# Modeling of Correlated Data and Multivariate Survival Data

Proefschrift voorgelegd tot het behalen van de graad van
Doctor in de Wetenschappen, richting Wiskunde
te verdedigen door

Fabián Santiago TIBALDI

Promotor : Prof. dr. Geert Molenberghs

*A mis padres*
*Susana y Santiago*

# Acknowledgements

This work would not have been possible without the support and help of many people. These few lines intend to thanks some of them.

First of all, I want to deeply thank to my promotor professor Geert Molenberhgs. Without Geert this work would not have been possible. It was for me a big privilege and a real pleasure to have had the supervision of Geert. I want to thank him not only for sharing with me all his passion for scientific research, but also for his friendship, guidance, support and patience during these last four years.

During my stay at the Limburgs Universitair Centrum I was also involved in consultancy projects and teaching in the Master of Sciences; thanks to Liesbeth Bruckers, to all researchers and to all students, I have learned a lot from them.

Many thanks to all the members of the Surrogate Endpoint Research Group: Ariel Alonso, Tomasz Burzykowski, Marc Buyse, José Cortiñas Abrahantes, Helena Geys, Geert Molenberghs, Dider Renard, Franz Torres Barbosa and Ziv Shkedy, parts of my research is the result of our collaboration.

I especially want to thank professor Paul Janssen and professor Marc Buyse for their reading of this manuscript and for many helpful suggestions. My thanks also go to professor Geert Verbeke and professor Yi Li for the many interesting discussions we have had.

Quiero dedicar unas lineas para agradecer a toda mi familia por el cariño de siempre, por el apoyo infinito y por haber aceptado siempre todas mis decisiones. Gracias a ustedes por haber estado siempre ahi, del otro lado del teléfono o esperándome en el aeropuerto durante todos estos años.

Finally, I gratefully acknowledge the financial support from an LUC Bijzonder Onderzoeksfonds grant that allowed me to stay at L.U.C. and carry out this work.

Fabián Tibaldi

*Diepenbeek*
*March 2004*

# Contents

# List of Abbreviations

| | |
|---|---|
| ACh | Adoptive Child |
| AF | Adoptive Father |
| AM | Adoptive Mother |
| ARMD | Age Related MAcular Degeneration |
| BF | Biological Father |
| BM | Biological Mother |
| EM | Expectation Maximization |
| GEE | Generalized Estimating Equations |
| HIS | Health Interview Survey |
| ML | Maximum Likelihood |
| MLE | Maximum Likelihood Estimator |
| NLPNRR | Non Linear Optimization by Newton-Raphson Ridge Method |
| NSCLC | Non Small Cell Lung Cancer |
| PE | Proportion Explained |
| PL | Pseudo-likelihood |
| RE | Relative Effect |
| REML | Restricted Maximum Likelihood |
| SD | Standard Deviation |
| SE | Standard Error |
| TTP | Time to Progresion |

# List of Tables

# List of Figures

# Chapter 1

# Introduction

## 1.1 Statistical Models and Correlated Data

In many areas of statistics the main goal or objective is to model the data in order to explain a response variable. However, sometimes the interest goes behind this objective and the aim is to study dependencies or correlation between them. This last situation corresponds to studies where the particular type of designs implies to gather the data in groups or clusters.

In the very last years there has been a growing interest in modelling different sort of correlated data. Models for correlated data are appearing as one of the most vital and exciting fields of statistical methodology. Random effects models and estimating equations are the two major schemes used for dealing with the impressive mass of potential relevant applications. These are ranging from spatial analysis up to designs using repeated measures on the same subject, with application in classical medical fields, genetic studies, clinical trials, public health and economics, up to political sciences and sociology.

Sometimes the complexity in the modelling process rises due to the fact that the responses are of different types, for example binary, continuous, survival times, etc. Therefore the methodology needs to be adapted to each of these particular cases.

Some progress have been made in the area of longitudinal data (Verbeke and Molenberghs, 2000) and also with clustered data (Aerts *et al.*, 2002). But these techniques were only developed for classical structures. When the data structures are more complex some problems can be faced and we intend, with this work to propose

new techniques to give alternative solutions for some of these specific problems.

In some areas, an extensive amount of work has been done. A clear example is the analysis of correlated binary data with important contributions made by Liang and Zeger (1986), Zeger and Liang (1986), Zeger, Liang and Albert (1988), who introduced the generalized estimating equation approach. The GEE1 approach makes only first order assumptions and it was extended by Zhao and Prentice (1990); and Liang, Zeger and Qaquish (1992) by incorporating second order assumptions. This introduced the GEE2 method.

Generally, interest in marginal models increases rapidly. These methods emphasize efficient estimation of the effect of covariates on the marginal probabilities.

In the framework of multivariate correlated binary data some work has been done by using pseudo-likelihood as an alternative estimation method (Geys, 1999). This is a non-likelihood method where the principal idea is to replace a numerically challenging joint density by a simpler function that is a suitable product of ratios of likelihoods of subset of the variables. For example, when a joint density contains a computationally intractable normalizing constant, one might calculate a suitable product of conditional densities which does not involve such a complicated function. While the method achieves important computational economies by changing the method of estimation, it does not affect model interpretation. Model parameters can be chosen in the same way as with full likelihood and retain their meaning. Pseudo-likelihood estimation for clustered binary outcomes and its relative merits assessed by means of some examples from developmental toxicity studies can be found in Geys (1999).

## 1.2 Data Structures

Many kinds of data, including observational data collected in the human and biological sciences, have correlated, clustered or hierarchical structure. For example, animal and human studies of inheritance deal with a natural hierarchy where offspring are grouped within families. Offsprings from the same parents tend to be more alike in their physical and mental characteristics than individuals chosen at random from a population at large. It is expected that children from the same family may all tend to be small, perhaps because their parents are small or because of a common impoverished environment.

Other examples of complex structures correspond to the area of surveys where the hierarchical structures are introduced by the sampling design itself. Therefore

in those cases the importance of the appropriate use of the sampling design aspects in producing valid estimates for survey data is very important and it can be easily done via software taking stratification and clustering into account, Tibaldi *et al.*(2003) provide details.

Many other designed experiments also create data hierarchies. In clinical trials carried out in several chosen groups of individuals, or centers, this introduces, in the same way, a kind of hierarchical structure.

In this work, we will deal with some specific settings. Within the surrogate markers evaluation field, correlation is present due to the fact that the clinical trials are run in different centers and patients are distributed over these centers. In addition to that, the type of responses can be of a different nature, depending on the outcome. This covers a wide spectrum: continuous, binary, categorical, survival data, etc. In this work we consider, in particular, the case of continuous normally distributed responses and we explore some potential uses of the multivariate survival situation.

Other examples of correlated data can be found in family studies where the correlation structure is introduced due to the fact that all family members share a common characteristics. We present two instances of this kind of data. An adoption study, where the association between the survival time of biological and adoptive family are modeled and a longevity study, with similar characteristics applied to a very large data set. The new model we develop is not restricted to these studies and to show this, we analyze data from a clinical trial where the correlation is due to the fact that the time-to-event responses come from the same patient.

To end, we present a particular case of structure that was motivated by a psychometric study. The particularity of this data poses an interesting challenge because students and items lead to crossed-random effects, as we will explain later, but only one response is observed at a cross classified level. The correlation appears then in two directions, students introduce dependency between the items (or responses) and times introduce dependency between students' responses. An extra complexity is added when the responses are continuous, categorical or binary. In particular, the binary case is explored at the end of this work and pseudo-likelihood ideas together with conditional logistic regression give an alternative solution for this problem.

## 1.3 Aim of the Thesis and Organization of Subsequent Chapters

The purpose of this thesis is to propose new strategies and techniques for the analysis of particular cases of complex data. Simplified methodologies, pseudo likelihood, conditional linear mixed models will be alternative approaches to tackle some of these problems. The application of such techniques has already begun to yield new and important insight in a number of areas as the examples in the following chapters illustrate. As software becomes always a point of attention when we want to implement the proposed methodology, we have developed our own routines with sufficient flexibility to make easier the fit of these models in a wide number of cases. Some of the developed SAS programs and macros are included in the corresponding chapters to illustrate the implementation of our strategies in typical real examples.

The focus of this thesis will be essentially on modeling of correlated survival data but we also be concerned by outcomes of different type: continuous (normally distributed data), discrete (mostly binary data) and event times with censoring indicator.

The topics covered will fall in three major areas: the first one will deal with modeling of data from clinical trials in order to study surrogacy as we will explain in Chapter 2, the second will deal with modeling of survival or time-to event data to be more general. Thus, in Chapter 3 and Chapter 4 we present the pseudolikelihood method and the copula functions, in particular we focus on the Plackett distribution and the corresponding Plackett copula. These ideas will be combined in order to construct the new model introduced in Chapter 5. Chapter 6 focuses on inference about the model parameters, and three tests are proposed.

In Chapter 7, we apply our model strategies to the problem of modelling associations between time-to-event responses in a pilot clinical trial.

Finally, Chapter 8 shows how a conditional linear mixed model approach can be used effectively in crossed (or non-nested) random-effects models for continuous and the extension to the binary case is presented in Chapter 9. Conclusions from this work and lines of research for the future are given in Chapter 10.

# Chapter 2

# Simplified Hierarchical Linear Models for the Evaluation of Surrogate Endpoints

## 2.1 Introduction

Repeated measures or data from meta-analyses are typical examples of continuous hierarchical data where the linear mixed-effects model (Verbeke and Molenberghs 2000) has become a standard tool. However, in certain situations the model does pose insurmountable computational problems. Precisely this has been the experience of Buyse *et al.* (2000a) who proposed an estimation- and prediction-based approach for evaluating surrogate endpoints. Their approach requires fitting linear mixed models to data from several clinical trials. In doing so, these authors built on the earlier, single-trial based, work by Prentice (1989), Freedman *et al.* (1992), and Buyse and Molenberghs (1998). While Buyse *et al.* (2000a) claim their approach has a number of advantages over the classical single-trial methods, a solution needs to be found for the computational complexity of the corresponding linear mixed model. In this chapter, we propose and study a number of possible simplifications. This is done by means of a simulation study and by applying the various strategies to data from three clinical

studies: Pharmacological Therapy for Macular Degeneration Study Group (1977), Ovarian Cancer Meta-analysis Project (1991) and Corfu-A Study Group (1995).

Prentice (1989) and Freedman *et al.* (1992) laid the foundations for the evaluation of surrogate endpoints in randomized clinical studies. Precisely, Prentice proposed a definition as well as a set of operational criteria. Freedman *et al.* (1992) supplemented these criteria with a quantity called *proportion explained* (PE). Buyse and Molenberghs (1998) proposed to use the *relative effect* (RE), linking the effect of treatment on both endpoints and an individual-level measure of agreement between both endpoints, after adjusting for the effect of treatment (*adjusted association*), instead of the PE. The adjusted association carries over when data are available on several randomized trials, while the RE can be extended to a trial-level measure of agreement between the effects of treatment of both endpoints. As observed by Molenberghs *et al.* (2002) and Alonso *et al.* (2002) there are serious issues surrounding the Prentice-Freedman framework. Let us briefly expand on this. It has been asserted that the criteria set out by Prentice are too stringent (Fleming *et al.* 1996) and neither necessary nor sufficient for his definition to be fulfilled, except in the special case of binary outcomes (Buyse and Molenberghs 1998). In addition, Freedman, Graubard and Schatzkin (1992) showed that these criteria were not straightforward to verify through statistical hypothesis tests. Therefore the PE was suggested but it is surrounded with difficulties, the most dramatic one being that it is not confined to the unit interval (Buyse *et al.* 2000a). Buyse *et al.* (2000a) argued that some fundamental criticisms towards the process of statistical validation can be overcome by combining evidence from several clinical trials, such as in a meta-analysis, rather than from a single study. To this end, they needed to formulate a bivariate hierarchical model, accommodating the surrogate and true endpoints in a multi-trial setting. In doing so, they carry over the relative effect and adjusted association to a trial-level $R^2$ and an individual-level $R^2$, respectively. Similar routes of meta-analytic thinking have been followed by Daniels and Hughes (1997) and Gail *et al.* (2000).

Of course, the switch to a meta-analytic framework does not solve all problems, surrounding surrogate marker validation, in a definitive way. First, one has to carefully reflect upon the question as to how broad the class of units, to be included in a validation study, can be. Clearly, the issue disappears when the same or similar treatments are considered across units (e.g., in multi-center or multi-investigator studies, or when data are used from a family of related study such as in a single drug development line). In a more loosely connected, meta-analytic setting it is important

to ensure that treatment assignments are logically consistent. This is possible, for example, when the same standard treatment is compared to members of a class of experimental therapies.

While the previous issue is relevant, this chapter is devoted to a different, very important, computationally-oriented issue. A result of the change to meta-analysis is that computationally rather involved statistical models have to be used. For the case of surrogates and true endpoints that are both normally distributed, Buyse *et al.* (2000a) employed linear mixed-effects models (Verbeke and Molenberghs 2000). Even in this case, which from a statistical modeling point of view can be considered a basic one, fitting such linear mixed models turns out to be surprisingly difficult. The thrust of their findings is that, when the between-trial variability is sufficiently large, little or no convergence problems occur except when the number of trials is very small.

Given the general importance of linear mixed models, going well beyond the surrogate marker validation case, it is necessary to study convergence properties in more detail, and to contrast the general linear mixed model, such as the one proposed by Buyse *et al.* (2000a), with alternative and/or simplified strategies. A number of such alternative strategies are proposed here and studied in terms of their statistical and numerical properties. To this end, a simulation study is considered, and the various methods are applied to the data studied in Buyse *et al.* (2000a).

The meta-analytic setting, to be used throughout the chapter, is introduced in Section 2.2. The simplified approaches, organized along three "dimensions", are presented in Section 2.3. Sections 2.4–2.6 are devoted to each of the three dimensions in turn. Case studies are introduced and analyzed in Section 2.7 and a simulation study is reported in Section 2.8.

## 2.2 Setting

As stated earlier, we will focus on normally distributed endpoints. Let us introduce a set of notation that will be used throughout the chapter. Let $T_{ij}$ and $S_{ij}$ be random variables denoting the true and the surrogate endpoints for subject $j = 1, \ldots n_i$ in trial $i = 1, \ldots N$. Further, let $Z_{ij}$ denote a binary treatment indicator.

The full random-effects model, as introduced by Buyse *et al.* (2000a) is

$$S_{ij} = \mu_S + m_{S_i} + \alpha Z_{ij} + a_i Z_{ij} + \varepsilon_{S_{ij}}, \tag{2.1}$$

$$T_{ij} = \mu_T + m_{T_i} + \beta Z_{ij} + b_i Z_{ij} + \varepsilon_{T_{ij}}, \tag{2.2}$$

where $\mu_S$ and $\mu_T$ are fixed intercepts, $m_{S_i}$ and $m_{T_i}$ are random intercepts for trial $i$, $\alpha$ and $\beta$ are fixed treatment effects and $a_i$ and $b_i$ are random treatment effects. The individual-specific error terms are $\varepsilon_{S_{ij}}$ and $\varepsilon_{T_{ij}}$.

The vector of random effects, $(m_{S_i}, m_{T_i}, a_i, b_i)'$, is assumed to be zero-mean normally distributed with covariance matrix

$$D = \begin{pmatrix} d_{SS} & d_{ST} & d_{Sa} & d_{Sb} \\ d_{ST} & d_{TT} & d_{Ta} & d_{Tb} \\ d_{Sa} & d_{Ta} & d_{aa} & d_{ab} \\ d_{Sb} & d_{Sa} & d_{ab} & d_{bb} \end{pmatrix}.$$

The individual-level error terms $(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}})'$ are also zero-mean normally distributed with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_{SS} & \sigma_{ST} \\ \sigma_{ST} & \sigma_{TT} \end{pmatrix}.$$

Parameter estimation can be based on, for example, maximum likelihood or restricted maximum likelihood (Verbeke and Molenberghs, 2000). Next, suppose we consider a new trial, $i = 0$ say, for which data are available on the surrogate endpoint but not on the true endpoint. We are interested in the estimated effect of $Z$ on $T$, given the effect of $Z$ on $S$ for this particular trial. Subscript all quantities pertaining to the particular trial under study with 0. It is easy to show (Buyse *et al.* 2000a) that $(\beta + b_0 | m_{S0}, a_0)$ follows a normal distribution with mean and variance:

$$E(\beta + b_0 | m_{S0}, a_0) = \beta + \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}' \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} \mu_{S0} - \mu_S \\ \alpha_0 - \alpha \end{pmatrix}, \quad (2.3)$$

$$\mathrm{Var}(\beta + b_0 | m_{S0}, a_0) = d_{bb} - \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}' \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}. \quad (2.4)$$

Related to prediction equations (2.3)–(2.4), a measure to assess the quality of the surrogate at the trial level is the coefficient of determination

$$R^2_{\text{trial}} = R^2_{b_i | m_{Si}, a_i} = \frac{\begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}' \begin{pmatrix} d_{SS} & d_{Sa} \\ d_{Sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{Sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}. \quad (2.5)$$

A good surrogate, *at the trial level*, would have (2.5) close to 1. Intuition can be gained by considering the simplified case where the prediction of $b_0$ is done independently of

the random intercept $m_{s0}$. The coefficient (2.5) then reduces to

$$R^2_{\text{trial(r)}} = R^2_{b_i|a_i} = \frac{d^2_{ab}}{d_{aa}d_{bb}}. \tag{2.6}$$

This formula is useful when the full random-effects model is hard to fit but a reduced version, excluding random intercepts, is easier to reach convergence. It is simply the square of the correlation between $\alpha_i$ and $\beta_i$. Note that $R^2_{\text{trial(r)}} = 1$ if the trial level treatment effects are simply multiples of each other.

## 2.3 Simplified Modelling Strategies

Buyse *et al.* (2000a) showed that fitting random-effects model (2.1)–(2.2) can be a surprisingly difficult task in a number of situations. This is particularly true when the number of trials or the number of patients per trial is small. Also, situations with extreme correlations pose problems. It is therefore imperative to explore approximate strategies with better computational properties. These authors studied one alternative approach in the sense that they replaced the random effects by their fixed-effect counterparts. Such a two-stage approach is very similar in spirit to the original proposal of Laird and Ware (1982). We will now embed this ad-hoc strategy in a more formally developed system of model simplifications.

Precisely, we consider three dimensions along which simplifications can be made:

**Trial dimension:** whether the trial-specific effects are treated as either random or fixed. A full random-effects is then distinguished from a two-stage approach.

**Endpoint dimension:** whether the surrogate and true endpoints are modelled as a bivariate outcome or two univariate ones. In the latter case the correlation between both endpoints is not incorporated into the modeling strategy, rendering the study of the individual-level surrogacy more involved. However, as stated earlier, throughout this chapter the focus is on trial-level surrogacy.

**Measurement error dimension:** whenever the full random-effects model is abandoned, one is confronted with measurement error since the treatment effects in the various trials are estimated with error. The magnitude of this error is likely to depend on several characteristics, such as trial size, which will vary across trials. We consider three ways to account for measurement error: unadjusted

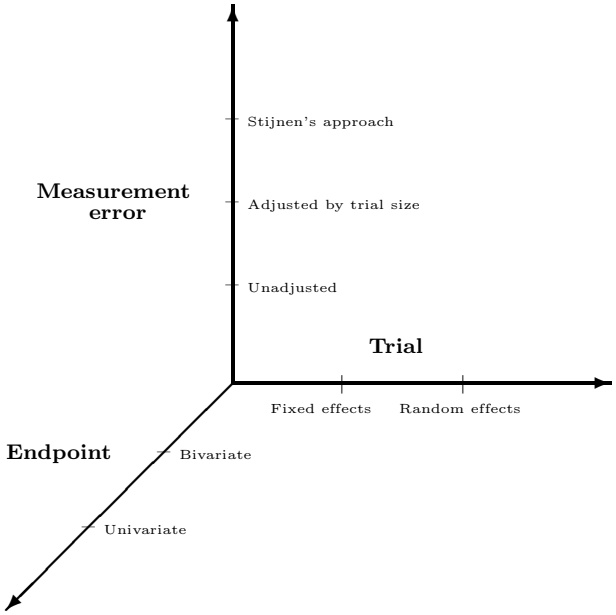


Figure 2.1: *Graphical representation of the three different approaches.*

(i.e., no correction at all), adjustment by trial size, and an approach suggested by T. Stijnen (Van Houwelingen *et al.* 2002) and explained in Section 2.5.

The combination of these three dimensions are graphically represented in Figure 2.1 and gives rise to twelve strategies. However, some do not have to be considered. For example, when one chooses for a bivariate (endpoint dimension) random-effects (trial dimension) approach, measurement error is automatically accounted for, whence explicit corrections are no longer needed. In the special case when sample size is constant across trials, further simplifications arise (see Section 2.8).

We will now discuss each of the three simplifying dimensions in turn.

## 2.4 The Trial Dimension

As stated before, the parameters of the full random-effects model (2.1)–(2.2) can be estimated by maximum likelihood or restricted maximum likelihood, using standard linear mixed model software such as the SAS procedure MIXED.

In case we treat the trial-level parameters as fixed, exactly as Buyse *et al.* (2000a), we can rewrite the model as

$$S_{ij} = \mu_{S_i} + \alpha_i Z_{ij} + \varepsilon_{S_{ij}}, \tag{2.7}$$

$$T_{ij} = \mu_{T_i} + \beta_i Z_{ij} + \varepsilon_{T_{ij}}, \tag{2.8}$$

where $\mu_{S_i}$, $\mu_{T_i}$, $\alpha_i$, and $\beta_i$ are trial-specific intercepts and treatment effects. The assumption about the error terms depends on the choice made on the *endpoint dimension* (Section 2.6). Indeed, when the univariate approach is opted for, both errors are assumed independent. Otherwise, a bivariate unstructured covariance matrix is considered.

At the second stage, a regression model is fitted to the treatment effects, estimated at the first stage, for example:

$$\widehat{\beta}_i = \lambda_0 + \lambda_1 \widehat{\mu}_{S_i} + \lambda_2 \widehat{\alpha}_i + \varepsilon_i. \tag{2.9}$$

This model can then be employed to assess trial-level surrogacy, using the $R^2_{\text{trial}}$ associated with this regression. Precisely, this is not calculated as in (2.5), but is merely the classical coefficient of determination found by regressing $\widehat{\beta}_i$ on $\widehat{\mu}_{S_i}$ and $\widehat{\alpha}_i$.

In case the trial-specific intercept from surrogate model (2.7) is not used, $\lambda_1$ would be dropped and an $R^2_{\text{trial(r)}}$ is obtained, similar in spirit to (2.6).

## 2.5 The Measurement Error Dimension

Recall that this dimension is irrelevant when the full random-effects model is assumed, but is crucial when a fixed-effects approach is selected on the *trial dimension* and/or when a univariate model is chosen on the *endpoint dimension*.

