University Hospital in Singapore. Data of blood glucose levels were monitored in 2012 at the National University Hospital for 6298 patients who were on capillary blood glucose monitoring. The patients were admitted in adult non-critical care wards (Tan et al.). The collected data contains information about the blood glucose level, the time of the measurement, the ward where the patient is admitted, the patients gender and date of birth.

These data have already been analysed by Tan et al. with the standard deviation of the glucose measurements in the first monitoring period as a response variable and the mean daily frequency of measurements as explanatory variable. The data was assumed to follow a linear model and different analysis approaches were applied. The confounders that were of interest to adjust for were age, sex and length of first monitoring period, where a monitoring period is defined as a period when there is no more than 48 hours between two subsequent measurements. First, Tan et al. fitted a linear model with normally distributed error terms, both an unadjusted model which only included the explanatory variable and an adjusted model which in addition included the potential confounders with and without interactions. Second, a linear model which assumes extreme value type I distributed errors, both unadjusted and adjusted. They also applied the rank-ordered logit model, where the population is matched on the confounders mentioned above. They found a positive association between the standard error and the mean frequency.

In this thesis we complement the analysis of the association between the standard deviation of the glucose measurements in the first monitoring period and the mean daily frequency of measurements with the between-within model as an alternative approach.

We use five different models on the data when we study the association. In all models we adjust for the potential confounders age, gender and length of the first monitoring period. The patients where categorized by age: 60 years or younger, between 60 and 75, and older than 75 years old. They where also categorized after the length of their first monitoring episode: less than 3 days, between 3 and 6 days

and more than 6 days.

We apply a linear model which assumes normally distributed error terms, where we only include the intercept, the covariate we are studying and the categorized confounders: age, sex and length of first monitoring period. We thus adjust for the confounders by modelling them. We fit the linear models in `R` with the `lm` function. The data are also analysed with linear regression models for extreme value type I distributed data with the same factors as the adjusted linear model, by maximizing the log likelihood in Equation (3.21) with the `optim` function. The estimate from this model is scaled in order to be comparable to the estimate of the rank-ordered logit model and the delta method is used in order to estimate the standard error of this estimate. The data were also analysed with the between-within and the rank-ordered logit models by dividing the data into clusters. With respect to the categories of age, length of first monitoring period and gender we obtained eighteen different clusters. For the between-within model we calculate the cluster average of the explanatory variable and fit the model assuming normally distributed errors and by including a random intercept with the `lmer` function. We also studied the between-within model when the error terms are assumed to follow an extreme value type I distribution. The rank-ordered logit model were fitted with the `coxph` function, where we included the explanatory variable in the linear predictor and used the matched clusters as strata. For all models we estimate the parameter of the explanatory variable that is of interest, together with a confidence interval of the estimate and the Wald p-value of the estimate.

Using similar criteria as Tan et al. to select patients for analysis, patients who were younger than 25 years old on the first day of their first monitoring period, whose first monitoring period was less than two days or whose average number of readings per day were less than four in the first monitoring period were excluded from the analysis. Our final study population consisted of 2113 patients.

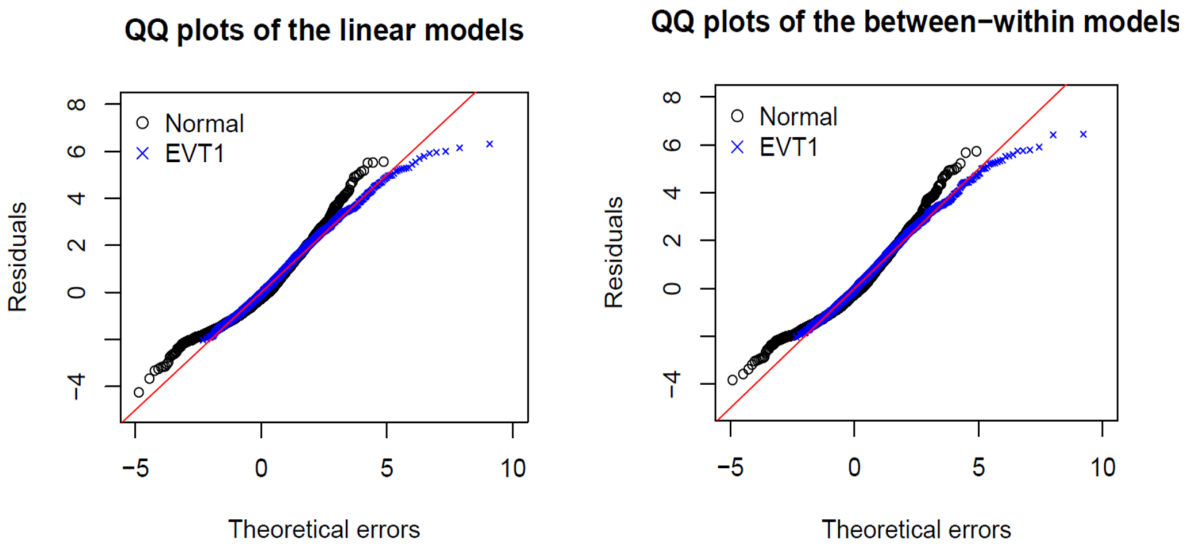
The estimates from the different analysis approaches are displayed in Table 5.1. We see that the estimates from the models assuming normally distributed error terms are similar and equal to 0.51. We fit the between-within model defined as in Equation (2.15), i.e. with a random intercept when we assume that the error distribution is normal. When we assume the errors to be extreme value type I distributed we do not include a random intercept. The estimates of the models that assumes extreme value type I distributed errors are equal to 0.253, 0.26 and 0.262, which are approximately equal. From the between-effect estimates from the between-within model we see that there are differences between the clusters.

Model	Estimate	Confidence interval 95%	P-value
Linear model, normal error	0.513	(0.431, 0.594)	< 0.001
Linear model, EVT error	0.253	(0.203, 0.304)	< 0.001
Rank-ordered logit model	0.26	(0.208, 0.312)	< 0.001
Between-within, normal error *	0.508 [−1.951]	(0.426 0.589) [−2.396, −1.515]	< 0.001 [< 0.001]
Between-within, EVT error *	0.262 [−1.792]	(0.21, 0.314) [−2.807, −1.095]	< 0.001 [< 0.001]

Table 5.1: The results from the blood glucose study with exposure the daily frequency of readings during the first monitoring period and response the standard deviation of the measured results the first monitoring period. The estimates from the linear models which assumes extreme value type I distributed error terms are scaled by the estimated scale parameter, which is equal to 1.088.
*The between estimate is presented within the brackets.

We study the residuals from the linear models in QQ plots against the theoretical quantiles as basis in the model choice. The residuals of the linear model are then used as a indicator of if we can assume that the error terms are extreme value type I or normally distributed.

The QQ plots for the association between the standard deviation of the measurements and the mean daily frequency are shown in Figure 5.1. In Figure 5.1a we see that for the linear models the QQ plots indicate a better fit for the model which assume an extreme value type I distribution for the errors. For



(a) QQ plot from the linear models. For the observed quantiles against the theoretical quantiles of the extreme value type I distribution and of the normal distribution, respectively.

(b) QQ plot of the between-within model. The residulas from the models which assumes normally and extreme value type I distributed error terms, respectively against the corresponding theoretical quantiles.

Figure 5.1: QQ plot from the models fitted to the association between the standard deviation of the measured results the first monitoring period and the daily frequency of readings during the first monitoring period.

the between-within models in Figure 5.1b, the residuals are better for the model with extreme value type I distributed errors than for the model when we assume normally distributed errors. With regard to the QQ plots, the models with extreme value type I distributed errors in Figure 5.1 seems to fit data better. Therefore, the estimates from the rank-ordered logit model are preferable to the estimates of the between-within model which assumes normally distributed error.

