Stockholms universitet

Handling Ties in the Rank Ordered Logit Model Applied in Epidemiological Settings

Angeliki Maraki

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2016:4
Matematisk statistik
September 2016

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2016:4**
http://www.math.su.se

Handling Ties in the Rank Ordered Logit Model Applied in Epidemiological Settings

Angeliki Maraki*

September 2016

Abstract

In epidemiological studies it is important to adjust for potential confounders when assessing the relationship between an explanatory variable and the outcome, in order to not obtain false, or miss true, statistically significant association between these two factors. Matching is a way to adjust for confounding having as a main advantage that it is not needed for the relationship between the outcome and the confounders to be specified, that way mis-specifications of the model can be avoided. In this Thesis, we consider continuous or ordinal outcomes and match away confounding using the Rank Ordered (RO) logit model by stratifying the cohort based on the confounders and ranking the outcomes within each stratum. When the underlying model is linear in parameters and the error terms have an Extreme Value Type I distribution the resulting likelihood is equivalent to the likelihood for the stratified Cox proportional hazards model. Consequently, the RO-logit model can be used for matching away possibly complex relationships between confounders and exposure/outcome by fitting stratified Cox-regressions. One challenge with the RO-logit model is ties. Similar to Cox-regression in survival analysis, the estimator assumes no ties in the outcome, but since the estimator has the same form as a stratified Cox-regression it is reasonable to assume that methodology for handling ties in survival analysis can be adopted in the RO-logit model. In this thesis we will investigate this by evaluating four methods for handling ties in Cox-regression, namely the Efron, Breslow, Discrete and Adding methods, by simulating scenarios with different degree of ties and different exposure and error distributions. We conclude that all four methods perform equivalently well, with the Adding method having a small advantage over the others. However, for some methods we found some bias. Moreover, we applied the RO-logit model on a data set from Maria Ungdom health clinic. The data set contains information about clients of the clinic, their family history, a selection of single nucleotide polymorphisms (SNP) and alcohol or drug abuse score of the individuals. The aim of the analysis was to identify significant relationships between some SNPs and alcohol or drug abuse score of the individuals participating in this study when the family history is adjusted for by matching on it. Significant associations between some of the SNPs and the alcohol and drug abuse score were detected. Nevertheless, some methods of handling ties in the model were biased and they are advised to be applied with caution.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: aggmara@gmail.com. Supervisor: Tom Britton.

Acknowledgements

I would like to express my deepest gratitude to my supervisor Nathalie Støer at the Department of Medical Epidemiology and Biostatistics at Karolinska Institute, for her guidance and support in my research. I really appreciate the possibility to work with this project. I would also like to express my gratitude to my supervisor Tom Britton at the Department of Mathematical Statistics at Stockholm University, for the advice, feedback and support during the writing of this thesis. Furthermore, I would like to thank the professor Marie Reilly at the Department of Medical Epidemiology and Biostatistics (MEB) at Karolinska Institute, for introducing me to the opportunity to work with Nathalie Støer and this project and all the help she provided in our weekly progress meetings.

Moreover, I would like to dedicate this thesis to my family for all the support and tolerance in my education. Your love, guidance and understanding throughout my life was priceless.

Finally, I would like to thank all of my friends in Stockholm and in Greece for being at my side, for your help and for loving me back.

Angeliki Maraki
Athens, August 15, 2016

Contents

List of Figures

List of Tables

Abbreviations

EVT1	**E**xtreme **V**alue **T**ype I
RO-logit model	**R**ank **O**rdered logit model
SNP	**S**ingle**N**ucleotide **P**olymorphism

To Sotiris and Gesthimani.

Chapter 1

Introduction

An essential part of an epidemiological study is to adjust for potential confounding when assessing the relationship between exposure and outcome. When the outcome is continuous, the confounders are traditionally adjusted for, by including them in a regression model. Matching is another way to adjust for confounders and one of its main advantages is that the relationship between the outcome and the confounders does not need to be specified and is thus robust to model misspecification. A method which can be used to match away confounding with continuous outcomes is the Rank Ordered (RO) logit model. This method was originally developed and applied in the econometrics literature, but transferred to epidemiological settings (Støer N. et. al., 2016 & Andersson M., 2015).

The idea is to stratify the cohort based on the confounders and then rank the outcomes within each stratum. The population is stratified into different stratums with respect to the confounders that we must adjust for. For each stratum all shared confounders are matched away, since all individuals within the stratum have similar or identical confounding profile. The underlying model is linear in the parameters and by assuming an extreme value type I distribution for the error term the resulting likelihood is on the same form as the likelihood for a stratified Cox-regression. Thus the RO-logit model opens up for the possibility of "matching away" possibly complex relationships between confounders and exposure/outcome by fitting stratified Cox-regressions.

One potential challenge with the RO-logit model is ties. For truly continuous data this is not a problem, but in practice even for continuous data some ties can be expected due to for instance rounding. Additionally, we also believe that the RO-logit model can be useful for ordinal data that somehow captures an underlying (unmeasured) continuous outcome, for instance different types of scores. Such data may however create a large

portion of ties, and learning how to deal with these ties will be an important step towards making the RO-logit model more applicable in epidemiology.

In this thesis we focused on handling ties when the outcome is ordinal. Handling ties in the RO-logit model was briefly mentioned by Allison and Christakis (1994), and as far as we know, this has never been fully investigated, neither numerically or theoretically. The aim of this project is therefore to thoroughly evaluate available tools for handling ties applied in Cox-regression in the RO-logit model.

By simulating scenarios with different exposure and error distributions and different fraction of ties we were able to evaluate the methods of handling ties.

We applied the model on a data set from Maria Ungdom (Maria Ungdom website `http://mariaungdom.se`), a health clinic in Stockholm, which contained information about clients of the clinic, their family history, a selection of single nucleotide polymorphisms (SNP) (described in detail in Section 5.1), and alcohol or drug abuse score of the individuals. The objective is to investigate whether it is a significant relationship between some single nucleotide polymorphisms (SNP), and alcohol or drug abuse score of the individuals participating in this study when the family history is adjusted for, by matching on it.

In Chapter 2, one can find the theory behind the Cox proportional hazards model along with the different methods of handling ties in the model and the Rank-Ordered logit model is presented in Chapter 3 . In Chapter 4 we simulated scenarios with different degree of ties and different exposure and outcome distributions. This Chapter also includes the results of the simulations and a description of the `R` packages used throughout this project. The data analysis is to be found in Chapter 5, where we apply the tools of handling ties on the data set from Maria Ungdom and look at the associations between a selection of SNPs and alcohol and drug abuse scores.

Chapter 2

Cox Proportional Hazards Model

2.1 Introduction to the Cox Proportional Hazards Model

In survival analysis, the data involve time to the occurrence of a certain event, such as time of death or time from diagnosis to remission, which is described by a random variable T. As described by Collette (Collett 2003, Chapter 1), when T is a continuous random variable, the cumulative distribution function (c.d.f) is $F(t) = P(T \leq t)$ and gives the probability that the event has occurred prior to time t , $t > 0$. The probability to experience the event of interest beyond a certain time t is given by the so called Survival function, $S(t)$, given by

$$S(t) = P(T > t) = 1 - F(t). \tag{2.1}$$

The hazard function, denoted by $\lambda(t)$[1], is defined as the probability that an individual has the event at time t, conditional on he or she not having the event before t (Collett 2003, Chapter 1). The hazard function ($\lambda(t)$) equals the limit, as Δt approaches zero, of a probability statement about survival, divided by Δt, where Δt denotes a small interval of time (Collett 2003, Chapter 1), i.e.

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{2.2}$$

For the Survival function, $S(t)$, for continuous time, the probability density function (p.d.f.) is

[1]Some books use $h(t)$, instead of $\lambda(t)$, as notation for the hazard function

$$f(t) = -\frac{dS(t)}{dt}, \tag{2.3}$$

and the hazard rate then can be written as

$$\lambda(t) = \frac{f(t)}{S(t)} \quad \text{(Fisher et al. 1993, Chapter 16).} \tag{2.4}$$

Cox (1972) defined the proportional hazards model for a set of explanatory variables $\boldsymbol{x}$, $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)'$, as

$$\lambda(t\ ;\ \boldsymbol{x}) = \lambda_0(t) \exp\{\boldsymbol{x}^T \boldsymbol{\beta}\}, \tag{2.5}$$

where, the function $\lambda_0(t)$ is called the baseline hazard function and is the hazard function for an individual for whom the values of all the explanatory variables (that make the vector $\boldsymbol{x}$) are zero (Collett 2003, Chapter 1).

Moreover, the general proportional hazards model can be expressed in the form

$$\log(\lambda(t\ ;\ \boldsymbol{x})) = \log(\lambda_0(t)) + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \tag{2.6}$$

and re-expressed in the form

$$\log\left(\frac{\lambda(t\ ;\ \boldsymbol{x})}{\lambda_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \tag{2.7}$$

where the $\frac{\lambda(t\ ;\ \boldsymbol{x})}{\lambda_0(t)} = \psi$, is defined as the hazard ratio or relative hazard, it is always non-negative and does not depend on time (Collett 2003, Chapter 3).

Hence, from Equation (2.7) one can see that the general proportional hazards model can be regarded as a linear model for the logarithm of the hazard ratio (Collett 2003, Chapter 3).

It is typical for survival data that not all individuals experience the event of interest by the end of the follow-up. The observation for an individual who has not experienced the event by the end of the study is said to be a censored observation (Collett 2003, Chapter 1). However, censoring is not relevant in our situation which will become evident in Chapter 3.

2.2 The Partial Likelihood Function for the Proportional Hazards Model with No Tied Event Times

In this section we will derive the partial likelihood function for the proportional hazard model following the derivation made by Collett (Collett 2003, Chapter 3). Suppose we have r ordered from smallest to largest and distinct event times, i.e. $t_{(1)} < \cdots < t_{(r)}$ and hence the order statistic is $\mathbf{O(t)}= [t_{(1)}, \ldots, t_{(r)}]$ (Kalbfleisch et. al. 1980, Chapter 4). We let this set of r, distinct and ordered event times $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$, be the observed event times with no ties. Moreover, the vector of explanatory variables for the individual j who has the event at some time $t_{(j)}$ is $\boldsymbol{x}_{(j)}$. We consider the probability that the j'th individual has the event at some time $t_{(j)}$ (given that he had not experienced the event before time $t_{(j)}$), conditional on $t_{(j)}$ being one of the event times, that is

$$P(\text{individual } j \text{ has the event at } t_{(j)} \mid \text{one event at } t_{(j)}). \tag{2.8}$$

Knowing that the probability of an event A conditional on an event B is given by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)},$$

Equation (2.8) can be written as

$$\frac{P(\text{individual } j \text{ has the event at } t_{(j)})}{P(\text{one event at } t_{(j)})}. \tag{2.9}$$

The numerator of the expression above is the hazard function, as given in Equation (2.5), i.e. the probability of the event to occur for the j'th individual at time $t_{(j)}$ and can be written as $\lambda(t_{(j)} \ ; \ \boldsymbol{x}_{(j)})$. Let $R(t_{(j)})$ denote the set of individuals at risk at time $t_{(j)}$. Then the denominator of Equation (2.9) equals the sum of the values of $\lambda(t_{(j)} \ ; \ \boldsymbol{x}_{(l)})$ over those individuals (whom we index by l) in the risk set at time $t_{(j)}$ (i.e. over all the individuals in $R(t_{(j)})$). Therefore, Equation (2.9) becomes

$$\frac{\lambda(t_{(j)} \ ; \ \boldsymbol{x}_{(j)})}{\sum_{l \in R(t_{(j)})} \lambda(t_{(j)} \ ; \ \boldsymbol{x}_{(l)})}.$$

Now, on using Equation (2.5), the baseline hazard ($\lambda_0(t_{(j)})$) in the numerator and the denominator cancels out leaving us with

$$\frac{\exp\{\boldsymbol{x}_{(j)}^T \boldsymbol{\beta}\}}{\sum_{l \in R(t_{(j)})} \exp\{\boldsymbol{x}_{(l)}^T \boldsymbol{\beta}\}}. \tag{2.10}$$

Finally, the partial likelihood function is the product of these conditional probabilities over all r event times, i.e.

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp\{\boldsymbol{x}_{(j)}^T \boldsymbol{\beta}\}}{\sum_{l \in R(t_{(j)})} \exp\{\boldsymbol{x}_{(l)}^T \boldsymbol{\beta}\}}. \tag{2.11}$$

As also stated in Collett (Collett 2003, Chapter 3), the likelihood function derived above considers probabilities for the individuals that have the event but does not directly consider probabilities for the censored individuals. That is the reason it is referred to as the partial likelihood function.

2.3 The Methods for Handling Tied Event Times

The proportional hazards model is based on several assumptions one of which concerns tied events, i.e. events with exactly the same survival time (Borucka 2014).

In this section we will present four methods for handling tied events in the Cox proportional hazards model, namely the Discrete, the Breslow the Efron and the Adding methods. Those four methods are approximations to the appropriate likelihood function in the presence of tied observations and provide computational advantages over the calculation of the exact likelihood especially when the proportion of ties is relatively large (Collett 2003, Chapter 3).

2.3.1 The Discrete Method

Cox (1972) proposed an approximation assuming that the time scale is discrete and hence the tied event times occurred at exactly the same time (Collett 2003, Chapter 3). In this section we will derive the Discrete approximation for the partial likelihood function for the proportional hazard model. We will follow the derivation for the partial likelihood function when there are no tied event times made by Collett (Collett 2003, Chapter 3) taking ties into consideration.

Suppose we have r ordered from smallest to largest and distinct event times, i.e. $t_{(1)} < \cdots < t_{(r)}$. We consider the probability that d_j individuals have the event of interest at some time $t_{(j)}$ (given that they had not experienced the event before time $t_{(j)}$), conditional on $t_{(j)}$ being one of the event times, where d_j number of events occur, that is

$$P(\text{individuals } j_1,\ j_2\ ,\ldots,j_{d_j} \text{ have the event at } t_{(j)})|\ d_j \text{ events at } t_{(j)}), \tag{2.12}$$

(Collett 2003, Chapter 3). Knowing that the probability of an event A conditional on an event B is given by

$$P(A \mid B) = \frac{P(A\ \cap\ B)}{P(B)},$$

Equation (2.12) can be written as

$$\frac{P(\text{individuals } j_1,\ j_2\ ,\ldots,j_{d_j} \text{ have the event at } t_{(j)} \text{ out of } \ R(t_{(j)})}{\sum_{l\in R(t_{(j)}\ ;\ d_j)} P(\text{individuals } l_1,\ l_2\ ,\ldots,l_{d_j} \text{ have the event at } t_{(j)} \text{ out of } \ R(t_{(j)})}, \tag{2.13}$$

where, as also stated in section 2.2, $R(t_{(j)})$ denotes the set of individuals at risk at time $t_{(j)}$. Furthermore, $R(t_{(j)}\ ;\ d_j)$ is a set of d_j possible individuals from $R(t_{(j)})$ and the summation in the denominator denotes the sum over all possible sets of d_j individuals from $R(t_{(j)})$ without replacement (Collett 2003, Chapter 3).

In our calculations, $\boldsymbol{x}_{j_m}$ denotes the vector of explanatory variables for the m'th individual who has the event at $t_{(j)}$ and $\boldsymbol{x}_{l_n}$ denotes the vector of explanatory variables for the n'th individual in the l'th set from $R(t_{(j)}\ ;\ d_j)$, (Collett 2003, Chapter 3). Thus, Equation (2.13) becomes

$$\frac{\lambda(t_{(j)}\ ;\ \boldsymbol{x}_{j_1})\cdot\ \lambda(t_{(j)}\ ;\ \boldsymbol{x}_{j_2})\cdot\ \ldots\ \cdot\lambda(t_{(j)}\ ;\ \boldsymbol{x}_{j_{d_j}})}{\sum_{l\in R(t_{(j)}\ ;\ d_j)}\lambda(t_{(j)}\ ;\ \boldsymbol{x}_{l_1})\cdot\ \lambda(t_{(j)}\ ;\ \boldsymbol{x}_{l_2})\cdot\ \ldots\ \cdot\lambda(t_{(j)}\ ;\ \boldsymbol{x}_{l_{d_j}})}.$$

Now, using Equation (2.5), the baseline hazard ($\lambda_0(t_{(j)})$) in the numerator and the denominator cancels out leaving us with

$$\frac{\exp\{\boldsymbol{x}_{j_1}^T\ \boldsymbol{\beta}\}\cdot\ \exp\{\boldsymbol{x}_{j_2}^T\ \boldsymbol{\beta}\}\cdot\ \ldots\ \cdot\exp\{\boldsymbol{x}_{j_{d_j}}^T\ \boldsymbol{\beta}\}}{\sum_{l\in R(t_{(j)}\ ;\ d_j)}\exp\{\boldsymbol{x}_{l_1}^T\ \boldsymbol{\beta}\}\cdot\ \exp\{\boldsymbol{x}_{l_2}^T\ \boldsymbol{\beta}\}\cdot\ \ldots\ \cdot\exp\{\boldsymbol{x}_{l_{d_j}}^T\ \boldsymbol{\beta}\}},$$

which can also be written as

$$\frac{\exp\{\sum_{m=1}^{d_j}\boldsymbol{x}_{j_m}^T\ \boldsymbol{\beta}\}}{\sum_{l\in R(t_{(j)}\ ;\ d_j)}\exp\{\sum_{n=1}^{d_j}\boldsymbol{x}_{l_n}^T\ \boldsymbol{\beta}\}}. \tag{2.14}$$

Let $\boldsymbol{s_j}$, $\boldsymbol{s_j} = \sum_{m=1}^{d_j}\boldsymbol{x}_{j_m}$, be the vector of sums of each of the p covariates for all the d_j individuals that have the event at the j'th event time, $t_{(j)}$, $j = 1,\ldots,r$ (Collett 2003,

Chapter 3). In addition, let $\boldsymbol{s_l}$, $\boldsymbol{s_l} = \sum_{n=1}^{d_j} \boldsymbol{x}_{l_n}$, be the vector of sums of each of the p covariates for all the d_j individuals in the l'th set drawn out of $R(t_{(j)}\ ;\ d_j)$ (Collett 2003, Chapter 3). Equation (2.14) is then equivalent to

$$\frac{\exp\{\boldsymbol{s_j}\boldsymbol{\beta}\}}{\sum_{l \in R(t_{(j)}\ ;\ d_j)} \exp\{\boldsymbol{s_l}\boldsymbol{\beta}\}}. \tag{2.15}$$

Then, the Discrete approximation of the partial likelihood is the product of these conditional probabilities over all r event times, i.e.

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp\{\boldsymbol{s_j}\boldsymbol{\beta}\}}{\sum_{l \in R(t_{(j)}\ ;\ d_j)} \exp\{\boldsymbol{s_l}\boldsymbol{\beta}\}}, \tag{2.16}$$

as also stated in Collett (Collett 2003, Chapter 3). The Discrete approximation can be used for a large proportion of tied events, and is preferable (Kalbfleisch et. al. 1980, Chapter 4).

2.3.2 The Breslow Method

Breslow (1974) suggested an approximation to the likelihood function when tied observations are present.

Suppose as above, that we have r ordered from smallest to largest and distinct event times, i.e. $t_{(1)} < \cdots < t_{(r)}$. As stated in Collett (Collett 2003, Chapter 3), let $\boldsymbol{s_j}$ be the vector of sums of each of the p covariates for all the individuals that have the event at the j'th event time, $t_{(j)}$, $j = 1, \ldots, r$. Moreover, suppose d_j individuals have the event of interest at $t_{(j)}$ and then x_{hjk} is the value of the h'th explanatory variable, $h = 1, \ldots, p$, for the k'th of the d_j individuals who have the event at time $t_{(j)}$ (Collett 2003, Chapter 3). Then we have that the h'th element of $\boldsymbol{s_j}$ is

$$s_{hj} = \sum_{k=1}^{d_j} x_{hjk},$$

where the d_j events at time $t_{(j)}$ are considered to occur sequentially and to be distinct (Collett 2003, Chapter 3). Furthermore, as also stated in section 2.2, $R(t_{(j)})$ denotes the set of individuals at risk at time $t_{(j)}$.

The approximation to the likelihood that Breslow suggested is

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp\{\boldsymbol{s_j}\boldsymbol{\beta}\}}{[\sum_{l \in R(t_{(j)})} \exp\{\boldsymbol{x_l}\boldsymbol{\beta}\}]^{d_j}}, \tag{2.17}$$

which, apart from a constant proportionality, is the same as the one Peto (1972) proposed (Collett 2003, Chapter 3). This approximation is quite easy to compute and is a fair approximation of the likelihood function, i.e. gives reasonably good estimates, when the number of tied events is not too large (Collett 2003, Chapter 3). However, the Breslow method can demonstrate a severe bias for a large proportion of tied event times (Kalbfleisch et. al. 1980, Chapter 4).