We allow for three possible choices. First, a simple linear model can be assumed to determine the relationship between $\beta_i$, $\alpha_i$, and $\mu_{S_i}$, whereby the errors in (2.9) are assumed to be zero-mean normally distributed with constant variance $\sigma^2$.

Clearly, this approach ignores the fact that the estimated treatment effects $\alpha_i$ and $\beta_i$ will typically come from trials with large variations in size. One way to address this issue is by weighing the contributions according to trial size, resulting in a weighted linear regression. Such an approach may account for some but not all of the heterogeneity in information content between trial-specific contributions. A nice way to overcome this is T. Stijnen's approach.

To this end, we introduce models for the estimated trial-specific treatment effects $(\widehat{\mu}_{S_i}, \widehat{\alpha}_i, \widehat{\beta}_i)'$, given the true trial-specific treatment effects $(\mu_{S_i}, \alpha_i, \beta_i)'$:

$$\begin{pmatrix} \widehat{\mu}_{S_i} \\ \widehat{\alpha}_i \\ \widehat{\beta}_i \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_{S_i} \\ \alpha_i \\ \beta_i \end{pmatrix}, C_i \right). \tag{2.10}$$

Here, $C_i$ is the variance-covariance matrix of the estimated treatment effects. In case we assume both treatment-effect estimates to be independent (which would result from a univariate choice on the *endpoint dimension*), $C_i$ would be assumed to be diagonal, even though this may be unrealistic.

Further, we assume a normal model for the true trial-specific treatment effects around the true overall treatment effects:

$$\begin{pmatrix} \mu_{S_i} \\ \alpha_i \\ \beta_i \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_S \\ \alpha \\ \beta \end{pmatrix}, \Sigma \right). \tag{2.11}$$

The resulting marginal model, combining (2.10) and (2.11), is:

$$\begin{pmatrix} \widehat{\mu}_{S_i} \\ \widehat{\alpha}_i \\ \widehat{\beta}_i \end{pmatrix} \sim N \left( \begin{pmatrix} \mu_S \\ \alpha \\ \beta \end{pmatrix}, \Sigma + C_i \right). \tag{2.12}$$

Maximum likelihood estimation for this model can be quite easily carried out by using mixed model software, provided the values for $C_i$ can be input and held fixed, as is the case in the SAS procedure MIXED. An example program is provided in the Appendix of this chapter.

## 2.6 Endpoint Dimension

It seems natural to assume both endpoints to be correlated. However, this assumption will almost always complicate modelling and corresponding parameter estimation. In addition, the bivariate nature of the outcome is related for the better part with individual-level surrogacy whereas our main goal is trial-level surrogacy. This suggests an additional simplification, i.e., by considering separate, independent models for each of the endpoints. It then remains to be seen inhowfar such a simplification hampers estimation of trial-level surrogacy.

We need to make a distinction between two cases, according to the corresponding choice on the *trial dimension*. In the random-effects approach, this simplification would lead to a pair of *univariate* hierarchical models, one for each endpoint. In the fixed-effects approach, one would fit a separate linear regression model per endpoint and per trial. It is easy to show that the parameter estimates as well as the estimated variances are identical to the ones obtained from fitting a fixed-effects *bivariate* model to each trial separately. This follows from standard multivariate normal theory (Johnson and Wichern 1992).

## 2.7 Case Studies

We consider three case studies, two of which were considered by Buyse *et al.* (2000). It permits us to compare their results with those obtained from a full set of computational approaches. Further, they cover three important but different therapeutic areas. Finally, by considering three case studies, we avoid the risk of running into results that are interesting but too specialized to a particular situation.

The first one, the Age Related Macular Degeneration Study, is an ophthalmologic study. The other two are from advanced colorectal and advanced ovarian cancer. These examples have been studied in Buyse *et al.* (2000a, 2000b). We will compare their results to the ones from the simplified approaches proposed in this chapter. Results are summarized in Table 2.1, following the three dimensions of Figure 2.1. The focus is on trial-level surrogacy, captured by $R^2_{\text{trial}}$. While, of course, the individual-level surrogacy is of interest when the focus is on predicting a particular patient's behavior and, in some contexts, can even be of primary interest (Alonso *et al.* 2001), it is fair to say that the clinical trialist will primarily be interested in the trial-level surrogacy. Further, since the inclusion of the individual-level surrogacy forces the models to have a bivariate nature, this comes at a computational cost.

In addition, we distinguish between "full" models where the trial level surrogacy $R^2_{\text{trial}}$ is calculated as in (2.5), and "reduced" models, where no random intercepts are included and hence $R^2_{\text{trial(r)}}$ as in (2.6) is used. Combining all possibilities on three dimensions and furthermore distinguishing between full and reduced models would, in principle, lead to 24 different approaches. However, the three bivariate random-effects approaches coincide. The columns for the full approaches are numbered for reference in the simulation study (Section 2.8).

Table 2.1: *Results of the trial-level surrogacy analysis for the three examples $R^2_{trial}$ (a − symbol indicates non-convergence).*

| | **Full Model** | | | | | |
|---|---|---|---|---|---|---|
| | Univariate Approach | | | | | |
| | Fixed-effects approach | | | Random-effects approach | | |
| | Unweighted | Weighted | Stijnen | Unweighted | Weighted | Stijnen |
| Study | 1 | 2 | 3 | 4 | 5 | 6 |
| ARMD | 0.692 | 0.693 | 0.689 | 0.664 | 0.801 | – |
| Colorectal | 0.473 | 0.488 | 0.466 | – | – | – |
| Ovarian | 0.939 | 0.917 | 0.937 | 0.911 | 0.905 | – |
| | Bivariate Approach | | | | | |
| | Fixed-effects approach | | | Random-effects approach | | |
| | Unweighted | Weighted | Stijnen | | | |
| Study | 7 | 8 | 9 | 10–12 | | |
| ARMD | 0.692 | 0.693 | 0.698 | – | | |
| Colorectal | 0.473 | 0.488 | 0.472 | – | | |
| Ovarian | 0.939 | 0.917 | 0.938 | – | | |
| | **Reduced Model** | | | | | |
| | Univariate Approach | | | | | |
| | Fixed-effects approach | | | Random-effects approach | | |
| Study | Unweighted | Weighted | Stijnen | Unweighted | Weighted | Stijnen |
| ARMD | 0.776 | 0.758 | 0.775 | 0.659 | 0.786 | 0.623 |
| Colorectal | 0.527 | 0.497 | 0.596 | – | – | – |
| Ovarian | 0.928 | 0.909 | 0.925 | 0.911 | 0.905 | 0.900 |
| | Bivariate Approach | | | | | |
| | Fixed-effects approach | | | Random-effects approach | | |
| Study | Unweighted | Weighted | Stijnen | | | |
| ARMD | 0.776 | 0.758 | 0.719 | – | | |
| Colorectal | 0.527 | 0.497 | 0.471 | – | | |
| Ovarian | 0.928 | 0.909 | 0.938 | 0.951 | | |

### 2.7.1 Age Related Macular Degeneration Study (ARMD)

These data arose from a randomized clinical trial comparing an experimental treatment (interferon-$\alpha$) to placebo in the treatment of patients with age-related macular

degeneration. The aim of the study was to compare placebo and the highest dose of interferon-$\alpha$. The treatment indicator is $Z_{ij} = 1$ for treatment and 0 for placebo. Since we have a single multi-centric trial, $i$ refers to *center* and $j$ to patient within center. The true endpoint in this study was the change in visual acuity at 12 months after starting the treatment. The surrogate endpoint considered is visual acuity at 6 months. Results from assessing the surrogate in terms of the Prentice-Freedman framework were reported in Buyse *et al.* (2000a) and are not repeated here.

Buyse *et al.* (2000a) experienced problems in fitting the full random-effects models, irrespective of whether standard statistical software or user-developed alternatives were used. Therefore, they entertained a (unweighted) fixed-effects approach instead. This produced a moderate trial-level surrogacy: $R^2_{\text{trial (f)}} = 0.692$ ($s.e.$ 0.087). The standard error has been calculated by means of a straightforward application of the delta method. Let us now compare their result to the ones obtained from the approaches described in Section 2.3.

As mentioned earlier, for the fixed-effects approaches, univariate and bivariate results values are equal. Of course, the univariate approach prohibits the assessment of individual-level surrogacy but, as mentioned earlier, in many trials the main interest is on trial-level surrogacy.

For the $R^2_{\text{trial}}$, Stijnen's approach is more difficult to fit in the sense that the random-effects values cannot be obtained.

The reduced-model values are generally higher than the full-model values, suggesting that the trial-specific intercept terms for the surrogate model does convey information and, if possible, full models should be used. Within the reduced-model approach, Stijnen's univariate random-effects approach yields a low value. This is in line with intuition, since it corrects for measurement error present in the estimated treatment effects. Simulations will have to weigh costs and benefits from this approach. In general computational terms, a choice for univariate models and/or fixed-effects approaches is less expensive.

### 2.7.2 Advanced Colorectal Cancer

We consider data from two randomized multicenter trials in colorectal cancer. These constitute the largest source of randomized data available in advanced colorectal cancer. All data were collected and checked by the Meta-Analysis Group In Cancer between 1990 and 1996 (Corfu-A Group, 1995; Greco *et al.* 1996) to confirm the

benefits of experimental fluoropyrimidine treatments with 5-fluorouracil (5FU) in advanced colorectal cancer. The principal investigators of all trials provided data for every patient, whether eligible or not, and whether properly followed-up or not. Previous publications provide full details on the trials included the treatments tested, the patient characteristics, and the therapeutic results (Burzykowski *et al.* 2001).

In this example, we will use $Z_{ij} = 0$ to denote 5FU plus interferon and for 5FU alone. The final endpoint $T_{ij}$ will be survival time in years. The surrogate endpoint $S_{ij}$ will be progression-free survival time, i.e., the years between the randomization to clinical progression of the disease or death. In agreement with previous analyses, only centers with at least 3 patients on each treatment arm are considered. The data include 48 centers, with a total sample size of 642 patients. Using the bivariate unweighted fixed-effects approach model proposed by Buyse *et al.* (2000a) we obtain $R^2_{\text{trial (f)}} = 0.473$ (*s.e.* 0.108), which is, of course, too low to be useful.

Results of fitting the various approaches and reported in Table 2.1 largely confirm the results from the ARMD study in terms of ease of convergence for the univariate and/or fixed-effects approaches. All coefficients are relatively close to each other, although the reduced versions tend to be a bit higher than the full versions.

### 2.7.3 Advanced Ovarian Cancer

These data arose from a meta-analysis of ovarian cancer (Ovarian Cancer Meta-Analysis Project, 1991). The comparison of two treatments was the principal aim of this study. We use $Z_{ij} = 0$ when cyclosphosphamide was applied and $Z_{ij} = 1$ when cyclosphosphamide plus cisplatin was applied. We considered survival time in years as final endpoint $T_{ij}$. The surrogate endpoint $S_{ij}$ is progression-free survival time. We used center as the unit of analysis given that the number of trials is insufficient to applied meta-analytic methods. The number of patients distributed over a total of 50 units varies from 2 to 254.

The bivariate fixed-effects approach used by Buyse *et al.* (2000a) produces $R^2_{\text{trial (f)}} = 0.917$ (*s.e.* 0.017), which is much higher than in the colorectal cancer case. Arguably, this is due to the relatively short time span that typically elapses between both endpoints. The difference between this result and those from the other approaches is even smaller than in the other two case studies. Further, the relative computational complexity, suggested by the other case studies, is confirmed here as well.

## 2.8 A Simulation Study

We studied performance of the various approaches, in terms of estimation (point and interval) of $R^2_{\text{trial}}$, and in terms of convergence through a simulation study. To make our results comparable with those from Buyse *et al.* (2000a), the same configuration setting is adopted. Precisely, model (2.1)–(2.2) is considered with $(m_{S_i}, m_{T_i}, a_i, b_i) \sim N(0, D)$, $\mu_S = 50$, $\mu_T = 45$, $\alpha = 5$, $\beta = 3$,

$$D = \sigma^2 \begin{pmatrix} 1 & 0.8 & 0 & 0 \\ 0.8 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho \\ 0 & 0 & \rho & 1 \end{pmatrix}, \tag{2.13}$$

with $\rho^2 = 0.5$ or $\rho^2 = 0.9$, and $(\varepsilon_{S_{ij}}, \varepsilon_{T_{ij}}) \sim N(0, \Sigma)$ with

$$\Sigma = 3 \begin{pmatrix} 1 & \sqrt{0.8} \\ \sqrt{0.8} & 1 \end{pmatrix}.$$

The parameter $\sigma^2$ was chosen to be either 3 or 10. Five hundred runs were completed for every setting, consisting of 25 trials each. The true $R^2$, following from (2.5) and (2.13) is set equal to either 0.5 or 0.9.

Results are presented in Tables 2.2–2.3. In all settings, convergence was 100%, which is slightly different from the analysis of the examples.

Stijnen's approach exhibits a small amount of bias. In case $R^2 = 0.9$ and $\sigma^2 = 3$, there is a hint of underestimation in column 3, 6, and somehow also 9. The situation is more dramatic in the case of $R^2 = 0.5$, where indeed we observe now overestimation in all but one columns, the exception being the full model (columns 10–12).

Table 2.2: *Means of the estimated trial-level surrogacy and 95% simulation-based confidence intervals for* $R^2 = 0.90$. *Column numbers refer to the columns of Table 2.1.*

| # Sub | 1, 2, 7, 8 | 3 | 4, 5 | 6 | 9 | 10–12 |
|---|---|---|---|---|---|---|
| | | | Variance 10 | | | |
| 50 | 0.898 (0.894;0.902) | 0.895 (0.890;0.900) | 0.898 (0.895;0.902) | 0.894 (0.890;0.898) | 0.898 (0.894;0.902) | 0.896 (0.892;0.900) |
| 60 | 0.900 (0.897;0.904) | 0.899 (0.896;0.903) | 0.901 (0.897;0.904) | 0.897 (0.893;0.900) | 0.900 (0.896;0.903) | 0.897 (0.894;0.901) |
| 70 | 0.898 (0.894;0.902) | 0.896 (0.892;0.901) | 0.898 (0.894;0.902) | 0.894 (0.890;0.899) | 0.897 (0.893;0.902) | 0.895 (0.891;0.900) |
| 80 | 0.899 (0.895;0.903) | 0.898 (0.894;0.902) | 0.899 (0.895;0.903) | 0.895 (0.891;0.899) | 0.898 (0.894;0.902) | 0.896 (0.892;0.900) |
| 90 | 0.900 (0.896;0.903) | 0.899 (0.895;0.902) | 0.900 (0.896;0.903) | 0.896 (0.892;0.899) | 0.899 (0.896;0.903) | 0.897 (0.893;0.901) |
| 100 | 0.901 (0.898;0.905) | 0.901 (0.897;0.904) | 0.901 (0.898;0.905) | 0.897 (0.894;0.901) | 0.901 (0.897;0.904) | 0.898 (0.895;0.902) |
| | | | Variance 3 | | | |
| 50 | 0.893 (0.889;0.897) | 0.889 (0.885;0.894) | 0.894 (0.890;0.898) | 0.892 (0.888;0.896) | 0.892 (0.888;0.896) | 0.896 (0.891;0.900) |
| 60 | 0.896 (0.893;0.900) | 0.893 (0.889;0.897) | 0.897 (0.893;0.901) | 0.896 (0.892;0.899) | 0.895 (0.892;0.899) | 0.897 (0.893;0.901) |
| 70 | 0.894 (0.890;0.898) | 0.890 (0.886;0.895) | 0.894 (0.890;0.898) | 0.891 (0.887;0.896) | 0.893 (0.889;0.897) | 0.895 (0.890;0.899) |
| 80 | 0.895 (0.891;0.899) | 0.892 (0.888;0.896) | 0.896 (0.892;0.900) | 0.894 (0.890;0.898) | 0.895 (0.891;0.899) | 0.896 (0.892;0.900) |
| 90 | 0.897 (0.893;0.900) | 0.894 (0.890;0.898) | 0.897 (0.894;0.901) | 0.893 (0.889;0.897) | 0.896 (0.893;0.900) | 0.897 (0.893;0.901) |
| 100 | 0.898 (0.895;0.902) | 0.896 (0.892;0.899) | 0.899 (0.895;0.902) | 0.895 (0.891;0.899) | 0.898 (0.894;0.901) | 0.898 (0.894;0.902) |

Table 2.3: *Means of the estimated trial-level surrogacy and 95% simulation-based confidence intervals for* $R^2 = 0.50$. *Column numbers refer to the columns of Table 2.1.*

| # Sub | 1, 2, 7, 8 | 3 | 4, 5 | 6 | 9 | 10–12 |
|---|---|---|---|---|---|---|
| | | | Variance 10 | | | |
| 50 | 0.527 (0.515;0.539) | 0.526 (0.514;0.538) | 0.528 (0.516;0.540) | 0.523 (0.511;0.535) | 0.526 (0.514;0.538) | 0.498 (0.485;0.510) |
| 60 | 0.532 (0.520;0.544) | 0.531 (0.519;0.543) | 0.533 (0.521;0.544) | 0.529 (0.517;0.540) | 0.531 (0.519;0.543) | 0.502 (0.490;0.515) |
| 70 | 0.525 (0.513;0.538) | 0.524 (0.512;0.537) | 0.526 (0.513;0.538) | 0.522 (0.509;0.535) | 0.525 (0.512;0.537) | 0.500 (0.487;0.513) |
| 80 | 0.522 (0.509;0.536) | 0.522 (0.509;0.535) | 0.523 (0.510;0.536) | 0.520 (0.506;0.533) | 0.522 (0.509;0.535) | 0.498 (0.484;0.511) |
| 90 | 0.524 (0.512;0.535) | 0.523 (0.511;0.535) | 0.524 (0.512;0.536) | 0.520 (0.509;0.532) | 0.523 (0.511;0.535) | 0.501 (0.488;0.513) |
| 100 | 0.526 (0.514;0.538) | 0.525 (0.513;0.538) | 0.527 (0.514;0.539) | 0.523 (0.510;0.535) | 0.525 (0.513;0.538) | 0.503 (0.490;0.516) |
| | | | Variance 3 | | | |
| 50 | 0.539 (0.527;0.551) | 0.535 (0.523;0.547) | 0.542 (0.530;0.554) | 0.534 (0.522;0.546) | 0.538 (0.526;0.550) | 0.496 (0.483;0.510) |
| 60 | 0.542 (0.531;0.554) | 0.539 (0.527;0.551) | 0.545 (0.534;0.557) | 0.538 (0.526;0.550) | 0.542 (0.530;0.553) | 0.501 (0.488;0.514) |
| 70 | 0.533 (0.521;0.546) | 0.530 (0.518;0.543) | 0.535 (0.522;0.547) | 0.528 (0.516;0.541) | 0.532 (0.520;0.545) | 0.497 (0.484;0.511) |
| 80 | 0.531 (0.517;0.544) | 0.529 (0.516;0.542) | 0.533 (0.519;0.546) | 0.527 (0.514;0.540) | 0.530 (0.517;0.543) | 0.497 (0.483;0.511) |
| 90 | 0.531 (0.519;0.542) | 0.529 (0.517;0.540) | 0.532 (0.520;0.544) | 0.527 (0.515;0.538) | 0.530 (0.518;0.542) | 0.500 (0.487;0.512) |
| 100 | 0.531 (0.519;0.544) | 0.530 (0.518;0.542) | 0.534 (0.521;0.546) | 0.528 (0.516;0.541) | 0.531 (0.519;0.543) | 0.502 (0.489;0.515) |

## 2.9 Conclusions

In this chapter, we have investigated several strategies to deal with the computational burden posed by using hierarchical linear models, primarily in the context of validating surrogate markers. These strategies are ordered following three choices: (1) whether trial-specific parameters are treated as random or fixed, (2) whether the endpoints are treated as correlated or not (bivariate versus univariate approach) and (3) the method of dealing with measurement error.

As a result of this, we recommend simplified computational methods for two main reasons. First, they are generally faster and easier to implement with standard software. Second, we showed, through simulations, that the simplified approaches often perform almost as good as the more advanced methods, and moreover enjoy much better convergence properties. In particular, opting for a fixed-effects approach over a full random-effects approach is very beneficial since there is at most a minor loss in statistical efficiency, the method has extremely good convergence properties, and is usually more than 10 times faster than the full approach. In addition, from the different simplifications proposed here, univariate approaches are the easiest to implement because they can be performed by means of linear regression by using any basic software. However, in case of using the Stijnen's correction a more powerful software is needed.

We re-analyzed the three case studies considered by Buyse *et al.* (2000), from three therapeutic areas: ophthalmology, advanced colorectal cancer, and advanced ovarian cancer. In agreement with the simulation study, the fixed-effects approaches have good convergence properties, but there are problems with the random-effects approaches. In particular, none of the fully bivariate random-effects models converged, while there were also problems with their univariate and/or reduced counterparts. While there are twelve versions of each fixed-effects approach, the results are generally very similar across these, except that there is a noticeable but not a dramatic difference between the full and reduced versions. Therefore, it is recommendable to use the full model version since, in doing so, full information is used towards estimation of the trial-level surrogacy.

# Appendix

```
/* First stage: bivariate fixed-effects model */

proc mixed data=mydata method=reml;
  class trial subj endpoint;
  model outcome=endpoint*trial endpoint*trial*treat
                / noint s covb ddfm=bw;
  repeated endpoint / subject=subj type=un r rcorr;
  make 'SolutionF' out=effects;
  make 'CovParms' out=covparms;
  make 'covb' out=covar;
run;

/*
** Assembling trial-specific covariance matrices of estimated
** fixed effects. There is one line per trial, each such line
** corresponding to a matrix.
*/

data cov0;
set covar;
 drop _row_ _effect_ trial endpoint;
run;

proc iml;
 use cov0;
 ntrial=25;
 read all into tempdat;
 dummy=j(ntrial,7,0);
 do i=1 to ntrial;
   dummy[i,1]=tempdat[2*i-1,2*ntrial+(2*i-1)];
   dummy[i,2]=i;
   dummy[i,3]=tempdat[2*i-1,2*ntrial+2*i];
   dummy[i,4]=tempdat[2*ntrial+2*i,2*ntrial+(2*i-1)];
   dummy[i,5]=tempdat[2*i-1,2*i-1];
   dummy[i,6]=tempdat[2*ntrial+(2*i-1),2*ntrial+(2*i-1)];
   dummy[i,7]=tempdat[2*ntrial+2*i,2*ntrial+2*i];
```

```
 end;
 nms={"cmsal","trial","cmsbe","calbe","varms","varal","varbe"};
 create cova0 from dummy [colname=nms];
 append from dummy;
quit;

data effects;
set effects;
 keep _EFFECT_ _EST_ _se_ trial endpoint order int surro main;
 int=0;
 surro=0;
 main=0;
 if _effect_='TRIAL*ENDPOINT' then do;
      if endpoint=1 then delete;
      if endpoint=0 then do;
         order=3;
         int=1;
      end;
 end;
 if _effect_='TREAT*TRIAL*ENDPOINT' then do;
      if endpoint=0 then do;
         order=1;
         surro=1;
      end;
      if endpoint=1 then do;
         order=2;
         main=1;
      end;
 end;
run;

proc sort data=effects;
 by trial order;
run;

data stijnen;
set effects;
 drop _est_;
```

```
 est=_est_;
run;

proc sort data=stijnen;
 by trial order;
run;

data row1;
set cova0;
 keep row col value trial;
 col=trial;
 row=1;
 value=varal;
run;

...

data row6;
set cova0;
 keep row col value trial;
 col=trial;
 row=6;
 value=varms;
run;

data matrix;
 set row1 row2 row3 row4 row5 row6;
run;

proc sort data=matrix;
 by col row;
run;

/* Second stage: Stijnen's regression */

proc mixed data=stijnen order=data method=real asycov scoring=2;
  class trial order;
  model est = order / solution noint ddfm=bw;
```

```
  random order / subject=trial group=trial type=un gdata=matrix;
  repeated order / subject=trial type=un;
  make 'CovParms' out=covparms noprint;
  make 'AsyCov' out=asycov noprint;
run;
```

# Chapter 3

# Multivariate Survival Models and Copulas

## 3.1 Introduction

The main purpose of this chapter is to review definitions, properties and concepts of copulas. We will consider a general copula approach to multivariate survival modelling. This summary is based on the monography by Nelsen (1999) where the proofs of theorems included in this chapter can be found. Copulas were used in survival analysis by Clayton (1978), Hougaard (1986), Marshall and Olkin (1988), Oakes (1989), Bagdonavicius, Malov and Nikulin (1999), Shih and Louis (1995), Burzykowski (2001), etc. Copulas provide a general framework, that could encompass many models generally presented without link between them.