From Table 5.1 we see that all three models which assume an extreme value type I error distribution provide estimates approximately equal to 0.26, provided we scale the estimates from the extreme value type I linear models using the estimated scale parameter. The estimated scale parameter in the regression model with extreme value type I distributed error terms is equal to 1.088, so it is actually close to 1. We can therefore interpret the coefficient of the rank-ordered logit model as in the linear model with extreme value type I distributed error. The p-values from the Wald tests reject the null hypothesis that the parameters are equal to zero for conventional significance levels, thus we have a statistically significant positive effect from the reading frequency on the standard deviation of the results. That is, if the blood glucose levels are unstable and thus vary a lot, the reading frequency goes up. A increase of 0.26 in standard deviation is associated with a one unit increase of mean daily reading frequency when all other factors remain unchanged.

5.2 Mammographic Density Study

The data in this study is from the Karolinska mammography project for risk prediction of breast cancer, or in short the "Karma study", where the study subjects attend one of the participating hospitals in Sweden for mammography screening or clinical mammography (KAR, 2015). In April 2015 the study had 70877 participants, who had filled out a comprehensive questionnaire, had their blood pressure measured and donated blood. The Karma project saves the images from the mammograms and calculates the density. The measurements are done automatically with the software Volpara (Brand et al., 2014). More

details of the study can be found at (KAR, 2015, http://karmastudy.org). We use a study population which consists of siblings, both half-siblings and full-siblings, who have complete information on the mammographic density measures of 3989 individuals. The aim of this study is to determine which risk factors that are associated with mammographic density by estimating the direct effect of the different risk factors on percent dense volume. We compare the results when we adjust for genetic and maternal factors which are shared by two sisters and when we study unrelated women in order to determine if the genetic factors influence the association.

As response variable we use the percent dense volume as the measurement of mammographic density, this is the quotient between absolute dense volume and the total volume of the breast. The risk factors that we study are the age at menarche, menopausal status, ever given birth, age at first birth, hormone replacement therapy status and if the woman have a history of benign breast disease. Menopause status has three different categories: pre-menopausal, peri-menopausal and post-menopausal, where menopause are defined by not having any periods during the last year, previous oophorectomy and older than or equal to 55 years old (Brand et al., 2014). In our analysis we merge the peri- and post menopausal women into one group and refer to them as post menopausal throughout this thesis. The variable hormone replacement therapy status has two different states: ever used or have never used. When we study the risk factor age at first birth we only investigate women who have given birth. Since the mammographic density is correlated with both body mass index (bmi) and age, we adjust for these confounders in the analysis (Vachon et al., 2007). The bmi is defined as the woman's weight in kilogram divided by the square of her length measured in meters. We compare the results from the analysis of unrelated women and pairs of sisters in order to estimate the direct effect of these factors, adjusted for confounding by shared childhood environment and genetic factors.

We use similar exclusion criteria as in the paper by Brand et al. (2014), namely we consider women younger than 74 and older than 40, exclude those with a history of any cancer (other than non-melanoma skin cancer), or of breast enlargement, breast reduction or other breast surgery. We also exclude those who did not have their period during the last year because of pregnancy or breast feeding. If data is missing for any of the explanatory variables or confounding variables of interest this individual is excluded from our analysis. For the analysis of siblings we use the reduced study population and we exclude both sisters if at least one of them are not fulfilling the criteria.

Our study population consists of 2960 women who fulfilled the criteria and including 869 pairs of sisters. By randomly excluding one of the sisters in each pair of siblings we have a study population of 2081 women who are assumed to be unrelated.

In Table 5.2 the characteristics of the variables in the data of all women who are fulfilling the criteria are presented. The characteristics of the subset with only unrelated women have similar characteristics and is therefore not presented.

For the subset of sisters we have a sample with characteristics similar to those presented in Table 5.2, it is of interest to investigate the differences or similarities between two sisters. The pairwise differences of the continuous characteristics are therefore presented in Table 5.3, where also the p-values of pairwise t-tests for the continuous variables are given. The difference we are studying is between the value of the risk factor for the older sister minus the value of the risk factor for the younger sister. With the t-tests we are testing the null hypothesis of the difference in mean is equal to zero. We test this against the alternative hypothesis of the difference in mean is not equal to zero. We see from the p-values that we cannot reject the null hypothesis for bmi and age at menarche. We can reject the null hypothesis for the factors percent dense volume, age and age at first birth. The mean age difference for two sisters are 4.983, furthermore, a younger sister is on average 0.787 years younger than her older sister when she has her first child. The mean difference in percent dense volume is also statistically significantly lower for the younger sister.

Variable	All women (n=2960)	
Percent dense volume*	8.923	(5.135 ; 11.387)
Age	54.676	(9.062)
Bmi	25.447	(4.368)
Age at menarche	13.176	(1.461)
Ever given birth, No. (%)		
Yes	2616	(88.378)
No	344	(11.622)
Age at first birth**	26.532	(5.033)
Menopause status, No. (%)		
Pre	1246	(42.095)
Post	1714	(57.905)
Hormone replacement therapy, No. (%)		
Never	2396	(80.946)
Ever	564	(0.191)
Benign breast disease, No. (%)		
Yes	509	(17.196)
No	2451	(82.804)

Table 5.2: Characteristics for the women in the data sample of all women that fulfilled the requirements, the mean and standard deviation are given.
*The 25% and 75% quantiles are given within the brackets.
**Based on the subset of women who have given birth (n^{**} =2616).

Variable	Sisters (n_1=1758)		
	Difference	Confidence interval	P-value
Percent dense volume	−1.035	(−1.415 ; −0.656)	< 0.001
Age	4.983	(4.743 ; 5.223)	< 0.001
Bmi	0.246	(−0.1 ; 0.591)	0.164
Age at menarche	0.058	(−0.062 ; 0.178)	0.341
Age at first birth**	−0.787	(−1.219 ; −0.355)	< 0.001

Table 5.3: The difference in characteristics among all pairs of sisters that fulfilled the requirements.
**We only look at the pair of siblings where both sisters have given birth (n_1^{**}=1364).

We log-transform the data in order to get a better normal approximation. The motivation behind this can be found in Appendix A, where we fit linear models with all covariates and present the QQ plots of the non-transformed and transformed data. We also present the QQ plots of the theoretical quantiles of the extreme value type I distribution against the residuals of the fitted linear regression of the non-transformed data when we assume extreme value type I distributed errors.

Our first analysis approach is to fit a linear model for each risk factor separately onto the data of unrelated women, in order to detect a possible association between the risk factor and the log-transformed measurement of mammographic density. We also compare the estimated effect of the risk factor in this univariate approach with the estimate of the risk factor when we adjust for the potential confounders age and bmi. The results of these initial analysis approaches are presented in Table 5.4.

We see in Table 5.4 that in the crude models all variables are significant except the indicator of ever given birth. We compare the crude and the adjusted models with a likelihood ratio test described in Section 4.1.2, the p-values from these test are very low ($\approx 10^{-200}$), which indicates that we should adjust for age and bmi. When we adjust for the confounders in the model we see that the age at menarche is no longer significant since it has a p-value equal to 0.425. This implies that we cannot reject the null hypothesis of the coefficient being equal to zero. We see from this initial analysis that menopause status is an important factor. The significance of the hormone replacement therapy is questionable, but since hormone replacement therapy is only of interest among women who are post menopause we continue the analysis by only considering post menopausal women. Furthermore, we see in Table 5.4 that the factor given birth is not significantly different from zero even when we adjust for confounders. Since women

Risk factor	Unrelated women (n_2=2081)					
	Crude model			Adjusted model		
	Estimate	(Std. Err.)	P-value	Estimate	(Std. Err.)	P-value
Age	−0.018	(0.001)	< 0.001	–	–	–
Bmi	−0.07	(0.002)	< 0.001	–	–	–
Age at Menarche	0.022	(0.008)	0.007	0.005	(0.006)	0.425
Hormone replacement therapy	−0.108	(0.03)	< 0.001	0.044	(0.025)	0.074
Benign breast disease	0.082	(0.03)	0.007	0.091	(0.023)	< 0.001
Menopause status	−0.351	(0.023)	< 0.001	−0.195	(0.029)	< 0.001
Given birth	−0.011	(0.037)	0.771	−0.041	(0.028)	0.147
Age at first birth**	0.024	(0.002)	< 0.001	0.007	(0.002)	< 0.001

Table 5.4: The estimates from the linear models where we fit the association between the outcome and each risk factor by an unadjusted (crude) model and a model where we also adjust for age and bmi.
**We only look at women who have given birth ($n_2^{**} = 1842$).

who are post menopause will not have any more children, the factor given birth is more fairly comparable in the sample of post menopausal women.