2.3.3 The Efron Method

Efron (1977) suggested

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp\{\boldsymbol{s_j}\boldsymbol{\beta}\}}{\prod_{k=1}^{d_j} [\sum_{l \in R(t_{(j)})} \exp\{\boldsymbol{s_l}\boldsymbol{\beta}\} - (k-1) d_k^{-1} \sum_{l \in D(t_{(j)})} \exp\{\boldsymbol{x_l}\boldsymbol{\beta}\}]} \tag{2.18}$$

as an approximation to the likelihood function for the proportional hazards model in presence of tied event times, where $D(t_{(j)})$ denotes the set of those individuals who have the event at time $t_{(j)}$ (Collett 2003, Chapter 3). This approximation is closer to the proper likelihood function than the Breslow approximation, but it is more challenging to compute. However, both Breslow and Efron methods often produce similar results. (Collett 2003, Chapter 3).

2.3.4 The Adding Method

Another method we used for handling tied event times was, as we will refer to it in this thesis, the Adding method. In practice this method deals with ties by adding a small, random number sampled from a Uniform distribution only to the values of the event times that are duplicated. Hence, there are no longer any ties and we can use the partial likelihood for a stratified Cox proportional hazards regression model with distinct and ordered event times $t_{(1)}, t_{(2)}, \ldots, t_{(r)}$, as described in section 2.2, i.e.

$$\boldsymbol{L}(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp\{\boldsymbol{x}_{(j)}^T \boldsymbol{\beta}\}}{\sum_{l \in R(t_{(j)})} \exp\{\boldsymbol{x}_{(l)}^T \boldsymbol{\beta}\}}. \tag{2.19}$$

In Equation (2.19), as also described in section 2.2, $\boldsymbol{x}_{(j)}$ denotes the vector of explanatory variables for the individual j who has the event at some time $t_{(j)}$ and $R(t_{(j)})$ denotes

the set of individuals at risk at time $t_{(j)}$. Since time is continuous, the event times are not truly tied and the Adding method is particularly reasonable when the ties are due to rounding. Moreover, it is quite simple and fast in computations.

Chapter 3

The Rank-Ordered Logit Model

3.1 Introduction to the Rank-Ordered Logit Model

Matching is a way to adjust for confounders and is common practice in epidemiology. The idea is to stratify the population into different clusters, where the individuals within a cluster have similar or identical confounding profile (i.e. they are similar with respect to the confounders that we want to adjust for). A benefit of matching compared to regular adjustment is that the confounders are matched away, thus the relationship between the outcome and the confounders does not need to be specified. Although matching is mainly used in case-control studies, it can be also applied in cohort designs through the matched cohort designs (Sjölander et al., 2013), where exposed subjects are matched to unexposed subjects on confounders (Greenland et al., 1990). An alternative to the matched cohort design, is for continuous outcomes, is the Rank Ordered (RO) logit model (Beggs et al., 1981).

The Rank-Ordered (RO) logit model, was originally developed and applied in econometrics (Beggs et al., 1981), for instance in marketing research. The data are generated by asking individuals to rank a set of items or services (Allison et. al, 1994) and an unobserved continuous utility or preference function governing these ranks is assumed. In an epidemiological setting each choice will represent a person and each person in the econometric setting can be seen as a matched set in an epidemiological setting, since all variables describing the person are the same for all choices that the individual makes.

The estimation is based on the ranks and not the observed outcomes since these are unobserved in the econometric setting. The idea is to stratify the cohort based on the confounders and then rank the outcome within each stratum. The underlying model is linear in the parameters and by assuming an extreme value type I distribution for the

error term, the likelihood of the ranks is on the same form as the likelihood for a stratified Cox-regression. Thus the RO-logit model opens up for the possibility of matching away possibly complex relationships between confounders and exposure/outcome by fitting stratified Cox-regressions.

Similar to Cox-regression in survival analysis, the estimator assumes no ties in the outcome, but since the estimator has the same form as a stratified Cox-regression it is reasonable to assume that methodology for handling ties in survival analysis can be adopted in the RO-logit model.

3.2 Extreme Value Type I Distribution

Extreme Value Type I distribution (EVT1), also referred to as the Gumbel distribution, is a specific example of the generalised extreme value distribution (Gorgoso et al., 2014) and is the most common of the three extreme value distributions (Forbes et al., 2010, Chapter 19). In probability theory and statistics, the Extreme Value Type I distribution (Gumbel, 1954) is used to model the maximum or the minimum values for a sample of independent and identically distributed continuous random variables (Gorgoso et al., 2014). The distribution function is defined as

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

where $a,\ a \in \mathbb{R}$, is the location parameter and $b,\ b > 0$, is the scale parameter (Forbes et al., 2010, Chapter 19). The probability density function is defined as

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

with the mean equal to $a - b\Gamma'(1)$, where $\Gamma'(1) = -0.57722$ is the first derivative of the gamma function $\Gamma(n)$ with respect to n at $n = 1$ and the variance equal to

$$\frac{b^2\pi^2}{6} \approx 1.645 \cdot b^2 \tag{3.3}$$

(Forbes et al., 2010, Chapter 19). In our study we will only use the special case when the location parameter is zero and the scale parameter is one ($a = 0$ & $b = 1$), known as the standard extreme value type I distribution (as $a = 0$ & $b = 1$ is called the standard Gumbel extreme value variate) (Forbes et al., 2010, Chapter 19). Throughout

this thesis, every reference to the Extreme value type I distribution or EVT1 will imply this special case.

Hence, for the standard extreme value type I distribution we can rewrite the distribution function as

$$F_X(x) = \exp\{-\exp\{-x\}\}, \tag{3.4}$$

as well as the probability density function as

$$f_X(x) = \exp\{-x\}\exp\{-\exp\{-x\}\}. \tag{3.5}$$

Hence, the mean of the standard extreme value type I distribution is $a - b\Gamma'(1) = -\Gamma'(1) = 0.57722$ and the variance is equal to

$$\frac{b^2\pi^2}{6} = \frac{\pi^2}{6}. \tag{3.6}$$

For illustration purposes only, using different variate relationships (Variate $\boldsymbol{V}$: a, b) we create figures 3.1 and 3.2 of the Extreme Value Type I distribution function and density function.

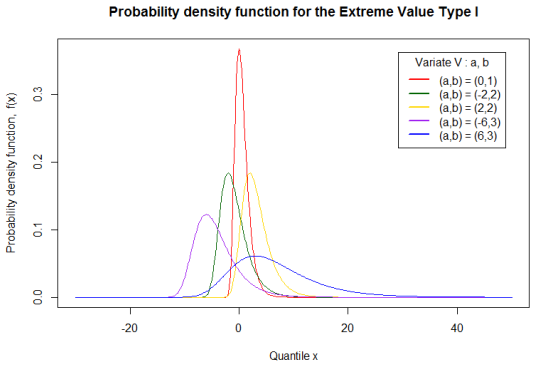

FIGURE 3.1: Extreme Value Type I probability density function, using different variate relationships V: (a, b).

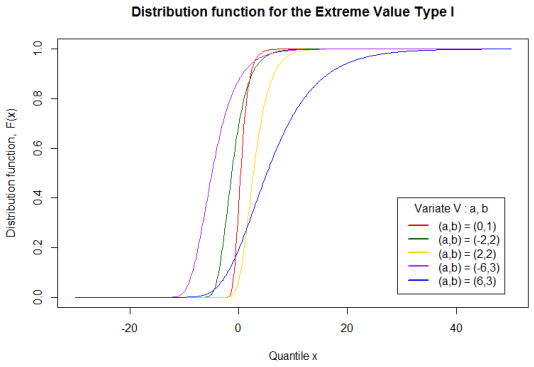

FIGURE 3.2: Extreme Value Type I distribution function, using different variate relationships V: (a, b).

3.3 Derivation of the Rank-Ordered Logit Model

Logit based models are often used in marketing and economics research to model the choices of individuals, and each individual is asked to rank a collection of elements

(Beggs et. al., 1981). Let Y_{ij} be the rank given to element j by the i'th individual. For J elements, Y_{ij} can take values form 1 to J, i.e. $j = 1, 2, \ldots, J$ (Allison et. al, 1994). We assume that i has utility U_{ij} for element j (Allison et. al, 1994), and that individual i gives a higher rank to element j than to element k (i.e. the utility of item j is greater than the utility of item k), when $U_{ij} > U_{ik}$ (Allison et. al, 1994). Moreover, U_{ij} is the sum of a deterministic component V_{ij} and a stochastic component ϵ_{ij}, that is $U_{ij} = V_{ij} + \epsilon_{ij}$, where ϵ_{ij} is assumed to be independent and standard extreme value type I distributed, namely $\epsilon_{ij} \overset{iid}{\sim}$ EVT1 (Beggs et. al., 1981). Without loss of generality if we assume the ordering $U_{i1} > U_{i2} > \cdots > U_{iJ}$ one can think of V_{i1} as a numerical quantity that shows to which degree individual i prefers element 1 over all elements and V_{i2} the numerical quantity to which extend individual i prefers element 2 over all remaining elements $(2, 3, \ldots, J)$ and so on (Allison et. al, 1994).

For our calculations, without loss of generality we can drop the person index, i, and have

$$U_j = V_j + \epsilon_j \tag{3.7}$$

(Beggs et. al., 1981). We will prove that if we assume a linear relationship between V_j and the regression coefficients, namely $V_j = \mathbf{z}^T\beta = \beta_1 z_{j1} + \beta_2 z_{j2} + \cdots + \beta_p z_{jp}$, and an extreme value type 1 distribution for the error term, the the resulting likelihood of the ranks is on the same form as the likelihood for a stratified Cox-regression.

As given by Beggs (Beggs et. al., 1981), the induced distribution for the U_j can be written as

$$H(U_j) = \exp\{-\exp\{-(U_j - V_j)\}\} \tag{3.8}$$

and we can also calculate

$$dH(U_j) = \exp\{-\exp\{-(U_j - V_j)\}\}\ \exp\{-(U_j - V_j)\}. \tag{3.9}$$

In this section we will derive the rank-ordered logit model following the derivation made by Beggs et. al. (Beggs et. al., 1981).

We can now compute the probability that an individual gives a higher rank to element j than to element k, i.e. $U_j > U_k$ by

$$
\begin{aligned}
P(U_j > U_k\ ,\ j \neq k) &= \int_{-\infty}^{\infty}\int_{-\infty}^{U_j} dH(U_k)\ dH(U_j) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{U_j} e^{-e^{-(U_j-V_j)}}\ e^{-(U_j-V_j)}\ e^{-e^{-(U_k-V_k)}}\ e^{-(U_k-V_k)}\ dU_k\, dU_j \\
&= \int_{-\infty}^{\infty} e^{-e^{-(U_j-V_j)}}\ e^{-(U_j-V_j)}\ \left[e^{-e^{-(U_k-V_k)}}\right]_{-\infty}^{U_j}\ dU_j \\
&= \int_{-\infty}^{\infty} e^{-e^{-(U_j-V_j)}}\ e^{-(U_j-V_j)}\ e^{-e^{-(U_j-V_k)}}\ dU_j \\
&= \int_{-\infty}^{\infty} \exp\left\{-e^{-(U_j-V_j)} - e^{-(U_j-V_k)}\right\} e^{-(U_j-V_j)}\ dU_j \\
&= e^{V_j}\ \int_{-\infty}^{\infty} \exp\left\{-e^{-U_j}\ (e^{V_j}+e^{V_k})\right\} e^{-U_j}\ dU_j \\
&= \left[\begin{array}{rc} \text{Substitutions :} & x = \exp\{-U_j\} \\ & dx = -\exp\{-U_j\}\ dU_j \\ & (-\infty \text{ to } \infty) \text{ becomes } (\infty \text{ to } 0) \end{array}\right] \\
&= -e^{V_j}\ \int_{\infty}^{0} \exp\left\{-x\ (e^{V_j}+e^{V_k})\right\} dx \\
&= -\frac{e^{V_j}}{e^{V_j}+e^{V_k}}\left[-\exp\left\{-x\ (e^{V_j}+e^{V_k})\right\}\right]_{\infty}^{0} \\
&= \frac{e^{V_j}}{e^{V_j}+e^{V_k}}. \qquad (3.10)
\end{aligned}
$$

The probability that the utility of choice j is greater than the utility of choice k for all $k \neq j$, $k \in (1\ ,\ J)$ can be calculated as

$$
\begin{aligned}
P(U_j > U_k\ ,\ \forall k \neq j) &= \int_{-\infty}^{\infty}\int_{-\infty}^{U_j} \prod_{k\neq j} dH(U_k)\ dH(U_j) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{U_j} \prod_{k\neq j} e^{-e^{-(U_j-V_j)}}\ e^{-(U_j-V_j)}\ e^{-e^{-(U_k-V_k)}}\ e^{-(U_k-V_k)}\ dU_k\, dU_j \\
&= \int_{-\infty}^{\infty} \prod_{k\neq j} e^{-e^{-(U_j-V_j)}}\ e^{-(U_j-V_j)}\ \left[e^{-e^{-(U_k-V_k)}}\right]_{-\infty}^{U_j}\ dU_j \\
&= \int_{-\infty}^{\infty} e^{-e^{-(U_j-V_j)}}\ e^{-(U_j-V_j)}\ e^{-\sum_{k\neq j} e^{-(U_j-V_k)}}\ dU_j
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} e^{-e^{-(U_j - V_j)}}\ e^{-(U_j - V_j)}\ e^{-e^{-U_j}\left(\sum_{k \neq j} e^{V_k}\right)}\ dU_j \\
&= \int_{-\infty}^{\infty} \exp\left\{-e^{-(U_j - V_j)} - e^{-U_j}\left(\sum_{k \neq j} e^{V_k}\right)\right\} e^{-(U_j - V_j)}\ dU_j \\
&= \int_{-\infty}^{\infty} \exp\left\{-e^{-U_j}\left(e^{V_j} + \sum_{k \neq j} e^{V_k}\right)\right\} e^{-(U_j - V_j)}\ dU_j \\
&= e^{V_j} \int_{-\infty}^{\infty} \exp\left\{-e^{-U_j}\left(\sum_{k=1}^{J} e^{V_k}\right)\right\} e^{-U_j}\ dU_j \\
&= \left[\begin{array}{rc} \text{Substitutions :} & x = \exp\{-U_j\} \\ & dx = -\exp\{-U_j\}\ dU_j \\ & (-\infty \text{ to } \infty) \text{ becomes } (\infty \text{ to } 0) \end{array}\right] \\
&= -e^{V_j} \int_{\infty}^{0} \exp\left\{-x\ \left(\sum_{k=1}^{J} e^{V_k}\right)\right\} dx \\
&= -\frac{e^{V_j}}{\left(\sum_{k=1}^{J} e^{V_k}\right)}\left[-\exp\left\{-x\ \left(\sum_{k=1}^{J} e^{V_k}\right)\right\}\right]_{\infty}^{0} \\
&= \frac{e^{V_j}}{\left(\sum_{k=1}^{J} e^{V_k}\right)}.
\end{aligned} \tag{3.11}
$$

We want to calculate the probability that U_j is less or equal than t given that the utility of choice j is larger than the utility of choice k, $U_j > U_k$, for $j \neq k$,

$$
P(U_j \leq t \mid U_j > U_k\ ,\ j \neq k) = \frac{P(U_j \leq t \bigcap U_j > U_k\ ,\ j \neq k)}{P(U_j > U_k\ ,\ j \neq k)}. \tag{3.12}
$$

The probability in the numerator of Equation (3.12) is equal to

$$
\begin{aligned}
&P(U_j \leq t \bigcap U_j > U_k\ ,\ j \neq k) = \\
&\quad = \int_{-\infty}^{t} \int_{-\infty}^{U_j} dH(U_k)\ dH(U_j) \\
&\quad = \int_{-\infty}^{t} \int_{-\infty}^{U_j} e^{-e^{-(U_j - V_j)}}\ e^{-(U_j - V_j)}\ e^{-e^{-(U_k - V_k)}}\ e^{-(U_k - V_k)}\ dU_k\, dU_j \\
&\quad = \int_{-\infty}^{t} e^{-e^{-(U_j - V_j)}}\ e^{-(U_j - V_j)}\ \left[e^{-e^{-(U_k - V_k)}}\right]_{-\infty}^{U_j}\ dU_j
\end{aligned}
$$

$$
\begin{aligned}
&= \int_{-\infty}^{t} e^{-e^{-(U_j - V_j)}}\, e^{-(U_j - V_j)}\, e^{-e^{-(U_j - V_k)}}\, dU_j \\
&= \int_{-\infty}^{t} \exp\{-e^{-(U_j - V_j)} - e^{-(U_j - V_k)}\}\, e^{-(U_j - V_j)}\, dU_j \\
&= e^{V_j} \int_{-\infty}^{t} \exp\{-e^{-U_j}\,(e^{V_j} + e^{V_k})\}\, e^{-U_j}\, dU_j \\
&= \left[\begin{array}{rc} \text{Substitutions}: & t = \exp\{-U_j\} \\ & dt = -\exp\{-U_j\}\, dU_j \\ & (-\infty \text{ to } t) \text{ becomes } (\infty \text{ to } e^{-U_j}) \end{array}\right] \\
&= -e^{V_j} \int_{-\infty}^{e^{-U_j}} \exp\{-t\,(e^{V_j} + e^{V_k})\}\, dt \\
&= -\frac{e^{V_j}}{e^{V_j} + e^{V_k}} \left[-\exp\{-t\,(e^{V_j} + e^{V_k})\} \right]_{\infty}^{e^{-U_j}} \\
&= \frac{e^{V_j}}{e^{V_j} + e^{V_k}}\, \exp\{-e^{-U_j}\,(e^{V_j} + e^{V_k})\} \\
&= \frac{e^{V_j}}{e^{V_j} + e^{V_k}}\, \exp\{-e^{-U_j}\, e^{\log(e^{V_j} + e^{V_k})}\} \\
&= \frac{e^{V_j}}{e^{V_j} + e^{V_k}}\, \exp\{-\exp\{-(U_j - \log(e^{V_j} + e^{V_k})\}\} \qquad (3.13)
\end{aligned}
$$

By inserting the probabilities in Equations (3.10) and (3.13) into Equation (3.12), we obtain

$$
\begin{aligned}
P(U_j \le t \mid U_j > U_k\ ,\ j \neq k) &= \frac{P(U_j \le t \bigcap U_j > U_k\ ,\ j \neq k)}{P(U_j > U_k\ ,\ j \neq k)} \\
&= \frac{\frac{e^{V_j}}{e^{V_j} + e^{V_k}}\, \exp\{-\exp\{-(U_j - \log(e^{V_j} + e^{V_k})\}\}}{\frac{e^{V_j}}{(e^{V_j} + e^{V_k})}} \\
&= \exp\{-\exp\{-(U_j - \log(e^{V_j} + e^{V_k})\}\}. \qquad (3.14)
\end{aligned}
$$

Consequently, we want to calculate the same conditional distribution of U_j, given that U_j is preferred over all U_k, for all $k \neq j$, i.e.

$$
P(U_j \le t \mid U_j > U_k\ ,\ \forall j \neq k) = \frac{P(U_j \le t \bigcap U_j > U_k\ ,\ \forall j \neq k)}{P(U_j > U_k\ ,\ \forall j \neq k)}. \qquad (3.15)
$$

From Equation (3.11) we have that the denominator equals $\frac{e^{V_j}}{(\sum_{k=1}^{J} e^{V_k})}$, thus we only need to compute the numerator, i.e.