The main differences between univariate and multivariate survival models is that the last models cover the field where independence between survival times cannot be assumed. The joint distribution of the survival times and the corresponding multivariate survival function need to be specified. This is usually done in two steps. In the first step we consider univariate separated models in order to characterize the margin distributions, afterwards a model for the joint model of the survival times is constructed by using the information obtained in the first step.

A multivariate survival function with given margins can be constructed using copulas. One of the first propositions in this area was done by Clayton (1978) who

used a bivariate association model for survival analysis. At that moment the concept of copula was not mentioned but implicitly a copula was used.

The aim of this chapter is to introduce some concepts that will be used along this work and to provide the reader with some flavor of the use of survival copulas and some measures of association related to the concept of copulas.

## 3.2 Definitions and Notation

Let us suppose that we have a survival time, or in a more general framework a time-to-event, $T$, with distribution $F$. The survival function is then given by

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

The density $f$ of $T$ can be obtained by taking the first derivate of $F$ or, equivalently, the first derivate of $-S(t)$.

When interested in modeling survival times, one of the main concepts is the so-called hazard rate, a function that can be interpreted as the instantaneous failure rate of an individual, given that it survived up to time $t$.

The hazard rate is defined as follows,

$$\lambda(t) = \lim_{\Delta \to 0^+} \frac{1}{\Delta} \Pr(t \leq T \leq t + \Delta | T \geq t).$$

Another equivalent expression for $\lambda$ is

$$\lambda(t) = -\frac{\partial S(t)}{\partial t} . \frac{1}{S(t)} = \frac{f(t)}{S(t)}.$$

Therefore, the integral between 0 and $t$ represents the cumulative hazard function $\Lambda(t)$;

$$\Lambda(t) = \int_0^t \lambda(x) dx.$$

There is a natural relationship between the cumulative hazard function and the survival function, given by

$$S(t) = \exp(-\Lambda(t)).$$

A baseline hazard function $\lambda_0$ can also be incorporated into the model. One way to incorporate explanatory variables ($X$) is by means of Cox's (1972) proportional hazards model. It takes the following expression:

$$\lambda(t) = \exp(\boldsymbol{\beta}' X) \lambda_0(t).$$

Up to now we presented some concepts of univariate survival analysis. These can be extended to the more general case of multivariate survival data.

If $T_1, \ldots, T_k$ are $k$ survival times, taking values in $\mathbb{R}^+$, the multivariate survival function $S(t)$ is defined as

$$S(t_1, \ldots, t_k) = \Pr(T_1 > t_1, \ldots, T_k > t_k).$$

These $k$ survival times have marginal survival functions $S_1(t_1), \ldots, S_k(t_k)$. Note that, we assume that the survival times are continuous and take values in $\mathbb{R}^+$. In general, a distribution function is defined by $F(t) = \Pr(T \leq t)$. This motivates the definition we used here for the survival function $S(t) = 1 - F(t) = \Pr(T > t)$, however, we can adopt the following definition $S(t) = \Pr(T \geq t)$.

As a consequence of that

$$\begin{aligned} S_n(t_n) &= \Pr(T_n > t_n) \\ &= \Pr(T_1 \geq 0, \ldots, T_{n-1} \geq 0, T_n > t_n, T_{n+1} \geq 0, \ldots, T_k \geq 0) \\ &= S(0, \ldots, 0, t_n, 0, \ldots, 0). \end{aligned}$$

Unfortunately the relationship between the multivariate survival distribution $F$ and the multivariate survival function $S$ is not a straightforward as in the univariate case because

$$S(t_1, \ldots, t_k) \neq 1 - F(t_1, \ldots, t_k).$$

However, if the survival function $S$ is absolutely continuous, the joint density function can be written as

$$f(t_1, \ldots, t_k) = \frac{\partial^k F(t_1, \ldots, t_k)}{\partial t_1 \ldots \partial t_k} = (-1)^k . \frac{\partial^k S(t_1, \ldots, t_k)}{\partial t_1 \ldots \partial t_k}.$$

By using similar ideas, we can find an expression for the multivariate hazard rate and hazard function as follows

$$\begin{aligned} &\lambda(t_1, \ldots, t_k) \\ &= \lim_{\max \Delta_k \to 0^+} \frac{1}{\Delta_1 \ldots \Delta_k} \\ &\quad \Pr(t_1 \leq T_1 \leq t_1 + \Delta_1, \ldots, t_k \leq T_k \leq t_k + \Delta_k | T_1 \geq t_1, \ldots, T_k \geq t_k) \\ &= \frac{f(t_1, \ldots, t_k)}{S(t_1, \ldots, t_k)} \\ &= \frac{(-1)^k}{S(t_1, \ldots, t_k)} \frac{\partial^k S(t_1, \ldots, t_k)}{\partial t_1 \ldots \partial t_k}, \end{aligned}$$

and

$$\Lambda(t_1, \ldots, t_k) = \int_0^{t_1} \cdots \int_0^{t_k} \lambda(x_1, \ldots, x_k) dx_1 \ldots dx_k.$$

In this case we cannot find a simple expression to link $S$ and $\Lambda$ as before and the survival function cannot be constructed easily. We focus our research on another approach called copula modelling. This is a marginal model. Many different options exist but for reasons explained later our work is based on the so-called Plackett-Dale copula. First some general ideas will be introduced.

## 3.3 Copulas

In this section we will introduce the concept of copulas and we will give some general definitions and known properties that we will use in subsequent chapters. The main topics are discussed here but a full explanation about copulas and related issues can be found in Nelsen (1999).

We will start by giving some ideas about bivariate copulas and we will extend them to the multivariate case with some emphasis on survival copulas.

**Definition 1** *A bivariate copula $C$ is a function from $[0,1] \times [0,1]$ into $[0,1]$ such that:*

1. *For every $u$, $v$ in $[0,1]$, $C(u,0) = C(0,v) = 0$, $C(u,1) = u$ and $C(1,v) = v$.*
2. *For every $u_1 \leq u_2$ and $v_1 \leq v_2$ in $[0,1]$,*

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$$

It immediately follows that a copula is a bivariate distribution function with uniform margins. When the margins are independent we obtain the so-called product copula:

$$C_P(u, v) = uv.$$

For every copula $C$ and every $(u, v)$ in $[0,1] \times [0,1]$ the following version of the Fréchet-Hoeffding bounds (1951) holds;

$$\max(u + v - 1, 0) \equiv W(u, v) \leq C(u, v) \leq M(u, v) \equiv \min(u, v),$$

where $W(u, v)$ and $M(u, v)$ are (degenerate) copulas.

The following theorem, due to Sklar (1959), is one of the main theorems in copulas theory.

**Theorem 1** *If we consider $F$ a bivariate joint distribution function with margins $F_1$ and $F_2$, there exists a copula $C$ such that for all $x_1, x_2$ in $[-\infty, +\infty]$:*

$$F(x_1, x_2) = C\{F_1(x_1), F_2(x_2)\}. \tag{3.1}$$

*In addition, $C$ is unique if $F_1(x_1)$ and $F_2(x_2)$ are continuous, otherwise $C$ is uniquely determined on $Ran(F_1) \times Ran(F_2)$, where $Ran(F_i)$ denotes the ranges of $F_i$. Conversely, given a copula $C$ and $F_1$ and $F_2$ univariate distribution functions, $F$ defined by (3.1) is a joint distribution with margins $F_1$ and $F_2$.*

Theorem 1 can be translated in terms of random variables and their respective distribution functions and we will do so in the next two theorems.

**Theorem 2** *Let $X_1$ and $X_2$ be random variables with $F$, $F_1$ and $F_2$ the joint distribution function and the marginals respectively. There exists a copula function $C_{X_1X_2}$ such that*

$$F(x_1, x_2) = C_{X_1X_2}\{F_1(x_1), F_2(x_2)\}.$$

*If $F_1$ and $F_2$ are continuous, then $C_{X_1X_2}$ is unique, otherwise $C_{X_1X_2}$ is uniquely determined on $Ran(F_1) \times Ran(F_2)$, as before.*

We can interpret a copula as a function that establishes a particular dependence structure on two given random margins. The following theorem formalizes this fact.

**Theorem 3** *Let $X_1$ and $X_2$ be continuous random variables with margins $F_1$ and $F_2$; respectively. The variables $X_1$ and $X_2$ are independent if and only if the corresponding copula $C_{X_1X_2}$ equals the product copula $C_{X_1X_2} \equiv C_P$.*

Before giving results for survival analysis we will extend these concepts to the multivariate situation. A multivariate copula is a continuous multivariate distribution function with uniform margins on the unit interval. Sklar (1959) demonstrated that any joint distribution $F(y_1, \ldots, y_n)$ with marginals $F_1(y_1), \ldots, F_n(y_n)$ could be written as $F(y_1, \ldots, y_n) = C[F_1(y_1), \ldots, F_n(y_n)]$, where $C$ is a $n-$dimensional copula, according to the following definition.

**Definition 2** *An $n$-copula is a function $C$ from $[0,1]^n$ into $[1,0]$, satisfying the conditions*

1. $C(1,\ldots,1,x_i,1,\ldots,1) = x_i$, *for* $i = 1,\ldots,n$ *and* $x_i \in I$.
2. $C(x_1,\ldots,x_n) = 0$ *if* $x_i = 0$ *for any* $i$.
3. $C$ *is n-increasing (in other words the* $C$*-volume of any n-dimensional interval is non-negative).*

From this last definition, it is clear that a copula in fact is a distribution function with uniform marginals. Therefore, a natural extension of the bivariate copula to a multivariate one is just by considering $F_1(x_1),\ldots,F_p(x_p)$. Then the function

$$C\{F_1(x_1),\ldots,F_p(x_p)\} = F(x_1,\ldots,x_p)$$

defines a multivariate distribution function evaluated at $x_1,\ldots,x_p$ with marginal distributions $F_1,\ldots,F_p$. When the $x_i$'s are continuous, $C$ is unique. In case of discrete data, the construction is only uniquely determined on the range of the margins. Reversely, any copula function with given margins generates a multivariate distribution having these margins. Other parametric families of copulas have been defined, including one or more parameters, to express a wide range of positive or negative dependence.

In the next section we will introduce some concepts related to survival analysis.

## 3.4 Survival Copulas

In this section we will extend the notions of copulas, as introduced before, to survival analysis.

**Definition 3** *If* $C_S$ *is a copula we can define a multivariate survival function* $S$ *as follows*

$$S(t_1,\ldots,t_k) = C_S(S_1(t_1),\ldots,S_k(t_k))$$

*where* $S_1,\ldots,S_k$ *are the marginal survival functions.*

From Nelsen (1999) it can be seen that $C_S$ couples the joint survival function and its univariate margins, analogous to the way of ordinary copula connects a joint distribution to its margins.

For survival distributions there is a version of the Sklar (1959) theorem where the proof is similar to the one for distribution functions.

**Theorem 4 (Sklar's canonical representation)** *Let $S$ be a $k$-dimensional survival function with margins $S_1, \ldots, S_k$. Then $S$ has a copula representation:*

$$S(t_1, \ldots, t_k) = C_S(S_1(t_1), \ldots, S_k(t_k)).$$

*If the margins are continuous, the copula $C_S$ it is unique, otherwise is uniquely determined on $Ran(S_1) \times \cdots \times Ran(S_k)$.*

This theorem is proved for the case $k = 2$ in Nelsen (1999) via distribution functions. In addition if $C$ is a copula function of $(T_1, T_2)$, then we write

$$\begin{aligned} S(t_1, t_2) &= \Pr(T_1 > t_1, T_2 > t_2) \\ &= 1 - F_1(t_1) - F_2(t_2) + F(t_1, t_2) \\ &= S_1(t_1) + S_2(t_2) - 1 + C(1 - S_1(t_1), 1 - S_2(t_2)) \\ &= C_S(S_1(t_1), S_2(t_2)), \end{aligned}$$

where $C_S(u_1, u_2) = u_1 + u_2 - 1 + C(1 - u_1, 1 - u_2)$.
We can easily verify that $C_S$ is a copula function:

1. The margins of $C_S$ are uniform:

$$\begin{aligned} C_S(u_1, 1) &= u_1 + C(1 - u_1, 0) = u_1 \\ C_S(1, u_2) &= u_2. \end{aligned}$$

2. $C_S$ verifies $C_S(u, 0) = C_S(0, u) = 0$

$$\begin{aligned} C_S(u, 0) &= C_S(0, u) \\ &= u - 1 + C(1, 1 - u) \\ &= u - 1 + 1 - u \\ &= 0. \end{aligned}$$

3. $C_S$ is an increasing function. If $(u_1, u_2) \in [0, 1] \times [0, 1]$ and $(v_1, v_2) \in [0, 1] \times [0, 1]$, such that $0 \leq u_1 \leq v_1 \leq 1$ and $0 \leq u_2 \leq v_2 \leq 1$. We define

$$H_{C_S}([u_1, v_1] \times [u_2, v_2]) = C_S(v_1, v_2) - C_S(v_1, u_2) - C_S(u_1, v_2) + C_S(u_1, u_2) \geq 0.$$

We can then translate this last equation in terms of the copulas function $C$ as follows

$$\begin{aligned} H_{C_S}([u_1, v_1] \times [u_2, v_2]) &= C(1 - v_1, 1 - v_2) - C(1 - v_1, 1 - u_2) \\ &\quad -C(1 - u_1, 1 - v_2) + C(1 - u_1, 1 - u_2). \end{aligned}$$

Using $\tilde{u}_i = 1 - u_i$ and $\tilde{v}_i = 1 - v_i$ we have

$$H_{C_S}([u_1, v_1] \times [u_2, v_2]) = H_C([\tilde{u}_1, \tilde{v}_1] \times [\tilde{u}_2, \tilde{v}_2]),$$

because $C$ is increasing, $0 \leq \tilde{v}_1 \leq \tilde{u}_1 \leq 1$ and $0 \leq \tilde{v}_2 \leq \tilde{u}_2 \leq 1$.

The results presented for the bivariate case can be extended to the general case but they will not be used in this work.

## 3.5 Examples of Copula Families

The definitions given in previous section can be used to obtain different copula families. In this section we present some examples of copula functions. We will start by reviewing some bivariate copulas with only one parameter and we will give some hints for the extension to higher order in case that this is possible.

From a long list of different type of copulas we will present here four examples of families, Frank (1979), Clayton (1978), Mardia (1970) and Plackett (1965). Of course many other copulas families exist and details can be found in other texts as Nelsen (1999). The expressions for these selected copulas are presented next together with some limit properties.

- Frank

$$C_\theta(u, v) = -\frac{1}{\theta} \ln\left[1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right]$$

with

$\theta \in I\!R \setminus \{0\}$
$C_0(u, v) = uv$
$C_{-\infty} = \max(u + v - 1, 0)$
$C_{+\infty} = \min(u, v)$.

- Clayton

$$C_\theta(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{1/\theta}$$

with

$\theta \in [-1, +\infty)$ and $\theta \neq 0$

$C_0(u, v) = uv$

$C_{-1} = \max(u + v - 1, 0)$

$C_{+\infty} = \min(u, v)$.

- Mardia

$$C_\theta(u, v) = \frac{\theta^2(1+\theta)}{2} \min(u, v) + (1-\theta)^2 uv + \frac{\theta^2(1-\theta)}{2} \max(u + v - 1, 0)$$

with

$\theta \in [-1, 1]$

$C_0(u, v) = uv$

$C_{-1} = \max(u + v - 1, 0)$

$C_{+1} = \min(u, v)$.

- Plackett

$$C_\theta(u, v) = \frac{[1 + (\theta - 1)(u + v)] - \sqrt{[1 + (\theta - 1)(u + v)]^2 - 4uv\theta(\theta - 1)}}{2(\theta - 1)}$$

with

$\theta \in (0, +\infty)$ and $\theta \neq 1$

$C_1(u, v) = uv$

$C_0 = \max(u + v - 1, 0)$

$C_{+\infty} = \min(u, v)$.

It can be seen from these last expressions that in all the cases the copulas enjoy the property of symmetry $C_\theta(u, v) = C_\theta(v, u)$, moreover the parameter $\theta$ models the association between the two margins. Concepts of positive and negative dependence can be defined based on copula functions. We can say that positive dependence is induced by a copula function if for all $u$ and $v$ in the unit interval $C_\theta(u, v) \geq uv$. In analogous way the concept of negative dependence is such as $C_\theta(u, v) \leq uv$ for all $u, v \in [0, 1]$.

In the families of Frank, Clayton and Plackett the parameter $\theta$ measures the strength of dependence between both margins. In addition to that Frank and Clayton

belong to the class of Archimedean copulas. This means that these functions can be generated by choosing appropriate $\phi$ functions such as

$$C_\theta(u, v) = \phi^{-1}[\phi(u) + \phi(v)]$$

with $\phi$ an strictly decreasing continuous function going from the unit interval to $\mathbb{R}_{\geq 0}$, and such that $\phi(1) = 0$ and $\phi(0) = +\infty$. Unfortunately Plackett copula does not belong to this family.

An extension of this last result to the $n$-variate case by using a $C^n$ function from $[0, 1] \times \cdots \times [0, 1]$ to $[0, 1]$ can be found in Kimberling (1974).

We will now turn our attention to the Plackett copula, in the next section we will present some results for this particular case.

## 3.6 Plackett Copula

In order to understand the characteristics and the properties of the Plackett copula we will start by introducing the bivariate Plackett distribution. Let $\boldsymbol{X} = (X_1, X_2)^T$ be a bivariate random variable. Suppose $F(x_1, x_2)$ is its joint distribution function with marginal distributions $F_j(x_j)$, $(j = 1, 2)$. The *global cross-ratio* at $(x_1, x_2)$ given by

$$\theta_{12}(x_1, x_2) = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{F(1 - F_1 - F_2 + F)}{(F_1 - F)(F_2 - F)}. \tag{3.2}$$

where $F_1 \equiv F_1(x_1)$ and $F_2 \equiv F_2(x_2)$ are the marginal cumulative density functions and $F \equiv F(x_1, x_2)$. The quantities $p_{ij}$ $(i, j = 1, 2)$ are the quadrant probabilities in $\mathbb{R}^2$ with vertex at $(x_1, x_2)$ as it can be seen from Figure 3.1. For a constant cross-ratio, $\theta_{12}(x_1, x_2) \equiv \theta$, the Plackett distribution is obtained (Plackett 1965, Mardia 1970). The values are found as one of the two solutions of the following second degree polynomial equation if the marginal distribution functions $F_1$ and $F_2$, and the cross-ratio $\theta_{12}$ are known:

$$\theta_{12}(F - F_1)(F - F_2) - F[F - (F_1 + F_2 - 1)] = 0. \tag{3.3}$$

Dale (1986) and Mardia (1970) gave an explicit solution for (3.3), $F(x_1, x_2)$ as

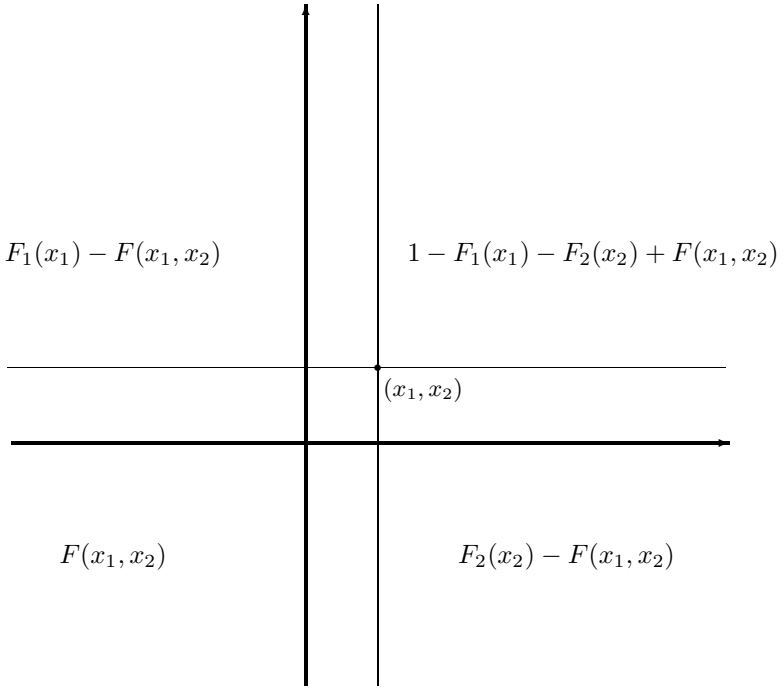


Figure 3.1: *Construction of Plackett's distribution.*

follows, for $\theta_{12}$ in $[0, +\infty]$, where:

$$F(x_1, x_2) = \begin{cases} \dfrac{1 + (F_2(x_2) + F_1(x_1))(\theta_{12} - 1) - H(x_2, x_1)}{2(\theta_{12} - 1)} & \text{if } \theta_{12} \neq 1, \\ F_2(x_2)F_1(x_1) & \text{if } \theta_{12} = 1, \end{cases} \quad (3.4)$$

with

$$H(x_1, x_2) = \sqrt{[1 + (\theta_{12} - 1)(F_1(x_1) + F_2(x_2))]^2 + 4\theta_{12}(1 - \theta_{12})F_1(x_1)F_2(x_2)}. \quad (3.5)$$

Mardia (1970) showed that $F(x_1, x_2)$ is always a bivariate copula. Here, $\theta_{12} = \theta_{12}(x_1, x_2)$ satisfies $0 \leq \theta_{12} \leq \infty$ when $F(x_1, x_2)$ satisfies the Fréchet-Hoeffding (1951) bounds.

$$\max(F_1 + F_2 - 1, 0) \leq F \leq \min(F_1, F_2).$$

At this point we can make some remarks, first notice that the lower bound of the last inequality is attained if $\theta_{12} = 0$. The upper bound is reached if $\theta_{12} = +\infty$. In

addition to that we observe that if $\theta_{12} \to 1$ then $F \to F_1 F_2$. Other consequence of the Fréchet bounds is that all four quadrant probabilities are all nonnegative.