From Table 5.2 we see that among all women 1714 are post menopausal. Among the subset of unrelated women we have 1226 post menopausal women. We repeat the adjusted analysis presented in Table 5.4 on our subsets of post menopausal women, both among all post menopausal women and among unrelated post menopausal women. The results are presented in Table 5.5.

Post menopause	Adjusted model					
	Unrelated women (n_2=1226)			All women (n=1714)		
Risk factor	Estimate	(Std. Err.)	P-value	Estimate	(Std. Err.)	P-value
Age at Menarche	0.003	(0.008)	0.672	0.009	(0.007)	0.168
Hormone replacement therapy	0.062	(0.026)	0.016	0.057	(0.022)	0.009
Benign breast disease	0.094	(0.028)	< 0.001	0.095	(0.025)	< 0.001
Given birth	−0.073	(0.035)	0.039	−0.048	(0.03)	0.113
Age at first birth**	0.006	(0.003)	0.011	0.006	(0.002)	0.005

Table 5.5: The estimates from the linear models where we fit the association between the outcome, log percent dense volume, and each risk factor in a model where we adjust for age and bmi among post menopausal women.
**We only look at women who have given birth ($n_2^{**} = 1082$ and $n^{**} = 1513$, respectively).

From Table 5.5 we can see that all risk factors but age at menarche are significantly different from zero for the post menopausal women when we adjust for age and bmi in the sample of unrelated women. When we include all post menopausal women the indicator for given birth is not significantly different from zero and in both samples the estimates of the risk factor are negative and small. Overall the estimated parameters are approximately equal in the two samples.

By comparing the estimates in the sample of unrelated women in Table 5.5 with the corresponding estimates in Table 5.4 we see that the magnitude of all risk factors are approximately equal in the two subsets, but the factor hormone replacement therapy is of greater magnitude among post menopausal women.

From these initial analyses we can conclude that we need to adjust for age and bmi when we study the other risk factors. The risk factors hormone replacement therapy, benign breast disease and age at first birth have a positive association with percent dense volume among post menopausal women. The exponential function of the coefficient for the risk factor gives us the estimated multiplicative effect on percent dense volume from a one unit increase of the risk factor when age and bmi remains constant. For example, a women who had had hormone replacement therapy have a multiplicative effect of exp(0.062)=1.064 on percent dense volume compared to a woman with the same age and bmi who never had hormone replacement therapy. Age at menarche does not have a statistically significant effect on the

outcome when we adjust for age and bmi. From the p-values of the indicator variable given birth we have a statistically significant negative effect among unrelated women but the same factor is not significant on the 5 % level in the whole sample, even though there is an indicated effect.

We now want to adjust for unmeasurable confounders with the between-within and rank-ordered logit models applied on pairs of sisters. In both models we also adjust for the confounders age and bmi by including them in the linear predictor. In the between-within model we calculate the average of the risk factor within the pair of sisters. Thus, we adjust both for the measurable confounders and the unmeasurable variables that are shared within a pair of siblings. The within- and the between-estimate of the between-within analysis are presented in Table 5.6, where we have fitted the model

$$\log(\text{PDV}_{ij}) = \beta_0 + \beta_W \text{Risk factor}_{ij} + \beta_B^* \overline{\text{Risk factor}}_i + \beta_2 \text{Age}_{ij} + \beta_3 \text{Bmi}_{ij} + \epsilon_{ij}$$

where i denotes the pair and j denotes the sister, $j = 1, 2$. From the theory presented in Section 2.3 we know that if $\beta_B^* = 0$, then we have a model with only an explanatory variable and no cluster specific effect.

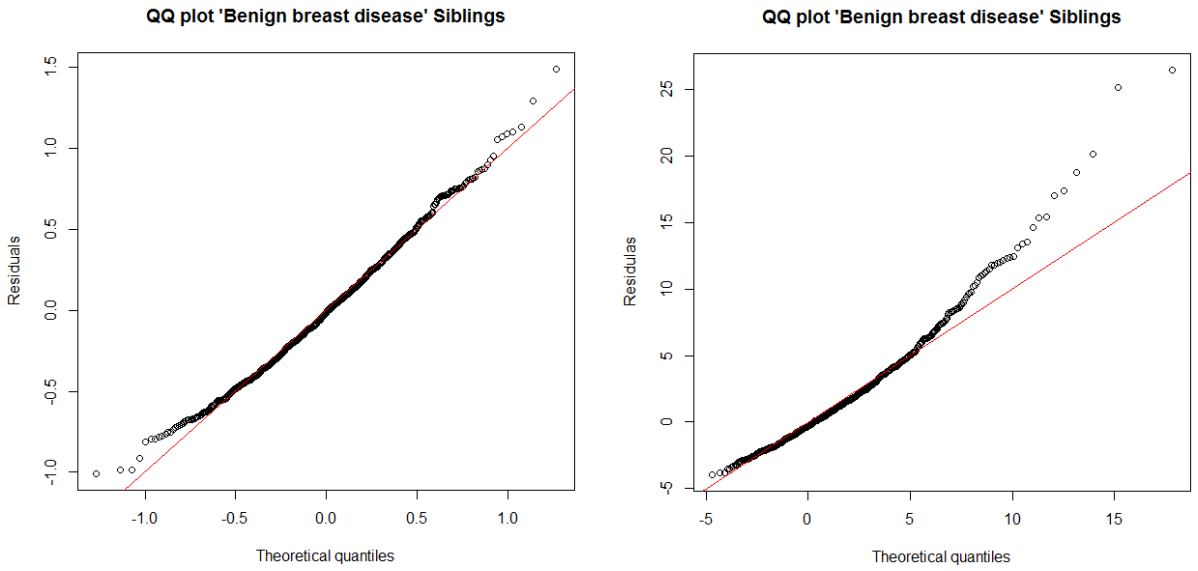
We present the results of the between-within model without a random intercept, since the estimates of the model with and the model without a random intercept generates approximately the same estimates. From Table 5.6 we can see that the between-effect is not significantly different from zero for any risk factor, this indicate that we only have a within-effect. This imply that the within-effect estimates of the between-within model should be similar to those from the adjusted linear model. We have to keep in mind that we use two different samples in Table 5.5 and Table 5.6. In the between-within model the risk factors age at menarche and the indicator of given birth are not significantly different from zero. This is similar to the results of the adjusted model of all women.

Sisters (n=792)	Between-within						Rank-ordered logit		
	β_W			β_B^*			β_1		
Risk factor	Est.	(Std. Err.)	P-val.	Est.	(Std. Err.)	P-val.	Est.	(Std. Err.)	P-val.
Age at men.	0.006	(0.016)	0.68	0.017	(0.02)	0.38	0.049	(0.064)	0.448
Hormone repl. therapy	0.101	(0.047)	0.03	−0.068	(0.061)	0.269	0.367	(0.194)	0.059
Benign breast disease	0.143	(0.051)	0.005	−0.082	(0.074)	0.266	0.205	(0.204)	0.315
Given birth	−0.003	(0.063)	0.959	−0.048	(0.086)	0.575	−0.083	(0.254)	0.743
Age at first birth**	0.011	(0.006)	0.042	−0.006	(0.007)	0.423	0.045	(0.024)	0.061

Table 5.6: The estimates from the between-within and rank-ordered logit models adjusted for bmi and age by including them in the linear model when we investigate the association between the outcome, log percent dense volume and each risk factor in a model among post menopausal women. HRT denotes hormone replacement therapy and BBD denotes benign breast disease.
The estimate is based on the population of sisters that both have given birth (n=616).