$$
\begin{aligned}
P(U_j \leq t \bigcap U_j > U_k \ , \ \forall j \neq k) &= \int_{-\infty}^{t} \int_{-\infty}^{U_j} \prod_{k \neq j} dH(U_k) \ dH(U_j) \\
&= \int_{-\infty}^{t} \int_{-\infty}^{U_j} \prod_{k \neq j} e^{-e^{-(U_j - V_j)}} \ e^{-(U_j - V_j)} \ e^{-e^{-(U_k - V_k)}} \ e^{-(U_k - V_k)} \ dU_k \, dU_j \\
&= \int_{-\infty}^{t} \prod_{k \neq j} e^{-e^{-(U_j - V_j)}} \ e^{-(U_j - V_j)} \ \left[e^{-e^{-(U_k - V_k)}} \right]_{-\infty}^{U_j} \ dU_j \\
&= \int_{-\infty}^{t} e^{-e^{-(U_j - V_j)}} \ e^{-(U_j - V_j)} \ e^{-\sum_{k \neq j} e^{-(U_j - V_k)}} \ dU_j \\
&= \int_{-\infty}^{t} e^{-e^{-(U_j - V_j)}} \ e^{-(U_j - V_j)} \ e^{-e^{-U_j} (\sum_{k \neq j} e^{V_k})} \ dU_j \\
&= \int_{-\infty}^{t} \exp \{ -e^{-(U_j - V_j)} - e^{-U_j} \left(\sum_{k \neq j} e^{V_k} \right) \} \ e^{-(U_j - V_j)} \ dU_j \\
&= \int_{-\infty}^{t} \exp \{ -e^{-U_j} \left(e^{V_j} + \sum_{k \neq j} e^{V_k} \right) \} \ e^{-(U_j - V_j)} \ dU_j \\
&= e^{V_j} \int_{-\infty}^{t} \exp \{ -e^{-U_j} \left(\sum_{k=1}^{J} e^{V_k} \right) \} \ e^{-U_j} \ dU_j \\
&= \left[\begin{array}{r} \text{Substitutions :} \qquad t = \exp \{ -U_j \} \\ dt = - \exp \{ -U_j \} \ dU_j \\ (-\infty \text{ to } t) \text{ becomes } (\infty \text{ to } e^{-U_j}) \end{array} \right] \\
&= -e^{V_j} \int_{\infty}^{\exp \{ -U_j \}} \exp \{ -t \left(\sum_{k=1}^{J} e^{V_k} \right) \} \ dt \\
&- \frac{e^{V_j}}{\left(\sum_{k=1}^{J} e^{V_k} \right)} \left[- \exp \{ -t \left(\sum_{k=1}^{J} e^{V_k} \right) \} \right]_{\infty}^{e^{-U_j}} \\
&- \frac{e^{V_j}}{\left(\sum_{k=1}^{J} e^{V_k} \right)} \left(- \exp \{ -e^{-U_j} \left(\sum_{k=1}^{J} e^{V_k} \right) \} \right) \\
&\frac{e^{V_j}}{\left(\sum_{k=1}^{J} e^{V_k} \right)} \left(\exp \{ -e^{-(U_j - \log(\sum_{k=1}^{J} e^{V_k}))} \} \right). \qquad (3.16)
\end{aligned}
$$

Hence, we have calculated the conditional distribution of U_j, given that U_j is preferred over all U_k, for all $k \neq j$, by inserting the probabilities of Equations (3.11) and (3.16) into Equation (3.15), i.e.

$$
P(U_j \leq t \mid U_j > U_k \ , \ \forall j \neq k) = \exp \{ -e^{-(U_j - \log(\sum_{k=1}^{J} e^{V_k}))} \}, \qquad (3.17)
$$

and we see that it is the distribution function of an extreme value type I distribution with location parameter $a = \log(\sum_{k=1}^{J} e^{V_k})$ and scale parameter $b = 1$. Furthermore, for a set of J elements, we will compute the conditional probability of U_1, the element with the highest rankling, being less than t, given that the ranking of the remaining alternatives is also known, i.e.,

$$\begin{aligned} P(U_1 \leq t \mid U_1 > U_2 > \cdots > U_J) &= \frac{P(U_1 \leq t \bigcap U_1 > U_2 > \cdots > U_J)}{P(U_1 > U_2 > \cdots > U_J)} \\ &= \frac{P(t \geq U_1 > U_2 > \cdots > U_J)}{P(U_1 > U_2 > \cdots > U_J)} \end{aligned} \tag{3.18}$$

The probability in the numerator of Equation (3.18) is equal to

$$\begin{aligned} &P(t \geq U_1 > U_2 > \cdots > U_J) = \\ &= \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-1}} dH(U_J)\, dH(U_{J-1}) \ \ldots \ dH(U_1) \\ &= \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-1}} \prod_{k=1}^{J} e^{-e^{-(U_k - V_k)}}\, e^{-(U_k - V_k)}\, dU_J\, dU_{J-1} \ \ldots \ dU_1 \\ &= \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-1}} \prod_{k=1}^{J-1} \underbrace{e^{-e^{-(U_k - V_k)}}\, e^{-(U_k - V_k)}}_{\mathrm{S}}\, e^{-e^{-(U_J - V_J)}}\, e^{-(U_J - V_J)}\, dU_J\, dU_{J-1} \ \ldots \ dU_1 \\ &= \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-2}} \prod_{k=1}^{J-1} S \left[e^{-e^{-(U_J - V_J)}} \right]_{-\infty}^{U_{J-1}} dU_{J-1}\, dU_{J-2} \ \ldots \ dU_1 \\ &= \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-2}} \prod_{k=1}^{J-1} S\, e^{-e^{-(U_{J-1} - V_J)}}\, dU_{J-1}\, dU_{J-2} \ \ldots \ dU_1 \\ &= \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-2}} \prod_{k=1}^{J-2} S\, e^{-e^{-(U_{J-1} - V_{J-1})}}\, e^{-(U_{J-1} - V_{J-1})} \cdot \\ &\qquad \cdot e^{-e^{-(U_{J-1} - V_J)}}\, dU_{J-1}\, dU_{J-2} \ \ldots \ dU_1 \\ &= \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-2}} \prod_{k=1}^{J-2} S\, \exp\{-e^{-U_{J-1}}(e^{V_J} + e^{V_{J-1}})\} \cdot \\ &\qquad \cdot e^{-e^{-(U_{J-1} - V_{J-1})}}\, dU_{J-1}\, dU_{J-2} \ \ldots \ dU_1 \end{aligned} \tag{3.19}$$

$$
\begin{aligned}
&= e^{V_{J-1}} \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-2}} \prod_{k=1}^{J-2} S \, \exp\{-e^{-U_{J-1}}(e^{V_J} + e^{V_{J-1}})\} e^{-U_{J-1}} \, dU_{J-1} \, dU_{J-2} \, \ldots \, dU_1 \\
&= \begin{bmatrix} \text{Substitutions :} & x = \exp\{-U_{J-1}\} \\ & dx = -\exp\{-U_{J-1}\} \, dU_{J-1} \\ & (-\infty \text{ to } U_{J-2}) \text{ becomes } (e^{-U_{J-2}} \text{ to } \infty) \end{bmatrix} \\
&= -e^{V_{J-1}} \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-3}} \int_{e^{-U_{J-2}}}^{\infty} \prod_{k=1}^{J-2} S \, \exp\{-x(e^{V_J} + e^{V_{J-1}})\} \, dx \, dU_{J-2} \, \ldots \, dU_1 \\
&= -\frac{e^{V_{J-1}}}{e^{V_J} + e^{V_{J-1}}} \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-3}} \prod_{k=1}^{J-2} S \, \left[-\exp\{-x(e^{V_J} + e^{V_{J-1}})\} \right]_{e^{-U_{J-2}}}^{\infty} \, dU_{J-2} \, \ldots \, dU_1 \\
&= -\frac{e^{V_{J-1}}}{e^{V_J} + e^{V_{J-1}}} \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-3}} \prod_{k=1}^{J-2} S \, \exp\{-e^{-U_{J-2}}(e^{V_J} + e^{V_{J-1}})\} \, dU_{J-2} \, \ldots \, dU_1 \\
&= -\frac{e^{V_{J-1}}}{e^{V_J} + e^{V_{J-1}}} \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-3}} \prod_{k=1}^{J-3} S \, e^{-e^{-(U_{J-2}-V_{J-2})}} \, e^{-(U_{J-2}-V_{J-2})} \cdot \\
&\qquad \cdot \exp\{-e^{-U_{J-2}}(e^{V_J} + e^{V_{J-1}})\} \, dU_{J-2} \, \ldots \, dU_1 \\
&= -\frac{e^{V_{J-1}}}{e^{V_J} + e^{V_{J-1}}} \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-3}} \prod_{k=1}^{J-3} S \, \exp\{-e^{-U_{J-2}}(e^{V_J} + e^{V_{J-1}} + e^{V_{J-2}})\} \cdot \\
&\qquad \cdot e^{-(U_{J-2}-V_{J-2})} \, dU_{J-2} \, \ldots \, dU_1
\end{aligned}
\tag{3.20}
$$

$$
\begin{aligned}
&P(t \geq U_1 > U_2 > \cdots > U_J) = \\
&\quad = -\frac{e^{V_{J-1}}}{e^{V_J} + e^{V_{J-1}}} \cdot e^{V_{J-2}} \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-3}} \prod_{k=1}^{J-3} S \, \exp\{-e^{-U_{J-2}}(e^{V_J} + e^{V_{J-1}} + e^{V_{J-2}})\} \cdot \\
&\qquad \cdot e^{-U_{J-2}} \, dU_{J-2} \, \ldots \, dU_1 \\
&\quad = \begin{bmatrix} \text{Substitutions :} & x = \exp\{-U_{J-2}\} \\ & dx = -\exp\{-U_{J-2}\} \, dU_{J-2} \\ & (-\infty \text{ to } U_{J-3}) \text{ becomes } (e^{-U_{J-3}} \text{ to } \infty) \end{bmatrix} \\
&\quad = \frac{e^{V_{J-1}}}{e^{V_J} + e^{V_{J-1}}} \cdot e^{V_{J-2}} \int_{-\infty}^{t} \int_{-\infty}^{U_1} \cdots \int_{-\infty}^{U_{J-3}} \prod_{k=1}^{J-3} S \cdot \\
&\qquad \cdot \exp\{-x(e^{V_J} + e^{V_{J-1}} + e^{V_{J-2}})\} \, dx \, dU_{J-3} \, \ldots \, dU_1 \\
&\quad = \cdots =
\end{aligned}
$$

$$= \frac{e^{V_{J-1}}}{e^{V_J}+e^{V_{J-1}}} \frac{e^{V_{J-2}}}{e^{V_J}+e^{V_{J-1}}+e^{V_{J-2}}} \int_{-\infty}^{t}\int_{-\infty}^{U_1}\cdots\int_{-\infty}^{U_{J-4}} \prod_{k=1}^{J-3} S \cdot$$
$$\cdot \exp\{-e^{-U_{J-3}}(e^{V_J}+e^{V_{J-1}}+e^{V_{J-2}})\}\, dU_{J-3}\ldots dU_1$$
$$= \cdots =$$
$$= \prod_{k=2}^{J-1}\left(\frac{e^{V_k}}{\sum_{n=k}^{J} e^{V_n}}\right) \int_{-\infty}^{t} \exp\{-e^{-U_1}(\sum_{f=1}^{J} e^{V_f})\}\, e^{-(U_1-V_1)}\, dU_1$$
$$= \cdots =$$
$$= \prod_{k=2}^{J-1}\left(\frac{e^{V_k}}{\sum_{n=k}^{J} e^{V_n}}\right) \frac{e^{V_1}}{\sum_{f=1}^{J} e^{V_f}} \exp\{-e^{-U_1}(\sum_{f=1}^{J} e^{V_f})\}$$
$$= \prod_{k=1}^{J-1}\left(\frac{e^{V_k}}{\sum_{n=k}^{J} e^{V_n}}\right) \exp\{-\exp\{-(U_1-\log(\sum_{f=1}^{J} e^{V_f}))\}\}$$
$$= \prod_{k=1}^{J}\left(\frac{e^{V_k}}{\sum_{n=k}^{J} e^{V_n}}\right) \exp\{-\exp\{-(U_1-\log(\sum_{f=1}^{J} e^{V_f}))\}\}. \tag{3.21}$$

Furthermore we can calculate the probability in the denominator of Equation (3.18) as

$$P(U_1 > U_2 > \cdots > U_J) =$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{U_1}\cdots\int_{-\inf^{\infty} y}^{U_{J-1}} dH(U_J)\, dH(U_{J-1}) \ldots dH(U_1)$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{U_1}\cdots\int_{-\infty}^{U_{J-1}} \prod_{k=1}^{J} e^{-e^{-(U_k-V_k)}}\, e^{-(U_k-V_k)}\, dU_J\, dU_{J-1}\ldots dU_1$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{U_1}\cdots\int_{-\infty}^{U_{J-1}} \prod_{k=1}^{J-1} \underbrace{e^{-e^{-(U_k-V_k)}}\, e^{-(U_k-V_k)}}_{\mathrm{S}}\, e^{-e^{-(U_J-V_J)}}\, e^{-(U_J-V_J)}\, dU_J\, dU_{J-1}\ldots dU_1$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{U_1}\cdots\int_{-\infty}^{U_{J-2}} \prod_{k=1}^{J-1} S\, \left[e^{-e^{-(U_J-V_J)}}\right]_{-\infty}^{U_{J-1}} dU_{J-1}\, dU_{J-2}\ldots dU_1$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{U_1}\cdots\int_{-\infty}^{U_{J-2}} \prod_{k=1}^{J-1} S\, e^{-e^{-(U_{J-1}-V_J)}}\, dU_{J-1}\, dU_{J-2}\ldots dU_1$$
$$\overset{\text{Eq.(3.21)}}{=} \cdots =$$

$$
\begin{aligned}
&= \prod_{k=2}^{J-1} \left(\frac{e^{V_k}}{\sum_{n=k}^{J} e^{V_n}} \right) \cdot e^{V_1} \int_{-\infty}^{\infty} \exp\{-e^{-U_1}(\sum_{f=1}^{J} e^{V_f})\}\, e^{-U_1}\, dU_1 \\
&= \begin{bmatrix} \text{Substitutions :} & x = \exp\{-U_1\} \\ & dx = -\exp\{-U_1\}\, dU_1 \\ & (-\infty \text{ to } \infty) \text{ becomes } (\infty \text{ to } 0) \end{bmatrix} \\
&= -\prod_{k=2}^{J-1} \left(\frac{e^{V_k}}{\sum_{n=k}^{J} e^{V_n}} \right) \cdot e^{V_1} \int_{-\infty}^{\infty} \exp\{-x \left(\sum_{f=1}^{J} e^{V_f} \right)\}\, dx \\
&= -\prod_{k=2}^{J-1} \left(\frac{e^{V_k}}{\sum_{n=k}^{J} e^{V_n}} \right) \frac{e^{V_1}}{\left(\sum_{f=1}^{J} e^{V_f} \right)} \left[-\exp\{-x \left(\sum_{f=1}^{J} e^{V_f} \right)\} \right]_{\infty}^{0} \\
&= \prod_{k=1}^{J-1} \left(\frac{e^{V_k}}{\sum_{n=k}^{J} e^{V_n}} \right) \\
&= \prod_{k=1}^{J} \left(\frac{e^{V_k}}{\sum_{n=k}^{J} e^{V_n}} \right). \qquad (3.22)
\end{aligned}
$$

Hence, by inserting the probabilities of Equation (3.21) and (3.22) into Equation (3.18) we obtain

$$
P(U_1 \leq t \mid U_1 > U_2 > \cdots > U_J) = \exp\{-\exp\{-(U_1 - \log(\sum_{f=1}^{J} e^{V_f}))\}\}, \qquad (3.23)
$$

which, together with Equation (3.17), indicates that the conditional distribution of U_1 given the ordering $U_1 > U_2 > \cdots > U_J$ is independent of the ranking. Thus, by assuming a linear in parameter form for V_{ij},

$$
V_{ij} = \mathbf{z}_{ij}^T \beta = \beta_1 z_{j1} + \beta_2 z_{j2} + \cdots + \beta_p z_{jp}, \qquad (3.24)
$$

and if the ranking of an individual's J choices is $R_i = (r_1, r_2, \ldots, r_J)$, the probability of the individual's observed ranking is $P(R_i)$. With the use of Equation (3.22) we can calculate this probability as

$$
\begin{aligned}
P(R_i) &= P(U_{r_1} > U_{r_2} > \cdots > U_{r_J}) \\
&= \prod_{j=1}^{J} \left(\frac{e^{\mathbf{z}_{r_j}^T \beta}}{\sum_{m=j}^{J} e^{\mathbf{z}_{r_m}^T \beta}} \right), \qquad (3.25)
\end{aligned}
$$

and we can conclude that the probability of the individual's observed ranking equals the likelihood of the Cox proportional hazards model, which was derived in Chapter 2 (Equation (2.11).

The likelihood for an independent sample of N individuals can be calculated as $L(\beta) = \prod_{i=1}^{N} P(R_i)$, thus the log likelihood equals

$$\begin{aligned} l(\beta) &= \sum_{i=1}^{N} \log(P(R_i)) \\ &= \sum_{i=1}^{N} \sum_{j=1}^{J} \left(\mathbf{z}_{ir_j}^T \beta \; - \; \log \left(\sum_{m=j}^{J} e^{\mathbf{z}_{ir_m}^T \beta} \right) \right). \end{aligned} \tag{3.26}$$

A unique maximum of the likelihood function exists, since the log likelihood is globally concave in β (Beggs et. al., 1981).

3.4 Usage of the Rank-Ordered Logit Model in an Epidemiological Setting

In an epidemiological setting we apply the rank-ordered logit model to data where the outcome is continuous and we stratify the cohort based on the confounders, hence the individuals within each stratum have equal or similar confounding profile. When the assumptions of the RO-logit model are fulfilled, the regression coefficient β can be interpreted as the per unit change of the outcome when the exposure is increased by one unit in the same way as a standard linear regression. However, when the assumptions are not fulfilled the coefficient can still be interpreted as the log odds ratio, i.e. the log odds of having a higher ranking when the exposure is increased by one unit. This can be seen by Equation 3.10, where we computed the probability that an individual gives a higher rank to element j than to element k, i.e. $U_j > U_k$. This probability was of the form $P(U_j > U_k \; , \; j \neq k) = \frac{e^{V_j}}{e^{V_j} + e^{V_k}}$, which imply that the model can be considered as a conditional logistic regression.

In order to use the RO-logit model the data is stratified based on confounders. For categorical variables one can take the levels of the variable to create strata, while for continuous variables some kind of categorisation is necessary.

In our data set (described in detail in Chapter 5) the outcome was the alcohol abuse score and the drug abuse score of the individuals which is not continuous but ordinal.

However, these scores can be seen as the observed realisation of the underlying continuous "strength of addiction" which can still be ranked, thus the RO-logit model can be used as an estimator.

The outcome may contain ties, especially with an ordinal outcome, and the aim of the thesis is to evaluate methods for handling ties.

Chapter 4

R Programming and Simulations

4.1 Introduction to the Simulation Study and R Programming

In order to evaluate the different methods for handling ties we performed simulations of scenarios with different degree of ties and different exposure and error distributions. For each scenario we simulated 1000 samples of 1000 individuals each and in this Chapter we will describe the different scenarios as well as the tools applied for handling ties. Moreover, the outcome of the simulations will be presented in detail and the packages and the functions used in R will be introduced as well as the reasons for using them.

4.2 Simulations using R

We simulated 4 different scenarios with different exposure and outcome distributions. For each of the four scenarios, we assumed different degree of ties. In the first and second scenario we simulated Extreme Value Type I distributed error terms, and the exposure was normally distributed in the first and binomial in the second. For scenario 3 and 4 we had normally distributed error terms and the exposure was normally distributed in the third and binomial in the fourth. The reason for choosing to simulate these four scenarios was to evaluate the different methods of handling ties (the Breslow, the Efron, the Discrete and the Adding methods) in various situations.

4.2.1 Scenario 1

$$\epsilon \sim \text{EVT1}(\ 0\ ,\ 1\ ) \qquad \& \qquad x \sim \text{N}(0.5 * c\ ,\ 1.5)$$

In our simulation study, we simulated 1000 samples of 1000 individuals each. The linear model that we used is

$$y_i = \beta_0 + \beta_1 \cdot x_i + \beta_2 \cdot c_i + \epsilon_i,$$

where i denotes the i-th individual ($i = 1, \ldots, 1000$), the x and c variables are the exposure and stratum variables respectively and ϵ the error terms.

For each sample, we created a data set of 1000 individuals, containing a c-variable for each person sampled from a vector with integers from 1 to 10, with equal probability. This way we obtained 10 categories (strata) and every individual belongs to one of those 10 groups. Moreover, the exposures of interest were denoted by x, have a Normal distribution, the mean of which depends on the c-variable in order to make c a confounder and more specifically the mean equals $0.5 * c$. In addition, the error terms were assumed to be independent and Extreme Value Type I distributed.

We specified $\beta_0 = 17$ and $\beta_2 = 1$, while for β_1 we considered 0 and 0.15. Furthermore, we created three different proportions of tied event times in our data set (5%, 20% and 60%). In order to do so, after sorting the outcome in increasing order, we specified a number of categories to which the sample of 1000 individuals was to be divided in and instead of the real value of outcomes in the group, all the individuals in the same category were given the mean of the outcomes in the group. For a small proportion of ties we chose 500 categories and in that way 5% of ties were created in the sample. For a medium and large proportion of ties we divided the individuals into 200 and 50 categories respectively.

Since the underlying model is linear in the parameters and by assuming an extreme value type I distribution for the error term the resulting likelihood is on the same form as the likelihood for a stratified Cox-regression. In addition, for us the event indicator in the Cox proportional hazards regression model (`d`) equals 1 for all individuals (see Equation (3.26).Consequently, we fitted a stratified Cox proportional hazards regression model using `coxph` (from survival package).

For the different degree of ties, we used three different methods for handling ties implemented in `coxph`, the Breslow, the Efron and the Discrete. In addition, as mentioned in Section 2.3.4, we added a small, random number sampled from a Uniform distribution to the values of the response variable that were duplicated, which is referred to as the Adding method.