In Mardia (1970) a detailed discussion on the so-called *contingency type distribution* can be found. One of the most interesting features is that most properties are independent of the choice for the distribution function. A thorough discussion can be found also in Schweizer and Sklar (1983).

Theorem 1, due to Sklar, connects the Plackett distribution to the Plackett copula. This fact implies that the Plackett distribution defined in (3.4) can be expressed as a one parameter family of bivariate copulas, $C_\theta$, with $\theta$ in $[0, +\infty]$:

$$C_\theta(x,y) = \begin{cases} \dfrac{1+(x+y)(\theta-1)-H_\theta(x,y)}{2(\theta-1)} & \text{if } \theta \neq 1, \\ xy & \text{if } \theta = 1, \end{cases} \tag{3.6}$$

where

$$H_\theta(x,y) = \sqrt{[1+(\theta-1)(x+y)]^2 + 4\theta(1-\theta)xy}. \tag{3.7}$$

Note that here $\theta = 0$ and $\theta = +\infty$ corresponds to the copulas $W_2(x,y)$ and $M_2(x,y)$, respectively, that are the so-called the Fréchet bounds.

It is known (Schweizer and Sklar, 1983) that for every copula $C$ the following inequalities hold

$$W_n(x_1,\ldots,x_n) \leq C(x_1,\ldots,x_n) \leq M_n(x_1,\ldots,x_n)$$

with $x_i \in I$ for all $i$, and

$$\begin{aligned} W_n(x_1,\ldots,x_n) &= \max\left(\sum_{i=1}^{n} x_i - n + 1, 0\right) \\ M_n(x_1,\ldots,x_n) &= \min_{i=1}^{n} x_i. \end{aligned} \tag{3.8}$$

where $M_n$ is a copula in any dimension and $W_n$ is a copula only for $n = 2$. The case $\theta = 1$ corresponds to the independence copula $C_P(x,y) = xy$. From Mardia (1970) it follows that $C_\theta$ is a 2-copula for every $\theta \in [0, +\infty]$. In an alternative way the bivariate Plackett copula can be seen as the only root of

$$\theta(C-a_1)(C-a_2) - (C-b_1)(C-b_2) = 0,$$

with

$$\begin{aligned} a_1 &= x, & b_1 &= 0, \\ a_2 &= y, & b_2 &= x+y-1, \end{aligned}$$

that satisfies

$$W_2(x, y) \leq C(x, y) \leq M_2(x, y).$$

This will be the basis for the extension of the case of $n$ variables, i.e., the construction of a Plackett $n$-copula. To define the $n$-dimensional Plackett distribution, choose a set of $2^n - 1$ generalized cross-ratios taking positive values

$$\begin{array}{ll} \theta_{ij}, & (1 \leq i < j \leq n), \\ \vdots & \\ \theta_{i_1 \ldots i_k}, & (1 \leq i_1 < \cdots < i_k \leq n), \\ \vdots & \\ \theta_{1 \ldots n}. & \end{array} \tag{3.9}$$

The generalized cross-ratios will be linked to a distribution in the same way we did it for the bivariate case. To create a unified description, we will introduce the odds of the marginals by

$$\theta_i = \frac{F_i}{1 - F_i}, \tag{3.10}$$

with $1 \leq i \leq n$. The bivariate associations are defined as in (3.2):

$$\theta_{ij} = \frac{F_{ij}(1 - F_i - F_j + F_{ij})}{(F_i - F_{ij})(F_j - F_{ij})}, \tag{3.11}$$

with $(1 \leq i < j \leq n)$. The joint distribution $F_{ij}$ can be calculated when $\theta_i$, $\theta_j$ and $\theta_{ij}$ are known. Note that the cross-ratios $\theta_{ij}$ can be written as the odds ratio of $\theta_i$ and $\theta_j$ defined by (3.10). The three-dimensional cross-ratios can be defined in a similar way by means of the two-dimensional cross-ratios and the distribution functions $F$'s.

$$\theta_{123} = \frac{p_{111}p_{122}p_{212}p_{221}}{p_{112}p_{121}p_{211}p_{222}} \tag{3.12}$$

where the orthant probabilities in the three dimensional case are defined as follows:

$$\begin{aligned}
p_{111} &= F_{123}, \\
p_{112} &= F_{12} - F_{123}, \\
p_{121} &= F_{13} - F_{123}, \\
p_{211} &= F_{23} - F_{123}, \\
p_{122} &= F_1 - F_{12} - F_{13} + F_{123}, \\
p_{212} &= F_2 - F_{12} - F_{23} + F_{123}, \\
p_{221} &= F_1 - F_{12} - F_{13} + F_{123}, \\
p_{222} &= 1 - F_1 - F_2 - F_3 + F_{12} + F_{13} + F_{23} - F_{123}.
\end{aligned} \tag{3.13}$$

This is only an example but a systematic derivation of this result can be found in Molenberghs (1992) and a $n$-way Plackett distribution can be constructed following these lines. There also some interesting properties of the corresponding Plackett copula were studied.

## 3.7 Dependence Measures and Related Concepts

In this section we will discuss some general ideas about dependence in particular in the field of survival analysis. The dependence between random variables is characterized entirely by the copula of the corresponding multivariate distribution as it was noted by Deheuvels (1978) and Schweizer and Wolff (1981). However, the direct comparison between survival copulas may not be obvious. An interesting issue could be to use a dependence measure, i.e., a single value to relate different survival functions. Correlation measures are standard, but for correlated survival times this concept is not always useful. Copulas then bring a natural way to study and measure dependence between random variables as we will show next.

The traditional way of evaluating dependence in a bivariate distribution is by means of the Pearson correlation coefficient. It is defined by (Hougaard, 2000)

$$\rho(T_1, T_2) = \frac{\text{cov}(T_1, T_2)}{\sqrt{\text{var}(T_1)\text{var}(T_2)}},$$

with

$$\text{cov}(T_1, T_2) = \int_0^\infty \int_0^\infty (S(t_1, t_2) - S_1(t_1)S_2(t_2))dt_1 dt_2,$$

and

$$\text{var}(T_i) = 2\int_0^\infty tS_i(t)dt - \left[\int_0^\infty S_i(t)dt\right]^2, \quad i = 1, 2.$$

The Pearson correlation is an appropriate measure of dependence when the random variables jointly have a multivariate normal distribution. Moreover, the standard correlation approach of dependency remains natural and unproblematic in the class of elliptical distributions as noted by Embrechts *et al.* (1999). When the distribution is not elliptical, the use of Pearson correlation may be problematic, as is the case in general in survival analysis (Lindeboom and Van Den Berg, 1994).

## 3.8 Dependence Measures

It is well known that the correlation is a relevant measure of dependence in a few special cases. More appropriate are measures of concordance (Nelsen (1999), p. 136). This is the case for Kendall's $\tau$ and Spearman's $\rho$. We will give here a summary of the main properties and definitions. A detailed explanation can be found in Chapter 5 of Nelsen (1999).

Kendall's $\tau$ can be seen as the difference between the probability of concordance and the probability of discordance of two realizations of $(T_1, T_2)$. This coefficient lies in the $[-1, 1]$ interval and a zero value implies independence between $T_1$ and $T_2$. There exists a relationship between Kendall's $\tau$ and $\theta$ for any copula $C(t_1, t_2, \theta)$ (Genest and MacKay, 1986):

$$\tau(\theta) = 4\int_0^1\int_0^1 C_{T_1T_2}(t_1, t_2, \theta)C_{T_1T_2}(dt_1, dt_2, \theta) - 1. \tag{3.14}$$

The marginal distributions of $T_1$ and $T_2$ do not affect (3.14), and hence it follows that $\tau$ only depends on the copula function $C_{T_1T_2}$ (Schweizer and Wolff, 1981). Such a relationship is very simple for the Clayton and Hougaard copulas (Burzykowski *et al.,* 2001). Precisely, one obtains $\tau = (\theta-1)/(\theta+1)$ for Clayton and $\tau = 1-\theta$ for Hougaard. Estimates and confidence intervals (using the delta method) are accordingly easily obtained.

Spearman's $\rho$ is a particular interesting measure, which enjoys the properties of being independent of marginal transformations and of being a non-parametric measure. The Spearman's $\rho$ is also based on concordance and discordance, independent of the marginal distributions as we said, and belongs to the interval $[-1, 1]$. The

relationship between Spearman's $\rho$ and the copula function is

$$\rho(\theta) = 12 \int_0^1 \int_0^1 C_{T_1 T_2}(t_1, t_2, \theta) dt_1 dt_2 - 3. \tag{3.15}$$

The following theorem gives the link between the expression of $\tau$ and $\rho$ for a copula and the corresponding survival copula.

**Theorem 5** *Kendall's tau and Spearman's rho of a survival copula $C_S$ are equal to the Kendall's tau and Spearman's rho of the associated copula $C$.*

**Proof** *Spearman's coefficient $\rho$ can be written as*

$$\begin{aligned}
\rho(C_S) &= 12 \int \int_{[0,1]^2} C_S(t_1, t_2) dt_1 dt_2 - 3 \\
&= 12[t_1^2 t_2 + t_1 t_2^2 - t_1 t_2]_0^1 + 12 \int \int_{[0,1]^2} C(1 - t_1, 1 - t_2) dt_1 dt_2 - 3 \\
&= \rho(C)
\end{aligned}$$

*and $\tau$ can be written as*

$$\begin{aligned}
\tau(C_S) &= 1 - 4 \int \int_{[0,1]^2} \frac{\partial C(t_1, t_2)}{\partial t_1} \frac{\partial C(t_1, t_2)}{\partial t_2} dt_1 dt_2 \\
&= 1 - 4 \int \int_{[0,1]^2} \left[1 - \frac{\partial C(1 - t_1, t_2)}{\partial t_1}\right] \left[1 - \frac{\partial C(1 - t_1, t_2)}{\partial t_2}\right] dt_1 dt_2 \\
&= \tau(C),
\end{aligned}$$

*therefore the coefficients corresponding to both copulas are exactly the same.*

Different copula models have different parameters and therefore the results cannot be compared in a direct way. However, the two measures of concordance proposed here can easily be obtained for every copula model, sometimes with a closed formula, sometimes using integration by means of expressions (3.14) and (3.15); making straightforward the comparison between two different copula models.

## 3.9 Statistical Inference

In this section, we discuss some methods to estimate multivariate survival functions. We do not provide an exhaustive review of all the methods. For example, we do not consider non-parametric estimation methods or counting process approaches. Moreover, the methodology will be presented for survival-times, but the results are valid for other type of time-to-event variables as well.

### 3.9.1 Maximum Likelihood Method

We consider the estimation problem for a vector of parameters $\boldsymbol{\phi}$ of the survival function $S$ with a copula structure, we have then,

$$S(t_1, t_2, \boldsymbol{\phi}) = C_S(S_1(t_1, \boldsymbol{\phi}^1), S_2(t_2, \boldsymbol{\phi}^2), \boldsymbol{\phi}^{12}),$$

with $\boldsymbol{\phi} = (\boldsymbol{\phi}^1, \boldsymbol{\phi}^2, \boldsymbol{\phi}^{12})$. In this case, $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$ are specific parameters of the univariate functions, whereas $\boldsymbol{\phi}^{12}$ is the parameter (in $\mathbb{R}^p$) of the survival copula function.

Let us assume that we have a sample of bivariate survival times $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n$ where $\boldsymbol{t}_i = (t_{1i}, t_{2i})$ with $i = 1, \ldots, n$.

Hence, the log-likelihood is

$$\ell(\boldsymbol{t}, \boldsymbol{\phi}) = \sum_{i=1}^{n} \ln f(\boldsymbol{t}_i, \boldsymbol{\phi}).$$

The ML estimate (MLE) corresponds then to

$$\widehat{\boldsymbol{\phi}}_{ML} = \arg\max_{\boldsymbol{\phi} \in \Phi} \ell(\boldsymbol{t}, \boldsymbol{\phi}),$$

with $\Phi$ the corresponding parameter space. However, dealing with survival times is often subject to additional complexity, because records on survival times are often incomplete. Censoring and truncation are often present in survival data and this has to be taken into account. We will consider a general situation with $T$ the survival time, $C^-$ the left censoring time, $C^+$ the right censoring time and $D$ the observed time.

We observe the triplet $(D, \Delta^-, \Delta^+)$ with

$$D = C^- I_{[T \leq C^-]} + T I_{[C^- < T \leq C^+]} + C^+ I_{[C^+ < T]}$$

and $(\Delta^-, \Delta^+) = (I_{[T \leq C^-]}, I_{[T > C^+]})$.

The three possible cases can be summarized as follows,

| Case | Observed | $\Delta^-$ | $\Delta^+$ |
|---|---|---|---|
| $C^- < T \leq C^+$ | $T$ | 0 | 0 |
| $T \leq C^-$ | $C^-$ | 1 | 0 |
| $C^+ < T$ | $C^+$ | 0 | 1 |

Moreover, we could consider a left truncation variable $W$. $D$ is observed only if $T > W$, but this is out of the scope of this work. In what follows, bivariate survival data correspond to a sample of the form

$$\boldsymbol{y} = \{\boldsymbol{y}_i = (d_{1i}, d_{2i}, \Delta_{1i}^-, \Delta_{2i}^-, \Delta_{1i}^+, \Delta_{2i}^+),\, i = 1, \ldots, n\}.$$

We will briefly explain how to estimate the parameters of this model. To do that we will assume that censoring times are independent of the survival times and that they are not informative, i.e., the censoring at time $t$ does not depend on the observed process up to time $t$. In process-dependent censoring an individual can be censored based on his history.

First we will calculate all the different contributions to the likelihood. We will tackle the case of censoring but extensions for left truncation also exist.

In the bivariate case, we observe $(D_1, D_2, \Delta_1^-, \Delta_2^-, \Delta_1^+, \Delta_2^+)$ and we further assume that survival times are independent of the censoring times. Let us suppose that we have a function $C_S$ with density $c_S$ and that $g_1^-$, $g_1^+$, $g_2^-$, $g_2^+$ are the density functions associated to $C_1^-$, $C_1^+$, $C_2^-$ and $C_2^+$, respectively. To simplify notation we will not write the parameter $\boldsymbol{\phi}$ in the survival functions.

We can distinguish the following cases of contributions to the likelihood:

1. $T_1$ and $T_2$ not censored. Then, $(\Delta_1^-, \Delta_2^-, \Delta_1^+, \Delta_2^+) = (0, 0, 0, 0)$ and

$$\begin{aligned}\Pr\{D_1 \le d_1, D_2 \le d_2\} &= \Pr\{T_1 \le d_1, T_2 \le d_2\} \\ &= 1 - S_1(d_1) - S_2(d_2) + C_S(S_1(d_1), S_2(d_2)).\end{aligned}$$

   Hence the contribution to the likelihood can be written as

$$c_S(S_1(d_1), S_2(d_2)) f_1(d_1) f_2(d_2).$$

2. $T_1$ is right censored and $T_2$ is not censored. Then, $(\Delta_1^-, \Delta_2^-, \Delta_1^+, \Delta_2^+) = (0, 0, 1, 0)$ and

$$\begin{aligned}\Pr\{D_1 \le d_1, D_2 \le d_2\} &= \Pr\{C_1^+ \le d_1, T_2 \le d_2, T_1 > C_1^+\} \\ &= \int\int\int I_{[c \le d_1, t_2 \le d_2, t_1 > c]} f(t_1, t_2) g_1^+(c) dt_1 dt_2 dc \\ &= \int_0^{d_1} \left[ \int_c^{\infty} \int_0^{d_2} f(t_1, t_2) dt_1 dt_2 \right] g_1^+(c) dc \\ &= \int_0^{d_1} \left[ S_1(c) - C_S(S_1(c), S_2(d_2)) \right] g_1^+(c) dc\end{aligned}$$

and the contribution to the likelihood can be written as

$$\partial_2 C_S(S_1(d_1), S_2(d_2)) f_2(d_2) g_1^+(d_1).$$

3. $T_2$ is right censored and $T_1$ is not censored, then $(\Delta_1^-, \Delta_2^-, \Delta_1^+, \Delta_2^+) = (0, 0, 0, 1)$ and the contribution to the likelihood can be written as

$$\partial_1 C_S(S_1(d_1), S_2(d_2)) f_1(d_1) g_2^+(d_2).$$

4. $T_1$ is left censored and $T_2$ is not censored. Then, $(\Delta_1^-, \Delta_2^-, \Delta_1^+, \Delta_2^+) = (1, 0, 0, 0)$,

$$\begin{aligned}
\Pr\{D_1 \le d_1, D_2 \le d_2\} &= \Pr\{C_1^- \le d_1, T_2 \le d_2, T_1 > C_1^-\} \\
&= \int\int\int I_{[c \le d_1, t_2 \le d_2, t_1 \le c]} f(t_1, t_2) g_1^-(c) dt_1 dt_2 dc \\
&= \int_0^{d_1} \left[ \int_0^c \int_0^{d_2} f(t_1, t_2) dt_1 dt_2 \right] g_1^-(c) dc \\
&= \int_0^{d_1} \left[ 1 - S_1(c) - S_2(d_2) + C_S(S_1(c), S_2(d_2)) \right] g_1^+(c) dc
\end{aligned}$$

and the contribution to the likelihood can be written as

$$\left(1 - \partial_2 C_S(S_1(d_1), S_2(d_2))\right) f_2(d_2) g_1^-(d_1).$$

5. $T_2$ is left censored and $T_1$ not censored, then $(\Delta_1^-, \Delta_2^-, \Delta_1^+, \Delta_2^+) = (0, 1, 0, 0)$ and, symmetrically we can write the contribution to the likelihood as follows

$$\left(1 - \partial_1 C_S(S_1(d_1), S_2(d_2))\right) f_1(d_1) g_2^-(d_2).$$

6. $T_1$ and $T_2$ are right censored. Then, $(\Delta_1^-, \Delta_2^-, \Delta_1^+, \Delta_2^+) = (0, 0, 1, 1)$, and we have

$$\begin{aligned}
&\Pr\{D_1 \le d_1, D_2 \le d_2\} \\
&= \Pr\{C_1^+ \le d_1, C_2^+ \le d_2, T_1 > C_1^+, T_2 > C_2^+\} \\
&= \int\int\int\int I_{[c_1 \le d_1, c_2 \le d_2, t_1 > c_1, t_2 > c_2]} f(t_1, t_2) g_1^+(c_1) g_2^+ dt_1 dt_2 dc_1 dc_2 \\
&= \int_0^{d_1} \int_0^{d_2} \left[ \int_{c_1}^\infty \int_{c_2}^\infty f(t_1, t_2) dt_1 dt_2 \right] g_1^+(c_1) g_2^+(c_2) dc_1 dc_2
\end{aligned}$$

and the contribution to the likelihood can be written as

$$C_S(S_1(d_1), S_2(d_2))g_1^+(d_1)g_2^+(d_2).$$

7. $T_1$ and $T_2$ are left censored. Then, $(\Delta_1^-, \Delta_2^-, \Delta_1^+, \Delta_2^+) = (1, 1, 0, 0)$, and doing similar calculations to the right censored case we obtain the contribution to the likelihood as

$$\left(1 - S_1(d_1) - S_2(d_2) + C_S(S_1(d_1), S_2(d_2))\right) g_1^-(d_1)g_2^-(d_2).$$

8. $T_1$ right censored and $T_2$ left censored. Then, $(\Delta_1^-, \Delta_2^-, \Delta_1^+, \Delta_2^+) = (0, 1, 1, 0)$, and we have the contribution to the likelihood written as

$$\left(S_1(d_1) - C_S(S_1(d_1), S_2(d_2))\right) g_1^+(d_1)g_2^-(d_2).$$

9. $T_1$ left censored and $T_2$ right censored is symmetric to the previous case, with $(\Delta_1^-, \Delta_2^-, \Delta_1^+, \Delta_2^+) = (1, 0, 0, 1)$. Then the contribution to the likelihood can be written as

$$\left(S_2(d_2) - C_S(S_1(d_1), S_2(d_2))\right) g_1^-(d_1)g_2^+(d_2).$$

Now we will integrate all the different cases assuming that the censoring times are independent of the survival times and not informative. Therefore,

$$\begin{aligned}
&\ln f(\boldsymbol{y}_i, \boldsymbol{\phi}) \\
&\propto (1-\Delta_{1i}^-)(1-\Delta_{2i}^-)(1-\Delta_{1i}^+)(1-\Delta_{2i}^+) \ln(C_S(S_1(d_{1i}, \boldsymbol{\phi}^1), S_2(d_{2i}, \boldsymbol{\phi}^2), \boldsymbol{\phi}^{12}) + \\
&(1-\Delta_{1i}^-)(1-\Delta_{1i}^+)(1-\Delta_{2i}^+\Delta_{2i}^-) \ln f_1(d_{1i}, \boldsymbol{\phi}^1) + \\
&(1-\Delta_{2i}^-)(1-\Delta_{2i}^+)(1-\Delta_{1i}^+\Delta_{1i}^-) \ln f_2(d_{2i}, \boldsymbol{\phi}^2) + \\
&\Delta_{1i}^-(1-\Delta_{2i}^-)(1-\Delta_{1i}^+)(1-\Delta_{2i}^+) \ln\left(1 - \frac{\partial C_S(S_1(d_{1i}, \boldsymbol{\phi}^1), S_2(d_{2i}, \boldsymbol{\phi}^2), \boldsymbol{\phi}^{12})}{\partial t_2}\right) + \\
&(1-\Delta_{1i}^-)\Delta_{2i}^-(1-\Delta_{1i}^+)(1-\Delta_{2i}^+) \ln\left(1 - \frac{\partial C_S(S_1(d_{1i}, \boldsymbol{\phi}^1), S_2(d_{2i}, \boldsymbol{\phi}^2), \boldsymbol{\phi}^{12})}{\partial t_1}\right) + \\
&(1-\Delta_{1i}^-)(1-\Delta_{2i}^-)\Delta_{1i}^+(1-\Delta_{2i}^+) \ln\left(\frac{\partial C_S(S_1(d_{1i}, \boldsymbol{\phi}^1), S_2(d_{2i}, \boldsymbol{\phi}^2), \boldsymbol{\phi}^{12})}{\partial t_2}\right) + \\
&(1-\Delta_{1i}^-)(1-\Delta_{2i}^-)(1-\Delta_{1i}^+)\Delta_{2i}^+ \ln\left(\frac{\partial C_S(S_1(d_{1i}, \boldsymbol{\phi}^1), S_2(d_{2i}, \boldsymbol{\phi}^2), \boldsymbol{\phi}^{12})}{\partial t_1}\right) +
\end{aligned}$$

$$\Delta_{1i}^{-}\Delta_{2i}^{-}(1-\Delta_{1i}^{+})(1-\Delta_{2i}^{+})$$

$$\ln\left(1-S_1(d_{1i},\boldsymbol{\phi}^1)-S_2(d_{2i},\boldsymbol{\phi}^2)+C_S(S_1(d_{1i},\boldsymbol{\phi}^1),S_2(d_{2i},\boldsymbol{\phi}^2),\boldsymbol{\phi}^{12})\right)+$$

$$\Delta_{1i}^{-}(1-\Delta_{2i}^{-})(1-\Delta_{1i}^{+})\Delta_{2i}^{+}\ln\left(S_2(d_{2i},\boldsymbol{\phi}^2)-C_S(S_1(d_{1i},\boldsymbol{\phi}^1),S_2(d_{2i},\boldsymbol{\phi}^2),\boldsymbol{\phi}^{12})\right)+$$

$$(1-\Delta_{1i}^{-})\Delta_{2i}^{-}\Delta_{1i}^{+}(1-\Delta_{2i}^{+})\ln\left(S_1(d_{1i},\boldsymbol{\phi}^1)-C_S(S_1(d_{1i},\boldsymbol{\phi}^1),S_2(d_{2i},\boldsymbol{\phi}^2),\boldsymbol{\phi}^{12})\right)+$$

$$(1-\Delta_{1i}^{-})(1-\Delta_{2i}^{-})\Delta_{1i}^{+}\Delta_{2i}^{+}\ln\left(C_S(S_1(d_{1i},\boldsymbol{\phi}^1),S_2(d_{2i},\boldsymbol{\phi}^2),\boldsymbol{\phi}^{12})\right).$$

Hence the log-likelihood can be constructed and the maximum likelihood estimator of the vector parameter $\boldsymbol{\phi}$ obtained as the solution of

$$\frac{\partial}{\partial\boldsymbol{\phi}}\ell(\boldsymbol{y},\boldsymbol{\phi})=0.$$

Let us call $\widehat{\boldsymbol{\phi}}$ the ML estimate of $\boldsymbol{\phi}_0$ the true parameter vector. Therefore, under regularity conditions, $\sqrt{n}(\widehat{\boldsymbol{\phi}}-\boldsymbol{\phi}_0)$ is asymptotically normally distributed with $N(0,I^{-1}(\boldsymbol{\phi}_0))$ and where $I(\boldsymbol{\phi}_0)$ is the called information matrix.