In Table 5.6 the rank-ordered logit estimates are presented. None of the estimates of the rank-ordered logit model are significant on the 5% level, but the factors hormone replacement therapy and age at first birth are significant on the 10% level.

In Figure 5.2a we show the QQ plots of the residuals of the between-within model plotted against the theoretical quantiles when we study the association between percent dense volume and benign breast disease. In order to see if it would be appropriate to assume that the error distribution is extreme value type I we fitted an adjusted linear model with the covariate benign breast disease, age and bmi with extreme value type I distributed errors against the theoretical quantiles in Figure 5.2b. The corresponding QQ plots of the other risk factors were approximately the same, they are therefore not included here. From the figures we can conclude that a normal approximation of the error terms are reasonable, while an extreme value type I distribution is not.



(a) QQ plot from the between-within model describing the association between log percent dense volume and benign breast disease among siblings, when we adjusted for age and bmi.

(b) QQ plot from the linear model describing the association between percent dense volume and benign breast disease with extreme value type I distributed errors, when we adjust for age and bmi.

Figure 5.2: QQ plots from the models fitted to the association between the percent dense volume and the risk factor benign breast disease when we adjust for age and bmi.

The risk factor benign breast disease is not significant in the rank-ordered logit model. That this factor is not significant might be a consequence of the fact that the errors do not seem to follow an extreme value type I distribution. Furthermore, when we have dichotomous variables only the discordant pairs contribute with any information. I.e. only the pairs where one sister has had benign breast disease and the other sister has not had benign breast disease is informative since otherwise the risk factor is eliminated in the model (see Equation (3.18)). In total we have 120 discordant pairs for the risk factor benign breast disease. From the adjusted linear model with extreme value type I distributed errors with the risk factor benign breast disease, the scale parameter was estimated to 2.40 while the scaled coefficient was 0.26 which is larger but still similar to the estimate of the rank-ordered logit model in Table 5.6.

If we look at the variable benign breast disease, the within-effect estimate is interpreted as a multiplicative effect. A person who had had benign breast disease has a multiplicative effect of exp(0.143)=1.154 on the percent dense volume compared to her sister if the sister never had benign breast disease, when all other factors are equal for the two siblings. The estimates of the rank-ordered logit model are scaled. Therefore, we cannot compare the size of the estimates of the rank-ordered logit model with the estimates of the between-within model. The estimate of the rank-ordered logit model can be interpreted as a log odds ratio. The conclusion we can make is that the odds that a woman with a history of benign breast disease has a higher percent dense volume than a woman of the same age and bmi who never had benign breast disease is $\exp(0.205) = 1.228$.

Since the results of the between-within analysis on siblings are telling us that the between-effect is not significantly different from zero in Table 5.6, this indicates that the genetic effect that we try to adjust for with the between-within model do not have a big impact on the outcome. We therefore compare the results of the sibling comparison to those when we pair two women at random in the population of unrelated women, to see if we get similar estimates. We adjust for age and bmi in the linear predictor. The results of these comparisons are presented in Table 5.7.

The results presented in Table 5.7 show that we do not have any between-effects when we study

Unrelated (n=1226)	Between-within						Rank-ordered logit		
	β_W			β_B^*			β_1		
Risk factor	Est.	(Std. Err.)	P-val.	Est.	(Std. Err.)	P-val.	Est.	(Std. Err.)	P-val.
Age at men.	0.014	(0.011)	0.204	−0.021	(0.015)	0.168	0.043	(0.045)	0.338
Hormone repl. therapy	0.031	(0.036)	0.381	0.062	(0.05)	0.217	0.092	(0.15)	0.536
Benign breast disease	0.041	(0.041)	< 0.001	0.105	(0.057)	0.063	0.209	(0.168)	0.212
Given birth	−0.117	(0.051)	0.022	0.084	(0.071)	0.233	−0.414	(0.216)	0.055
Age at first birth**	0.009	(0.004)	0.013	−0.003	(0.005)	0.556	0.027	(0.016)	0.094

Table 5.7: The estimates from the between-within and the rank-ordered logit models on a paired design, where the women who represent a pair is chosen at random. In these models we have adjusted for bmi and age by including them in the linear model when we investigate the association between the outcome, log percent dense volume, and each risk factor in a model among post menopausal women.
**The estimate is based on the population of unrelated women where both women within a pair have given birth (n^{**}=960).

two randomly chosen individuals, as expected. The between-within model estimates are from the model representation without a random intercept since the estimates and standard errors were similar when we included and when we excluded the random intercept. The QQ plots of the between-within model and the adjusted linear model with extreme value type I errors are shown in Appendix A.2 for the risk factor benign breast disease. Just as in the sample of sisters the assumption of normally distributed errors seem reasonable but the assumption of extreme value type I distributed errors do not. The difference in results between the study of siblings and the study of random individuals is that we in Table 5.7 do not get a significant result of the factor hormone replacement therapy but instead we find the factor given birth significant in the between-within model. By performing a likelihood ratio test between the models in Table 5.5 and the between-within models in Table 5.7 we cannot reject the null hypothesis of the between-effect being equal to zero on the 5 % level since the p-values are equal to 0.167, 0.216, 0.063 and 0.232 for the first four factors presented in the same order as in the table and equal to 0.555 for the factor age at first birth. It is therefore better to study the results from the linear adjusted model than the between-within model for unrelated women.

The rank-ordered logit model show that none of the risk factors are significantly different from zero. The standard errors of the estimates are greater than from the between-within model. Since the underlying assumptions of the model are not fulfilled, i.e. the error distribution does not follow an extreme value type I distribution, we put more trust in the results from the between-within model in this analysis.

From this analysis we have shown that the results of the adjusted linear model for unrelated women are approximately the same as those from the sibling analysis when we adjust for genetic factors.

We also apply a stratified analysis where we match on the confounders age and bmi. In this setting we do not need to include the specific values in the model fitting, we do not need to model the functional form of the confounders we match upon neither. From the simulations presented in Section 4.2 we saw that if there are confounders that are not completely shared between sisters, a stratified approach is less biased.

We divide persons into different risk groups based on their bmi value, one way of doing this is to use the following groups: 'Underweight' (bmi$<$ 18.5), 'Normal weight' ($18.5 \leq$ bmi ≤ 24.9), 'Pre-obesity' ($25 \leq$ bmi ≤ 29.9), 'Obesity class 1' ($30 \leq$ bmi ≤ 34.9), 'Obesity class II' ($35 \leq$ bmi ≤ 39.9) and 'Obesity class III' ($40 \leq$ bmi) which are defined by WHO (2015). Furthermore, we divide the population into different age groups, where one group corresponds to a certain decade. The post menopausal women are between 40 and 74 years old and are divided into four different age groups: $40-49$, $50-59$, $60-69$ and $70-74$ years old. One combination of age group and bmi group contained 0 observations and therefore this partition resulted in 23 different groups, each containing at least 2 observations. For the analysis of