We will now present the `R` code for the the three different methods of handling ties. The Efron method can be implemented as

```
coxph(Surv(y, d) ~ x + strata(c), dataframe, ties = "efron") ,
```

the Breslow method as

```
coxph(Surv(y, d) ~ x + strata(c), dataframe, ties = "breslow") ,
```

and the Discrete method as

```
coxph(Surv(y, d) ~ x + strata(c), dataframe, ties = "exact") .
```

The Adding method was implemented by adding a small number from a uniform distribution to the tied outcomes as

```
dataframe\$y[dup] <- runif(xn, min = 0.00001, max = 0.0001) + dataframe\$y[dup]   ,
```

where `xn` is the number of tied outcomes and `dup` stands for "duplicated" and gives the indices of the vector of outcomes (`y`) that are tied). Thus, in `R` it suffice to fit a stratified Cox proportional hazards regression model using `coxph` with the new outcome vector with no ties.

The results of this simulation is found in Tables 4.1 in Section 4.2.5.

4.2.2 Scenario 2

$$\epsilon \sim \text{EVT1}(\,0\,,\,1\,) \qquad \& \qquad x \sim \text{Bin}(1\,,\,p)$$

For Scenario 2 we followed the steps described above, but instead of a Normally distributed exposure variable, it is now sampled from a Binomial distribution, $x \sim \text{Bin}(1\ ,\ p)$, where

$$p_i = \frac{\exp\{c_i \cdot 0.1\}}{1 + \exp\{c_i \cdot 0.1\}},$$

where i stands for the i-th individual. The results are found in Table 4.2 in Section 4.2.5.

4.2.3 Scenario 3

$$\epsilon \sim \text{N}(\ 0\ ,\ \pi^2/6\ ) \qquad \& \qquad x \sim \text{N}(0.5 * c\ ,\ 1.5)$$

This situation is similar to Scenario 1, where the exposure variable x is normally distributed and the mean of the Normal distribution depends on the c-variable. However, the error terms ϵ_i,are normally distributed, with mean 0 and standard deviation $\pi/\sqrt{6}$, $\epsilon \sim \text{N}(\ 0\ ,\ \pi^2/6\ )$.

The reason for the specific choice of standard deviation is to ensure equal variance for the Extreme Value Type I and normally distributed error terms (as shown in Equation (3.6) the Variance for EVT1 is $\frac{\pi^2}{6}$). Table B.1 in Appendix B, displays the results of this simulation.

4.2.4 Scenario 4

$$\epsilon \sim \text{N}(\ 0\ ,\ \pi^2/6\ ) \qquad \& \qquad x \sim \text{Bin}(1\ ,\ p)$$

Finally, in Scenario 4 we have the normally distributed error terms as above, $\epsilon \sim \text{N}(\ 0\ ,\ \pi^2/6\ )$, and the exposure variable has a Binomial distribution, $x \sim \text{Bin}(1\ ,\ p)$, as described in Scenario 2. See Table B.2 in Appendix B for the results of this simulation.

4.2.5 Results of Simulations

The results of our simulations are summarised in the two following tables (and the two tables in Appendix B, Tables B.1 and B.2). For every scenario we calculated the average coefficient, i.e. the mean of the estimated parameter, the average variance, i.e. the mean of the estimated variances of the coefficients, the empirical variance, i.e. the variance of the estimated coefficients, the coverage, i.e. the percentage of the times the 95% confidence interval covers the true value of β_1 and the power, i.e. the probability of correctly rejecting the null hypothesis, H_0, of the coefficient being 0. Our goal was to evaluate the four available methods of handling ties in Cox-regression for the RO-logit model, for different proportions of ties in the outcome and different exposure and error distributions.

			$\epsilon \sim$ EVT1(0 , 1) & $x \sim$ N				
True β_1	Method	Prop of Ties ($\simeq$)	Avg Coeff ($*10^{-3}$)	Avg Var ($*10^{-3}$)	Emp Var ($*10^{-3}$)	Cov	Type I Error
0	NoTies	-	0.9832	0.4710	0.4459	0.957	0.043
	Efron	5%	0.1700	0.4700	0.4877	0.940	0.060
		20%	1.6220	0.4696	0.4940	0.945	0.055
		60%	1.1022	0.4671	0.4272	0.960	0.040
	Breslow	5%	0.1734	0.4699	0.4816	0.944	0.056
		20%	1.6236	0.4691	0.4727	0.948	0.052
		60%	0.9724	0.4657	0.3639	0.970	0.030
	Discrete	5%	0.1761	0.4761	0.4944	0.942	0.058
		20%	1.7099	0.4926	0.5212	0.942	0.058
		60%	1.1812	0.5706	0.5463	0.954	0.046
	Adding	5%	0.1539	0.4705	0.4898	0.940	0.060
		20%	1.5498	0.4709	0.4958	0.947	0.053
		60%	1.4128	0.4710	0.4427	0.959	0.041
True β_1	Method	Prop of Ties ($\simeq$)	Avg Coeff	Avg Var ($*10^{-3}$)	Emp Var ($*10^{-3}$)	Cov	Pow
0.15	NoTies	-	0.1514	0.4938	0.4748	0.959	1
	Efron	5%	0.1503	0.4925	0.5057	0.946	1
		20%	0.1508	0.4921	0.5081	0.944	1
		60%	0.1468	0.4889	0.4613	0.960	1
	Breslow	5%	0.1494	0.4923	0.4996	0.949	1
		20%	0.1476	0.4913	0.4865	0.953	1
		60%	0.1342	0.4860	0.3880	0.910	1
	Discrete	5%	0.1514	0.4992	0.5122	0.948	1
		20%	0.1551	0.5180	0.4344	0.942	1
		60%	0.1652	0.6051	0.5844	0.911	1
	Adding	5%	0.1504	0.4929	0.5071	0.950	1
		20%	0.1512	0.4935	0.5088	0.947	1
		60%	0.1478	0.4933	0.4806	0.954	1

TABLE 4.1: Scenario I of simulations, where 1000 samples of 1000 individuals were simulated, the exposures of interest were normally distributed and the error terms extreme value type I distributed. The true parameter, β_1, was assumed to take values 0 and 0.15 and three different degrees of ties (proportion of ties) were considered and handled by three different approximations (methods). The results are the average coefficient (Avg Coeff), the average variance (Avg Var), the empirical variance (Emp Var), the coverage (Cov), the type I error (Type I Error) and the power (Pow).

When β_1 equals 0, Table 4.1 shows that for a small proportion of ties in the data (5%), the estimated coefficient is closer to the true value of β_1 when using the Adding method compared to the others. Moreover, none of the methods returns biased estimates, in the sense that all the, somewhat crudely calculated, 95% confidence intervals cover the true value of β_1. However, the Type I error is above 0.05, indicating that we could incorrectly reject the null hypothesis (false positive) with a probability more than the commonly accepted 5%. In addition, the Adding method also performs better for the medium proportion of ties (20%), the Efron and Breslow are fairly good and the Discrete tends to overestimate β_1. However, for all four methods we get somewhat biased estimates but the type I errors are closer to 5%. In the case where the proportion of ties is large (60%), Breslow seems to give a better approximation than the other three. Moreover, the 95% confidence intervals for the coefficients cover the true value of β_1 except of the case of the Adding method, which returned a biased estimate. In addition, Type I error was below 0.05 for all four methods. For all three degrees of ties, confidence interval coverage for all three estimators fell in the range of $94 - 97\%$. Moreover, our approach is considered conservative, since the variance is in most of the cases overestimated.

In the case where β_1 was equal to 0.15, and for a small and medium proportion of ties (5 & 20%), Efron seems to be the best approximation, since using that method we get the closest average coefficient to the true value of β_1. Moreover, when handling ties with the other 3 available methods, the true β_1 is estimated quite well and empirical variance is close to the average variance. However, there are some biased estimates when using the Discrete for small proportion of ties (5%) and the Breslow and Discrete for medium proportion of ties in the data (20%) respectively. For the large proportion of ties the Adding method, performs better than the other three. The power is 1 for all three degrees of ties and all methods of handling ties, while the confidence interval coverage ranges from 91 to 96%.

			$\epsilon \sim$ EVT1(0 , 1) & $x \sim$ Bin				
True β_1	Method	Prop of Ties ($\simeq$)	Avg Coeff ($*10^{-3}$)	Avg Var ($*10^{-2}$)	Emp Var ($*10^{-2}$)	Cov	Type I Error
0	NoTies	-	0.6303	0.4615	0.4559	0.956	0.044
	Efron	5%	-0.4663	0.4645	0.4338	0.953	0.047
		20%	-1.4723	0.4604	0.4454	0.958	0.042
		60%	0.2696	0.4575	0.4315	0.964	0.036
	Breslow	5%	-0.5047	0.4613	0.4286	0.952	0.048
		20%	-1.3423	0.4601	0.4264	0.960	0.040
		60%	0.3253	0.4563	0.3643	0.980	0.020
	Discrete	5%	-0.5080	0.4676	0.4403	0.951	0.049
		20%	-1.3830	0.4836	0.4711	0.955	0.045
		60%	0.4827	0.5604	0.5486	0.961	0.039
	Adding	5%	-0.4568	0.4619	0.4355	0.953	0.047
		20%	-1.5498	0.4618	0.4534	0.956	0.044
		60%	-0.0373	0.4616	0.4569	0.960	0.040
True β_1	Method	Prop of Ties ($\simeq$)	Avg Coeff	Avg Var ($*10^{-2}$)	Emp Var ($*10^{-2}$)	Cov	Pow
0.15	NoTies	-	0.1509	0.4647	0.4538	0.961	0.612
	Efron	5%	0.1496	0.4646	0.4362	0.953	0.603
		20%	0.1478	0.4633	0.4465	0.957	0.572
		60%	0.1451	0.4598	0.4258	0.966	0.588
	Breslow	5%	0.1487	0.4644	0.4306	0.954	0.601
		20%	0.1447	0.4629	0.4274	0.961	0.557
		60%	0.1332	0.4584	0.3581	0.968	0.514
	Discrete	5%	0.1508	0.4709	0.4424	0.954	0.603
		20%	0.1522	0.4868	0.4726	0.956	0.574
		60%	0.1639	0.5642	0.5436	0.960	0.601
	Adding	5%	0.1498	0.4651	0.4381	0.951	0.609
		20%	0.1483	0.4648	0.4528	0.955	0.575
		60%	0.1465	0.4641	0.4495	0.959	0.592

TABLE 4.2: Scenario II of simulations, where 1000 samples of 1000 individuals were simulated, the exposures of interest had a Binomial distribution and the error terms extreme value type I distributed. The true parameter, β_1, was assumed to take values 0 and 0.15 and three different degrees of ties (proportion of ties) were considered and handled by three different approximations (methods). The results are the average coefficient (Avg Coeff), the average variance (Avg Var), the empirical variance (Emp Var), the coverage (Cov), the type I error (Type I Error) and the power (Pow).

From Table 4.2 we can see that when β_1 equals 0, and for a small and large proportion of ties (5 & 60%), the Adding method of handling ties estimates the true value of β_1 better, in the sense that the estimates for β_1 are closest to the true value. For the medium proportion of ties Breslow seems to give a better estimate for the true parameter than all the other methods. The the mean of the estimated variances of the coefficients (the average variance) is for all cases almost equal to the variance of the estimated coefficients (the empirical variance) and the coverage ranges from $95.2-98\%$. Furthermore, the type I error is on average around 5%. Also, the 95% confidence intervals for the coefficients show that there is no bias, since in all case the confidence intervals cover the true value of β_1.

In the second half of Table 4.2, we see that when $\beta_1 = 0.15$, and for all proportion of ties the Efron and Adding methods preform better. The power is 1 for all cases and all proportion of ties and the coverage falls between 91 and 95.9%. However, for 60% of ties Efron and Discrete give biased estimates, while Breslow returns biased estimates for both medium and large proportion of ties. The power ranges from 51.4% to 60.9% and the variance is in all cases overestimated.

In Appendix B one can find the results of the simulation of Scenarios 3 (Table B.1) and 4 (Table B.2). For Scenario 3, when β_1 equals 0, there is no bias, since in all cases the confidence intervals cover the true value of β_1. For a small proportion of ties the Breslow method performs better, while for medium and large proportion of ties Adding method gives estimates closer to the true β_1. However, when β_1 equals 0.15, the estimates are biased for all methods and degrees of ties, which could be due to the normally distributed error terms. Moreover, for all degrees of ties the Discrete method performs better than the other three.

For Scenario 4 (Table B.2), when β_1 equals 0, there are no biased estimates for all methods and degrees of ties. For a small proportion of ties the Efron method performs better, while for medium and large proportion of ties Adding method gives estimates closer to the true β_1. For $\beta_1 = 0.15$, Discrete method is the best approximation although, similarly with Scenario 3, there is bias for all methods and degrees of ties.

4.3 The R Packages Used

4.3.1 Package 'foreign'

In order to perform the real data analysis, we used the package 'foreign' (cran.r-project.org) to read the data that were provided in Stata format into a data frame in R.

4.3.2 Package 'SpatialExtremes'

We used the function `rgev` of the `SpatialExtremes` package (`cran.r-project.org`) in order to generate Extreme value type I distributed error terms in our simulations. We specified a location parameter 0 and scale parameter 1 in all cases.

4.3.3 Package 'stats'

The Package 'stats' (`R Documentation`) was used in multiple occasions throughout our simulations and real data analysis. The function `rnorm` was used for the random generation for the normal distribution. Moreover, The (S3) generic function `density` was used to compute kernel density estimates and `kruskal.test` was used in order to performs multiple Kruskal-Wallis rank sum tests.

4.3.4 Package 'survival'

In order to fit the Cox proportional hazards regression models needed in our analysis, we used the function `coxph` of the package 'survival' (`cran.r-project.org`). This function also provided us with three possible options for handling tied event times, the Breslow approximation, the Efron approximation and the "discrete" option. Although, as described below the options are `ties=c("efron","breslow","exact")`, the "exact" options as described in `cran.r-project.org`, stands for the Discrete approximation for handling ties. More specifically, it is written in the description of the methods for handling ties that "Using the "exact partial likelihood" approach the Cox partial likelihood is equivalent to that for matched logisitic regression. (The clogit function uses the coxph code to do the fit.) It is technically appropriate when the time scale is discrete and has only a few unique values, and some packages refer to this as the "discrete" option. There is also an "exact marginal likelihood" due to Prentice which is not implemented here" (`https://cran.r-project.org/web/packages/survival/survival.pdf`).

4.3.5 Package 'xlsx'

As mentioned above, when performing the real data analysis it was needed to read the data that were provided in Stata format into a data frame in `R`. Eventually, we used the function `write.xlsx`, of the package 'xlsx' (`cran.r-project.org`), to write a `a data.frame` to an Excel workbook. This way it was easier to import data into `R` and after "clearing" the data set to re-import them when needed.

Chapter 5

Real Data Analysis

Maria Ungdom is a family planning and sexual health clinic in Stockholm, Sweden, where children, adolescents and young adults (between the ages of 16 and 25 years) who have problems with alcohol or drugs, can find help. Moreover, this clinic is part of the Addiction Centre Stockholm and the Stockholm County Council, and it also interacts with, among others, social services and the police (Maria Ungdom website `http://mariaungdom.se`).

We were provided with a data set from Maria Ungdom, which we will describe in this Chapter, along with the data analysis that we performed. We will use this data set to investigate whether it is a significant relationship between some single nucleotide polymorphisms (SNP) (described in detail in Section 5.1) included in the data set, and alcohol or drug abuse score of the individuals participating in this study. Alcohol and drug abuse score, which is our outcomes, indicate the size or graveness of the addiction, and the higher the score, the greater the addiction. This is a very suitable data set for this thesis since it is a nominal outcome with potential for a large degree of ties.

5.1 Description of the data set

The data set provided by Maria Ungdom, consisted of 180 individuals, clients of the clinic who were addicted to alcohol and/or drugs, as well as their parents and siblings. The data included information on alcohol and drug addiction both represented by a score, `AUDIT` (Alcohol Disorders Identification Test) and `DUDIT` (Drug Use Disorders Identification Test) respectively (the greater the score the more serious the abuse), which in our data took values from 0 to 35 and from 0 to 42 respectively. Both the variables of `AUDIT` and `DUDIT` (Bergman a. 2012, Chapter 2 page 20) are tools for identifying problems

with alcohol and drugs. Both AUDIT and DUDIT are calculated by short questionnaires with 10 and 11 questions respectively (Bergman a. 2012, Chapter 2 page 20). AUDIT was developed by WHO (World Health Organization http://www.who.int/en/) in the late 1980s and DUDIT was developed by the Karolinska Institute in the early 2000s, and the calculation of the response scores of both instruments is done similarly (Bergman a. 2012, Chapter 2 page 20). For illustration purposes we include Table 5.1 (Bergman a. 2012, Chapter 2 page 21), where the response scores for AUDIT are presented.

Risk Level	AUDIT Points	Interpretation
Zone I	Men 0-7 Women 0-5	Not risky alcohol habits
Zone II	Men 8-15 Women 6-13	Heavy alcohol consumption but not necessarily an abuse dependence
Zone III	Men 16-19 Women 14-17	Problematic alcohol use, it is likely that there is an alcohol-related diagnosis
Zone IV	Men 20+ Women 18+	Very problematic alcohol consumption, it is likely that there is an alcohol-related diagnosis

TABLE 5.1: Response scores for AUDIT and interpretation.

In addition, and also for illustration purposes, we plot the Kernel density function of the alcohol and drug abuse score of the data set. Figures 5.1 and 5.2 show that neither alcohol nor drug abuse score are normally distributed and the skewed distribution indicate that an EVT1 distribution may be more suitable than a normal distribution.

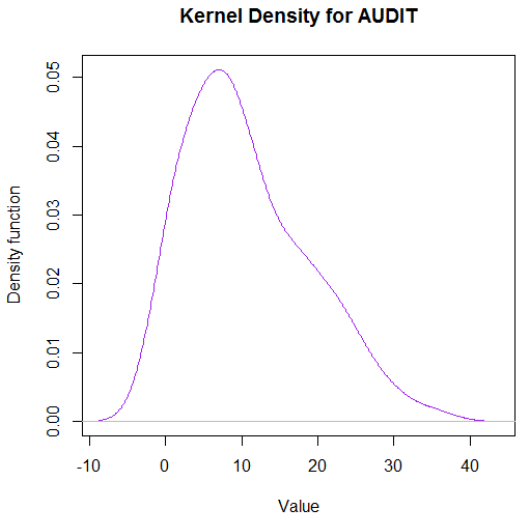

FIGURE 5.1: Kernel density function for alcohol abuse score (AUDIT).

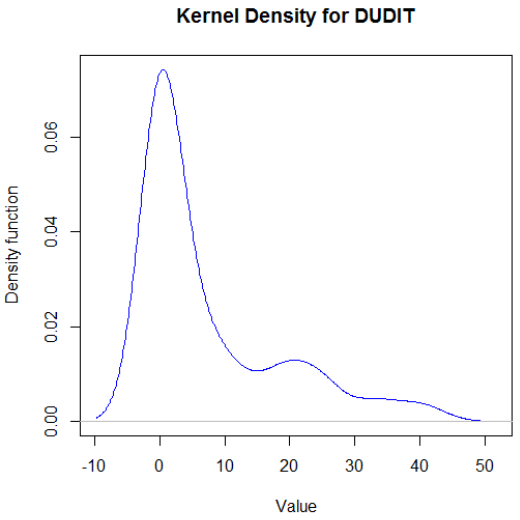

FIGURE 5.2: Kernel density function for drug abuse score (DUDIT).

In addition, our data set consisted of information about the family history (including the client), such as whether the parents or the siblings were also drug or alcohol abusers, whether the family received welfare, whether there was depression in the family or history of physical or sexual abuse. We consider these variables to be possible confounders and aim to match them away using the RO-logit model.

Finally, the data set provided information about 64 single nucleotide polymorphisms (SNPs, pronounced *'snips'*). "A SNP is a variation in DNA occurring when a single nucleotide (A, T, C or G) at a particular site in the genome differs between members of a species" (Kirch 2008, p. 1305). For example, for two different individuals we obtain two stretches of a DNA sequence at the same site, which after they are sequenced, are: TTGC**T**ATT for the first and TTGC**A**ATT for the second. Hence, there is a difference in a single nucleotide, and if both A allele and T allele at the specific site were frequent enough on the chromosomes in the population of interest, the variants at this genomic site would be called a biallelic SNP (Kirch 2008, p. 1305).

Furthermore, SNPs are determined by the unordered combinations of the two nucleotides observed at the same site "AA", "Aa" or "aa", where "AA" and "aa" are said to be homozygotes and "Aa" heterozygotes and "A" and "a" are the alleles A, T, C or G (Hommersom et al.2015, Chapter 9). We considered the SNPs as numeric variables with three levels, 0, 1 and 2. For example for a chosen SNP we could have that it is of the form G:G, G:A and A:A. Then the most common variant out of the homozygotes, say G:G for example, was chosen as reference i.e. the variable takes the value 0 and the

least common (A:A), takes the value 2. The heterozygotes would always take the value 1. Hence for the specific SNP the explanatory variable would be

$$X_i = \begin{cases} 0 \text{ if the SNP is G:G} \\ 1 \text{ if the SNP is G:A} \\ 2 \text{ if the SNP is A:A} \end{cases}$$

Our final data set contained 61 SNPs, since 3 out of the 64 of the initial data set were removed due to missing values.