It is important to point out that in this approach, the vector of the parameters is estimated simultaneously. It means that we perform a full likelihood estimation. This procedure can be computationally very expensive, therefore Shih and Louis (1995) proposed the so-called two stage ML method. As its name suggests this method performs the estimations in two steps. In the first step $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$ are estimated by using the log-likelihood of the univariate survival distributions. And in the second step the copula $\boldsymbol{\phi}^{12}$ parameters are estimated, given the values obtained in the first step. There is also another alternative that we will discuss in the next section.

### 3.9.2 Semi-parametric Estimation

This approach is called semi-parametric in the sense that we will use non-parametric estimators of the survival functions $\widehat{S}_i$, while the copula parameters are estimated by maximizing the log-likelihood

$$
\begin{aligned}
&\ln f(\boldsymbol{y}_i, \boldsymbol{\phi}^{12}) \\
&\propto (1-\Delta_{1i}^{-})(1-\Delta_{2i}^{-})(1-\Delta_{1i}^{+})(1-\Delta_{2i}^{+}) \ln(C_S(\widehat{S}_1(d_{1i}), \widehat{S}_2(d_{2i}), \boldsymbol{\phi}^{12}) + \\
&\Delta_{1i}^{-}(1-\Delta_{2i}^{-})(1-\Delta_{1i}^{+})(1-\Delta_{2i}^{+}) \ln\left(1 - \frac{\partial C_S(\widehat{S}_1(d_{1i}), \widehat{S}_2(d_{2i}), \boldsymbol{\phi}^{12})}{\partial t_2}\right) + \\
&(1-\Delta_{1i}^{-})\Delta_{2i}^{-}(1-\Delta_{1i}^{+})(1-\Delta_{2i}^{+}) \ln\left(1 - \frac{\partial C_S(\widehat{S}_1(d_{1i}), \widehat{S}_2(d_{2i}), \boldsymbol{\phi}^{12})}{\partial t_1}\right) + \\
&(1-\Delta_{1i}^{-})(1-\Delta_{2i}^{-})\Delta_{1i}^{+}(1-\Delta_{2i}^{+}) \ln\left(\frac{\partial C_S(\widehat{S}_1(d_{1i}), \widehat{S}_2(d_{2i}), \boldsymbol{\phi}^{12})}{\partial t_2}\right) + \\
&(1-\Delta_{1i}^{-})(1-\Delta_{2i}^{-})(1-\Delta_{1i}^{+})\Delta_{2i}^{+} \ln\left(\frac{\partial C_S(\widehat{S}_1(d_{1i}), \widehat{S}_2(d_{2i}), \boldsymbol{\phi}^{12})}{\partial t_1}\right) + \\
&\Delta_{1i}^{-}\Delta_{2i}^{-}(1-\Delta_{1i}^{+})(1-\Delta_{2i}^{+}) \\
&\ln\left(1 - \widehat{S}_1(d_{1i}) - \widehat{S}_2(d_{2i}) + C_S(\widehat{S}_1(d_{1i}), \widehat{S}_2(d_{2i}), \boldsymbol{\phi}^{12})\right) + \\
&\Delta_{1i}^{-}(1-\Delta_{2i}^{-})(1-\Delta_{1i}^{+})\Delta_{2i}^{+} \ln\left(\widehat{S}_2(d_{2i}) - C_S(\widehat{S}_1(d_{1i}), \widehat{S}_2(d_{2i}), \boldsymbol{\phi}^{12})\right) + \\
&(1-\Delta_{1i}^{-})\Delta_{2i}^{-}\Delta_{1i}^{+}(1-\Delta_{2i}^{+}) \ln\left(\widehat{S}_1(d_{1i}) - C_S(\widehat{S}_1(d_{1i}), \widehat{S}_2(d_{2i}), \boldsymbol{\phi}^{12})\right) + \\
&(1-\Delta_{1i}^{-})(1-\Delta_{2i}^{-})\Delta_{1i}^{+}\Delta_{2i}^{+} \ln\left(C_S(\widehat{S}_1(d_{1i}), \widehat{S}_2(d_{2i}), \boldsymbol{\phi}^{12})\right).
\end{aligned}
$$

Notice that in this last expression the parameters $\boldsymbol{\phi}^1$ and $\boldsymbol{\phi}^2$ vanish because no parametric form is assumed for the marginal survival distributions.

This estimation method has been introduced by Genest *et al.* (1995) and Shih and Louis (1995), who both show that this semi-parametric estimator is consistent and asymptotically normally distributed.

## 3.10 Conclusions

In this chapter, we have introduced a general copula approach to multivariate modelling. The main theorems of the copula theory have been presented and survival copulas have been defined. We showed how copulas can be used in survival analysis. We discussed some measures of concordance for copula models and we stated the basic general properties that link the vector of parameters of a copula model with Kendall's and Spearman's coefficients. In that sense copula models give a powerful

tool to estimate the strength of the association between two random variables, in particular dealing with survival times. However, the methods to estimate the parameters of these models can be very expensive in computational terms. In the last section of this chapter we presented the different expressions for these log-likelihood functions whether a parametric method or semi-parametric estimation is used to estimate the marginal distributions. We mentioned other alternative methods as well.

It is clear, just by looking at the expression of the log-likelihood function for the bivariate case, that the extension to a multivariate case will be a non-trivial task for some specific copula families. In some cases, where for example the copula structure has special characteristics, it can be easier but still poses non trivial numerical complexities. In the next chapter we will introduce ideas of pseudo-likelihood estimation to tackle the issue of the multivariate estimation that at a very low cost in efficiency reduces the amount of numerical problems and gives estimates for all parameters in the copula model at once.

# Chapter 4

# Pseudo-likelihood Estimation: Definitions and Properties

## 4.1 Introduction

Pseudo-likelihood (PL) method is an alternative estimation method when the maximum likelihood method becomes prohibitive for different reasons. We present a formal definition of this technique together with the most relevant properties and asymptotic results based on the work of Aerts *et al.* (2002). The PL method is a non-likelihood method where the principal idea is to replace a numerically complex joint density by a simpler function that is a suitable product of ratios of likelihoods of subsets of the variables. This method becomes important to estimate de parameters of the models we will develop in next chapters. One important feature of this model is that while the method achieves important computational economies by changing the estimation strategy, it does not have any impact on the model interpretation. In other words, model parameters can be chosen in the same way as with the full likelihood and they retain their meaning. But the most attractive characteristic of this method is that it converges quickly with only minor efficiency loses, especially for a range of realistic parameter setting. In next chapters we will apply this methodology to obtain estimates of the parameters of a multivariate survival model.

## 4.2 Pseudo-likelihood Definition

In this section, we will formally introduce the pseudo-likelihood method in a general framework but we will focus on a particular case that we will call pairwise pseudo-likelihood as we will explain in Section 4.4. For convenience we use the definition given by Arnold and Strauss (1991) and Geys (1999) that was adopted by Renard (2002) in the context of multilevel data. We assume the response vector $\boldsymbol{Y}_i$ for subject $i$ with $i = 1, \ldots, N$ to have a constant length $L$, however the extension to variable lengths of $\boldsymbol{Y}_i$ is not problematic and can be easily performed.

To formally introduce pseudo-likelihood let us start by defining $S$ as the set of all $2^L - 1$ vectors of length $L$, consisting solely of zeros and ones, in addition to that we consider vectors having at least one non zero entry. Let us denote by $\boldsymbol{y}_i^{(s)}$ the subvector of $\boldsymbol{y}_i$ corresponding to the components of $s$ that are not zero. The associated joint density is $f_s(\boldsymbol{y}_i^{(s)}, \boldsymbol{\phi})$. We define now a pseudo-likelihood function by choosing a set $\delta = \{\delta_s | s \in S\}$ of real numbers such as there is at least one none zero component.

The log of the pseudo-likelihood function is defined as

$$\ln p\ell = \sum_{i=1}^{N} p\ell_i, \tag{4.1}$$

with

$$p\ell_i = \sum_{s \in S} \delta_s \ln f_s(\boldsymbol{y}_i^{(s)}, \boldsymbol{\phi}).$$

By setting $\delta_s = 1$ if $s$ is the vector consisting solely of ones, and $\delta_s = 0$ otherwise, the classical log-likelihood function is found.

An example where pseudo-likelihood is very relevant was studied by Geys (1999) within the framework of exponential family models. The pseudo-likelihood there is found by replacing the joint density by the product of univariate full conditional densities. Other types of pseudo-likelihood functions were considered in many different situations (Aerts *et al.*, 2002).

Similarly to maximum likelihood estimation the pseudo-likelihood estimator can be obtained by maximizing the pseudo-likelihood expression (4.1). Therefore this value can be calculated by differentiating (4.1) and setting the derivate to zero.

It means,

$$\begin{aligned}\frac{\partial}{\partial\phi_r}\ln p\ell &= \frac{\partial}{\partial\phi_r}\sum_{i=1}^{N}\sum_{s\in S}\delta_s \ln f_s(\boldsymbol{y}_i^{(s)},\boldsymbol{\phi}) \\ &= \sum_{i=1}^{N}\sum_{s\in S}\delta_s \frac{\partial f_s(\boldsymbol{y}_i^{(s)},\boldsymbol{\phi})/\partial\boldsymbol{\phi}_r}{f_s(\boldsymbol{y}_i^{(s)},\boldsymbol{\phi})} \\ &= 0\end{aligned} \tag{4.2}$$

for $r = 1, \ldots, p$.

Adequate regularity conditions have to be assumed to ensure that (4.1) can be maximized, in addition to that the pseudo-likelihood estimator enjoys attractive asymptotic properties that we will show in next sections.

## 4.3 Pseudo-likelihood Estimator Properties

We will start with a series of required assumptions on the density functions $f_s(\boldsymbol{y}_i^{(s)},\boldsymbol{\phi})$.

**(A1)** The parameter space $\Omega$ contains an open region $\omega$ of which the true parameter value $\boldsymbol{\phi}_0$ is an interior point and where $\omega$ is such that for all $s \in S$ and for almost all $y^{(s)}$, the density $f(\boldsymbol{y}_i^{(s)},\boldsymbol{\phi})$ admits third derivates

$$\frac{\partial^3}{\partial\phi_k\partial\phi_l\partial\phi_m} f(\boldsymbol{y}_i^{(s)},\boldsymbol{\phi})$$

for all $\boldsymbol{\phi} \in \omega$.

**(A2)** The first and the second order logarithmic derivates of $f_s$ satisfy the following equations for all $s \in S$

$$E_{\boldsymbol{\phi}}\left[\frac{\partial \ln f_s(\boldsymbol{y}_i^{(s)},\boldsymbol{\phi})}{\partial\phi_r}\right] = 0, \qquad r = 1, \ldots, p, \tag{4.3}$$

and

$$E_{\boldsymbol{\phi}}\left[\frac{\partial \ln f_s(\boldsymbol{y}_i^{(s)},\boldsymbol{\phi})}{\partial\phi_r}\frac{\partial \ln f_s(\boldsymbol{y}_i^{(s)},\boldsymbol{\phi})}{\partial\phi_l}\right] = E_{\boldsymbol{\phi}}\left[\frac{-\partial^2 \ln f_s(\boldsymbol{y}_i^{(s)},\boldsymbol{\phi})}{\partial\phi_r\partial\phi_l}\right] < \infty, \tag{4.4}$$

with $r, l = 1, \ldots, p$.

**(A3)** The matrix $J(\boldsymbol{\phi})$, defined by

$$J_{rl}(\boldsymbol{\phi}) = -\sum_{s\in S} \delta_s E_{\boldsymbol{\phi}} \left[ \frac{\partial^2 \ln f_s(\boldsymbol{y}^{(s)}, \boldsymbol{\phi})}{\partial \phi_r \partial \phi_l} \right], \tag{4.5}$$

is positive definite for all $\boldsymbol{\phi} \in \omega$.

**(A4)** There exist functions $M_{klr}$ such that

$$\left| \sum_{s\in S} \delta_s E_{\boldsymbol{\phi}} \left[ \frac{\partial^3 \ln f_s(\boldsymbol{y}^{(s)}, \boldsymbol{\phi})}{\partial \phi_k \partial \phi_l \partial \phi_r} \right] \right| \leq M_{klr}(\boldsymbol{y}) \qquad \text{for all } \boldsymbol{\phi} \in \omega,$$

with $m_{klr} = E_{\boldsymbol{\phi}}[M_{klr}(\boldsymbol{y})] < \infty$.

The following theorem proven by Arnold and Strauss (1991) ensures the existence of at least one solution to the pseudo-likelihood equations (4.1), which is consistent and asymptotically normal distributed.

**Theorem 6 (Consistency and Asymptotic Normality)**
*Let us assume that a vector* $(\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$ *have a common density that depends on* $\boldsymbol{\phi}$. *Then under the assumptions(A1)–(A4), the pseudo-likelihood estimator* $\widehat{\boldsymbol{\phi}}_N$, *defined as the maximizer of (4.1), has the following properties*

1. *the pseudo-likelihood estimator of* $\boldsymbol{\phi}_0$, $\widehat{\boldsymbol{\phi}}_N$, *converges in probability to the true parameter, in other words,* $\widehat{\boldsymbol{\phi}}_N$ *is consistent for estimating* $\boldsymbol{\phi}_0$.

2. $\sqrt{N}(\widehat{\boldsymbol{\phi}}_N - \boldsymbol{\phi}_0)$ *converges in distribution to*

$$N_q(\mathbf{0}, J(\boldsymbol{\phi}_0)^{-1} K(\boldsymbol{\phi}_0) J(\boldsymbol{\phi}_0)^{-1}) \tag{4.6}$$

*with* $J(\boldsymbol{\phi})$ *defined by (4.5) and* $K(\boldsymbol{\phi})$ *by*

$$K_{rl} = \sum_{s,t\in S} \delta_s \delta_t E_{\boldsymbol{\phi}} \left[ \frac{\partial \ln f_s(\boldsymbol{y}_i^{(s)}, \boldsymbol{\phi})}{\partial \phi_r} \frac{\partial \ln f_t(\boldsymbol{y}_i^{(t)}, \boldsymbol{\phi})}{\partial \phi_l} \right]. \tag{4.7}$$

The proofs are closely related to the classical proofs for maximum likelihood estimator, for more details see Lehmann (1983, p. 429-434).

It is important to point out the connection existing between pseudo-likelihood and estimating equations. Many researchers have used them to model correlated data because they have the advantage of no requiring knowledge about the whole

distribution of the response vector. Therefore they avoid the need of prohibited calculations for the likelihood. Estimating equations replace then the classical score equations and they have the advantage that are much easier to evaluate. An estimator can be defined as a solution of the estimating equation $g(\boldsymbol{y}, \phi) = 0$. Examples in literature are, among others, the approach to model correlated discrete data used by Liang and Zeger (1986).

However, a major advantage of pseudo-likelihood over other estimating equation approach is that we face an optimization problem, i.e., maximizing the pseudo-likelihood function. A further advantage of the PL approach is the close connection of pseudo-likelihood with likelihood, enabling one to construct pseudo-likelihood ratio and pseudo-score test statistics that have easy-to-compute expressions and intuitively appealing distributions (Aerts *et al.* 2002). These tests will be extended in the framework of multivariate survival data as we will show in Chapter 6.

Similar in spirit to generalized estimating equations (Liang and Zeger 1986), this asymptotic normality result provides an easy way to estimate consistently the asymptotic covariance matrix. Indeed, the matrix $J$ is found from evaluating the second derivate of the log $p\ell$ function at the PL estimate. The expectation in $K$ can be replaced by the cross-product of the observed scores. We will refer to $J^{-1}$ as the model based variance estimator, which should not be used as such because it overestimates precision; to $K$ as the empirical correction; and to $J^{-1}KJ^{-1}$ as the empirically corrected variance estimator.

As discussed by Arnold and Strauss (1991), the Cramèr-Rao inequality implies that $J^{-1}KJ^{-1}$ is greater than the inverse of $I$, corresponding to the Fisher information matrix for the maximum likelihood case, in the sense that $J^{-1}KJ^{-1} - I^{-1}$ is positive semidefinite. Therefore, a PL estimator is always less efficient than the corresponding ML estimator. In other words, maximum likelihood estimator will be, in general, more efficient than maximum pseudo-likelihood estimators. Aerts *et al.* (2002) show that in many realistic settings efficiency losses are minor. Sacrificing some efficiency is therefore the price to pay for computational simplicity. The covariance matrix of the pseudo-likelihood estimator $\widehat{\boldsymbol{\phi}}_N$ in practice can be easily estimated (consistently) by

$$\widehat{\Sigma}_N = J_N^{-1} K_N J_N^{-1}, \tag{4.8}$$

with

$$J_N = -\sum_{i=1}^{N}\sum_{s\in S}\delta_s \frac{\partial^2 \ln f_s(\boldsymbol{y}_i^{(s)}, \widehat{\boldsymbol{\phi}}_N)}{\partial\boldsymbol{\phi}\partial\boldsymbol{\phi}^T} \tag{4.9}$$

and

$$K_N = \sum_{i=1}^{N}\sum_{s,t\in S}\delta_s\delta_t \frac{\partial \ln f_s(\boldsymbol{y}_i^{(s)}, \widehat{\boldsymbol{\phi}}_N)}{\partial\boldsymbol{\phi}} \frac{\partial \ln f_t(\boldsymbol{y}_i^{(t)}, \widehat{\boldsymbol{\phi}}_N)}{\partial\boldsymbol{\phi}^T}. \tag{4.10}$$

The expression (4.8) is known as a "sandwich" estimator, similar in sprit to the robust variance estimate of Liang and Zeger (1986). Some properties of this estimator together with examples and some applications are discussed by Royall (1986). In order to estimate $J(\boldsymbol{\phi}_0)$ we can use (4.4), which does not require evaluation of any second order derivate and then we can write

$$J_N = \sum_{i=1}^{N}\sum_{s\in S}\delta_s \frac{\partial \ln f_s(\boldsymbol{y}_i^{(s)}, \widehat{\boldsymbol{\phi}}_N)}{\partial\boldsymbol{\phi}} \frac{\partial \ln f_s(\boldsymbol{y}_i^{(s)}, \widehat{\boldsymbol{\phi}}_N)}{\partial\boldsymbol{\phi}^T} \tag{4.11}$$

## 4.4 Pairwise Pseudo-likelihood

We have introduced the concept of pseudo-likelihood in a very general framework. The principal idea, as mentioned before, is to replace a complicated joint density by a simpler function, for example the product of the conditional distributions. We will restrict this work to a particular form of pseudo-likelihood, that we call pairwise pseudo-likelihood. We will combine this technique with survival multivariate models, and even if this methodology have not been used very much in practical situations until quite recently, the main advantages were showed in the spatial data context (Hjort, 1993), correlated multivariate data (Geys, 1999), multilevel modeling (Renard, 2002), etc.

Other examples of pseudo-likelihood are the works of Le Cessie and Van Houwelingen (1994) who used this technique to fit a model with logistic marginal responses probabilities, using the odds ratio or tetrachoric correlation as a measure of association.

Geys, Molenberghs and Lipsitz (1998) compared pairwise pseudo-likelihood with other approaches in marginally specified odds ratio models with exchangeable association structure and Kuk and Nott (2000) used pairwise pseudo-likelihood in a model with a more general specification for the association structure.

Technically, the pairwise likelihood consists in replacing the likelihood contribution $f(y_{1j}, \ldots, y_{n_j j})$ by the product of all possible pairwise densities. The logarithm of the pairwise likelihood for the response vector $\boldsymbol{y}_j$ can be expressed as

$$p\ell_j(\boldsymbol{\phi}) = \sum_{i=1}^{n_j} \sum_{i'>i} \ln f_{y_k y_l}(y_{i'j}, y_{ij}, \boldsymbol{\phi}).$$

Notice first that the pairwise pseudo-likelihood and the classical likelihood functions are identical if all clusters have size 2. Second, the marginal pairwise densities are all marginal and not conditional.

The procedures to obtain estimates of the parameters remain the same as described in previous section. The covariance matrix of $\boldsymbol{\phi}$ is constructed as shown in Section 4.3 with a "sandwich type" estimator.