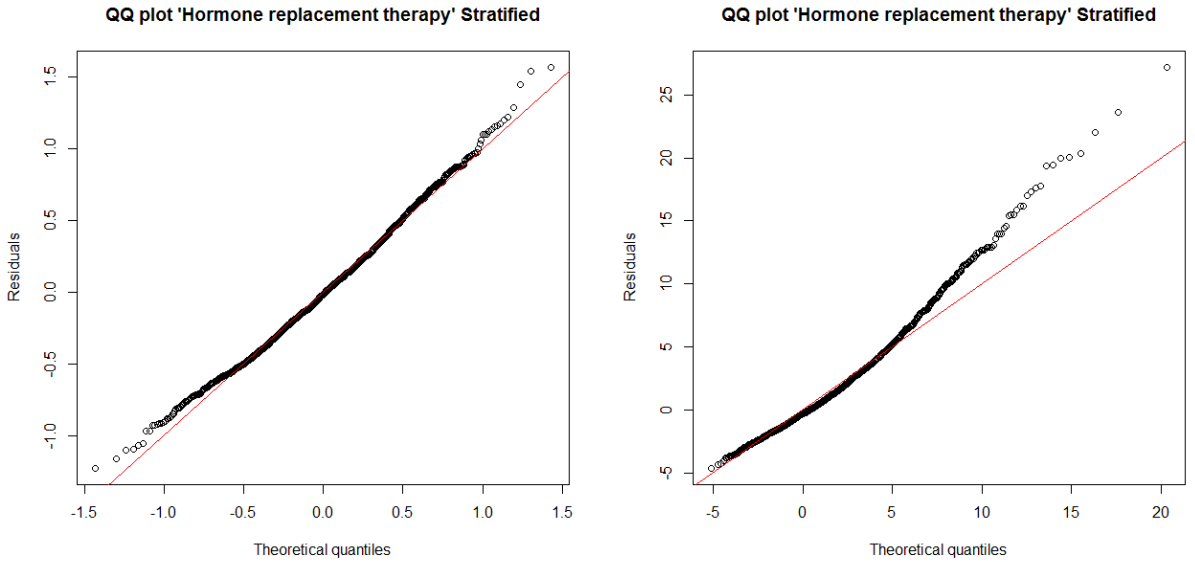
age at first birth the population were divided into 22 different clusters.

All women (n=1714)	Stratified Between-within						Stratified Rank-ordered logit		
	β_W			β_B^*			β_1		
Risk factor	Est.	(Std. Err.)	P-val.	Est.	(Std. Err.)	P-val.	Est.	(Std. Err.)	P-val.
Age at men.	0.012	(0.007)	0.078	0.331	(0.097)	0.002	0.022	(0.017)	0.193
Hormone repl. therapy	0.055	(0.023)	0.015	0.873	(0.735)	0.247	0.115	(0.055)	0.037
Benign breast disease	0.11	(0.026)	< 0.001	-0.239	(0.833)	0.777	0.247	(0.063)	< 0.001
Given birth	-0.046	(0.032)	0.143	-0.423	(0.703)	0.554	-0.074	(0.077)	0.336
Age at first birth**	0.007	(0.002)	0.004	0.117	(0.04)	0.01	0.012	(0.006)	0.031

Table 5.8: The estimates from the between-within and the rank-ordered logit models when we have stratified groups, i.e. we divide the population into groups of women with respect to their bmi group and age. We investigate the association between the outcome, log percent dense volume, and each risk factor in a model among all post menopausal women. HRT denotes hormone replacement therapy and BBD denotes benign breast disease.
The estimate is based on the population of women that have given birth (n=1512).

In Table 5.8 the results of the stratified analysis are presented. The applied between-within model have a random intercept. For the risk factor age at menarche and age at first birth the between-effect is significant. Thus, the risk factors have different effects on the outcome in different combinations of age and bmi. The risk factors whose within-effect is significant is similar in the two models. The within-effects of hormone replacement therapy, benign breast disease and age at first birth are significant and positive on the 5 % level in both the between-within model and in the rank-ordered logit model. We see that the same risk factors are significant in the stratified between-within model and in the between-within model on the sample of siblings.

In Figure 5.3 the QQ plots of the between-within and the corresponding linear model with extreme



(a) QQ plot from the between-within model describing the association between log percent dense volume and hormone replacement therapy in the stratified analysis, with normally distributed errors.

(b) QQ plot from the linear model describing the association between percent dense volume and hormone replacement therapy when we adjust for age group and bmi group, with extreme value type I distributed errors.

Figure 5.3: QQ plots from the models fitted to the association between the percent dense volume and the risk factor hormone replacement therapy when we adjust for the groups of age and bmi, where we use the same groups as in the stratified analysis.

value type I distributed errors where we adjust for the same categorical groups of confounders by including them in the linear model are shown. In the linear model with extreme value type I distributed errors the percent dense volume is not log-transformed since such a transformation did not improve the fit of the model. In the figure the plots are from the studied association between percent dense volume and the risk factor hormone replacement therapy, the QQ plots of the other risk factors are similar. We can conclude that the error terms in these data can not be considered extreme value type I distributed. In the association between hormone replacement therapy and percent dense volume the scaled estimate of the linear model with extreme value type I distributed errors were equal to 0.094 and the scale parameter where estimated to 2.485. The scaled estimate is similar to the one in the stratified rank-ordered logit model.

From all these analyses we have seen that as long as we adjust for age and bmi we do not gain any further information about the associations studied by adjusting for shared confounders within a pair of siblings. The conclusions made from the adjusted linear model, the between-within model applied on sisters and the stratified analyses are similar. The risk factors hormone replacement therapy, benign breast disease and age at first birth have a statistically significant positive association with the percent dense volume in these models. However, the estimates of the rank-ordered logit model are larger than those from the models in which we assume normally distributed errors. Since the error distribution does not seem to follow an extreme value type I distribution and are scaled the estimated magnitude of the parameters in the rank-ordered logit model are not reliable.

The significant risk factors within-effects, i.e. the effects of hormone replacement therapy, benign breast disease and age at first birth are approximately equal in the stratified between-within model, the sibling analysis and in the adjusted model. Thus, we can use anyone of these models in order to determine the importance of the risk factors.

Chapter 6

Discussion and Conclusion

In this thesis we have compared the between-within and the rank-ordered logit regression models used to adjust for potential confounders when we study the association between a continuous outcome and an explanatory variable. The advantage of these models is that we can adjust for shared confounders that are not measurable by dividing the population into different clusters and adjust for confounders without making any assumptions of their functional form. In Chapter 2 we defined the between-within model, which is a generalized linear mixed model and a common analysis method in the epidemiological field. We have focused on the model when we have an identity link and in most applications we have assumed that the error terms follows a normal distribution. The rank-ordered logit model we described in Chapter 3, is developed in econometrics literature where the interest could be to investigate the demand for a new product on the market, see for example Beggs et al. (1981). In this kind of applications, we cannot measure the outcome which is the utility of a certain choice, instead we observe the ranked outcomes. By assuming that the outcome can be explained by a linear predictor and an unmeasurable error term, the rank-ordered logit model is derived under the assumption that the error terms have a standard extreme value type I distribution. With this model it is possible to estimate the effects of the different covariates in the linear predictor.

We have seen in the example presented in Section 3.1.1 that *analysis approaches that assumes normally distributed errors are more robust than models which assumes extreme value type I distributed errors.* Thus, models which assumes that the errors have a normal distribution have an advantage against models which assumes extreme value type I distributed errors. For the linear model in the example, the approach with errors from the extreme value type I distribution suffered from underestimated standard error of the estimates, when the error in fact was normally distributed.

In order to investigate the strengths and weaknesses of the between-within and the rank-ordered logit models we performed simulation studies which we presented in Chapter 4.2. We studied the models ability to estimate the coefficient of the explanatory variable we were interested in under different circumstances.

When we have a random effect that influences the outcome but is independent of the exposure and this effect is completely shared within a cluster, both models performed well when the error terms were extreme value type I distributed. When the error terms followed a normal distribution, the rank-ordered logit model generated slightly biased estimates with lower coverage and power than the between-within model. In this simple set up we saw that we could just as well have used a linear model with the exposure variable as the only effect and the same holds for the mixed model where we in addition allowed a random intercept.