Moreover, we created four binary variables, `welfare`, `depression`, `sexual_abuse` and `physical_abuse` which took value 1 if at least one of the members of the family had what the variable names describe or else 0. Our final data set consisted of 93 individuals, out of the 180 included in the initial data set, after removing individuals with missing values on SNPs, alcohol and drug addiction score. We discarded completely the cases with missing values in one of those variables, since we assumed that they occurred at random. Interaction of these variables formed 12 matched sets with sizes ranging from 1 to 15. However, the subject in the stratum of size 1 was removed, since there was no effect of his presence in the data set and it does not contribute to the estimation in stratified Cox-regressions.

The `AUDIT` and `DUDIT` contained 67% and 72% of ties respectively.

5.2 Results of real data analysis

The purpose of our study was to determine if there are any significant associations between a selection of SNPs and alcohol and drug abuse scores when the confounders "family receiving welfare", "depression in the family", "history of physical abuse" and "history of sexual abuse" are matched away. We applied the four methods of handling ties in Cox-regression in the RO-logit model on the data set from Maria Ungdom. Moreover, we performed the Kruskal-Wallis test on the data, a non parametric test used to compare three or more independent samples, which gives us the crude or unadjusted p-values (Theodorsson-Norheim E., 1986). The significant SNPs (when ignoring multiple testing) are given in Tables 5.2 and 5.3, while the full tables are given in the Appendix B (Tables B.3, B.4, B.5, B.6, B.7, B.8, B.9, B.10, B.11, B.12, B.13, B.14, B.15, B.16, B.17, B.18, B.19 and B.20).

Alcohol abuse score				
Method of handling ties : Efron				
SNP	Odds ratio	Confidence Interval		p-value
rs1800497	0.5184608	(0.2873493 ,	0.9354523)	0.02913901
rs2740204N	1.3974518	(1.0148388 ,	1.9243170)	0.04034082
rs2770378N	1.4125413	(1.0018537 ,	1.9915813)	0.04877354
rs2290045N	1.8200281	(1.1229623 ,	2.9497895)	0.01506892
Method of handling ties : Breslow				
SNP	Odds ratio	Confidence Interval		p-value
rs1800497	0.5397309	(0.3006418 ,	0.9689587)	0.03886413
rs2770378N	1.4243156	(1.0092177 ,	2.0101460)	0.04419878
rs2290045N	1.7835376	(1.1002613 ,	2.8911372)	0.01889030
Method of handling ties : Discrete				
SNP	Odds ratio	Confidence Interval		p-value
rs6277N	0.6825601	(0.4696162 ,	0.9920618)	0.04530897
rs1800497	0.5065882	(0.2730594 ,	0.9398380)	0.03102259
rs2740204N	1.4021167	(1.0070647 ,	1.9521400)	0.04531774
rs2770378N	1.4694843	(1.0243549 ,	1.1080429)	0.03655477
rs2290045N	1.8772316	(1.1338843 ,	3.1078996)	0.01434558
Method of handling ties : Adding				
SNP	Odds ratio	Confidence Interval		p-value
rs1800497	0.5285183	(0.2943200 ,	0.9490744)	0.03276118
rs2740204N	1.4077230	(1.0214635 ,	1.9400438)	0.03663809
rs2290045N	1.7676635	(1.0879853 ,	2.8719452)	0.02141676
Kruskal-Wallis Test				
SNP				p-value
rs1799971				0.05038
rs2740204N				0.04431

TABLE 5.2: The significant results of applying the RO-logit model and the Kruskal-Wallis test on the real data, with exposure the particular SNP and outcome the alcohol abuse score. The methods for handling ties used are the Efron, Breslow, Discrete and Adding.

As shown in Table 5.2, using all four methods of handling ties (the Efron, Breslow, Discrete and Adding), revealed that the SNPs rs2290045N and rs1800497 have a statistically significant relationship with alcohol abuse score on the 5% level.For rs2290045N we get a p-value around 0.017 (it varies from 0.01434558 to 0.021416) and the odds

ratio was for all methods around 1.8. This imply that having the SNP rs2290045N is associated with an increased probability of high alcohol abuse score. For rs1800497 the p-value was approximately 0.033 and the odds ratio below 1 indicating that having the SNP rs1800497 decreases the probability of a high alcohol abuse score. However, doing a crude test like the Kruskal-Wallis test, where there was no adjustment for the confounders these SNPs were not significant with p-values 0.415 and 0.663 respectively, Table B.11, Appendix B.

Moreover, the use of Efron, Discrete and Adding method resulted in a statistically significant effect of the SNP rs2740204N. In all three cases the p-value was less than 0.05, the 95% confidence interval was above 1 and the odds ratio was approximately 1.4 (OR$\simeq$ 1.4). The Breslow method had a p-value of 0.05217597, and the odds ratio was 1.3728806 (Table B.6). That specific SNP was also found by Kruskal-Wallis test with a p-value of 0.04431.

Furthermore, the odds of having a high alcohol abuse score was higher for the individuals who had the SNP rs2770378, as shown in Table 5.2. With the methods Efron, Breslow and Discrete, the p-value for that SNP was less than 0.05 (varying from 0.037 to 0.044) and the odds ratio was approximately 1.43. When using the Adding method of handling ties the p-value was 0.059 and the odds ratio 1.39, indicating that we could still consider it a significant SNP. Once more, the Kruskal-Wallis test did not reveal the significance of this SNP (p-value= 0.197, Table B.11).

Another SNP which was associated with alcohol abuse score, was rs6277N and when using the Discrete method for handling ties the p-value was equal to 0.0453. When using the Efron, Breslow and Adding the p-values were 0.0552, 0.0561 and 0.071 respectively, while the odds ratio for all four methods was below 1, approximately 0.7, indicating that individuals who have this SNP is less likely to have a high alcohol abuse score. For this SNP, Kruskal-Wallis test resulted in a p-value of 0.674, as shown in Table B.11.

In Table 5.3 below, the significant SNPs found when we investigated their relationship with the drug abuse score of the data set are presented.

Drug abuse score				
Method of handling ties : Efron				
SNP	Odds ratio	Confidence Interval		p-value
rs53576N	1.4631316	(1.0239376 ,	2.0907076)	0.03662662
rs237880N	0.6173289	(0.4189563 ,	0.9096294)	0.01473112
Method of handling ties : Discrete				
SNP	Odds ratio	Confidence Interval		p-value
rs53576N	1.7623717	(1.0457907 ,	2.9699574)	0.03332460
rs237880N	0.5631450	(0.3478131 ,	0.9117891)	0.01951121
Method of handling ties : Adding				
SNP	Odds ratio	Confidence Interval		p-value
rs237880N	0.6085972	(0.4110998 ,	0.9009748)	0.01310276
Kruskal-Wallis Test				
SNP				p-value
rs6190N				0.01288
rs521674				0.04125
rs602618				0.0406
rs237880N				0.03812
rs1488467N				0.01669

TABLE 5.3: The significant results of applying the RO-logit model and the Kruskal-Wallis test on the real data, with exposure the particular SNP and outcome the drug abuse score. The methods for handling ties used are the Efron, Breslow, Discrete and Adding.

The SNP rs237880N has a statistically significant relationship with drug abuse score at the 5% level, for the Efron, Discrete and Adding methods. The p-value varies from 0.0131 to 0.0195 and the odds ratio is below 1, around 0.57. Moreover, when using the Breslow method we found a p-value of 0.0673 and the odds ratio equal to 0.699, also below 1. Hence, the probability that an individual who has the SNP to have a high drug abuse score is smaller than the probability of an individual who has not. The Kruskal-Wallis test also resulted in a p-value of 0.0381 for that SNP.

Furthermore, one can see that when using Efron and Discrete methods for handling ties the SNP rs53576N appears to have a statical significant relationship at the 5% level (p-value 0.0366 and 0.0333 for each methods respectively). The odds ratio is above 1 for both methods which indicates that is more likely to have a high drug abuse score an individual who has the particular SNP. However, when using the Breslow or Adding method does not result in a p-value less than 0.05 (p-value 0.119 and 0.216 for each of

the methods respectively). Moreover, the Kruskal-Wallis test, as shown in Table B.20 of the Appendix B, resulted in a p-value of 0.656.

Finally, it is interesting that Kruskal-Wallis test resulted in four significant SNPs (rs6190N, rs521674, rs602618 & rs1488467N) for the drug abuse score that were not identified in the rank-ordered logit model. It could be due to confounding, since there was no adjusting for confounders, which can lead to inaccurate results.

From the analysis we see that there are SNPs that can affect both the alcohol and drug abuse score. Moreover, adjusting for confounders seems to be important, since the Kruskal-Wallis test failed to identify the significant relationship of some SNPs with the alcohol or drug abuse score, while in other cases possibly wrongly identified an effect of some SNPs when the RO-logit model did not.

Chapter 6

Discussion and Conclusion

The aim of this thesis was to evaluate available tools for handling ties in Cox-regression in the RO-logit model. In Chapter 2, we derived the partial likelihood of Cox proportional hazards model with no tied event times and we presented four methods of handling tied event times in the model. In Chapter 3, the Rank-Ordered logit model, a model originally developed and applied in econometrics (Beggs et. al, 1981), was presented. In Section 3.3, we proved that for the RO-logit model, when the underlying model is linear in the parameters and by assuming an extreme value type I distribution for the error term the resulting likelihood is on the same form as the likelihood for a stratified Cox-regression. Similar to Cox-regression in survival analysis, the estimator assumes no ties in the outcome, but since the estimator has the same form as a stratified Cox-regression it was reasonable to assume that methodology for handling ties in survival analysis can be adopted in the RO-logit model.

In Chapter 4, we simulated different scenarios with different degree of ties and different exposure and error terms distributions. For the first scenario we had Extreme Value Type I distributed error terms and normally distributed exposure ($\epsilon \sim$ EVT1(0 , 1) & $x \sim$ N). When β_1 equals 0 and for small and medium (5 and 20%) proportion of ties the estimated coefficient is closer to the true value of $beta_1$ when using the Adding method compared to the others, while for large (60%) proportion of ties the Breslow has the best performance. When β_1 was equal to 0.15 and for small and medium (5 and 20%) proportion of ties the Efron method performs better and for large the Adding is better. For the second scenario we had binomially distributed exposure ($\epsilon \sim$ EVT1(0 , 1) & $x \sim$ Bin)and when β_1 was 0 for small and large proportion of ties Adding was the best method, while for medium ties Breslow was best. When β_1 was equal to 0.15 Efron and Adding methods gave the best estimates. For the third scenario we assumed a Normal distribution for the error terms and the exposure

($\epsilon \sim$ N(0 , $\pi^2/6$) & $x \sim$ N) and when β_1 equals 0 for medium and large proportion of ties the Adding method performed better, while for small proportion of ties the Breslow method performed better. In the case where β_1 was assumed to be 0.15 and for all degrees of ties the Discrete method performs better than the other three. For the fourth scenario we had normally distributed error terms and binomially distributed exposure ($\epsilon \sim$ N(0 , $\pi^2/6$) & $x \sim$ Bin) and for β_1 equal to 0 and for medium and large proportion of ties Adding method gave estimates closer to the true β_1, while for small Efron. In the case where β_1 was assumed to be 0.15 and for all degrees of ties the Discrete method performs better than the other three. Furthermore, our approach is considered conservative, since the variance was in most cases overestimated. Generally, it is preferable to have overestimated variance compared to underestimated, which can lead to false conclusions. In addition, the proportion of ties corresponds to the whole data set, hence the proportion of ties in each stratum could vary. However, we expect that the proportion of ties will be more or less the same in each stratum.

In Chapter 5, we apply the tools of handling ties on the data set from Maria Ungdom and looked at the associations between a selection of SNPs and alcohol and drug abuse scores. The data set consisted of 180 clients of the clinic who were addicted to alcohol and/or drugs, as well as their parents and siblings. The data included information on alcohol and drug addiction both represented by a score, information about the family history (including the client), such as whether the parents or the siblings were also drug or alcohol abusers, whether the family received welfare, whether there was depression in the family or history of physical or sexual abuse. We consider these variables to be possible confounders and we aimed to match them away using the RO-logit model. In addition, the data set provided information about 64 single nucleotide polymorphisms (SNPs). After excluding clients with missing values on SNPs, alcohol or drug abuse score, we were left with 93 individuals. We divided the data into different strata, where subjects within a stratum had similar family history and represented a matched set. However, since the data set in the end was very small the confounding profiles were quite crude. Consequently, as shown in Section 5.2, we applied the RO-logit model (using all four methods of handling ties, Efron, Breslow, Discrete and Adding), and we performed the Kruskal-Wallis test on the real data, with the particular SNP as exposure and alcohol or drug abuse score as the outcome.

When the outcome was considered to be the alcohol abuse score, as shown in Table 5.2, using the methods of handling ties (the Efron, Breslow, Discrete and Adding), revealed that the SNPs rs2290045N, rs1800497, rs2740204N, rs2770378N and rs6277N have a statistically significant relationship with alcohol abuse score on the 5% level. These relationships have not been discovered before. When doing a crude test like the Kruskal-Wallis test, where there was no adjustment for the confounders, only rs2740204N

was sognificant. Moreover, the SNP rs1799971 was found by Kruskal-Wallis test, while it was not found by the RO-logit model.

When the drug abuse score was considered to be the outcome, from Table 5.3, we found that using the RO-logit model and the four methods of handling ties, resulted in two significant SNPS, rs53576N and rs237880N. The Kruskal-Wallis test found four significant SNPs (rs6190N, rs521674, rs602618 & rs1488467N) for the drug abuse score that were not identified in the rank-ordered logit model. This could be due to confounding, since there was no adjusting for confounders, which can lead to inaccurate results. However, it also found one of the two SNPs we found by applying the RO-logit model on data (rs237880N).

From the analysis we see that there are SNPs that can affect both the alcohol and drug abuse score. Furthermore, adjusting for confounders seems to be important, since the Kruskal-Wallis test failed to identify the significant relationship of some SNPs with the alcohol or drug abuse score, while in other cases possibly wrongly identified an effect of some SNPs when the RO-logit model did not. Our findings suggest that the four methods of handling ties are equivalent since they produced similar results.

One limitation of the RO-logit model is that the error terms must have an extreme value type I distribution, in order to obtain a likelihood of the same form as the likelihood for a stratified Cox-regression, which may not often be the case. Another limitation concerns matching, as a way of adjusting for confounders. The advantage of this method is that taking into account the relationship between the outcome and the potential confounders is not needed and when the explanatory variable (exposure) is continuous we can categorise it in order to create strata. However, when the sample is small, like in our case, the groups or strata turn out to be quite crude and the confounding profiles quite broad.

To conclude, in this thesis we focused on handling ties when the outcome was ordinal and we used the RO-logit model by "matching away" possibly complex relationships between confounders and exposure/outcome by fitting stratified Cox-regressions. Similar to Cox-regression in survival analysis, the estimator assumes no ties in the outcome, however, handling ties in the same way as for Cox-regression in the RO-logit model, turned out to be possible. However, this model should be used with caution since for some methods of handling ties there was some bias. More specifically, when extreme value type I distribution was assumed for the error terms in the simulations we performed, for Scenario 1 and for β_1 equal to 0, all four methods were biased for medium proportions of ties, while for large proportion of ties the Adding returned biased estimates. When β_1 was equal to 0.15, the Breslow method was biased for small proportion of ties, while the Discrete was biased for both small and medium proportion of ties. Furthermore,

for Scenario 2 and for β_1 equal to 0.15 the Efron and the Discrete methods returned biased estimates for large proportion of ties, while the Breslow for both medium and large. Finally, the Adding method seems to have an advantage over the others, however in general all methods produced similar results.

Appendix A

The R Packages

A.1 The R Packages and The Functions Used In Simulations and Real Data Analysis

For this thesis we used R in order to perform simulations and the real data analysis. In this Appendix we will describe the packages, as well as the functions we used, how they can be used and which are their arguments.

A.1.1 Package 'foreign'

In order to perform the real data analysis, we used the package 'foreign' (cran.r-project.org) to read the data that were provided in Stata format into a data frame in R.

A.1.1.1 Usage

```
read.dta(file, convert.dates = TRUE, convert.factors = TRUE,
         missing.type = FALSE,
         convert.underscore = FALSE, warn.missing.labels = TRUE)
```

A.1.1.2 Arguments

`file`	a filename or URL as a character string.
`convert.dates`	Convert Stata dates to `Date` class, and date-times to `POSIXct` class?
`convert.factors`	Use Stata value labels to create factors? (Version 6.0 or later).
`missing.type`	For version 8 or later, store information about different types of missing data?
`convert.underscore`	Convert "_" in Stata variable names to "." in R names?
`warn.missing.labels`	Warn if a variable is specified with value labels and those value labels are not present in the file.

A.1.2 Package '`SpatialExtremes`'

We used the function `rgev` of the `SpatialExtremes` package (`cran.r-project.org`) in order to generate Extreme value type I distributed error terms in our simulations. We specified a location parameter 0 and scale parameter 1 in all cases.

A.1.2.1 Usage

```
rgev(n, loc = 0, scale = 1, shape = 0)
```

A.1.2.2 Arguments

`n`	number of observations.
`loc`	vector of the location parameters.
`scale`	vector of the scale parameters.
`shape`	a numeric of the shape parameter.

A.1.3 Package '`stats`'

The Package '`stats`' (`R Documentation`) was used in multiple occasions throughout our simulations and real data analysis. The function `rnorm` was used for the random generation for the normal distribution. Moreover, The (S3) generic function `density` was used to compute kernel density estimates and `kruskal.test` was used in order to performs multiple Kruskal-Wallis rank sum tests.

A.1.3.1 Usage

Function 1

```
rnorm(n, mean = 0, sd = 1)
```

Function 2

```
density(x, bw = "nrd0", adjust = 1,
        kernel = c("gaussian", "epanechnikov", "rectangular",
                   "triangular", "biweight",
                   "cosine", "optcosine"),
        weights = NULL, window = kernel, width,
        give.Rkern = FALSE,
        n = 512, from, to, cut = 3, na.rm = FALSE, ...)
```

Function 3

```
kruskal.test(formula, data, subset, na.action, ...)
```

A.1.3.2 Arguments

Function 1

`n`	number of observations. If `length(n) > 1`, the length is taken to be the number required.
`mean`	vector of means.
`sd`	vector of standard deviations.

Function 2

x	the data from which the estimate is to be computed.
bw	the smoothing bandwidth to be used. The kernels are scaled such that this is the standard deviation of the smoothing kernel. (Note this differs from the reference books cited below, and from S-PLUS.)
adjust	the bandwidth used is actually adjust*bw. This makes it easy to specify values like 'half the default' bandwidth.
kernel, window	a character string giving the smoothing kernel to be used. This must partially match one of "gaussian", "rectangular", "triangular", "epanechnikov", "biweight", "cosine" or "optcosine", with default "gaussian".
weights	numeric vector of non-negative observation weights, hence of same length as x. The default NULL is equivalent to weights = rep(1/nx, nx) where nx is the length of (the finite entries of) x[].
width	this exists for compatibility with S; if given, and bw is not, will set bw to width if this is a character string, or to a kernel-dependent multiple of width if this is numeric.
give.Rkern	logical ; if TRUE, no density is estimated, and the 'canonical bandwidth' of the chosen kernel is returned instead.
n	the number of equally spaced points at which the density is to be estimated.
from,to	the left and right-most points of the grid at which the density is to be estimated; the defaults are cut * bw outside of range(x).
cut	by default, the values of from and to are cut bandwidths beyond the extremes of the data. This allows the estimated density to drop to approximately zero at the extremes.
na.rm	logical; if TRUE, missing values are removed from x. If FALSE any missing values cause an error.
...	further arguments for (non-default) methods.

Function 3

x	a numeric vector of data values, or a list of numeric data vectors. Non-numeric elements of a list will be coerced, with a warning.
g	a vector or factor object giving the group for the corresponding elements of x. Ignored with a warning if x is a list.
formula	a formula of the form response ∼ group where response gives the data values and group a vector or factor of the corresponding groups.
data	an optional matrix or data frame (or similar: see model.frame) containing the variables in the formula formula. By default the variables are taken from environment(formula).
subset	an optional vector specifying a subset of observations to be used.
na.action	a function which indicates what should happen when the data contain NAs. Defaults to getOption("na.action").
...	further arguments to be passed to or from methods.