## 4.5 Conclusions

We have introduced the method of PL as an alternative to ML. The asymptotic properties together with the advantages have been discussed. We comment on the different advantages of this method over the ML, especially in the case of multivariate data. We have been proposed a "sandwich" type estimator for the variance of the parameters of the model. At the end we focused on the pairwise version of the PL method that will be used in next chapter to fit a multivariate marginal survival model.

# Chapter 5

# Pseudo-likelihood Estimation for a Marginal Multivariate Survival Model

## 5.1 Introduction

This chapter is devoted to the development of a new multivariate model for survival outcomes based on the Plackett-Dale distribution (Dale 1986). The pseudo-likelihood method for the estimation of the parameters introduced in Chapter 4 and these ideas are applied to two case studies.

The chapter is organized as follows. Section 5.2 motivates the problem through two case studies. Section 5.3.1 gives a description of the Plackett-Dale model (Molenberghs and Lesaffre 1994) for survival data in the bivariate case. Section 5.3.2 describes an extension of the model to the case of $k$ correlated survival times and proposes a pseudo-likelihood approach for the estimation of the parameters of the model. Section 5.4 gives a summary of the association measures and Section 5.5 contains the analysis of the case studies. The first study is in AIDS where overall survival time and different opportunistic infections in HIV-infected patients are studied. The second study is on adoption data where the association of the survival times within families is modeled, applying the proposed methodology in the context of population genetics.

Survival models have been used intensively during the past two decades, across a number of application areas. Medical researchers used them extensively but in many other fields, where the main interest is in time-to-event, they became an important tool as well (Fleming and Harrington 1991). The effect of one or more covariates on the patient's survival can be modeled via the Cox model (Cox 1972), but we should recall that independence of survival times from one observation to the other is one of the basic assumptions of this model. However, in the last years there has been an increasing interest in frameworks where two or more events per patient or per statistical unit are observed. These statistical units can refer to clusters and hence multivariate survival models should be used, taking into account within-cluster dependencies. The former phenomenon is observed in groups of patients that share common characteristics, such as in family studies where the members share genetic and environmental factors. There are several issues one should take into account when extending the Cox model or any other univariate survival model, to the situation where the association needs to be modeled, which is the topic of the current chapter. The key idea is to introduce a model that allows for a full association structure between the times to event pertaining to a given unit while, due to an appropriate use of pseudo-likelihood ideas, keeping the computational burden under control.

## 5.2 Motivating Cases

In this section, we will introduce two different studies for which the proposed methodology is of use. The AIDS case study deals with intrasubject correlation, i.e., multiple events per subjects are recorded. The adoption study is an example of a study where clustering, i.e., within-cluster dependencies, are present.

### 5.2.1 The AIDS Study

These data arise from randomized clinical trials. A total of 1530 patients who participated in two clinical trials sponsored by the AIDS Clinical Trials Group (ACTG): ACTG 116A (Dolin *et al.* 1995) and 116B/117 (Kahn *et al.* 1992) were randomized to compare zidovudine (ZDV) and two doses of didanosine (ddI). Participants either had a diagnosis of AIDS or AIDS related complex (ARC) and/or had CD4 counts of 300 or fewer. The primary outcomes of interest for this analysis were survival and appearance of new or recurrent AIDS-defining events. Patients were randomly assigned

to receive one of the following three treatments: ddI 750 mg per day, ddI 500 mg per day, or ZDV 600 mg per day. These studies enrolled patients between October 1989 and April 1991; patients were followed for a median of 65 weeks and a maximum of 132 weeks. For illustration, ZDV is compared to any dose of ddI; therefore we use a binary indicator variable for treatment effect. Measures of CD4 for individual patients are included in the model. This choice is supported by the work of Saah *et al.* (1994), who found that CD4 was a laboratory measure in a Cox proportional hazards model which predicted survival after AIDS. There has been some debate in the literature as to whether a dichotomization of CD4 can be justified or not. We will use a continuous version of this variable but any other categorization can be considered without substantially having to modify the methodology. Molenberghs, Williams, and Lipsitz (2002) studied the joint modeling of survival and CD4 count on these data.

### 5.2.2 The Adoption Study

This study, presented in Sørensen *et al.* (1988), was carried out to analyze the impact of environmental and genetic factors on survival of adult adoptees. To this end, dependencies between the survival times of children and biological parents, and between children and adoptive parents are the focus of interest. In this study, families with adoptive children, born between 1924 and 1926, were analyzed. The basic idea is that association between survival times of biological parents and children can be assigned to some extent to genetic factors, while associations between children and adoptive parents can be due only to environmental factors.

These data were studied by Nielsen *et al.* (1992) who proposed a shared gamma frailty model and by Parner (2001) who proposed a composite likelihood method for the estimation of the frailty parameters and the standard deviations. We propose to use a Plackett-Dale model for correlated survival times data with Weibull margins, as will be described next.

## 5.3 Model Description

### 5.3.1 Bivariate Plackett-Dale Model for Survival Data

In this section, we will introduce the Plackett-Dale model for two survival outcomes. Assume that $T_1$ and $T_2$ are correlated survival times, then the joint survival function

of $(T_1, T_2)$ can be written as

$$\begin{aligned} S_{T_1T_2}(t_1,\, t_2) &= P(T_1 \geq t_1,\, T_2 \geq t_2) \\ &= C_{\theta_{12}}\{S_{T_1}(t_1),\, S_{T_2}(t_2)\}, \quad t_1,\, t_2 \geq 0, \end{aligned} \tag{5.1}$$

where $S_{T_1}$ and $S_{T_2}$ denote marginal survival functions and $C_{\theta_{12}}$ is a copula. An attractive feature of model (5.1) is that the margins do not depend on the choice of the copula function.

In principle, in model (5.1) any copula function can be used. For simplicity, we consider primarily one-parameter families; hence the use of a single parameter $\theta_{12}$ in (5.1). Some possible options are the Clayton, Hougaard, and Plackett copulas. Burzykowski *et al.* (2001) studied them in detail within the framework of surrogate endpoints. For the Clayton and Hougaard copulas, model (5.1) reduces to a proportional frailty model (Oakes 1989) with frailties generated, respectively, by the gamma and the positive stable distributions.

To model the effect of specific covariates on the marginal distributions of $T_1$ and $T_2$ in (5.1) we propose to use the proportional hazard model:

$$S_{T_k}(t_k) = \exp\left\{-\int_0^{t_k} h_{T_k}(x)\exp(\boldsymbol{\beta}_{T_k} Z_k)dx\right\}, \qquad k = 1, 2, \tag{5.2}$$

where $h_{T_1}$ and $h_{T_2}$ are marginal baseline hazard functions and $\boldsymbol{\beta}_{T_1}$ and $\boldsymbol{\beta}_{T_2}$ are vectors of unknown regression parameters corresponding to the covariates $Z_1$ and $Z_2$. The hazard functions can be specified parametrically or can be left unspecified as in the classical model proposed by Cox (1972). When the hazard functions are specified, maximum likelihood estimates of the parameters for joint model (5.1)–(5.2) can be obtained (Lehmann 1983). Alternatively, the two-stage parametric procedure proposed by Shih and Louis (1995) can be used, in which parameters of the marginal survival functions $S_{T_1}$ and $S_{T_2}$ are estimated first (assuming independence), and then $\theta_{12}$ is estimated conditional on the estimated values of the marginal parameters.

This one-parameter family is closely related to the Plackett family of bivariate distributions (Plackett 1965). In this case the dependence can be defined using a *global cross-ratio* at $(t_1, t_2)$ which, given the marginal cumulative density functions $F_{T_1}$ and $F_{T_2}$, is given by

$$\theta_{12}(t_1, t_2) = \frac{F(t_1, t_2)[1 - F_{T_1}(t_1) - F_{T_2}(t_2) + F(t_1, t_2)]}{[F_{T_1}(t_1) - F(t_1, t_2)][F_{T_2}(t_2) - F(t_1, t_2)]}. \tag{5.3}$$

Note that "global" refers to the fact that, at every point, the bivariate space is divided into four quadrants. Then, the probability over each quadrant is calculated and these four quantities are then used to compute the odds ratio. In chapter 3 we presented a graphical representation of the construction of the Plackett's distribution.

The components in (5.3) are the quadrant probabilities in $\mathbb{R}^2$ with vertex at $(t_1, t_2)$. Specifically, in the survival setting this parameter can be expressed for two survival times $T_1$ and $T_2$ as

$$\theta_{12} = \frac{P(T_1 > t|T_2 > k)P(T_1 \leq t|T_2 \leq k)}{P(T_1 \leq t|T_2 > k)P(T_1 > t|T_2 \leq k)}. \tag{5.4}$$

and therefore is naturally interpreted as the ratio of the odds for surviving beyond time $t$ given response higher than $k$ to the odds of surviving beyond time $t$ given response at most $k$. Mardia (1970) showed that $F_{T_1,T_2}(t_1, t_2)$ is always a bivariate copula, with $\theta_{12}$ in $[0, +\infty]$.

Based upon this distribution function, we can derive a bivariate Plackett *density* function $f_{T_1T_2}(t_1, t_2)$ for two survival times by calculating $\partial F_{T_1T_2}(t_1, t_2)/\partial t_1 \partial t_2$ in an appropriate way taking censoring into account.

The parameters of this model and their standard deviations can be estimated by means of the maximum likelihood method. We provide the expression for the log likelihood function, together with the derivatives of the distribution function $F$ in the next paragraphs.

Let $(T_1, T_2)$ denote paired failures times and $(S_1, S_2)$, $(f_1, f_2)$ the corresponding marginal survival and density functions. Then, the joint survival and density functions of $(T_1, T_2)$ are given by

$$\begin{aligned} S(t_1, t_2) &= F_{T_1T_2}(S_{T_1}(t_1), S_{T_2}(t_2)), \\ f(t_1, t_2) &= \frac{\partial^2 S(t_1, t_2)}{\partial t_1 \partial t_2} f_{T_1}(t_1) f_{T_2}(t_2), \end{aligned} \tag{5.5}$$

respectively, with $t_1,\ t_2 \geq 0$.

Let us denote by $(C_1, C_2)$ the paired censoring times. For $i = 1, \ldots, n$, assume that $(T_{i1}, T_{i2})$ and $(C_{i1}, C_{i2})$ are independent. For each $i$ we observe $T_{ij} = \min(D_{ij}, C_{ij})$ $j = 1, 2$ then $\Delta_{ij} = I\{D_{ij} = T_{ij}\}$, i.e., indicates whether the lifetime is observed $(\Delta_{ij} = 1)$ or not $(\Delta_{ij} = 0)$.

In Chapter 3 we have defined a general expression for the log-likelihood function using any copula structure and with different possible kind of censoring. We will write

this general log-likelihood for the case of the Plackett copula with right censored times. We can write the log likelihood function by combining four different situations in one expression as follows

- Case 1 : $\Delta_{i1} = 1$ and $\Delta_{i2} = 1$
- Case 2 : $\Delta_{i1} = 1$ and $\Delta_{i2} = 0$
- Case 3 : $\Delta_{i1} = 0$ and $\Delta_{i2} = 1$
- Case 4 : $\Delta_{i1} = 0$ and $\Delta_{i2} = 0$

Therefore,

$$\begin{aligned}\log \ell = \quad & \sum_{i=1}^{n} \Delta_{1i}\Delta_{i2} \log(f(t_{i1}, t_{i2})) + \Delta_{i1}(1-\Delta_{i2}) \log\left(-\frac{\partial S(t_{i1}, t_{i2})}{\partial t_1}\right) \\ & +(1-\Delta_{i1})\Delta_{i2} \log\left(-\frac{\partial S(t_{i1}, t_{i2})}{\partial t_2}\right) \\ & +(1-\Delta_{i1})(1-\Delta_{i2}) \log(S(t_{i1}, t_{i2}))\end{aligned} \tag{5.6}$$

where $S(t_1, t_2)$ and $f(t_1, t_2)$ were defined in (5.5).

The estimated parameters of this bivariate model can be obtained via ML method by maximizing the expression (5.6). The distribution function together with its derivates for the case $\theta \neq 1$ are displayed in the appendix of this chapter. The case $\theta = 1$ is trivial and it corresponds to the independence case.

### 5.3.2 Multivariate Plackett-Dale Model for Survival Data with Pseudo-likelihood Estimation

While the model described in Section 5.3.1 suffices to analyze bivariate time-to-event outcomes, an extension is needed for applications with more than two times. To this end, consider an experiment involving $N$ subjects or clusters of $k$ time-to-event measurements.

The principal idea can be laid out in three steps. First, we construct a model for these $k$ times by considering univariate models for every time-to-event separately. It is evident that covariates can be included in these parametric marginal models. Second, we consider bivariate models for every possible pair that can be formed from the $k$ times and of which the univariate marginal models are the ones already considered; in other words, Plackett-Dale models will be considered for every possible pair. Third,

in order to avoid the full multivariate specification of the model, while nevertheless properly accounting for the full association structure, pseudo-likelihood ideas are used to obtain valid point estimates as well as valid precision estimates.

This approach is similar in spirit to the one proposed by Parner (2001) in the sense that both are marginal models for multivariate survival data and both use pseudo-likelihood related ideas. However, the actual copulas chosen are different, enabling a comparison of the results from both, for example. Since there is no unambiguous choice as to what the best model would be for multivariate survival data, a more ample choice of models is desirable and can lead up to a sensitivity analysis.

Suppose that we also observe a vector of covariates $Z$. A Weibull distribution is assumed for each time $T_j$ with $\lambda_{T_j}$ and $p_{T_j}$ the scale and shape parameters, respectively. While we focus on Weibull marginals, different researchers may choose to use different univariate marginal survival distributions, implying only relatively small adaptations of the methodology. The information concerning subject $i$ can be expressed in vector format as $(T_{i1}, \ldots, T_{ik}, \Delta_{i1}, \ldots, \Delta_{ik}, z_{i1}, \ldots, z_{in_k})$, with $n_k$ the number of covariates, so that $\boldsymbol{W}_{ij} = (T_{ij}, \Delta_{ij}, Z_i)$ are the values for a particular subject $i$ and time point $j$.

While a full multivariate formulation of the Plackett-Dale model has been done in the context of ordinal data (Molenberghs and Lesaffre 1994, 1999), it poses non-trivial computational complexities. Instead, marginal pseudo-likelihood ideas will be used to keep the amount of computation under control, while enabling to answer relevant research questions (le Cessie and Van Houwelingen 1994; Geys, Molenberghs and Lipsitz 1998; Geys, Molenberghs and Ryan 1999).

In Chapter 4 the pseudo-likelihood method was introduced with all its different versions. In the same chapter we commented on the potential advantages in some areas where it eliminates the difficulties due to strong distributional assumptions or intensive computations. In the particular case of a multivariate Plackett model PL avoids the need to find the zeros of a polynomial of a high degree and to compute numerical implicit derivates.

The idea behind our (pairwise) pseudo-likelihood function is based on considering all possible pairs $(\boldsymbol{W}_{ir}, \boldsymbol{W}_{il})$ of outcomes on an individual, producing $f_{T_rT_l}(\boldsymbol{W}_{ir}, \boldsymbol{W}_{il})$, rather than the full multivariate density, and then taking the product over them. The resulting function will be denoted by $PL$ and its log by

$$\ln p\ell(\boldsymbol{\phi}) = \sum_{i=1}^{N} p\ell_i, \tag{5.7}$$

with

$$p\ell_i = \sum_{(s,t)\in S} \ln f_{T_sT_t}(\boldsymbol{W}_{is}, \boldsymbol{W}_{it}, \boldsymbol{\phi}),$$

where $S$ is the set of indices with all possible pairs of outcomes of interest, $f_{T_sT_t}$ is the value of the function defined in Section 5.3.1 evaluated in the respective outcomes for subject $i$, and $\boldsymbol{\phi}$ is the vector of parameters. Specifically $\boldsymbol{\phi}' = (\boldsymbol{\theta}', \boldsymbol{\beta}_T', \boldsymbol{\lambda}_T', \boldsymbol{p}_T')$ with $\boldsymbol{\theta}$ the subvectors of association parameters, $\boldsymbol{\beta}_T$ the subvector of coefficients corresponding to the covariates $\boldsymbol{z}$ and, $\boldsymbol{\lambda}_T$ and $\boldsymbol{p}_T$ subvector of parameters from the Weibull distribution.

The pseudo-likelihood estimator $\widehat{\boldsymbol{\phi}}$ is defined as the maximizer of (5.7) as explained in Chapter 4. Consistency also holds for the pairwise version of this method (see Chapter 4) where $\widehat{\boldsymbol{\phi}}$ converges in probability to $\boldsymbol{\phi}_0$, the true parameter value and $\sqrt{N}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)$ converges in distribution to

$$N_q(\mathbf{0}, J(\boldsymbol{\phi}_0)^{-1} K(\boldsymbol{\phi}_0) J(\boldsymbol{\phi}_0)^{-1}), \tag{5.8}$$

with $J(\boldsymbol{\phi})$ defined by

$$J_{rl} = -\sum_{(s,t)\in S} E_{\boldsymbol{\phi}} \left( \frac{\partial^2 \ln f_{T_sT_t}(t_{is}, t_{it})}{\partial \phi_r \partial \phi_l} \right) \tag{5.9}$$

and $K(\boldsymbol{\phi})$ by

$$K_{rl} = -\sum_{(s,t)\in S} E_{\boldsymbol{\phi}} \left( \frac{\partial \ln f_{T_s}(t_{is}, t_{it})}{\partial \phi_r} \frac{\partial \ln f_{T_t}(t_{is}, t_{it})}{\partial \phi_l} \right). \tag{5.10}$$

The main difference with the general theory already introduced is that the set of indices $S$ contains all different pairs of observations. The asymptotic covariance matrix can be estimated in an easy way. Indeed, the matrix $J$ is found from evaluating the second derivate of the log $p\ell$ function at the PL estimate and $K$ can be replaced by the cross-product of the observed scores.

## 5.4 Association Measures

The Plackett-Dale model allows us to estimate and interpret the strength of the association between a pair of survival times via global cross ratios (the $\theta$ parameters in the model). Therefore, $\theta$ may be considered a natural candidate for the measure of

association. However, some researchers may find it is hard to get a feel for it because it ranges throughout the entire real line. Further, different copulas (like the Clayton and Hougaard copulas; Hougaard, 2000) carry different and less straightforward association parameters. In such a situation it would be easier to work with a transformation of $\theta$ that has the interpretational properties of a correlation coefficient, such as Kendall's $\tau$ or Spearman's $\rho$. These will be reviewed in turn.

### 5.4.1 Kendall's $\tau$

This measure defined in Chapter 3 for any pair of random variables and for any copula can be seen as the difference between the probability of concordance and the probability of discordance of two realizations of $(T_1, T_2)$, lies in the $[-1, 1]$ interval and a zero value implies independence between $T_1$ and $T_2$. For any copula $C(t_1, t_2, \theta)$ (Genest and MacKay, 1986) found the following expression:

$$\tau(\theta) = 4 \int_0^1 \int_0^1 C_{T_1 T_2}(t_1, t_2, \theta) C_{T_1 T_2}(dt_1, dt_2, \theta) - 1. \tag{5.11}$$

One interesting feature is that the marginal distributions of $T_1$ and $T_2$ do not affect (5.11), and hence it follows that $\tau$ only depends on the copula function $C_{T_1 T_2}$ For the Plackett copula there is no closed form for Kendall's $\tau$ and an estimate has to be obtained directly from (5.11). Confidence intervals via the delta method can be constructed, we have developed a SAS IML 8.02 macro to obtain $\widehat{\tau}$ and confidence intervals.

Kendall's $\tau$ is very useful in the sense that it measures the association between both time points after adjustment for the covariates used in the model.

### 5.4.2 Spearman's $\rho$

Spearman's $\rho$ is also based on concordance and discordance, it can be shown that it equals Pearson's product-moment for grades of a pair of continuous random variables. The relationship between Spearman's $\rho$ and the Plackett copula function is

$$\begin{aligned} \rho(\theta) &= 12 \int_0^1 \int_0^1 C_{T_1 T_2}(t_1, t_2, \theta) dt_1 dt_2 - 3 \\ &= \frac{\theta + 1}{\theta - 1} - \frac{2\theta \ln \theta}{(\theta - 1)^2}. \end{aligned} \tag{5.12}$$

As we can see from (5.12) there is a closed-form expression for the Plackett copula and in addition to that an estimate follows from $\rho = \rho(\widehat{\theta})$, with delta-method variance

$$\mathrm{Var}(\widehat{\rho}) = \left[\frac{-4(\widehat{\theta}-1) + 2(\widehat{\theta}+1)\ln\widehat{\theta}}{(\widehat{\theta}-1)^3}\right]^2 \mathrm{Var}(\widehat{\theta}).$$

The asymptotic properties are analogous to these for the general case, i.e., $\rho(\theta) \to 0$ when $\theta \to 1$, $\rho(\theta) \to -1$ when $\theta \to 0$ and $\rho(\theta) \to 1$ when $\theta \to \infty$.

## 5.5 Case Studies

We are now in a position to analyze the data from Sections 5.2.1 and 5.2.2. Pseudo-likelihood estimates were obtained using Newton-Raphson with analytical first derivatives and numerical second derivatives, implemented in SAS IML 8.02 and using routine NLPNRR (SAS Institute Inc. 1999–2001). Standard errors of the parameters were calculated using the inverse of the observed matrix of second derivatives. Although in these two examples a trivariate model is considered, the methodology is fully generally applicable to longer sequences of time-to-event endpoints. Indeed, the structure of the SAS programs allows us to fit any model and any number of outcomes with only minor changes. Using a flexible design matrix structure, a large class of model specifications is possible.

### 5.5.1 Analysis of the Adoption Study

We first consider bivariate analyses, selecting pairs out of the three possible survival times of interest. The first aim is to describe the biological associations between mother, father and child, and then to study the environmental effect, e.g., correlations with the adoptive parents. In each case, a trivariate analysis is envisaged. We will start with bivariate analyses and compare these results with those obtained from modeling the trivariate data directly. We will use the abbreviations BM, BF and ACh for biological mother, biological father and adoptive child in the biological models, replacing BM with AM and BF with AF in the adoptive models. The corresponding subscripts are 1, 2, and 3 in each case. All results for the biological families are presented in Table 5.1, while Table 5.2 presents estimates for the adoptive families. The marginal distributions are all assumed to be Weibull with parameters $\lambda_j$ and $p_j$, $j = 1, 2, 3$, and we consider three different parameters $\beta_1$, $\beta_2$, and $\beta_3$ to adjust for

the sex of the child as it was done by Parner (2001). All association parameters are assumed to be constant.