In the next model under investigation we converted the random effect into a confounder, we still looked at the situation when the confounder was completely shared within a cluster. We saw once again that neither the between-within nor the rank-ordered logit model had any trouble estimating the effect of

the exposure variable when the error terms was standard extreme value type I distributed. The estimates of the rank-ordered logit model was larger than the corresponding errors of the between-within model which resulted in a lower power for the rank-ordered logit model. When the error terms had a normal distribution, the rank-ordered logit model acted in the same way as when we studied the first simulation model. Neither the unadjusted model nor the mixed model performed well.

When we looked at the more generalized setting and allowed the confounder to be correlated within the cluster, but not completely shared, we noticed that the models had trouble handling this. Even though the adjusted models were less biased than the crude and the mixed model, the estimates were very biased and had low coverage. When the correlation was equal to 0.9 the confidence intervals of the between-within model covered the true value in approximately 40 % of the cases and when the correlation was even lower and equal to 0.7 the model never managed to cover the true value. The rank-ordered logit model performed somewhat better in terms of bias and coverage when the correlation was below one. Both models overestimates the coefficients and even though the rank-ordered logit model performs better than the between-within, none of the models are able to cope with the situation. This is analogue to what has been shown by Frisell et al. (2012), that when the correlation of the confounder is stronger than the correlation of the explanatory variable within a pair, the between-within model is less biased than the crude model. From our simulations we can conclude that the same hold for the rank-ordered logit model. Thus, both the between-within and the rank-ordered logit models works well when the confounder is completely shared within the cluster, but suffers from bias when it is not. *The between-within is more robust against the underlying distributional assumptions than the rank-ordered logit model.*

We have looked at the performance of these two models when we are not able to measure the random effect or confounder. If it is measurable we can include it in the linear model and thus adjust for it by modelling or we could also divide the population into different clusters with respect to these covariates and apply either the between-within or the rank-ordered logit model. In this latter approach we do not need to know the exact measurements since it is the cluster structure that is of interest and we do not have to assume a functional form either. In Section 4.3.2 we applied these models on the same data as in the earlier analyses where we assumed that the covariates were unmeasurable. The linear model that is adjusted for the random covariates is the best model, since it includes all information about the covariates. *The stratified between-within and rank-ordered logit models have a worse performance than the corresponding paired models when the random effect or confounder is completely shared within a pair (but not within a group in the stratified approach). When the confounder is correlated within a pair and the correlation is below one the stratified models performed much better than the paired models in terms of bias, coverage and power.* Therefore, based on the results of the simulations we can conclude that if we have a paired structure but we do not think that the confounders are completely shared within a pair, a stratified approach will generate better estimates than a paired approach.

These models were applied to two different datasets, the first data was collected in Singapore and contained measurements of blood glucose levels for patients admitted to the hospital. After excluding patients that did not meet the criteria we were left with a population of size 2113. We were interested in the possible association between the variation of measurement levels and the daily reading frequency during the first monitoring period. We analysed the data with respect to the measurable potential confounders: age, gender and length of first monitoring period. By fitting linear models with normally and extreme value type I distributed error terms, respectively, we noticed that the distribution of the residuals were well approximated with an extreme value type I distribution and the estimated scale parameter was approximately 1. Thus the rank-ordered logit model should be a good analysis approach. The estimates of the extreme value type I error models are approximately equal to 0.26 and significantly different from zero. Since the scale parameter is approximately 1 we can also interpret the effect of the rank-ordered logit model as in the linear models with extreme value type I distributed errors. Thus,

we can conclude that an one unit increase the daily reading frequency is associated with a 0.26 increase in variation of measurement levels, when all other covariates remains constant. The residuals were not well approximated with a normal distribution, the estimate from the linear model and the within-effect estimate of the between-within model are approximately equal to 0.51.

We have also analysed data from the Karma study, which contain information of the mammographic density measure percent dense volume, among women who have a sister who also, at some point of time, participated in the study. We wanted to determine which of the risk factors age, bmi, age at menarche, hormone replacement therapy, benign breast disease, given birth and age at first birth that were significantly associated with the percent dense volume. After excluding women without complete information and who did not meet our criteria we had a population of 2960 women, among these women there were 879 pairs of sisters and thus 2081 unrelated women. For the application of linear models and for the between-within model we choose to log transform the percent dense volume in order to obtain normally approximated data. From the QQ plots of the data we could also conclude that the normal approximation fits better than the extreme value type I distribution. From our first crude analysis we concluded that age, bmi and menopause status were strongly associated with mammographic density. We therefore only considered women post menopause and adjusted for age and bmi by including the variables in the linear predictor in all analysis approaches. In the adjusted linear model we studied the impact of the other risk factors and found that age at menarche is the only factor that is not a significant risk factor for percent dense volume. If a women had given birth this was significant in the sample of unrelated sample but not in the sample of all women who fulfilled the criteria, even though it is an indication of an effect also in the large sample.

From the results of the between-within model on the subset of sisters, we could conclude that the between-effect was not significant for any risk factor and that the within-effect estimates were similar to the estimates of the adjusted linear model. Since the error distribution did not seem to follow an extreme value type I distribution we put more trust in the models which assumes normally distributed errors, i.e. the adjusted linear model or the between-within model on the sample of sisters. From these models we can conclude that the factors hormone replacement therapy, benign breast disease and the factor age at first birth are significant have a positive and significant effect on percent dense volume. Furthermore, the factor given birth is significant in the adjusted model in the sample of unrelated women. If we look at the estimates of the adjusted model among unrelated women, a post menopausal woman who has had hormone replacement therapy is expected to have a $\exp(0.06)=1.06$ times higher percent dense volume than woman with the same age and bmi who has not had hormone replacement therapy. The corresponding effect of benign breast disease is equal to 1.1. The effect of given birth is negative and thus has a decreasing effect of percent dense volume. A woman who has given birth has 0.93 times lower percent dense volume than a woman who had never given birth in this population. Every additional year in age at first birth has a multiplicative effect of 1.01 on percent dense volume.

We compared the estimates with the stratified approach of the between-within and rank-ordered logit models, where we divide the population into different clusters with regard to their bmi and age group. We applied the models on the data of all women who fulfilled the criteria and noticed that the two models generate similar results, even though the estimates of the rank-ordered logit model are scaled, and the conclusions agree with the other models.

We can therefore conclude that the risk factors hormone replacement treatment and benign breast disease have a statistically significant positive effect on percent dense volume among post menopausal women. Age at first birth is significant among women who have given birth, but the factor of given birth is only significant in the adjusted model in the population of unrelated women and not a statistically significant effect in the whole sample. We could use a sample of unrelated women and apply an adjusted model or a stratified between-within model, or use a sample of siblings and apply the between-within

model to estimate these effects, and none of them were better than the other. We did not gain anything by adjusting for genetic factors.

To summarize we have shown that the rank-ordered logit model is not as robust as the between-within model with respect to the underlying distributional assumptions. An advantage of the between-within model compared to the rank-ordered logit is that it is more easily interpreted. In the rank-ordered logit model we do not care about how big the difference in outcome between two individuals is, we only take into the calculations the information about which one is larger or smaller. Our estimates of the rank-ordered logit model are scaled and in order to determine the magnitude of the effect we need to estimate a scale parameter by fitting a linear model with extreme value type I distribution where we need to include the factors we want to adjust for in the linear predictor. Furthermore, the rank-ordered logit model only gives an estimate of the within-effect, while the between-within model also give an estimated effect of the difference between clusters. Both models perform well when all confounders are completely shared within a cluster. The rank-ordered logit model has performed slightly better than the between-within model when the confounders are not completely shared. We conclude that the rank-ordered logit model should be used with caution and only when there are reasons to believe that the error distribution is standard extreme value type I. The restriction of the model makes it less preferable than the between-within model and it is more useful when the outcome is not observed as in the econometrics literature than when we have a measurable continuous outcome. The between-within model is preferable since it is not as complicated and more robust towards the assumptions. However, this model should be used with caution since it only adjust for confounders that are completely shared within a cluster.