A.1.4 Package 'survival'

In order to fit the Cox proportional hazards regression models needed in our analysis, we used the function coxph of the package 'survival' (cran.r-project.org). This function also provided us with three possible options for handling tied event times, the Breslow approximation, the Efron approximation and the "discrete" option. Although, as described below the options are ties=c("efron","breslow","exact"), the "exact" options as described in cran.r-project.org, stands for the discrete approximation for handling ties. More specifically, it is written in the description of the methods for handling ties that "Using the "exact partial likelihood" approach the Cox partial likelihood is equivalent to that for matched logisitic regression. (The clogit function uses the coxph code to do the fit.) It is technically appropriate when the time scale is discrete and has only a few unique values, and some packages refer to this as the "discrete" option. There is also an "exact marginal likelihood" due to Prentice which is not implemented here" (https://cran.r-project.org/web/packages/survival/survival.pdf).

A.1.4.1 Usage

```
coxph(formula, data=, weights, subset,
      na.action, init, control,
      ties=c("efron","breslow","exact"),
      singular.ok=TRUE, robust=FALSE,
```

```
      model=FALSE, x=FALSE, y=TRUE, tt, method, ...)
```

A.1.4.2 Arguments

`formula` a formula object, with the response on the left of a ∼ operator, and the terms on the right. The response must be a survival object as returned by the `Surv` function.

`data` a data.frame in which to interpret the variables named in the `formula`, or in the subset and the weights argument.

`weights` vector of case weights. For a thorough discussion of these see the book by Therneau and Grambsch.

`subset` expression indicating which subset of the rows of data should be used in the fit. All observations are included by default.

`na.action` a missing-data filter function. This is applied to the model.frame after any subset argument has been used. Default is `options()$na.action`.

`init` vector of initial values of the iteration. Default initial value is zero for all variables.

`control` Object of class `coxph.control` specifying iteration limit and other control options. Default is `coxph.control(...)`.

`ties` a character string specifying the method for tie handling. If there are no tied death times all the methods are equivalent. Nearly all Cox regression programs use the Breslow method by default, but not this one. The Efron approximation is used as the default here, it is more accurate when dealing with tied death times, and is as efficient computationally. The "exact partial likelihood" is equivalent to a conditional logistic model, and is appropriate when the times are a small set of discrete values.

`singular.ok` logical value indicating how to handle collinearity in the model matrix. If `TRUE`,the program will automatically skip over columns of the X matrix that are linear combinations of earlier columns. In this case the coefficients for such columns will be `NA`, and the variance matrix will contain zeros. For ancillary calculations, such as the linear predictor, the missing coefficients are treated as zeros.

`robust` this argument has been deprecated, use a cluster term in the model instead. (The two options accomplish the same goal - creation of a robust variance - but the second is more flexible).

model	logical value: if TRUE, the model frame is returned in component model.
x	logical value: if TRUE, the x matrix is returned in component x
y	logical value: if TRUE, the response vector is returned in component y.
tt	optional list of time-transform functions.
method	alternate name for the ties argument.
...	Other arguments will be passed to coxph.control

A.1.5 Package 'xlsx'

As mentioned above, when performing the real data analysis it was needed to read the data that were provided in Stata format into a data frame in R. Eventually, we used the function write.xlsx, of the package 'xlsx' (cran.r-project.org), to write a a data.frame to an Excel workbook.

A.1.5.1 Usage

```
write.xlsx(x, file, sheetName="Sheet1",
        col.names=TRUE, row.names=TRUE, append=FALSE, showNA=TRUE)
```

A.1.5.2 Arguments

x	a a data.frame to write to the workbook.
file	the path to the output file.
sheetName	a character string with the sheet name.
col.names	a logical value indicating if the column names of x are to be written along with x to the file.
row.names	a logical value indicating whether the row names of x are to be written along with x to the file.
append	a logical value indicating if x should be appended to an existing file. If TRUE the file is read from disk.
showNA	a logical value. If set to FALSE, NA values will be left as empty cells.

Appendix B

Tables

B.1 Results of The Simulations

In this Section we include the Tables that contain the results of the simulation scenarios 3 and 4 as described in Section 4.

			$\epsilon \sim$ N(0 , $\pi^2/6$) & $x \sim$ N				
True β_1	Method	Prop of Ties ($\simeq$)	Avg Coeff ($*10^{-3}$)	Avg Var ($*10^{-3}$)	Emp Var ($*10^{-3}$)	Cov	Type I Error
0	NoTies	-	0.4619	0.4712	0.4723	0.949	0.051
	Efron	5%	0.8863	0.4722	0.4542	0.948	0.052
		20%	-0.4602	0.4698	0.4983	0.947	0.053
		60%	0.6712	0.4680	0.4551	0.953	0.047
	Breslow	5%	0.8840	0.4721	0.4503	0.948	0.052
		20%	-0.4587	0.4695	0.4809	0.947	0.053
		60%	0.6559	0.4680	0.4551	0.969	0.031
	Discrete	5%	0.8940	0.4770	0.4599	0.948	0.052
		20%	-0.4806	0.4884	0.5202	0.945	0.055
		60%	0.7806	0.5513	0.5497	0.950	0.050
	Adding	5%	0.9163	0.4725	0.4548	0.949	0.051
		20%	-0.4521	0.4709	0.5022	0.844	0.056
		60%	0.4864	0.4711	0.4605	0.951	0.049
True β_1	Method	Prop of Ties ($\simeq$)	Avg Coeff	Avg Var ($*10^{-3}$)	Emp Var ($*10^{-3}$)	Cov	Pow
0.15	NoTies	-	0.1090	0.4811	0.4990	0.529	0.998
	Efron	5%	0.1091	0.4821	0.4854	0.548	0.998
		20%	0.1079	0.4795	0.5299	0.502	1
		60%	0.1072	0.4776	0.4782	0.500	0.998
	Breslow	5%	0.1087	0.4819	0.4809	0.545	0.998
		20%	0.1065	0.4791	0.5119	0.481	1
		60%	0.1011	0.4763	0.4119	0.380	0.997
	Discrete	5%	0.1099	0.4874	0.4910	0.568	0.998
		20%	0.1110	0.5000	0.5548	0.557	1
		60%	0.1203	0.5704	0.5832	0.759	0.999
	Adding	5%	0.1091	0.4824	0.4871	0.552	0.998
		20%	0.1082	0.4806	0.5353	0.503	1
		60%	0.1080	0.4810	0.4962	0.515	0.999

TABLE B.1: Scenario III of simulations, where 1000 samples of 1000 individuals were simulated, the exposures of interest and the error terms were normally distributed. The true parameter, β_1, was consider to take values 0 and 0.15 and three different degrees of ties (proportion of ties) were considered and handled by three different approximations (methods). The results are the average coefficient (Avg Coeff), the average variance (Avg Var), the empirical variance (Emp Var), the coverage (Cov), and the power (Pow).

			$\epsilon \sim$ N(0 , $\pi^2/6$) & $x \sim$ Bin				
True β_1	Method	Prop of Ties ($\simeq$)	Avg Coeff ($*10^{-2}$)	Avg Var ($*10^{-2}$)	Emp Var ($*10^{-2}$)	Cov	Type I Error
0	NoTies	-	0.0597	0.4623	0.4317	0.962	0.038
	Efron	5%	-0.1577	0.4617	0.4861	0.950	0.050
		20%	-0.1728	0.4613	0.4568	0.952	0.048
		60%	0.0601	0.4590	0.4169	0.959	0.041
	Breslow	5%	-0.1594	0.4616	0.4818	0.950	0.050
		20%	-0.1714	0.4610	0.4422	0.956	0.044
		60%	0.0650	0.4581	0.3640	0.969	0.031
	Discrete	5%	-0.1609	0.4666	0.4921	0.950	0.050
		20%	-0.1782	0.4799	0.4790	0.952	0.048
		60%	0.0799	0.5421	0.5092	0.957	0.043
	Adding	5%	-0.1691	0.4621	0.4877	0.949	0.051
		20%	-0.1691	0.4624	0.4598	0.954	0.046
		60%	0.0523	0.4622	0.4378	0.953	0.047
True β_1	Method	Prop of Ties ($\simeq$)	Avg Coeff	Avg Var ($*10^{-2}$)	Emp Var ($*10^{-2}$)	Cov	Pow
0.15	NoTies	-	0.1084	0.4633	0.4334	0.911	0.351
	Efron	5%	0.1060	0.4626	0.4818	0.893	0.353
		20%	0.1056	0.4623	0.4546	0.902	0.347
		60%	0.1066	0.4598	0.4147	0.916	0.339
	Breslow	5%	0.1056	0.4625	0.4775	0.894	0.351
		20%	0.1044	0.4620	0.4397	0.899	0.339
		60%	0.1010	0.4589	0.3617	0.914	0.297
	Discrete	5%	0.1067	0.4675	0.4878	0.895	0.352
		20%	0.1086	0.4809	0.4765	0.903	0.351
		60%	0.1196	0.5430	0.5070	0.938	0.371
	Adding	5%	0.1060	0.4630	0.4844	0.896	0.354
		20%	0.1060	0.4635	0.4569	0.898	0.350
		60%	0.1072	0.4631	0.4331	0.909	0.342

TABLE B.2: Scenario IV of simulations, where 1000 samples of 1000 individuals were simulated, the exposures of interest had a Binomial distribution and the error terms normally distributed. The true parameter, β_1, was consider to take values 0 and 0.15 and three different degrees of ties (proportion of ties) were considered and handled by three different approximations (methods). The results are the average coefficient (Avg Coeff), the average variance (Avg Var), the empirical variance (Emp Var), the coverage (Cov), and the power (Pow).

B.2 Results of The Real Data Analysis

B.2.1 Alcohol abuse score

Method of handling ties : Efron				
SNP	Odds ratio	Confidence Interval		p-value
rs1176744	1.2141502	(1.8212708 ,	1.7949753)	0.33063486
rs12529	0.9368151	(0.6915426 ,	1.2690793)	0.67344528
rs1799836	0.9402567	(0.7032193 ,	1.2571933)	0.67766365
rs1799971	0.7600798	(0.3590440 ,	1.6090544)	0.47341049
rs4680	0.9299015	(0.6602652 ,	1.3096507)	0.67742647
rs6265	1.0206384	(0.5939154 ,	1.7539579)	0.94105098
rs36020	1.2258399	(0.7277732 ,	2.0647688)	0.44399469
rs36029	0.8863914	(0.6453791 ,	1.2174080)	0.45633720
rs6277N	0.7050693	(0.4932411 ,	1.0078694)	0.05523789
rs6190N	0.7729107	(0.3333407 ,	1.7921335)	0.54828385
rs6196N	1.0860293	(0.6563105 ,	1.7971064)	0.74808430
rs242938	0.8366821	(0.4203389 ,	1.6654109)	0.61166608
rs244465	0.7632601	(0.4216074 ,	1.3817736)	0.37231814
rs521674	0.9037226	(0.6171267 ,	1.3234148)	0.60294799
rs602618	0.9394968	(0.6520233 ,	1.3537159)	0.73770414
rs53576N	1.0701231	(0.7786401 ,	1.4707223)	0.67612939
rs1800497	0.5184608	(0.2873493 ,	0.9354523)	0.02913901
rs1876831	0.8098933	(0.4793170 ,	1.3684621)	0.43077078
rs1978340	1.2456302	(0.8821206 ,	1.7589369)	0.21218757
rs3219151	1.3555047	(0.9589100 ,	1.9161265)	0.08499542
rs3782025	1.1774281	(0.8107893 ,	1.7098610)	0.39084885
rs6943555	1.1332093	(0.7256751 ,	1.7696120)	0.58237063
rs7590720	0.9180267	(0.6498530 ,	1.2968672)	0.62751529
rs9939609	1.4046884	(0.9508331 ,	2.0751796)	0.08786395
rs237887N	0.8766991	(0.6242393 ,	1.2312606)	0.44760584
rs237889N	0.9353512	(0.6711920 ,	1.3034747)	0.69305241
rs237880N	0.8387399	(0.5772539 ,	1.2186743)	0.35625070
rs237898N	0.9504303	(0.6483132 ,	1.3933356)	0.79448767
rs110402N	1.0871093	(0.7519812 ,	1.5715907)	0.65692477
rs242924N	1.0294282	(0.7165248 ,	1.4789751)	0.87533506
rs7632287N	1.0113093	(0.6874637 ,	1.4877097)	0.95446221

TABLE B.3: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the alcohol abuse score. The method for handling ties used is the Efron method (Part 1).

Method of handling ties : Efron				
SNP	Odds ratio	Confidence Interval		p-value
rs2268491N	0.9737489	(0.5702376 ,	1.6627924)	0.92237777
rs4561970N	0.6176241	(0.3072928 ,	1.2413553)	0.17606723
rs4686302N	0.7619235	(0.4304711 ,	1.3485864)	0.35061155
rs1042778N	0.9767273	(0.6865166 ,	1.3896187)	0.89585152
rs1488467N	0.4677649	(0.1944980 ,	1.1249677)	0.08969762
rs2268490N	0.7936006	(0.4651606 ,	1.3539452)	0.39633021
rs2740204N	1.3974518	(1.0148388 ,	1.9243170)	0.04034082
rs3800373N	0.8714082	(0.6025955 ,	1.2601359)	0.46454090
rs4813625N	1.3089668	(0.9300652 ,	1.8422299)	0.12254425
rs2770378N	1.4125413	(1.0018537 ,	1.9915813)	0.04877354
rs6770632N	0.9069187	(0.5994081 ,	1.3721896)	0.64377314
rs2268493N	1.1004236	(0.7742154 ,	1.5640764)	0.59371932
rs2268498N	1.0476830	(0.7312802 ,	1.5009838)	0.79954755
rs2290045N	1.8200281	(1.1229623 ,	2.9497895)	0.01506892
rs1900586N	1.5447004	(0.9465299 ,	2.5208916)	0.08184290
rs9296158N	0.9042189	(0.6300367 ,	1.2977208)	0.58492487
rs7748266N	0.8107696	(0.4748633 ,	1.3842877)	0.44214877
rs1360780N	0.9042189	(0.6300367 ,	1.2977208)	0.58492487
rs9394309N	0.8956407	(0.6258550 ,	1.2817220)	0.54670216
rs9470080N	0.8976686	(0.6426951 ,	1.2537964)	0.52656559
rs4792887N	0.8414087	(0.4913128 ,	1.4409733)	0.52928998
rs7209436N	1.0200251	(0.7215621 ,	1.4419429)	0.91061481
rs1344694N	0.9465283	(0.6664734 ,	1.3442636)	0.75881134
rs13316193N	1.1282211	(0.7894423 ,	1.6123825)	0.50783153
rs13125511N	1.4260934	(0.9319593 ,	2.1822223)	0.10197865
rs11131149N	1.1604660	(0.8112909 ,	1.6599241)	0.41513481
rs13273672N	1.0379260	(0.6921503 ,	1.5564400)	0.85709782
rs41423247N	1.1063341	(0.7454236 ,	1.6419859)	0.61594403
rs35369693N	1.1087636	(0.5851638 ,	2.1008761)	0.75152421
rs16859448N	1.2460725	(0.7359269 ,	2.1098519)	0.41290484

TABLE B.4: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the alcohol abuse score. The method for handling ties used is the Efron method (Part 2).

Method of handling ties : Breslow				
SNP	Odds ratio	Confidence Interval		p-value
rs1176744	1.1941116	(0.8095916 ,	1.7612614)	0.37094256
rs12529	0.9289206	(0.6846102 ,	1.2604157)	0.63582124
rs1799836	0.9401329	(0.7034175 ,	1.2565083)	0.67657972
rs1799971	0.7704845	(0.3639291 ,	1.6312143)	0.49565985
rs4680	0.9169334	(0.6517823 ,	1.2899504)	0.61849929
rs6265	0.9890704	(0.5752117 ,	1.7006959)	0.96830063
rs36020	1.1987197	(0.7110352 ,	2.0208970)	0.49638015
rs36029	0.8917723	(0.6486648 ,	1.2259921)	0.48059677
rs6277N	0.7061836	(0.4941687 ,	1.0091598)	0.05614102
rs6190N	0.7710488	(0.3324995 ,	1.7880218)	0.54459979
rs6196N	1.1030650	(0.6672176 ,	1.8236215)	0.70213966
rs242938	0.8537325	(0.4286226 ,	1.7004686)	0.65283535
rs244465	0.7495968	(0.4132734 ,	1.3596214)	0.34274743
rs521674	0.9196637	(0.6281184 ,	1.3465316)	0.66682461
rs602618	0.9545216	(0.6627564 ,	1.3747306)	0.80252892
rs53576N	1.0702207	(0.7789399 ,	1.4704243)	0.67543563
rs1800497	0.5397309	(0.3006418 ,	0.9689587)	0.03886413
rs1876831	0.8065057	(0.4774363 ,	1.3623837)	0.42143486
rs1978340	1.2467503	(0.8837428 ,	1.7588673)	0.20908179
rs3219151	1.3478812	(0.9546117 ,	1.9031651)	0.08986795
rs3782025	1.1776951	(0.8116668 ,	1.7087870)	0.38910379
rs6943555	1.1405945	(0.7321205 ,	1.7769696)	0.56086764
rs7590720	0.9160160	(0.6489378 ,	1.2930135)	0.61792189
rs9939609	1.3719304	(0.9281712 ,	2.0278511)	0.11271208
rs237887N	0.8894203	(0.6337015 ,	1.2483297)	0.49805803
rs237889N	0.9467569	(0.6797156 ,	1.3187113)	0.74622575
rs237880N	0.8394446	(0.5777290 ,	1.2197195)	0.35857482
rs237898N	0.9403646	(0.6411589 ,	1.3791989)	0.75301198
rs110402N	1.1061211	(0.7656156 ,	1.5980656)	0.59107299
rs242924N	1.0482159	(0.7297417 ,	1.5056787)	0.79883784
rs7632287N	1.0036944	(0.6827628 ,	1.4754794)	0.98503346

TABLE B.5: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the alcohol abuse score. The method for handling ties used is the Breslow method (Part 1).

Method of handling ties : Breslow				
SNP	Odds ratio	Confidence Interval		p-value
rs2268491N	0.9819869	(0.5757160 ,	1.6749548)	0.94680247
rs4561970N	0.6365854	(0.3170868 ,	1.2780130)	0.20403860
rs4686302N	0.7735950	(0.4374025 ,	1.3681887)	0.37755514
rs1042778N	0.9590495	(0.6740483 ,	1.3645549)	0.81622952
rs1488467N	0.4935071	(0.2055940 ,	1.1846126)	0.11392759
rs2268490N	0.8043608	(0.4725133 ,	1.3692657)	0.42249104
rs2740204N	1.3728806	(0.9970142 ,	1.8904457)	0.05217597
rs3800373N	0.8716654	(0.6033208 ,	1.2593640)	0.46439875
rs4813625N	1.3012864	(0.9243750 ,	1.8318825)	0.13121779
rs2770378N	1.4243156	(1.0092177 ,	2.0101460)	0.04419878
rs6770632N	0.9020333	(0.5961331 ,	1.3649035)	0.62561835
rs2268493N	1.0867332	(0.7650789 ,	1.5436173)	0.64227345
rs2268498N	1.0537996	(0.7357502 ,	1.5093350)	0.77496732
rs2290045N	1.7835376	(1.1002613 ,	2.8911372)	0.01889030
rs1900586N	1.5255605	(0.9368514 ,	2.4842092)	0.08954756
rs9296158N	0.9067888	(0.6323717 ,	1.3002889)	0.59467291
rs7748266N	0.8170764	(0.4785589 ,	1.3950506)	0.45918778
rs1360780N	0.9067888	(0.6323717 ,	1.3002889)	0.59467291
rs9394309N	0.8999244	(0.6302328 ,	1.2850234)	0.56179560
rs9470080N	0.9016184	(0.6464949 ,	1.2574202)	0.54169574
rs4792887N	0.8667410	(0.5068022 ,	1.4823140)	0.60141849
rs7209436N	1.0369933	(0.7334831 ,	1.4660940)	0.83709533
rs1344694N	0.9456354	(0.6665656 ,	1.3415429)	0.75406662
rs13316193N	1.1105595	(0.7777711 ,	1.5857395)	0.56391406
rs13125511N	1.4132759	(0.9255098 ,	2.1581065)	0.10924746
rs11131149N	1.1395695	(0.7970870 ,	1.6292054)	0.47373811
rs13273672N	1.0328821	(0.6900344 ,	1.5460758)	0.87508176
rs41423247N	1.1060696	(0.7478732 ,	1.6358253)	0.61361415
rs35369693N	1.0788072	(0.5700703 ,	2.0415466)	0.81569041
rs16859448N	1.2415818	(0.7338010 ,	2.1007402)	0.41998128

TABLE B.6: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the alcohol abuse score. The method for handling ties used is the Breslow method (Part 2).