It is clear from the way in which PL is defined that ML estimates are exactly the same when only two outcomes are considered. Although model-based standard errors and empirically corrected standard errors (i.e., those based on (5.8)) are numerically different, they are of similar magnitudes and no clear ordering is seen between them. The tables reveal that the model based standard errors calculated by means of the information matrix and the empirically corrected ones differ only slightly. Common parameters estimated using two different bivariate models are similar since all models are of a marginal type. By "marginal type" we mean that the univariate marginal parameters in a bivariate model have exactly the same meaning as their counterparts in the corresponding univariate model. For example, $\widehat{\beta}_1 = -0.085$ in model BM–BF as opposed to $\widehat{\beta}_1 = -0.086$ in BM–ACh.

Tables 5.1 and 5.2 include all three types of association parameters: not only the odds ratios $\theta$ but also Kendall's $\tau$ and Spearman's $\rho$, as introduced in Section 5.4. We observe the association is not very strong but nevertheless significantly different from zero in some cases. The $\tau$ and $\rho$ parameters are relatively similar but, in spite of them ranging on the same scale, they have a different meaning and they are not directly comparable.

Let us now turn attention to the trivariate situation. Let us consider a model with different association parameters for each pair of outcomes $\theta_{12}$, $\theta_{13}$, and $\theta_{23}$ and with different parameters for the covariates corresponding to each outcome $\beta_1$, $\beta_2$, and $\beta_3$. Specific Weibull distributions with different scale and shape parameters for each outcome were used to model the marginals, i.e., $p_1$, $p_2$, $p_3$, $\lambda_1$, $\lambda_2$, and $\lambda_3$. Effectively, this is the trivariate version of the previous bivariate ones. For the trivariate models, only empirically corrected standard errors are given in Tables 5.1 and 5.2, since the model-based ones ignore the fact that in using all pairs out of three survival times on a cluster, all outcomes are used twice, leading to an exaggerated precision.

Therefore, model-based standard errors are useless, even if all marginal and association models are correctly specified. We like to point out this feature since it is different from the GEE setting. Other than being a disadvantage, it is merely a "side effect" of the way marginal pseudo-likelihood works. Let us add that obtaining convergence was not difficult and using different sets of starting values showed stability of the process.

Parameters retain their meaning they had in the bivariate models, with two ad-

vantages. First, using the data in a trivariate model is more statistical efficient than using them in three separate models. Second, one avoids the occurrence of double estimates for the marginal parameters ($\beta$, $\lambda$, and $p$ parameters), in spite of them being not too different between various bivariate models. The same model was applied to the biological and adoptive families, enabling to contrast both sets of dependencies.

Comparisons of our association parameters with the ones given by Parner (2001) cannot be made directly, since they are expressed on different scales. The association in our case is the global odds ratio, while Parner's quantity is based on the mean and variance of the assumed Gamma distribution. Therefore, both sets of association parameters are transformed to Kendall's $\tau$ and Spearman's $\rho$. There is a close agreement between both methods (see Table 5.2), and both enable consideration of multivariate models.

According to Parner's conclusions, the environmental association between the adoptive child and the mother was significant and negative; the environmental association between adoptive father and the adoptive mother was significant. In our case, we can see from Table 5.2 that the estimated Kendall's coefficients are, $\tau_{13} = -0.036$ with a 95% confidence interval $(-0.060, -0.012)$ and $\tau_{12} = 0.052$ with a 95% confidence interval $(0.041, 0.063)$ respectively. These results suggest that the longevity of the mother and the adoptive child were negatively correlated. Thus, we arrived to the same conclusions. The estimates are similar to the estimates obtained using Parner's model as shown in Table 5.2. We could also test for equal environmental effects and genetic effects using a Wald type test, but this is not the main goal of this work; details can be found in Parner (2001).

Table 5.1: *Adoption study: Model for the biological families. Maximum likelihood estimates (model based standard errors; empirically corrected standard errors) of bivariate survival times and pseudo-likelihood estimates (standard errors) for trivariate model. For Kendall's $\tau$ and Spearman's $\rho$, estimates and 95% confidence intervals are given.*

| Par. | BM-BF | BM-ACh | BF-ACh | BM–BF–ACh |
|---|---|---|---|---|
| $\theta_{12}$ | 1.076(0.128;0.128) | | | 1.076(0.127) |
| $\theta_{13}$ | | 1.164(0.193;0.187) | | 1.164(0.187) |
| $\theta_{23}$ | | | 1.176(0.194;0.202) | 1.175(0.201) |
| $\beta_1$ | -0.085(0.086;0.077) | -0.086(0.086;0.077) | | -0.084(0.069) |
| $\beta_2$ | -0.009(0.078;0.072) | | -0.010(0.078;0.072) | -0.004(0.036) |
| $\beta_3$ | | -1.066(0.164;0.159) | -1.060(0.164;0.159) | -1.063(0.137) |
| $p_1$ | 0.220(0.017;0.015) | 0.219(0.017;0.015) | | 0.220(0.013) |
| $p_2$ | 0.279(0.011;0.010) | | 0.279(0.011;0.010) | 0.279(0.006) |
| $p_3$ | | 0.086(0.054;0.063) | 0.085(0.054;0.063) | 0.086(0.054) |
| $\lambda_1$ | 3.818(0.146;0.178) | 3.817(0.146;0.179) | | 3.818(0.155) |
| $\lambda_2$ | 5.568(0.179;0.201) | | 5.568(0.179;0.200) | 5.568(0.174) |
| $\lambda_3$ | | 2.312(0.175;0.290) | 2.313(0.176;0.291) | 2.313(0.252) |
| $\tau_{12}$ | 0.016(0.003,0.029) | | | 0.016(0.003,0.029) |
| $\tau_{13}$ | | 0.034(0.016,0.051) | | 0.034(0.016,0.051) |
| $\tau_{23}$ | | | 0.036(0.018,0.054) | 0.036(0.017,0.054) |
| (Parner) $\tau_{12}$ | 0.035(0.024,0.045) | | | |
| (Parner) $\tau_{13}$ | | 0.050(0.036,0.064) | | |
| (Parner) $\tau_{23}$ | | | 0.037(0.023,0.050) | |
| $\rho_{12}$ | 0.024(-0.053,0.102) | | | 0.024(-0.053,0.102) |
| $\rho_{13}$ | | 0.051(-0.054,0.155) | | 0.051(-0.054,0.155) |
| $\rho_{23}$ | | | 0.054(-0.054,0.162) | 0.054(-0.058,0.165) |
| (Parner) $\rho_{12}$ | 0.052(-0.010,0.113) | | | |
| (Parner) $\rho_{13}$ | | 0.075(-0.010,0.165) | | |
| (Parner) $\rho_{23}$ | | | 0.055(-0.027,0.137) | |

Table 5.2: *Adoption study: Model for the biological and adoptive families. Maximum likelihood estimates (model based standard errors; empirically corrected standard errors) of bivariate survival times and pseudo-likelihood estimates (standard errors) for trivariate model. For Kendall's $\tau$ and Spearman's $\rho$, estimates and 95% confidence intervals are given.*

| Par. | AM-AF | AM-ACh | AF-ACh | AM–AF–ACh |
|---|---|---|---|---|
| $\theta_{12}$ | 1.265(0.132;0.127) | | | 1.265(0.127) |
| $\theta_{13}$ | | 0.844(0.138;0.133) | | 0.849(0.133) |
| $\theta_{23}$ | | | 1.237(0.200;0.198) | 1.240(0.197) |
| $\beta_1$ | -0.015(0.077;0.072) | -0.012(0.077;0.072) | | -0.029(0.064) |
| $\beta_2$ | 0.078(0.075;0.074) | | 0.077(0.075;0.074) | 0.025(0.034) |
| $\beta_3$ | | -1.066(0.164;0.159) | -1.064(0.164;0.158) | -1.068(0.137) |
| $p_1$ | 0.210(0.009;0.009) | 0.210 (0.009;0.009) | | 0.211(0.008) |
| $p_2$ | 0.235(0.008;0.008) | | 0.235(0.008;0.008) | 0.241(0.005) |
| $p_3$ | | 0.085(0.054;0.063) | 0.085(0.054;0.063) | 0.086(0.055) |
| $\lambda_1$ | 6.402(0.203;0.218) | 6.406(0.203;0.219) | | 6.405(0.189) |
| $\lambda_2$ | 7.223(0.210;0.220) | | 7.228(0.210;0.220) | 7.222(0.191) |
| $\lambda_3$ | | 2.312(0.176;0.290) | 2.311(0.176;0.291) | 2.312(0.252) |
| $\tau_{12}$ | 0.052(-0.045,0.150) | | | 0.052(0.041,0.063) |
| $\tau_{13}$ | | -0.038(-0.184,0.108) | | -0.036(-0.060,-0.012) |
| $\tau_{23}$ | | | 0.047(0.030,0.065) | 0.048(0.030,0.065) |
| (Parner) $\tau_{12}$ | 0.051(0.040,0.061) | | | |
| (Parner) $\tau_{13}$ | | -0.069(-0.085,-0.052) | | |
| (Parner) $\tau_{23}$ | | | 0.041(0.027,0.054) | |
| $\rho_{12}$ | 0.078(-0.501,0.657) | | | 0.078(0.013,0.143) |
| $\rho_{13}$ | | -0.057(-0.931,0.818) | | -0.055(-0.198,0.089) |
| $\rho_{23}$ | | | 0.071(-0.033,0.175) | 0.072(-0.034,0.177) |
| (Parner) $\rho_{12}$ | 0.076(0.013,0.140) | | | |
| (Parner) $\rho_{13}$ | | -0.103(-0.202,-0.004) | | |
| (Parner) $\rho_{23}$ | | | 0.061(-0.021,0.143) | |

### 5.5.2 Analysis of the AIDS Study

In this section, we analyze the data described in Section 5.2.1. In the original paper by Finkelstein *et al.* (1996) the pattern of the development of opportunistic infections in HIV-infected patients was evaluated, based on a cohort of 1530 patients.

The more common AIDS-defining opportunistic infections are *Pneumocystis carinii* pneumonia (PCP), *Mycobacterium avium* complex (MAC), cytomegalovirus (CMV) and systemic mycosis. These authors performed all the analyses adjusted for CD4 count. Without loss of generality, we perform the analysis for three time-to-event outcomes: PCP, CMV and the overall survival time of the AIDS patients (DTH).

The main objective is to describe the association between all three outcomes after adjusting by CD4 count and treatment effect.

Parameters are subscripted with 1, 2, and 3 to refer to CMV, DTH, and PCP, respectively. For the sake of illustration, consider $\beta_T$ to be the *common* treatment effect and $\beta_1$, $\beta_2$, and $\beta_3$ the outcome-specific parameters associated with the CD4 count. We will assume a Weibull distribution with parameters $p_1$, $p_2$, $p_3$, $\lambda_1$, $\lambda_2$, and $\lambda_3$. Therefore, the vector of parameters to be estimated has 13 components:

$$\boldsymbol{\phi} = (\theta_{12}, \theta_{13}, \theta_{23}, \beta_T, \beta_1, \beta_2, \beta_3, p_1, p_2, p_3, \lambda_1, \lambda_2, \lambda_3), \tag{5.13}$$

where $\theta_{12}$, $\theta_{13}$ and $\theta_{23}$ are the global cross ratios. Using generalized linear models technology, it is straightforward to construct the overall design matrix $\boldsymbol{X}$, consisting of 13 columns (as many as there are parameters), and $3 \times 7 \times N$ rows. The calculation of the number of rows follows because there are 3 pairs to be formed out of three outcomes, for each pair (i.e., for each bivariate model), there are 7 "natural" parameters (an association parameter, and then a $\beta$, $\lambda$, and $p$ parameter for each component of the pair).

Let us exemplify the construction of a design matrix for the this case study. The contribution of a single individual can be seen in our case as the contribution of three pseudo-likelihood individuals. Thus, $\boldsymbol{X}$ can be written as $N$ blocks,

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_N \end{pmatrix},$$

where the block corresponding to subject $i$ is expressed as:

$$\boldsymbol{X}_i = \begin{pmatrix} \boldsymbol{X}_{i12} \\ \boldsymbol{X}_{i13} \\ \boldsymbol{X}_{i23} \end{pmatrix},$$

where

$$\boldsymbol{X}_{i12} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & trt_i & cd4_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & trt_i & 0 & cd4_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix},$$

$$\boldsymbol{X}_{i13} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & trt_i & cd4_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & trt_i & 0 & 0 & cd4_i & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

and

$$\boldsymbol{X}_{i23} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & trt_i & 0 & cd4_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & trt_i & 0 & 0 & cd4_i & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Generalization to more than three outcomes is straightforward and the SAS macro we developed carries the general situation. Parameter estimates are summarized in Table 5.3.

Parameters in common between different bivariate models are generally fairly close, with the exception of $\beta_T$, which is even changing signs. While not significant, this is a clear indication that the trivariate model is the more appealing one,

Table 5.3: *AIDS study: Maximum likelihood estimates (model based standard errors; empirically corrected standard errors) of bivariate survival times and pseudo-likelihood estimates (standard errors) for trivariate model. For Kendall's $\tau$ and Spearman's $\rho$, estimates and 95% confidence intervals are given.*

| Par. | CMV–DTH | CMV–PCP | DTH–PCP | CMV–DTH–PCP |
|---|---|---|---|---|
| $\theta_{12}$ | 5.165(2.570;2.401) | | | 4.369(1.165) |
| $\theta_{13}$ | | 4.434(1.850;2.182) | | 4.466(1.446) |
| $\theta_{23}$ | | | 3.943(1.023;0.959) | 3.691(0.865) |
| $\beta_T$ | -0.054(0.020;0.020) | 0.183(0.032;0.033) | -0.014(0.019;0.019) | 0.016(0.111) |
| $\beta_1$ | 1.708(1.816;1.681) | 1.504(1.892;1.547) | | 1.579(1.095) |
| $\beta_2$ | 2.160(0.706;0.752) | | 2.010(0.696;0.703) | 2.069(0.732) |
| $\beta_3$ | | 2.037(1.570;1.845) | 2.168(1.487;1.838) | 2.109(1.169) |
| $p_1$ | -0.240(0.137;0.142) | -0.657(0.193;0.184) | | -0.451(0.350) |
| $p_2$ | 0.341(0.033;0.038) | | 0.353(0.032;0.035) | 0.338(0.164) |
| $p_3$ | | -1.147(0.257;0.331) | -0.807(0.203;0.270) | -0.958(0.469) |
| $\lambda_1$ | 1.606(0.033;0.030) | 1.406(0.023;0.022) | | 1.487(0.136) |
| $\lambda_2$ | 1.941(0.015;0.017) | | 1.933(0.015;0.016) | 1.940(0.111) |
| $\lambda_3$ | | 1.117(0.012;0.014) | 1.215(0.014;0.018) | 1.161(0.108) |
| $\tau_{12}$ | 0.352(0.307,0.397) | | | 0.318(0.292,0.345) |
| $\tau_{13}$ | | 0.321(0.272,0.370) | | 0.323(0.291,0.355) |
| $\tau_{23}$ | | | 0.297(0.273,0.322) | 0.284(0.260,0.308) |
| $\rho_{12}$ | 0.503(0.269,0.736) | | | 0.459(0.318,0.599) |
| $\rho_{13}$ | | 0.462(0.204,0.721) | | 0.464(0.295,0.634) |
| $\rho_{23}$ | | | 0.430(0.298,0.563) | 0.412(0.283,0.541) |

in spite of a larger standard error. Note that for some, but not all, parameters the standard error produced by the trivariate model is smaller. The global cross ratios $\theta$ are quite large, showing a strong association between all pairs of outcomes. Also here, Kendall's $\tau$ and Spearman's $\rho$ are calculated to get a better grip on the association. Based on the correlation parameters $\rho$, a consistent picture of a correlation around 0.5 emerges.

## 5.6 Conclusions

In this chapter, we have extended the Plackett-Dale model for survival data to the multivariate case and we have shown that pseudo-likelihood estimation, in the sense of Arnold and Strauss (1991), is a viable and attractive alternative to maximum likelihood in case of multivariate survival data. Maximum likelihood becomes prohibitive for large sequences of times, due to computational requirements. In contrast, the pseudo-likelihood procedure gives quite satisfactory results. In addition, we proposed other association measures and we have shown the link of Spearman's $\rho$ and Kendall's $\tau$ to the association parameter of the Plackett-Dale model $\theta$. The method yields consistent and asymptotically normal estimates of the parameters of interest and the computational complexity is manageable.

The choice of the Plackett-Dale model was motivated by the fact that the association parameter $\theta$, has a natural interpretation for this copula. However, other copulas can be considered (Oakes 1989; Shih and Louis 1995; Joe 1997; Nelsen 1999). To this end, checking the goodness of fit of copulas to bivariate survival data can be done by using the method proposed by Wang and Wells (2000) and an adaptation of this method to our framework is a topic for future research. It is also worth noting that, while in this work we considered Weibull marginal distributions, it is possible to use other distributional assumptions, or even use a semi-parametric approach with unspecifed baseline hazard functions (Shih and Louis 1995).

The approach we presented gives a flexible tool for modeling any kind of time-to-event data accounting for the association between two or more outcomes. To illustrate our findings we have applied the proposed method in two different situations. Also, we have shown how the standard errors of the parameters need to be corrected in order to account for the lack of independence introduced by the fact that the information of a single subject is used more than once.

# Appendix

Distribution function and its derivatives for the case $\theta \neq 1$.

$$
\begin{aligned}
F(u,v,\theta) &= \frac{1}{2(\theta-1)} + \frac{u+v}{2} - \frac{H(u,v,\theta)}{2(\theta-1)} \\
H(u,v,\theta) &= \sqrt{(1+(\theta-1)(u+v))^2 - 4\theta(\theta-1)uv}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial H}{\partial u} &= \frac{(\theta-1)}{H(u,v,\theta)}[(1+(\theta-1)(u+v)) - 2\theta v] \\
\frac{\partial H}{\partial v} &= \frac{(\theta-1)}{H(u,v,\theta)}[(1+(\theta-1)(u+v)) - 2\theta u] \\
\frac{\partial H}{\partial \theta} &= \frac{(1+(\theta-1)(u+v))(u+v) - 2uv(2\theta-1)}{H(u,v,\theta)} \\
\frac{\partial^2 H}{\partial u^2} &= \frac{\left[(\theta-1)^2 - (\frac{\partial H}{\partial u})^2\right]}{H(u,v,\theta)} \\
\frac{\partial^2 H}{\partial v^2} &= \frac{\left[(\theta-1)^2 - (\frac{\partial H}{\partial v})^2\right]}{H(u,v,\theta)} \\
\frac{\partial^2 H}{\partial \theta^2} &= \frac{\left[(u-v)^2 - (\frac{\partial H}{\partial \theta})^2\right]}{H(u,v,\theta)} \\
\frac{\partial^2 H}{\partial u \partial \theta} &= \frac{\partial H}{\partial u}\left[\frac{1}{\theta-1} - \frac{1}{H(u,v,\theta)}\frac{\partial H}{\partial \theta}\right] + \frac{(\theta-1)(u-v)}{H(u,v,\theta)} \\
\frac{\partial^2 H}{\partial v \partial \theta} &= \frac{\partial H}{\partial v}\left[\frac{1}{\theta-1} - \frac{1}{H(u,v,\theta)}\frac{\partial H}{\partial \theta}\right] + \frac{(\theta-1)(v-u)}{H(u,v,\theta)} \\
\frac{\partial^2 H}{\partial u \partial v} &= -\frac{1}{H(u,v,\theta)}\left[\frac{\partial H}{\partial u}\frac{\partial H}{\partial v} + (\theta-1)(\theta+1)\right] \\
\frac{\partial F}{\partial u} &= \frac{1}{2}\left[1 - \frac{1}{\theta-1}\frac{\partial H}{\partial u}\right] \\
\frac{\partial F}{\partial v} &= \frac{1}{2}\left[1 - \frac{1}{\theta-1}\frac{\partial H}{\partial v}\right] \\
\frac{\partial F}{\partial \theta} &= -\frac{H(u,v,\theta)}{\theta-1} + \frac{1}{2(\theta-1)}\left[u+v-\frac{\partial H}{\partial \theta}\right]
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 F}{\partial u^2} &= -\frac{1}{2(\theta-1)}\left(\frac{\partial H}{\partial u}\right)^2 \\
\frac{\partial^2 F}{\partial v^2} &= -\frac{1}{2(\theta-1)}\left(\frac{\partial H}{\partial v}\right)^2 \\
\frac{\partial^2 F}{\partial \theta^2} &= -\frac{1}{\theta-1}\left[2\frac{\partial F}{\partial \theta}+\frac{1}{2}\frac{\partial^2 H}{\partial \theta^2}\right] \\
\frac{\partial^2 F}{\partial u \partial v} &= -\frac{1}{2(\theta-1)}\left[\frac{\partial H}{\partial u}\frac{\partial H}{\partial v}+(\theta-1)(\theta+1)\right] \\
\frac{\partial^2 F}{\partial u \partial \theta} &= \frac{1}{2H(u,v,\theta)(\theta-1)}\left[\frac{\partial H}{\partial u}\frac{\partial H}{\partial \theta}-(\theta-1)(u-v)\right] \\
\frac{\partial^2 F}{\partial v \partial \theta} &= \frac{1}{2H(u,v,\theta)(\theta-1)}\left[\frac{\partial H}{\partial v}\frac{\partial H}{\partial \theta}-(\theta-1)(v-u)\right] \\
\frac{\partial^3 F}{\partial u^3} &= \frac{1}{H(u,v,\theta)}\frac{\partial H}{\partial u}\left[-\frac{\partial^2 F}{\partial u^2}+\frac{1}{\theta-1}\frac{\partial^2 H}{\partial u^2}\right] \\
\frac{\partial^3 F}{\partial v^3} &= \frac{1}{H(u,v,\theta)}\frac{\partial H}{\partial v}\left[-\frac{\partial^2 F}{\partial v^2}+\frac{1}{\theta-1}\frac{\partial^2 H}{\partial v^2}\right] \\
\frac{\partial^3 F}{\partial u^2 \partial v} &= \frac{1}{H(u,v,\theta)}\left[-\frac{\partial^2 F}{\partial u^2}\frac{\partial H}{\partial v}+\frac{1}{\theta-1}\frac{\partial^2 H}{\partial v \partial u}\frac{\partial H}{\partial u}\right] \\
\frac{\partial^3 F}{\partial u \partial v^2} &= \frac{1}{H(u,v,\theta)}\left[\frac{\partial^2 F}{\partial v^2}\frac{\partial H}{\partial u}-\frac{1}{\theta-1}\frac{\partial^2 H}{\partial u \partial v}\frac{\partial H}{\partial v}\right] \\
\frac{\partial^3 F}{\partial u^2 \partial \theta} &= \frac{1}{\theta-1}\frac{\partial^2 F}{\partial u^2}+\frac{1}{H(u,v,\theta)}\left[-\frac{\partial^2 F}{\partial u^2}\frac{\partial H}{\partial \theta}+\frac{1}{\theta-1}\frac{\partial H}{\partial u}\frac{\partial^2 H}{\partial u \partial \theta}\right] \\
\frac{\partial^3 F}{\partial u \partial \theta^2} &= -\frac{1}{\theta-1}\frac{\partial^2 F}{\partial u \partial \theta}-\frac{1}{H(u,v,\theta)}-\frac{\partial^2 F}{\partial u \partial \theta}\frac{\partial H}{\partial \theta} \\
&\quad +\frac{1}{2H(u,v,\theta)(\theta-1)}\left[\frac{\partial^2 H}{\partial u \partial \theta}\frac{\partial H}{\partial \theta}+\frac{\partial H}{\partial u}\frac{\partial^2 H}{\partial \theta^2}-(u-v)\right] \\
\frac{\partial^3 F}{\partial u \partial v \partial \theta} &= -\frac{1}{\theta-1}\frac{\partial^2 F}{\partial u \partial v}-\frac{1}{H(u,v,\theta)}-\frac{\partial^2 F}{\partial u \partial v}\frac{\partial H}{\partial \theta} \\
&\quad +\frac{1}{2H(u,v,\theta)(\theta-1)}\left[\frac{\partial^2 H}{\partial u \partial \theta}\frac{\partial H}{\partial v}+\frac{\partial H}{\partial u}\frac{\partial^2 H}{\partial v \partial \theta}+2\theta\right]
\end{aligned}
$$

# Chapter 6

# Multivariate Plackett-Dale Inference to Study the Inheritance of Longevity in a Belgian Village

## 6.1 Introduction

In Chapter 5 we have proposed a marginal multivariate survival model and we have shown how pseudo-likelihood methods can be applied to obtain estimates of the parameters. In this chapter we propose a series of tests to perform inferences of these model parameters. In addition, we apply our methodology to study familial transmittance of longevity. We focus on associations between mother, father and first child, and therefore we deal with family clusters of equal size. The methodology is applied to a demographic database of a Flemish village ($18^{\text{th}}$-$20^{\text{th}}$ century). We investigate familial transmission mainly via the mother and we explore the impact of such other factors as censoring, gender effect, age at death, etc. This work complements the results of Matthijs *et al.* (2002) and suggests further analyses to better understand the precise mechanisms behind these associations.