Bibliography

General-purpose optimization, June 2015. URL https://stat.ethz.ch/R-manual/R-patched/library/stats/html/optim.html.

The karma study, 2015. URL http://karmastudy.org/.

Body mass index - bmi, June 2015. URL http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi.

A. Agresti. *Categorical Data Analysis*. John Wiley & Sons, Inc, second edition, 2002.

I. F. Alves and C. Neves. Extreme value distributions. In *International Encyklopedia of Statistical Science*, chapter Extreme Value Distributions, pages 493–496. Springer, 2011.

A. D. Bates, M. Maechler, B. Bolker, S. Walker, R. H. B. Christensen, H. Singmann, and B. Dai. *Linear mixed-effects models using Eigen and S4*, March 5 2015. URL http://cran.r-project.org/web/packages/lme4/lme4.pdf.

D.M. Bates. *lme4: Mixed-effects modeling with R*. Springer, June 2010. URL http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf.

M. D. Begg and M. K. Parides. Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data. *Statistics in Medicine*, 22:2591–2602, 2003.

S. Beggs, S. Cardell, and J. Hausman. Assessing the potential demand for electric cars. *Journal of Econometrics*, 16:1–19, 1981.

J.S. Brand, K. Humphreys, D.J. Thompson, J. Li, M. Eriksson, P. Hall, and K. Czene. Volume mammographic density: Heritability and association with breast cancer susceptibility loci. *Journal of the National Cancer Institute*, 106, 2014. Issue 12.

N. E. Breslow and N. E. Day. *Statistical Methods in Cancer Research. Volume 1 - The Analysis of Case-Control Studies*. Lyon, International Agency for Research on Cancer (IARC Scientific Publications No.32), 1980.

J. B. Carlin, L. C. Gurrin, J. A. C. Sterne, R. Morely, and T. Dwyer. Regression models for twin studies a critical review. *International Journal of Epidemiology*, 34:1089–1099, 2005.

C. Cox. Delta method. In *Encyclopedia of Biostatistics*. John Wiley & Sons, Ltd, 2005.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

J.S. Cramer. *Logit Models from Economics and Other Fields*. Cambridge University Press, 2003.

L. Fahrmeir, T. Kneib, S. Lang, and B. Marx. *Regression Models, Methods and Applications*. Springer, 2013.

G.M. Fitzmaurice, N.M. Laird, and J.H. Ware. *Applied Longitudinal Analysis*. John Wiley & Sons, second edition, 2011.

R.H. Fletcher and S.W. Fletcher. *Clinical Epidemiology: The Essentials*. Lippincott Williams & Wilkins, fourth edition, 2005.

C. Forbes, M. Evans, and N. Hastings. *Statistical Distributions*. John Wiley & Sons, Inc, fourth edition, 2010.

J. Fox. Cox proportional-hazards regression for survival data: An appendix to an r and s-plus companion to applied regression, 2002. URL http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-cox-regression.pdf.

T. Frisell, S. Öberg, R. Kuja-Halkola, and A. Sjölander. Sibling comparison designs: Bias from non-shared confounders and measurement error. *Epidemiology*, 23:713–720, 2012.

T. Horthorn, A. Zeileis, R.W. Farebrother, C. Cummins, G. Millo, and D. Mitchell. Testing linear regression models, June 2015. URL http://cran.r-project.org/web/packages/lmtest/lmtest.pdf.

J.P. Klein and M.L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 1997.

A. Kuznetsova, P. B. Brockhoff, and R.H.B. Christensen. Tests in linear mixed effects models, June 2015. URL http://cran.r-project.org/web/packages/lmerTest/lmerTest.pdf.

J. M. Lachin. *Biostatistical Methods: The Assessment of Relative Risks*. John Wiley & Sons, Inc, 2 edition, 2011. Chapter 9.

D. McFadden. Conditional logit analysis of quantitative choice behavior. In *Frontiers in Econometrics*. Academic Press, INC, 1974.

J. M. Neuhaus and J. D. Kalbfleisch. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics*, 54:638–645, 1998.

J. M. Neuhaus and C. E. McCulloch. Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 68:859–872, 2006.

A. Sjölander, P. Lichtenstein, H. Larsson, and Y. Pawitan. Between-within models for survival analysis. *Statistics in Medicine*, 32:3067–3076, 2013.

A. Stephenson. Functions for extreme value distributions, February 2015. URL http://cran.r-project.org/web/packages/evd/evd.pdf.

C. S. Tan, N. Støer, Y. Chen, E.-S. Tai, H. L. Wee, E. Y. H. Khoo, S. L. Kao, and M. Reilly. Matched designs for continuous outcome.

T. M. Therneau. Survival analysis, February 2015. URL http://cran.r-project.org/web/packages/survival/survival.pdf.

T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000.

K. E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2003.

L. Fahrmeir, and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, 1994.

C.M. Vachon, C.H. van Gils, T.A. Sellers, K. Ghosh, S. Pruthi, K.Rk Brandt, and V.S. Pankratz. Mammographic density, breast cancer risk and risk prediction. *Breast Cancer Research*, 9(6):217, 2007.

P.M. Visscher, W.G. Hill, and N.R. Wray. Heritability in the genomics era - concepts and misconceptions. *Nature Reviews Genetics*, 9:255–266, 2008.

B. T. West, K. B. Welch, and A. T. Gałecki. *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman & Hall/CRC, 2007.

Appendix A

Mammographic Density Data

A.1 Transformation of the outcome

In order to see if data are approximately normal distributed we fit a linear model with all risk factors. First, we estimate the linear model

$$\begin{aligned}\text{PDV} = \beta_0 + \beta_1\text{Age} + \beta_2\text{Bmi} + \beta_3\text{AgeM} + \beta_4\text{Birth} + \beta_5\text{Birth} \times \text{AgeBirth} \\ + \beta_6\text{Meno} + \beta_7\text{HRT} + \beta_8\text{BBD} + \epsilon_{ij}\end{aligned} \tag{A.1}$$

where PDV denotes percent dense volume, AgeM denotes age at menarche, Birth is an indicator variable and equal to 1 if the woman have given birth, Birth×AgeBirth is a interaction variable of the indicator of a woman ever have been given birth and her age at this occurrence, Meno is a indicator variable and equal to one if the women are peri och post menopause, HRT is an indicator variable and equal to 1 if the woman ever have taken any hormone replacement therapy, BBD is an indicator variable which is equal to 1 if the woman ever been diagnosed with benign breast disease and ϵ_{ij} is the error term. In Figure A.1 we look at the residuals of the model in a QQ plot against the corresponding theoretical normal quantiles in order to determine if the observations can be assumed to be normally distributed.

From the Figure we can conclude that the normal approximation is not appropriate, we therefore log-transform the data. We fit the model

$$\begin{aligned}\text{lPDV} = \beta_0 + \beta_1\text{Age} + \beta_2\text{Bmi} + \beta_3\text{AgeM} + \beta_4\text{Birth} + \beta_5\text{Birth} \times \text{AgeBirth} \\ + \beta_6\text{Meno} + \beta_7\text{HRT} + \beta_8\text{BBD} + \epsilon_{ij}\end{aligned}$$

where we used the same notation as in equation (A.1) but with log(Percent Dense Volume) as response variable, denoted by lPDV. The QQ plot of this model is shown in Figure A.2.

Furthermore, in order to investigate if we can assume the data to be extreme value type I distributed, we fit the model in Equation (A.1) by finding the maximum likelihood estimates of the likelihood function of this distribution. The residuals of the model is plotted against the theoretical extreme value type I quantiles.