Method of handling ties : Discrete				
SNP	Odds ratio	Confidence Interval		p-value
rs1176744	1.2104527	(0.8089766 ,	1.8111720)	0.35291356
rs12529	0.9224326	(0.6703609 ,	1.2692893)	0.62004915
rs1799836	0.9353737	(0.6919338 ,	1.2644620)	0.66401383
rs1799971	0.7611088	(0.3538510 ,	1.6370919)	0.48481780
rs4680	0.9111932	(0.6399439 ,	1.2974154)	0.60597410
rs6265	0.9879516	(0.5591261 ,	1.7456679)	0.96670971
rs36020	1.2166667	(0.7058438 ,	2.0971746)	0.48020530
rs36029	0.8829311	(0.6336089 ,	1.2303607)	0.46206153
rs6277N	0.6825601	(0.4696162 ,	0.9920618)	0.04530897
rs6190N	0.7607524	(0.3215912 ,	1.7996269)	0.53363786
rs6196N	1.1123705	(0.6582073 ,	1.8799067)	0.69079183
rs242938	0.8458073	(0.4164582 ,	1.7177958)	0.64317146
rs244465	0.7321481	(0.3960461 ,	1.3534809)	0.31997843
rs521674	0.9123450	(0.6119826 ,	1.3601260)	0.65250621
rs602618	0.9504517	(0.6491444 ,	1.3916140)	0.79391366
rs53576N	1.0757701	(0.7735393 ,	1.4960861)	0.66426178
rs1800497	0.5065882	(0.2730594 ,	0.9398380)	0.03102259
rs1876831	0.7924173	(0.4598750 ,	1.3654259)	0.40198563
rs1978340	1.2700347	(0.8866254 ,	1.8192442)	0.19232926
rs3219151	1.3716193	(0.9613630 ,	1.9569501)	0.08138750
rs3782025	1.1906447	(0.8106985 ,	1.7486583)	0.37355668
rs6943555	1.1599764	(0.7234933 ,	1.8597897)	0.53779374
rs7590720	0.9098522	(0.6363444 ,	1.3009166)	0.60453531
rs9939609	1.4149880	(0.9395425 ,	2.1310275)	0.09661267
rs237887N	0.8808215	(0.6190692 ,	1.2532467)	0.48060743
rs237889N	0.9425009	(0.6677005 ,	1.3303987)	0.73632418
rs237880N	0.8278349	(0.5616917 ,	1.2200832)	0.33968398
rs237898N	0.9342749	(0.6250419 ,	1.3964976)	0.74026176
rs110402N	1.1148027	(0.7607101 ,	1.6337171)	0.57728914
rs242924N	1.0522476	(0.7220556 ,	1.5334348)	0.79095688
rs7632287N	1.0039496	(0.6740686 ,	1.4952704)	0.98452652

TABLE B.7: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the alcohol abuse score. The method for handling ties used is the Discrete method (Part 1).

Method of handling ties : Discrete				
SNP	Odds ratio	Confidence Interval		p-value
rs2268491N	0.9805452	(0.5628962 ,	1.7080749)	0.94468673
rs4561970N	0.6091087	(0.2947801 ,	1.2586110)	0.18062256
rs4686302N	0.7568513	(0.4184388 ,	1.3689551)	0.35686067
rs1042778N	0.9555036	(0.6613434 ,	1.3805037)	0.80843201
rs1488467N	0.4615582	(0.1864323 ,	1.1426991)	0.09460416
rs2268490N	0.7891497	(0.4532802 ,	1.3738901)	0.40253609
rs2740204N	1.4021167	(1.0070647 ,	1.9521400)	0.04531774
rs3800373N	0.8631062	(0.5899985 ,	1.2626343)	0.44815254
rs4813625N	1.3338158	(0.9341604 ,	1.9044530)	0.11292382
rs2770378N	1.4694843	(1.0243549 ,	1.1080429)	0.03655477
rs6770632N	0.8952731	(0.5828686 ,	1.3751194)	0.61339755
rs2268493N	1.0974579	(0.7567306 ,	1.5916020)	0.62390859
rs2268498N	1.0581701	(0.7285847 ,	1.5368481)	0.76650258
rs2290045N	1.8772316	(1.1338843 ,	3.1078996)	0.01434558
rs1900586N	1.5894711	(0.9518647 ,	2.6541781)	0.07649091
rs9296158N	0.9008565	(0.6209293 ,	1.3069803)	0.58237182
rs7748266N	0.8029224	(0.4600544 ,	1.4013221)	0.43981938
rs1360780N	0.9008565	(0.6209293 ,	1.3069803)	0.58237182
rs9394309N	0.8940063	(0.6193056 ,	1.2905539)	0.54971442
rs9470080N	0.8962585	(0.6366418 ,	1.2617445)	0.53022982
rs4792887N	0.8590742	(0.4949026 ,	1.4912195)	0.58930064
rs7209436N	1.0396503	(0.7265487 ,	1.4876811)	0.83157019
rs1344694N	0.9417916	(0.6555199 ,	1.3530807)	0.74564427
rs13316193N	1.1225747	(0.7722777 ,	1.6317628)	0.54458747
rs13125511N	1.4520177	(0.9339822 ,	2.2573828)	0.09759486
rs11131149N	1.1548088	(0.7935390 ,	1.6805516)	0.45209711
rs13273672N	1.0360316	(0.6792283 ,	1.5802663)	0.86947138
rs41423247N	1.1137648	(0.7431767 ,	1.6691480)	0.60167394
rs35369693N	1.0843840	(0.5605658 ,	2.0976818)	0.80982896
rs16859448N	1.2545103	(0.7313834 ,	2.1518073)	0.41012807

TABLE B.8: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the alcohol abuse score. The method for handling ties used is the Discrete method (Part 2).

Method of handling ties : Adding				
SNP	Odds ratio	Confidence Interval		p-value
rs1176744	1.2090505	(0.8187592 ,	1.7853882)	0.33981425
rs12529	0.9450363	(0.6962579 ,	1.2827052)	0.71683746
rs1799836	0.9454144	(0.7065450 ,	1.2650409)	0.70560580
rs1799971	0.7637754	(0.3608266 ,	1.6167127)	0.48120800
rs4680	0.9360937	(0.6626716 ,	1.3223314)	0.70787737
rs6265	1.0090718	(0.5881638 ,	1.7311947)	0.97384022
rs36020	1.2708890	(0.7492862 ,	2.1555966)	0.37385994
rs36029	0.8774748	(0.6363872 ,	1.2098955)	0.42516777
rs6277N	0.7197362	(0.5037936 ,	1.0282391)	0.07076385
rs6190N	0.7436553	(0.3204149 ,	1.7259596)	0.49052643
rs6196N	1.1128135	(0.6699579 ,	1.8484055)	0.67969630
rs242938	0.8497817	(0.4275311 ,	1.6890676)	0.64234062
rs244465	0.7660627	(0.4203925 ,	1.3959620)	0.38406582
rs521674	0.9143070	(0.6261578 ,	1.3350584)	0.64276024
rs602618	0.9636943	(0.6707277 ,	1.3846255)	0.84147867
rs53576N	1.0873370	(0.7900677 ,	1.4964564)	0.60734250
rs1800497	0.5285183	(0.2943200 ,	0.9490744)	0.03276118
rs1876831	0.8132498	(0.4796044 ,	1.3790017)	0.44293477
rs1978340	1.2645195	(0.8943764 ,	1.7878486)	0.18409924
rs3219151	1.3515830	(0.9555588 ,	1.9117364)	0.08856190
rs3782025	1.1693163	(0.8047655 ,	1.6990051)	0.41189526
rs6943555	1.1610879	(0.7430032 ,	1.8144271)	0.51197689
rs7590720	0.9206001	(0.6513012 ,	1.3012483)	0.63937789
rs9939609	1.4102806	(0.9547186 ,	2.0832225)	0.08413272
rs237887N	0.8672419	(0.6180872 ,	1.2168323)	0.40977514
rs237889N	0.9360024	(0.6696350 ,	1.3083254)	0.69869361
rs237880N	0.8510845	(0.5857916 ,	1.2365230)	0.39752735
rs237898N	0.9719977	(0.6592504 ,	1.4331117)	0.88598935
rs110402N	1.0839979	(0.7513429 ,	1.5639349)	0.66626409
rs242924N	1.0299488	(0.7157130 ,	1.4821506)	0.87374645
rs7632287N	1.0119083	(0.6866621 ,	1.4912115)	0.95228442

TABLE B.9: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the alcohol abuse score. The method for handling ties used is the Adding method (Part 1).

Method of handling ties : Adding				
SNP	Odds ratio	Confidence Interval		p-value
rs2268491N	0.9525428	(0.5578903 ,	1.6263731)	0.85861948
rs4561970N	0.6124919	(0.3047743 ,	1.2308988)	0.16863024
rs4686302N	0.7838656	(0.4421775 ,	1.3895898)	0.40447026
rs1042778N	0.9642269	(0.6773039 ,	1.3726976)	0.83980060
rs1488467N	0.4581880	(0.1905603 ,	1.1016790)	0.08121837
rs2268490N	0.7959805	(0.4681153 ,	1.3534806)	0.39952463
rs2740204N	1.4077230	(1.0214635 ,	1.9400438)	0.03663809
rs3800373N	0.8660072	(0.5977972 ,	1.2545533)	0.44680005
rs4813625N	1.3030277	(0.9243448 ,	1.8368483)	0.13080666
rs2770378N	1.3929863	(0.9870498 ,	1.9658692)	0.05931700
rs6770632N	0.8969442	(0.5940020 ,	1.3543874)	0.60496732
rs2268493N	1.0608278	(0.7466496 ,	1.5072071)	0.74174829
rs2268498N	1.0546106	(0.7378752 ,	1.5073057)	0.77044112
rs2290045N	1.7676635	(1.0879853 ,	2.8719452)	0.02141676
rs1900586N	1.5642257	(0.9580084 ,	2.5540508)	0.07369434
rs9296158N	0.9096142	(0.6323145 ,	1.3085228)	0.60961424
rs7748266N	0.8235811	(0.4818432 ,	1.4076900)	0.47789854
rs1360780N	0.8996068	(0.6254216 ,	1.2939951)	0.56839818
rs9394309N	0.8959209	(0.6255450 ,	1.2831597)	0.54874185
rs9470080N	0.8965873	(0.6414187 ,	1.2532669)	0.52293323
rs4792887N	0.8435825	(0.4925855 ,	1.4446862)	0.53545649
rs7209436N	1.0202901	(0.7231249 ,	1.4395743)	0.90895023
rs1344694N	0.9641147	(0.6788539 ,	1.3692449)	0.83821084
rs13316193N	1.1389474	(0.7958353 ,	1.6299870)	0.47685197
rs13125511N	1.4060868	(0.9190931 ,	2.1511206)	0.11616441
rs11131149N	1.1686312	(0.8152902 ,	1.6751077)	0.39625955
rs13273672N	1.0562369	(0.7010621 ,	1.5913517)	0.79360330
rs41423247N	1.1516251	(0.7728585 ,	1.7160198)	0.48782245
rs35369693N	1.0941311	(0.5764331 ,	2.0767767)	0.78321188
rs16859448N	1.2282028	(0.7254675 ,	2.0793242)	0.44414005

TABLE B.10: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the alcohol abuse score. The method for handling ties used is the Adding method (Part 2).

Kruskal-Wallis Test			
SNP	p-value	SNP	p-value
rs1176744	0.20733530	rs2268491N	0.66236956
rs12529	0.47307913	rs4561970N	0.28942730
rs1799836	0.83219216	rs4686302N	0.33156496
rs1799971	0.05038266	rs1042778N	0.37913136
rs4680	0.32637823	rs1488467N	0.22539886
rs6265	0.22799036	rs2268490N	0.19213746
rs36020	0.30220461	rs2740204N	0.04431439
rs36029	0.30476617	rs3800373N	0.71743941
rs6277N	0.67404768	rs4813625N	0.42052835
rs6190N	0.19902116	rs2770378N	0.19744653
rs6196N	0.50716306	rs6770632N	0.14968130
rs242938	0.20136852	rs2268493N	0.17318190
rs244465	0.36955172	rs2268498N	0.72799896
rs521674	0.89178915	rs2290045N	0.41488267
rs602618	0.93721337	rs1900586N	0.15175623
rs53576N	0.93073838	rs9296158N	0.64334596
rs1800497	0.66299333	rs7748266N	0.73366628
rs1876831	0.33526236	rs1360780N	0.75567537
rs1978340	0.51602796	rs9394309N	0.94566100
rs3219151	0.82775828	rs9470080N	0.91562832
rs3782025	0.32089642	rs4792887N	0.29295767
rs6943555	0.50712396	rs7209436N	0.33875606
rs7590720	0.51281278	rs1344694N	0.46346618
rs9939609	0.20813247	rs13316193N	0.18406731
rs237887N	0.08549422	rs13125511N	0.27186695
rs237889N	0.24396498	rs11131149N	0.19217724
rs237880N	0.46029852	rs13273672N	0.92244822
rs237898N	0.97372721	rs41423247N	0.76214710
rs110402N	0.33543725	rs35369693N	0.19349334
rs242924N	0.38039510	rs16859448N	0.24300662
rs7632287N	0.21406434		

TABLE B.11: The results of applying the Kruskla-Wallis test on the real data, with exposure the particular SNP and outcome the alcohol abuse score.

B.2.2 Drug abuse score

Method of handling ties : Efron				
SNP	0. Odds ratio	Confidence Interval		p-value
rs1176744	0.9213345	(0.6503429 ,	1.3052456)	0.64477813
rs12529	1.1800818	(0.8725602 ,	1.5959849)	0.28238401
rs1799836	0.8356492	(0.6270282 ,	1.1136812)	0.22048381
rs1799971	0.8533755	(0.4336053 ,	1.6795223)	0.64623816
rs4680	1.0361229	(0.7611779 ,	1.4103808)	0.82155551
rs6265	1.1756432	(0.7034103 ,	1.9649086)	0.53691595
rs36020	1.0714105	(0.6196608 ,	1.8524982)	0.80498474
rs36029	0.8909188	(0.6377724 ,	1.2445447)	0.49825046
rs6277N	0.9199127	(0.6543493 ,	1.2932535)	0.63100237
rs6190N	0.7854136	(0.3410458 ,	1.8087733)	0.57035582
rs6196N	0.7517105	(0.4556691 ,	1.2400855)	0.26379004
rs242938	1.0528781	(0.5215854 ,	0.1253513)	0.88567273
rs244465	0.9500475	(0.5093055 ,	1.7721979)	0.87201853
rs521674	0.8504745	(0.5745297 ,	1.2589547)	0.41834064
rs602618	0.8673846	(0.5985032 ,	1.2570626)	0.45233520
rs53576N	1.4631316	(1.0239376 ,	2.0907076)	0.03662662
rs1800497	0.8140243	(0.4826000 ,	1.3730534)	0.44045855
rs1876831	0.9594609	(0.5619489 ,	1.6381654)	0.87948470
rs1978340	1.0118787	(0.7193983 ,	1.4232706)	0.94590954
rs3219151	1.1366782	(0.8273628 ,	1.5616333)	0.42920730
rs3782025	0.8352636	(0.5757533 ,	1.2117434)	0.34299953
rs6943555	0.7778985	(0.4951571 ,	1.2220890)	0.27581425
rs7590720	1.0010599	(0.7120899 ,	1.4072955)	0.99513633
rs9939609	0.9893974	(0.6760855 ,	1.4479046)	0.95624466
rs237887N	1.0044063	(0.7363027 ,	1.3701321)	0.97785980
rs237889N	1.0160633	(0.7391341 ,	1.3967488)	0.92180971
rs237880N	0.6173289	(0.4189563 ,	0.9096294)	0.01473112
rs237898N	0.8967842	(0.6203132 ,	1.2964770)	0.56239060
rs110402N	1.2724222	(0.8664599 ,	1.8685900)	0.21912053
rs242924N	1.2303758	(0.8417962 ,	1.7983266)	0.28433244
rs7632287N	1.1853143	(0.8268382 ,	1.6992079)	0.35486054

TABLE B.12: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the drug abuse score. The method for handling ties used is the Efron method (Part 1).

Method of handling ties : Efron				
SNP	Odds ratio	Confidence Interval		p-value
rs2268491N	1.3052359	(0.7968283 ,	2.1380274)	0.29006529
rs4561970N	0.7829107	(0.4024824 ,	1.5229213)	0.47095186
rs4686302N	0.8421137	(0.4919966 ,	1.4413829)	0.53086662
rs1042778N	0.8611594	(0.6067167 ,	1.2223093)	0.40284917
rs1488467N	0.6518915	(0.2977353 ,	1.4273169)	0.28455699
rs2268490N	1.3281838	(0.7970773 ,	2.2131757)	0.27597124
rs2740204N	0.8703576	(0.6364701 ,	1.1901931)	0.38453231
rs3800373N	0.9754620	(0.6855738 ,	1.3879265)	0.89017764
rs4813625N	0.9849474	(0.7166636 ,	1.3536634)	0.92551585
rs2770378N	0.7724666	(0.5336040 ,	1.1182537)	0.17136655
rs6770632N	1.2212259	(0.8200136 ,	1.8187416)	0.32536344
rs2268493N	0.9701420	(0.6773262 ,	1.3895454)	0.86865846
rs2268498N	1.1894403	(0.8176895 ,	1.7302024)	0.36423302
rs2290045N	1.1632202	(0.7298782 ,	1.8538453)	0.52489357
rs1900586N	1.0782038	(0.6810522 ,	1.7069522)	0.74802954
rs9296158N	1.0509785	(0.7399327 ,	1.4927786)	0.78123266
rs7748266N	1.3430080	(0.7962270 ,	2.2652716)	0.26886778
rs1360780N	1.0673618	(0.7531355 ,	1.5126910)	0.71404807
rs9394309N	1.1236339	(0.7984115 ,	1.5813313)	0.50372608
rs9470080N	1.0249145	(0.7408343 ,	1.4179281)	0.88186841
rs4792887N	1.2056258	(0.6925926 ,	2.0986847)	0.50847670
rs7209436N	1.2084742	(0.8226785 ,	1.7751891)	0.33447500
rs1344694N	1.0127464	(0.7202722 ,	1.4239828)	0.94192938
rs13316193N	1.0088104	(0.7029463 ,	1.4477613)	0.96204073
rs13125511N	1.1861882	(0.8133507 ,	1.7299334)	0.37513379
rs11131149N	1.0267848	(0.7141129 ,	1.4763590)	0.88655650
rs13273672N	1.3819974	(0.9319740 ,	2.0493243)	0.10750299
rs41423247N	0.9283814	(0.6451764 ,	1.3359014)	0.68898418
rs35369693N	1.2343384	(0.6354203 ,	2.3977694)	0.53429985
rs16859448N	1.3187637	(0.8218840 ,	2.1160378)	0.25141495

TABLE B.13: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the drug abuse score. The method for handling ties used is the Efron method (Part 2).

Method of handling ties : Breslow				
SNP	Odds ratio	Confidence Interval		p-value
rs1176744	0.9491055	(0.6710604 ,	1.342355)	0.76773766
rs12529	1.1364760	(0.8357632 ,	1.545387)	0.41458216
rs1799836	0.8491963	(0.6348071 ,	1.135990)	0.27084562
rs1799971	0.8278707	(0.4259399 ,	1.609077)	0.57744379
rs4680	1.0472098	(0.7678860 ,	1.428140)	0.77072438
rs6265	1.2104433	(0.7238578 ,	2.024117)	0.46657203
rs36020	1.0092346	(0.5805126 ,	1.754578)	0.97401124
rs36029	0.9233847	(0.6610055 ,	1.289913)	0.64024377
rs6277N	0.9745534	(0.6966421 ,	1.363332)	0.88037753
rs6190N	0.8683238	(0.3746165 ,	2.012688)	0.74201553
rs6196N	0.8511988	(0.5177917 ,	1.399287)	0.52525301
rs242938	0.9388251	(0.4682607 ,	1.882269)	0.85882497
rs244465	0.9101647	(0.4938153 ,	1.677550)	0.76286137
rs521674	0.8860463	(0.6031991 ,	1.301524)	0.53743597
rs602618	0.8984932	(0.6249516 ,	1.291764)	0.56335553
rs53576N	1.3238125	(0.9302923 ,	1.883794)	0.11910441
rs1800497	0.8890237	(0.5312344 ,	1.487786)	0.65433114
rs1876831	0.9913745	(0.5863467 ,	1.676181)	0.97420847
rs1978340	1.0188308	(0.7285707 ,	1.424730)	0.91316782
rs3219151	1.0886950	(0.7937240 ,	1.493286)	0.59813053
rs3782025	0.8777854	(0.6101283 ,	1.262861)	0.48241914
rs6943555	0.8336230	(0.5310798 ,	1.308517)	0.42890267
rs7590720	0.9967982	(0.7107020 ,	1.398064)	0.98517619
rs9939609	1.0083134	(0.6912514 ,	1.470805)	0.96571624
rs237887N	1.0362242	(0.7578407 ,	1.416869)	0.82359842
rs237889N	1.0407877	(0.7564010 ,	1.432096)	0.80606355
rs237880N	0.6991074	(0.4764293 ,	1.025863)	0.06732568
rs237898N	0.9744211	(0.6766265 ,	1.403280)	0.88925467
rs110402N	1.1837868	(0.8201042 ,	1.708748)	0.36761296
rs242924N	1.1653106	(0.8087357 ,	1.679101)	0.41169542
rs7632287N	1.1422457	(0.7924600 ,	1.646424)	0.47585751

TABLE B.14: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the drug abuse score. The method for handling ties used is the Breslow method (Part 1).