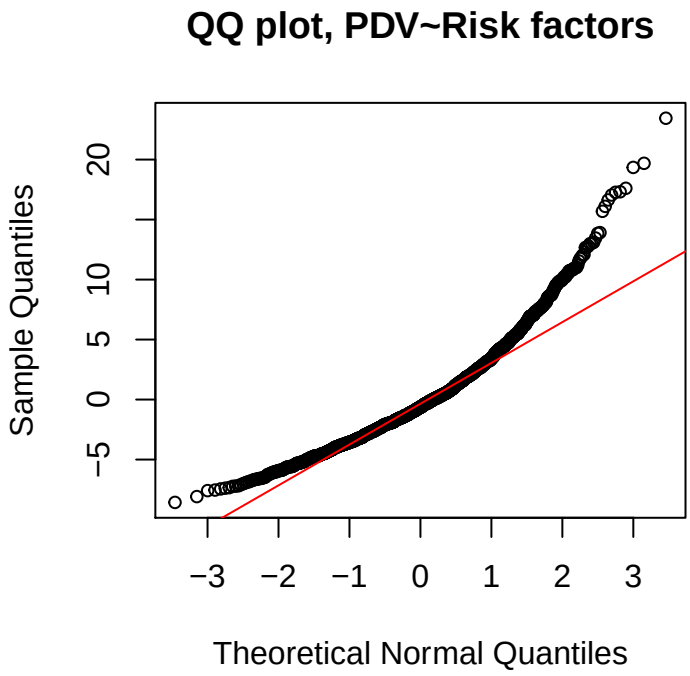


Figure A.1: QQ plots of the residuals from the fitted linear model with percent dense volume as response variable and all risk factors included in the linear predictor.

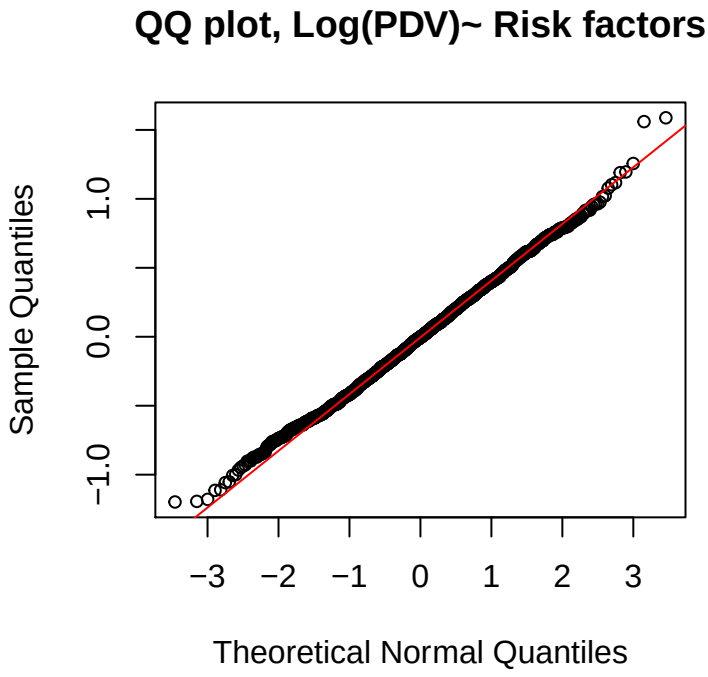


Figure A.2: QQ plots of the residuals from the fitted linear model with log percent dense volume as response variable and all risk factors included in the linear predictor.

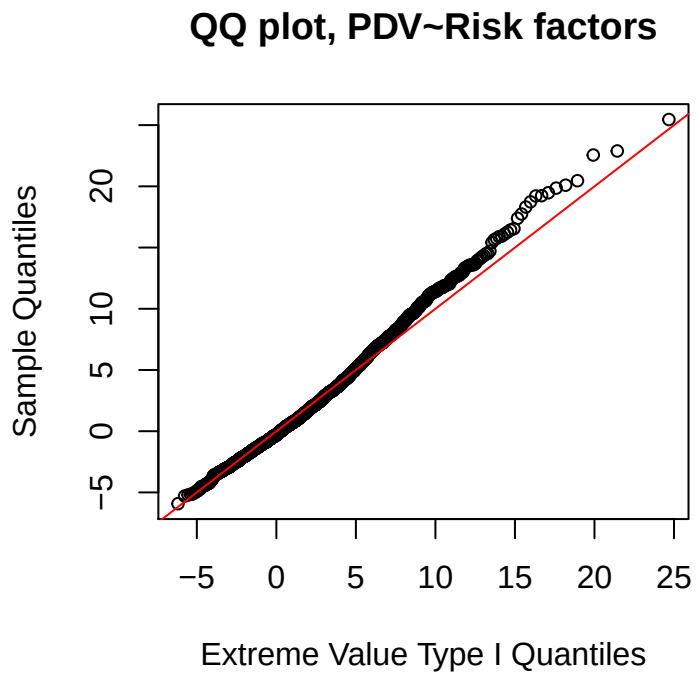
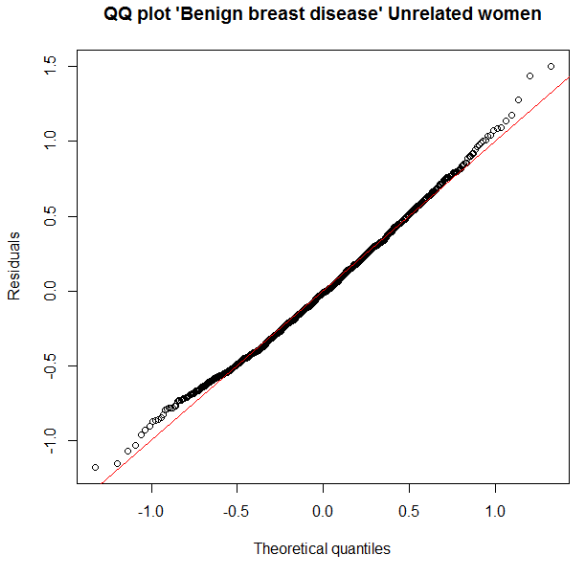


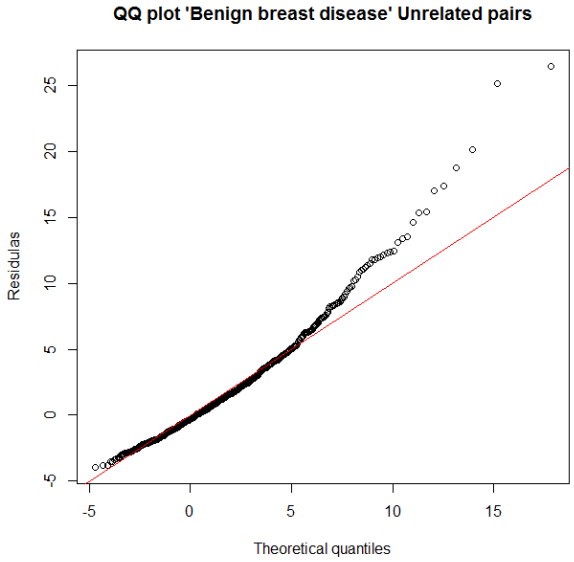
Figure A.3: QQ plots of the residuals from the fitted linear model with log percent dense volume as response variable and all risk factors included in the linear predictor.

A.2 QQ-plots Associated With Table 5.7

In Section 5.2 we fitted the between-within and the rank-ordered logit model to pairs of unrelated women. We adjusted for age and bmi in these models. To determine whether or not the underlying assumption of the error distributions are fulfilled fit the between-within model and the adjusted linear model when we assume extreme value type I distributed errors. In Figure A.4 we present the QQ plots of the residuals from these models when we study the association between percent dense volume and hormone replacement therapy. We do not present the corresponding plots of the other risk factors since they behaves similar.



(a) QQ plot of the from the between-within model when we study the association between log percent dense volume and benign breast disease, against the theoretical quantiles.



(b) QQ plot of the from the adjusted linear model with extreme value type I distributed errors, when we study the association between percent dense volume and benign breast disease, against the theoretical quantiles.

Figure A.4: QQ-plots of the residuals from the fitted linear models with percent dense volume and the risk factor benign breast disease.