Method of handling ties : Breslow				
SNP	Odds ratio	Confidence Interval		p-value
rs2268491N	1.2600048	(0.7658294 ,	2.073062)	0.36294211
rs4561970N	0.7913868	(0.4089312 ,	1.531537)	0.48732943
rs4686302N	0.8026311	(0.4695345 ,	1.372033)	0.42154958
rs1042778N	0.8423916	(0.5928015 ,	1.197068)	0.33873325
rs1488467N	0.6607798	(0.3018231 ,	1.446642)	0.30001927
rs2268490N	1.2741439	(0.7658972 ,	2.119661)	0.35084238
rs2740204N	0.8922209	(0.6514210 ,	1.222033)	0.47733867
rs3800373N	0.9897711	(0.6957872 ,	1.407969)	0.95440202
rs4813625N	0.9776902	(0.7100810 ,	1.346153)	0.89002254
rs2770378N	0.8559612	(0.5926646 ,	1.236230)	0.40694744
rs6770632N	1.1518061	(0.7735881 ,	1.714940)	0.48647829
rs2268493N	0.9133622	(0.6407041 ,	1.302053)	0.61640349
rs2268498N	1.1302639	(0.7808092 ,	1.636119)	0.51641713
rs2290045N	1.1857570	(0.7468295 ,	1.882651)	0.47007229
rs1900586N	1.1392261	(0.7144660 ,	1.816512)	0.58397960
rs9296158N	1.0482409	(0.7369869 ,	1.490948)	0.79323364
rs7748266N	1.2720144	(0.7584700 ,	2.133269)	0.36174186
rs1360780N	1.0647302	(0.7502941 ,	1.510942)	0.72541688
rs9394309N	1.1382114	(0.8076102 ,	1.604147)	0.45962055
rs9470080N	1.0348738	(0.7464694 ,	1.434705)	0.83705010
rs4792887N	1.0959970	(0.6358946 ,	1.889007)	0.74137951
rs7209436N	1.1274735	(0.7818572 ,	1.625868)	0.52061198
rs1344694N	1.0233080	(0.7268435 ,	1.440694)	0.89497451
rs13316193N	0.9578267	(0.6730654 ,	1.363065)	0.81082349
rs13125511N	1.1802432	(0.8076720 ,	1.724678)	0.39183086
rs11131149N	0.9675944	(0.6778754 ,	1.381137)	0.85601894
rs13273672N	1.2237107	(0.8312915 ,	1.801375)	0.30613153
rs41423247N	0.9813280	(0.6849051 ,	1.406041)	0.91818038
rs35369693N	1.3006786	(0.6794359 ,	2.489955)	0.42750900
rs16859448N	1.2437527	(0.7780570 ,	1.988184)	0.36206934

TABLE B.15: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the drug abuse score. The method for handling ties used is the Breslow method (Part 2).

Method of handling ties : Discrete				
SNP	Odds ratio	Confidence Interval		p-value
rs1176744	0.9247166	(0.6047949 ,	1.4138691)	0.71787876
rs12529	1.2317397	(0.8303669 ,	1.8271234)	0.30019324
rs1799836	0.7619458	(0.5246879 ,	1.1064891)	0.15318552
rs1799971	0.7523487	(0.3398408 ,	1.6655702)	0.48281026
rs4680	1.0708186	(0.7335903 ,	1.5630691)	0.72290873
rs6265	1.4199049	(0.7036510 ,	2.8652411)	0.32769484
rs36020	1.0130731	(0.5249737 ,	1.9549879)	0.96911005
rs36029	0.8817938	(0.5791223 ,	1.3426529)	0.55758499
rs6277N	0.9563837	(0.6152222 ,	1.4867309)	0.84294674
rs6190N	0.8263136	(0.3133576 ,	2.1789615)	0.69976119
rs6196N	0.7825600	(0.4245876 ,	1.4423409)	0.43190517
rs242938	0.9171373	(0.4069477 ,	2.0669504)	0.83472809
rs244465	0.8673462	(0.4128023 ,	1.8223967)	0.70714438
rs521674	0.8297003	(0.5147978 ,	1.3372290)	0.44329064
rs602618	0.8423376	(0.5329207 ,	1.3314040)	0.46261017
rs53576N	1.7623717	(1.0457907 ,	2.9699574)	0.03332460
rs1800497	0.7798616	(0.3713618 ,	1.6377132)	0.51128692
rs1876831	0.9859085	(0.5033307 ,	1.9311671)	0.96699849
rs1978340	1.0323010	(0.6663662 ,	1.5991888)	0.88680080
rs3219151	1.1491631	(0.7663902 ,	1.7231117)	0.50114252
rs3782025	0.8107796	(0.5102783 ,	1.2882451)	0.37460086
rs6943555	0.7408494	(0.4144800 ,	1.3242082)	0.31139354
rs7590720	0.9947327	(0.6444240 ,	1.5354691)	0.98097693
rs9939609	1.0143578	(0.6180486 ,	1.6647909)	0.95502615
rs237887N	1.0595435	(0.7108179 ,	1.5793531)	0.77641723
rs237889N	1.0758648	(0.6978011 ,	1.6587609)	0.74061043
rs237880N	0.5631450	(0.3478131 ,	0.9117891)	0.01951121
rs237898N	0.9526949	(0.5786420 ,	1.5685477)	0.84892145
rs110402N	1.3732189	(0.8237081 ,	2.2893186)	0.22388413
rs242924N	1.3392373	(0.8045854 ,	2.2291686)	0.26117404
rs7632287N	1.2514291	(0.7729361 ,	2.0261375)	0.36159585

TABLE B.16: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the drug abuse score. The method for handling ties used is the Discrete method (Part 1).

Method of handling ties : Discrete				
SNP	Odds ratio	Confidence Interval		p-value
rs2268491N	1.7861425	(0.7750981 ,	4.1160016)	0.17324174
rs4561970N	0.6851302	(0.2999794 ,	1.5647856)	0.36949996
rs4686302N	0.7216281	(0.3773003 ,	1.3801926)	0.32409447
rs1042778N	0.7408059	(0.4652139 ,	1.1796582)	0.20625505
rs1488467N	0.5481519	(0.2186907 ,	1.3739519)	0.19971459
rs2268490N	1.6693484	(0.7824856 ,	3.5613741)	0.18499630
rs2740204N	0.8196788	(0.5411647 ,	1.2415320)	0.34789213
rs3800373N	0.9844896	(0.6377584 ,	1.5197286)	0.94374042
rs4813625N	0.9622705	(0.6337395 ,	1.4611121)	0.85677178
rs2770378N	0.7515768	(0.4575184 ,	1.2346336)	0.25944689
rs6770632N	1.2693064	(0.7539649 ,	2.1368882)	0.36954167
rs2268493N	0.8546659	(0.5371382 ,	1.3598992)	0.50750470
rs2268498N	1.2359988	(0.7583730 ,	2.0144348)	0.39521905
rs2290045N	1.3359381	(0.7320898 ,	2.437857)	0.34527202
rs1900586N	1.2844019	(0.6694827 ,	2.4641238)	0.45148445
rs9296158N	1.0733887	(0.6958866 ,	1.6556768)	0.74875147
rs7748266N	1.4444520	(0.7563275 ,	2.7586482)	0.26529195
rs1360780N	1.1012450	(0.7120750 ,	1.7031080)	0.66462941
rs9394309N	1.2794611	(0.7916901 ,	2.0677544)	0.31429927
rs9470080N	1.0597195	(0.6921708 ,	1.6224399)	0.78953076
rs4792887N	1.1443573	(0.5877653 ,	2.2280212)	0.69160751
rs7209436N	1.2508667	(0.7538707 ,	2.0755118)	0.38627088
rs1344694N	1.0405595	(0.6637834 ,	1.6312010)	0.86238357
rs13316193N	0.9247824	(0.5754910 ,	1.4860743)	0.74660633
rs13125511N	1.4013911	(0.8025145 ,	2.4471792)	0.23542975
rs11131149N	0.9422729	(0.5845910 ,	1.5188024)	0.80713193
rs13273672N	1.4749434	(0.8516974 ,	2.5542617)	0.16542386
rs41423247N	0.9680345	(0.6038750 ,	1.5517958)	0.89266363
rs35369693N	1.7693332	(0.6568958 ,	4.7656566)	0.25901221
rs16859448N	1.4817873	(0.7752185 ,	2.8323544)	0.23415842

TABLE B.17: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the drug abuse score. The method for handling ties used is the Discrete method (Part 2).

Method of handling ties : Adding				
SNP	Odds ratio	Confidence Interval		p-value
rs1176744	0.9033193	(0.6315987 ,	1.2919373)	0.57755746
rs12529	1.2145548	(0.9004920 ,	1.6381527)	0.20288861
rs1799836	0.8712459	(0.6515938 ,	1.1649428)	0.35240489
rs1799971	0.8449992	(0.4225162 ,	1.6899323)	0.63388638
rs4680	1.0237722	(0.7516851 ,	1.3943466)	0.88150892
rs6265	1.0829696	(0.6418621 ,	1.8272198)	0.76519970
rs36020	1.0946353	(0.6293115 ,	1.9040273)	0.74884573
rs36029	0.8961872	(0.6405139 ,	1.2539174)	0.52243316
rs6277N	0.9105341	(0.6374119 ,	1.3006855)	0.60647087
rs6190N	0.7690425	(0.3340852 ,	1.7702863)	0.53700557
rs6196N	0.7894034	(0.4796843 ,	1.2990996)	0.35214486
rs242938	0.9828265	(0.4878013 ,	1.9802079)	0.96134391
rs244465	0.9888222	(0.5333312 ,	1.8333246)	0.97153226
rs521674	0.8669415	(0.5824834 ,	1.2903159)	0.48159572
rs602618	0.8366785	(0.5727290 ,	1.2222725)	0.35648033
rs53576N	1.2524658	(0.8769695 ,	1.7887402)	0.21571176
rs1800497	0.6008266	(0.3388228 ,	1.0654319)	0.08130965
rs1876831	0.8798093	(0.5063087 ,	1.5288388)	0.64967693
rs1978340	1.0767672	(0.7666928 ,	1.5122453)	0.66949758
rs3219151	1.1360304	(0.8174912 ,	1.5786898)	0.44744238
rs3782025	0.8319900	(0.5758741 ,	1.2020118)	0.32716738
rs6943555	0.8112532	(0.5166483 ,	1.2738488)	0.36355372
rs7590720	0.9261669	(0.6507807 ,	1.3180861)	0.67009529
rs9939609	0.9524435	(0.6496276 ,	1.3964133)	0.80290643
rs237887N	0.9344669	(0.6772500 ,	1.2893739)	0.67986289
rs237889N	0.9650472	(0.6939462 ,	1.3420579)	0.83253391
rs237880N	0.6085972	(0.4110998 ,	0.9009748)	0.01310276
rs237898N	0.9940270	(0.6823488 ,	1.4480714)	0.97510161
rs110402N	1.3372699	(0.9138621 ,	1.9568496)	0.13458571
rs242924N	1.2512179	(0.8474419 ,	1.8473787)	0.25959701
rs7632287N	1.1407008	(0.7930899 ,	1.6406694)	0.47776862

TABLE B.18: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the drug abuse score. The method for handling ties used is the Adding method (Part 1).

Method of handling ties : Adding				
SNP	Odds ratio	Confidence Interval		p-value
rs2268491N	1.3063093	(0.7797956 ,	2.1883220)	0.31005485
rs4561970N	0.7202440	(0.3710445 ,	1.3980843)	0.33217148
rs4686302N	0.8345359	(0.4867087 ,	1.4309382)	0.51086728
rs1042778N	0.9399128	(0.6575101 ,	1.3436083)	0.73392720
rs1488467N	0.6234970	(0.2844295 ,	1.3667658)	0.23810509
rs2268490N	1.3463214	(0.7831080 ,	2.3145995)	0.28208011
rs2740204N	0.8088235	(0.5807758 ,	1.1264163)	0.20927458
rs3800373N	0.9559933	(0.6677190 ,	1.3687243)	0.80584783
rs4813625N	1.0202776	(0.7296461 ,	1.4266730)	0.90657661
rs2770378N	0.7604057	(0.5240685 ,	1.1033229)	0.14923032
rs6770632N	1.1176498	(0.7508172 ,	1.6637086)	0.58369020
rs2268493N	0.9965116	(0.6876874 ,	1.4440214)	0.98526778
rs2268498N	1.1775285	(0.8070065 ,	1.7181687)	0.39660114
rs2290045N	1.0262934	(0.6324738 ,	1.6653311)	0.91630711
rs1900586N	1.0864921	(0.6812957 ,	1.7326766)	0.72755988
rs9296158N	1.0453622	(0.7325472 ,	1.4917564)	0.80682121
rs7748266N	1.4021937	(0.8103150 ,	2.4263989)	0.22696088
rs1360780N	1.1132132	(0.7807917 ,	1.5871630)	0.55341587
rs9394309N	1.1025210	(0.7742022 ,	1.5700710)	0.58843040
rs9470080N	0.9709336	(0.6972114 ,	1.3521180)	0.86141226
rs4792887N	1.1840874	(0.6592118 ,	2.1268779)	0.57175427
rs7209436N	1.0519648	(0.7185205 ,	1.5401509)	0.79450860
rs1344694N	1.0298456	(0.7269056 ,	1.4590369)	0.86858063
rs13316193N	0.9494020	(0.6574256 ,	1.3710512)	0.78184014
rs13125511N	1.1443550	(0.7850611 ,	1.6680846)	0.48309207
rs11131149N	1.1022986	(0.7645578 ,	1.5892353)	0.60181711
rs13273672N	1.4134016	(0.9536437 ,	2.0948118)	0.08478778
rs41423247N	0.9464047	(0.6487289 ,	1.3806719)	0.77496486
rs35369693N	1.1971642	(0.6076943 ,	2.3584262)	0.60292706
rs16859448N	1.2034791	(0.7605015 ,	1.9044828)	0.42899435

TABLE B.19: The results of applying the RO-logit model on the real data, with exposure the particular SNP and outcome the drug abuse score. The method for handling ties used is the Adding method (Part 2).

Kruskal-Wallis Test			
SNP	p-value	SNP	p-value
rs1176744	0.15393621	rs2268491N	0.81329724
rs12529	0.10434891	rs4561970N	0.16645468
rs1799836	0.38563688	rs4686302N	0.58969790
rs1799971	0.14105728	rs1042778N	0.77006155
rs4680	0.62781523	rs1488467N	0.01669361
rs6265	0.45270376	rs2268490N	0.40006526
rs36020	0.69537178	rs2740204N	0.36584371
rs36029	0.16714798	rs3800373N	0.70865275
rs6277N	0.73399392	rs4813625N	0.75387036
rs6190N	0.01288415	rs2770378N	0.40809464
rs6196N	0.37086837	rs6770632N	0.31543295
rs242938	0.33408269	rs2268493N	0.51313924
rs244465	0.26226806	rs2268498N	0.77427436
rs521674	0.04124652	rs2290045N	0.75313004
rs602618	0.04060413	rs1900586N	0.75619734
rs53576N	0.65562310	rs9296158N	0.60721975
rs1800497	0.34602560	rs7748266N	0.48438300
rs1876831	0.22306606	rs1360780N	0.70211659
rs1978340	0.76213775	rs9394309N	0.74273544
rs3219151	0.57381320	rs9470080N	0.82090817
rs3782025	0.29783387	rs4792887N	0.55936844
rs6943555	0.12603501	rs7209436N	0.53064064
rs7590720	0.53694224	rs1344694N	0.55170293
rs9939609	0.37047460	rs13316193N	0.65900812
rs237887N	0.55411170	rs13125511N	0.59485728
rs237889N	0.77945353	rs11131149N	0.66671571
rs237880N	0.03812366	rs13273672N	0.69970238
rs237898N	0.56123642	rs41423247N	0.90125796
rs110402N	0.45371398	rs35369693N	0.91236219
rs242924N	0.37487800	rs16859448N	0.13095838
rs7632287N	0.32215626		

TABLE B.20: The results of applying the Kruskla-Wallis test on the real data, with exposure the particular SNP and outcome the drug abuse score.

Bibliography

Agresti, A. *Analysis of Ordinal Categorical Data*. John Wiley & Sons, Inc, 2nd Edition, 2010.

Agresti, A. *Categorical Data Analysis*. John Wiley & Sons, Inc, 2nd Edition, 2002.

Allison, P.D. & Christakis, N.A. Logit models for sets of ranked items. *Sociological Methodology*, 24, 199-228, 1994.

Andersson, M. *Comparison of the Rank-Ordered Logit and Between-Within Regression Models*. Unpublished Manuscript, 2015.

Berman, A. *AUDIT & DUDIT*. GOTHIA, Edition 1.2, 2012.

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

Borucka, J. *Methods For Handling Tied Events in the Cox Proportional Hazard Model*. Studia Oeconomica Posnaniensia, Vol. 2, No. 2 (263), 2014.

Breslow, N.E. Covariance Analysis of Censored Survival Data. *Biometrics*, 30, 89-99, 1974.

Breslow, N.E. & Crowley, J. A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. *Annals of Statistics*, Vol. 2, No. 3, 437-453, 1974.

Collett, D. *Modelling Survival Data in Medical Research*. Chapman & Hall / CRC Texts in Statistical Science, Second Edition, 2003.

Cox, D.R. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, B, 74, 187-220, 1972.

Cox, D.R. Partial likelihood. *Biometrica*, 62, 269-276, 1975.

Efron, B. The Efficiency of Cox's Likelihood Function for Censored Data. *Journal of the American Statistical Association*, Vol. 72, Issue 359, 1977

Fisher, L.D. & van Belle, G. *Biostatistics - A Methodology for the Health Sciences.* John Wiley & Sons, Inc., 1993.

Forbes, C., Evans, M., Hastings, N. and Peacock, B. *Extreme Value (Gumbel) Distribution, in Statistical Distributions.* John Wiley & Sons, Inc., Fourth Edition, 2010.

Gorgoso-Varela, J. J., & Rojo-Alboreca, A. *Use of Gumbel and Weibull functions to model extreme values of diameter distributions in forest stands.* INRA and Springer-Verlag France, 2014.

Greenland, S. & Morgenstern, H. Matching and efficiency in cohort studies. *American Journal of Epidemiology.*131(1),151-159, 1990.

Gumbel, E.J. *Statistical theory of extreme values and some practical applications.* Applied Mathematics Series, 33. U.S. Department of Commerce, National Bureau of Standards, 1954.

Hommersom, A. & Lucas, P. *Foundations of Biomedical Knowledge Representation: Methods and Applications.* Springer International Publishing, Switzerland, 2015.

Kalbfleisch, J.D. & Pentice, R.L. *The Statistical Analysis of Failure Time Data.* Wiley, New York, 1980.

Kirch, W. Encyclopedia of Public Health. *Springer Netherlands*, Vol. 2, 2008.

Knuth: Computers and Typesetting,
`http://www-cs-faculty.stanford.edu/~uno/abcde.html`

Maria Ungdom website,
`http://mariaungdom.se`

Peto, R. Contribution to the Discussion of a paper of D.R. Cox. *Journal of the Royal Statistical Society*, B, 34, 205-207, 1972.

Sjölander, A. & Greenland, S. Ignoring the matching variable in cohort studies - When is it valid and why? *Statistics in Medicine*, 32,4696-4708, 2013.
*øerø*er

Støer, N. , Seng, C. & Reilly, M. *Matched Designs for Continuous Outcomes.* Unpublished Manuscript, 2016.

Theodorsson-Norheim, E. Kruskal-Wallis test: BASIC computer program to perform nonparametric one-way analysis of variance and multiple comparisons on ranks of several independent samples. *Computer Methods and Programs Biomedicine*, Aug;23(1):57-62, 1986.

R Core Team (2015). The R Package 'foreign' : Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ... R package version 0.8-66, https://CRAN.R-project.org/package=foreign.

Ribatet, M. (2015). The R Package 'SpatialExtremes' : Modelling Spatial Extremes. R package version 2.0-2, https://CRAN.R-project.org/package=SpatialExtremes

R Core Team (2015). The R Package 'stats' : A Language and Environment for Statistical Computing. https://www.R-project.org/

Therneau, T. (2015). The R Package 'survival' : A Package for Survival Analysis in S. R package version 2.38, http://CRAN.R-project.org/package=survival.

Dragulescu, A. A. (2014). The R Package 'xlsx' : Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.5.7, https://CRAN.R-project.org/package=xlsx.