The main aim of this work is to propose a set of inferential tools for the parameters of a multivariate marginal survival model. We will explain the methodology and apply it to the study of associations between longevity of family members in a small Flemish village. Each family is treated as a cluster and we will use a multivariate Dale model for survival data combined with pseudo-likelihood ideas already introduced in Chapter 4. In the next sections the statistical model will be described shortly and the methodology for performing inferences will be presented in detail. The main substantive is differences in the influences of fathers and mothers on the female offspring's longevity. This issue is closely linked to the discussion on the mitochondrial theory of ageing, which expects a relatively strong influence of the mother on the offspring's longevity.

It has frequently been claimed that longevity has a familial component, namely that longevity of parents is associated with longevity of offspring (Gavrilova *et al.* 1998a; Gudmundsson *et al.* 2000; Korpelainen 1999; Matthijs *et al.* 2002). We limited the research to a relatively small number of observations and we will therefore concentrate on a specific aspect of the longevity problem. Apart from the discussion whether this association is biological rather than social, many issues remain under debate, such as the gender-specificity of the patterns. In this chapter we address the effect of parental longevity on female mortality. Matthijs *et al.* (2002) found that for females born in the early $19^{\text{th}}$ century, parental longevity had a relatively strong effect, while for the male offspring, the effect is only present some decades later. The precise mechanisms underlying this association remained however unexplored.

The mitochondrial theory of ageing emphasises the adverse role of life-long mutation accumulation in the maternally inherited mitochondrial DNA (Korpelainen 1999). This should cause a greater maternally inherited genetic component in human life span. Analysing genealogies of a sample of noble families, Korpelainen (1999) indeed found such a maternal inherited longevity. Tanaka *et al.*'s (1998) research on Japanese centenarians support the concept that to carry an mtDNA genotype predisposing resistance to adult-onset diseases is one of the genetic factors for longevity. However, in a sample of mainly Russian noble families, Gavrilova *et al.*(1998b) did not find such a pattern (1998). They suggest a contradictory pattern of paternally dominated transmission, to be expected as a result of hemizygosity of genes on sex chromosomes in males.

Yet, not only genetic factors are important. Gavrilova *et al.*(1998c), for instance, points to the strong maternal-child interaction during in utero development and

later during the formative years of the child. Also "social" factors might be important. For instance, the intergenerational transmission of a weak position within the intra-household resource competition could cause a correlation of mother-daughter longevity. That is, in those households with strong resource competition (very poor families), one can expect negative effects on female longevity. If such a household's characteristic is transmitted from generation to generation, which is not improbable, then we could expect that especially females' longevity is correlated.

To explore this issue further, we need to address two specific questions. The first and most important question is whether we effectively find differences in the influence of fathers and mothers on the transmission of longevity. The second question is whether the transmission of mortality is age-related. Oftentimes, the analysis of longevity is limited to persons over age 50. The reason for this is that elimination of "phenotypic variation due to contagious diseases, accidents and war, and environmental maternal effect during early childhood" (Korpelainen, 1999) is necessary in order to detect the potentially maternally (or paternally) inherited genetic component in life span. Indeed, causes of death such as accidents and pregnancy-related diseases dominate the mortality pattern under that age and therefore may disturb the measurement. Furthermore, genetic variability for survival is expected to increase with age following the evolutionary theory of ageing and the mutation accumulation hypothesis in particular (Gavrilova *et al.*, 1998a). Of course, there are situations were a different choice of age cutoff is warranted. For example, Gavrilova *et al.* (1998abc) motivate the use of age 30 as a relevant cutoff.

If we find that the association is age-related, that is, if it is only visible at a later age, this will give credit to the views that there is a strong difference between adult and old age mortality. Yet, in a study of a population of French farmers, Cournil *et al.* (2000) found that the parent-daughter association is already visible at a relatively young age. Such a finding that there is no age-relatedness of association in longevity, will not exclude the existence of mother-child association at later age (e.g., due to mitochondrial processes). Nevertheless, if mortality of young adult females is mainly associated with the causes proposed by Korpelainen (i.e., accidents, contagious diseases, etc.), this finding gives some support for the view that alternative, social explanations are not to be excluded.

This chapter is organized as follows. In Section 6.2 we introduce the data and describe the village (Moerzeke) within which they were gathered. In Section 6.3 we present the statistical model used in the analysis. Proposals for statistical tests are

made in Section 6.4. Data from the Moerzeke study are analysed in Section 6.5. Section 6.6 is devoted to the impact of censoring on the analysis. Conclusions are formulated in Section 6.7.

## 6.2 Context of the Study

Moerzeke is a small village in the center of Flanders, the Dutch speaking part of Belgium, within the province of East Flanders (*Oost-Vlaanderen*). It is a geographical isolate as it is almost completely surrounded by the river Scheldt. Moerzeke was mainly populated by farmers until well into the $20^{\text{th}}$ century (De Ridder, 1984). During the second half of the $19^{\text{th}}$ century the rural textile industry gradually became more important. Before the First World War almost every inhabitant worked in agriculture (De Beule, 1962). After the War some (modest) industrial activity came to the village, but in 1947, sixty percent of the employed males were still involved in farming (De Beule, 1962).

The population of Moerzeke rose from approximately 2000 in 1761 to 4706 in 1950 (De Beule, 1962). Not surprisingly for a farmers' community, the mean age at first marriage was rather high, i.e., 31.3 years for men and 28.0 for women in the $19^{\text{th}}$ century. It rose from 1760 onwards (De Ridder, 1984) and peaked in the mid $19^{\text{th}}$ century. Also, fertility was traditionally high (De Ridder, 1984) and dropped at the beginning of the $20^{\text{th}}$ century. In the $18^{\text{th}}$ century major mortality crises (mainly dysentery) occurred, 24.8 percent of the children born in Moerzeke died within the first year, but these became less severe as the $18^{\text{th}}$ century progressed. Infant and childhood mortalities were strikingly high. Infant mortality did not drop until the first decades of the $20^{\text{th}}$ century. The life expectancy at age 50 steadily rose for those born in the $19^{\text{th}}$ century, reaching a peak at the end of the observation period (those born after 1850). For the group under study, the mean age at death for those who were born and deceased in Moerzeke was 71.9 years for men and 71.7 for women, respectively. In addition, the upper 10% percentiles for the lifespan are 83.3, 84.2, 84.8 and 84.4 years, for mothers, fathers, sons, and daughters, respectively.

The information in the Moerzeke database is drawn from church and civil registers. In Belgium, these sources are of good quality and appropriate for populations studies. The database contains all individuals who were born, married or died in Moerzeke.

## 6.3 Statistical Model

In this section, we will review the multivariate Dale model for survival times. This particular model augments standard univariate survival distributions for each of the family members (mother, father, and child) separately, with global odds ratios to describe the association between pairs of longevity outcomes. The main advantage of this modelling approach is that the univariate distributions derived from such a joint distribution are exactly equal to those that would be obtained were univariate analyses done on each outcome separately. In contrast, in the frailty model case (Clayton 1978, Hougaard 1986) the marginal distribution does not readily follow, implying that deriving, for example, the father's, mother's or children's marginal longevity distribution would be awkward.

Thus, a very attractive feature of this approach is the elegant way in which the association between the various longevities is modeled. This is important when one is interested in a separation between social and genetic aspects of longevity. Regarding the former, it is a strong asset that a number of covariates describing social and demographic aspects can be incorporated in the models (e.g., gender, parity, etc.). Social explanations often have empirical implications in terms of gender and parity differences. For example, if inheritance of material products, such as a farm, is gender and birth order related, then this must be reflected in the gender and parity pattern of longevity inheritance.

As indicated by Molenberghs and Lesaffre (1994), the multivariate Dale distribution can in principle be specified for any number of outcomes, using two-way and higher-order odds ratios to specify the associations. Such a specification is unavoidable should one choose for full maximum likelihood inference. However, calculations quickly become very cumbersome, not just in the binary or ordinal cases studied by these authors, but a fortiori so in the (possibly censored) survival time situation considered here. Therefore, we follow an alternative route, obviating the need to specify associations beyond the second order. Indeed, pseudo-likelihood ideas will be used to estimate the parameters. Building on Tibaldi *et al.* (2003), the procedure is detailed here while, in addition, a number of inferential tools are proposed. The methods will be applied to the Moerzeke study, where we consider the survival times $T_j$ of mother, father, and first child ($j = 1, 2, 3$) of 457 families with complete information on dates of death and the censored observations will be included in a second stage.

Suppose that, in addition, we observe a vector of covariates $Z$ and assume marginal

Weibull distributions for each survival time $T_j$, with density

$$f_{T_j}(t) = \lambda_{T_j} p_{T_j} (\lambda_{T_j} t)^{p_{T_j}-1} \exp(-(\lambda_{T_j} t)^{p_{T_j}}),$$

with $\lambda_{T_j}$, and $p_{T_j}$ the scale and shape parameters, respectively. The corresponding distribution for $T_j$ is

$$F_{T_j}(t) = 1 - \exp(-(\lambda_{T_j} t)^{p_{T_j}} \exp(\boldsymbol{z}\boldsymbol{\beta})) \tag{6.1}$$

where $\boldsymbol{\beta}$ is the vector of coefficients corresponding to covariates $\boldsymbol{z}$. While we focus on Weibull margins, different choices of univariate marginal survival distributions can be made, implying only relatively small modifications to the methodology, typically without major impact on the numerical values of the association parameters. Let us consider the individual information of family $i$ expressed in vector format as $(T_{i1}, T_{i2}, T_{i3}, \Delta_{i1}, \Delta_{i2}, \Delta_{i3}, z_{i1}, \ldots, z_{in_3})$ so that $\boldsymbol{W}_{ij} = (\boldsymbol{T}_i, \boldsymbol{\Delta}_i, Z_i)$ are the values for a particular cluster $i$ and survival time $j$ within cluster. The indicator $\Delta_{ij}$ indicates whether the lifetime is observed or not.

The pseudo-likelihood function to estimate the parameters of this model is constructed by considering all three possible pairs of outcomes on an family $(\boldsymbol{W}_{1r}, \boldsymbol{W}_{2\ell})$ $(\boldsymbol{W}_{1r}, \boldsymbol{W}_{3\ell})$ and $(\boldsymbol{W}_{2r}, \boldsymbol{W}_{3\ell})$. Those pairs produce $f_{T_rT_\ell}(\boldsymbol{W}_{ir}, \boldsymbol{W}_{i\ell})$ with $r < \ell$, $r = 1, 2, 3$ and $\ell = 1, 2, 3$, where $f_{T_rT_\ell}$ is the density function of the Plackett-Dale distribution defined in Chapter 5

Also in this case, the dependency can be defined using a *global cross-ratio* at $(t_r, t_\ell)$ given by $\theta_{r\ell}(t_r, t_\ell)$ and the Plackett distribution is obtained for constant cross-ratio $\theta_{r\ell}(t_r, t_\ell) \equiv \theta$. Based upon this distribution function, we can derive a bivariate Plackett *density* function $f_{T_rT_\ell}(t_r, t_\ell)$ for two survival times by calculating $\partial F_{T_rT_\ell}(t_r, t_\ell)/\partial t_r \partial t_\ell$ as follows

$$\begin{aligned}
& f_{T_rT_\ell}(t_r, t_\ell) \\
&= \frac{\partial^2 S(t_r, t_\ell)}{\partial t_r \partial t_\ell} f_{T_r}(t_r) f_{T_\ell}(t_\ell) \\
&= \Delta_r \Delta_\ell \log(f(t_r, t_\ell)) + \Delta_r(1-\Delta_\ell) \log\left(-\frac{\partial S(t_r, t_\ell)}{\partial t_r}\right) \\
&+ (1-\Delta_r)\Delta_\ell \log\left(-\frac{\partial S(t_r, t_\ell)}{\partial t_\ell}\right) + (1-\Delta_r)(1-\Delta_\ell)\log(S(t_r, t_\ell)),
\end{aligned} \tag{6.2}$$

where $S(t_r, t_\ell)$ is the joint survivorship function and $\Delta_i$ equals 1 if the survival time is observed and 0 otherwise. Next, we can use pseudo-likelihood methodology to obtain

estimates of $\boldsymbol{\phi}$ the vector of parameters. Specifically, $\boldsymbol{\phi}' = (\boldsymbol{\theta}', \boldsymbol{\beta}_T', \boldsymbol{\lambda}_T', \boldsymbol{p}_T')$ with $\boldsymbol{\theta}$ the subvector of association parameters, $\boldsymbol{\beta}_T$ the subvector of coefficients corresponding to the covariates $\boldsymbol{z}$ and, $\boldsymbol{\lambda}_T$, and $\boldsymbol{p}_T$ a subvector of parameters from the Weibull distribution.

The pseudo-likelihood estimator $\widehat{\boldsymbol{\phi}}$ is defined as the maximiser of $\ln p\ell(\boldsymbol{\phi})$. Consistency and asymptotic normality results provide an easy way to consistently estimate the asymptotic covariance matrix. Because $\widehat{\boldsymbol{\phi}}$ converges in probability to the true parameter value $\boldsymbol{\phi}_0$, and $\sqrt{N}(\widehat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)$ converges in distribution to normal distribution with covariance matrix $J(\boldsymbol{\phi}_0)^{-1}K(\boldsymbol{\phi}_0)J(\boldsymbol{\phi}_0)^{-1}$ where $J(\boldsymbol{\phi})$ is defined by (5.9) and $K(\boldsymbol{\phi})$ by (5.10). Indeed, the matrices $J$ and $K$ are found, as we showed in Chapter 5, from expression (5.11).

The Plackett-Dale model allows us to estimate and interpret the strength of the association between a pair of survival times $T_r$ and $T_\ell$, via global cross ratios (the $\theta$ parameters in the model). Nevertheless, it is often easier to work with a transformation of $\theta$ such as Spearman's $\rho$ or Kendall's $\tau$, which can be interpreted similar to Pearson's correlation coefficient. Kendall's $\tau$ ranges within the $[-1, 1]$ interval and a zero value implies independence between $T_r$ and $T_\ell$. There exists a relationship between Kendall's $\tau$ and $\theta$ for any copula $C(t_r, t_\ell, \theta)$ and independent of the marginal distributions as it can be seen in Chapter 3.

Kendall's $\tau$ thus measures the association between both time points after adjustment for the covariates used in the model. Estimates and confidence intervals, using the delta method, are accordingly easily obtained. There is no closed form for Kendall's $\tau$ in the Plackett-Dale case and an estimate has to be obtained directly using numerical integration. We have developed a SAS IML 8.02 macro to this effect.

Spearman's $\rho$ is also independent of the margins, and belongs to the unit interval. The relationship between Spearman's $\rho$ and $\theta$ is

$$\rho(\theta) = \frac{\theta + 1}{\theta - 1} - \frac{2\theta \ln \theta}{(\theta - 1)^2}. \tag{6.3}$$

An estimate follows from $\widehat{\rho} = \rho(\widehat{\theta})$, with variance estimated using the delta method. Figure 6.1 graphically displays the relationships between these three quantities ($\theta$, $\rho$, and $\tau$). The $(\rho, \tau)$ plot shows an almost linear relationship. Depending on the context, one can choose one of these three quantities to study association. By comparing the expressions to compute Spearman's and Kendall's coefficients, we observe that the computation of $\tau$ is more complex given that this involves numerical integration. In contrast, $\rho$ is very easy to obtain by plugging the estimated value of $\theta$ in formula (6.3).

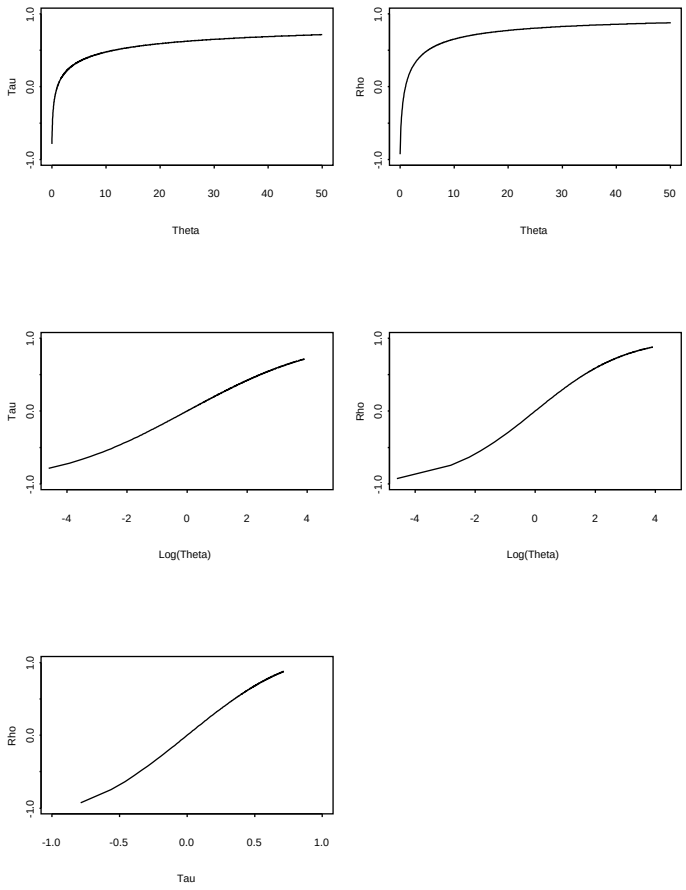


Figure 6.1: *Relationship between* $\theta$, $\log(\theta)$, $\tau$ *and* $\rho$ *plotted in pairs.*

From the shape of the curves represented in Figure 6.1 we observe that for values of $\theta$ larger than 10 the curves are quite stable and one really needs extreme values for $\theta$ to get values of $\tau$ and $\rho$ close to 1. Of course, the independence case ($\tau = 0$ and $\rho = 0$) corresponds to $\theta = 1$, in line with expectation.

Even though the use of marginal copula models, like the one proposed here, can be very much motivated by substantive considerations as given at the start of the section, one may want to reflect on the adequacy of model fit. Wang and Wells (2000)

developed such a method, confined though to Archimedean copulas. In principle, such an approach could be extended to copulas of the non-Archimedean type such as the one considered here. However, this is outside of the scope of this work. At any rate, an advantage of the Dale approach is that the choice for the univariate marginal distributions is independent of the association model choice. For example, should one want to replace the Weibull model with a different one, such as the Gompertz-Makeham distribution chosen by Gavrilova and Gavrilov (1991), then this is perfectly possible within this framework.

## 6.4 Test Statistics

The association between longevity of family members can be estimated with the pseudo-likelihood method presented in Section 6.3. In the case of maximum likelihood estimation, several tools can be used to test the parameters of the model such as Wald, score, or likelihood ratio tests. However, those tests need to be extended in our case as it was done by Geys *et al.* (1999) in the presence of clustered multivariate binary data. While point estimation and asymptotic normality have already been established in Section 6.3, the following subsections are devoted to the construction of the pseudo-likelihood counterparts to classical inferential tools such as ratio test statistics and score test statistics. Particularly, to perform a test for the association parameters of the model, we need to extend the Wald, score, and likelihood ratio test statistics to the pseudo-likelihood framework. The strategies proposed here are not restricted to those parameters and it can be applied to any other model parameter.

Association parameters $\theta_{ij}$ equaling one indicate independence between $T_i$ and $T_j$. This can be translated in terms of hypotheses such as

$$\mathrm{H}_0 : \theta_{r\ell} = 1 \qquad \theta_{r\ell} \in I\!R_{\geq 0} \qquad r, \ell = 1, 2, 3.$$

More generally, let us assume we are interested in an hypothesis of the type $\mathrm{H}_0 : \ \varphi = \varphi_0$ where $\varphi$ denotes a $q$-dimensional subvector of the $p$-dimensional vector of regression parameters $\boldsymbol{\phi}$ and write $\boldsymbol{\phi} = (\varphi', \beta')'$. We will discuss a number of tests in turn, whereafter they will be applied.

### 6.4.1 Wald Test Statistics

To construct this test we will use the asymptotic normality properties of the pseudo-likelihood estimators. We use the following result

$$W^* = N(\widehat{\varphi} - \varphi_0)' \Sigma^{-1}_{\varphi\varphi} (\widehat{\varphi} - \varphi_0) \sim \chi^2_q.$$

In this expression, $\Sigma_{\varphi\varphi}$ denotes the $q \times q$ submatrix of $\Sigma = J^{-1}KJ$. The matrices $J$ and $K$ were defined according to (5.9) and (5.10). The matrix $\Sigma$ can be estimated by using the pseudo-likelihood estimate $\widehat{\boldsymbol{\phi}}$. Thus, the Wald statistic is very easy to obtain and the more convienient one in cases where model fitting is very time consuming. However, it is highly sensitive to changes in parameterization as it was noted by Fears *et al.* (1996). We can see, via the delta method, that the value of the Wald statistic used for the hypothesis $\mathrm{H}_0 : \varphi = 0$ doubles the one corresponding to $\mathrm{H}_0 : \varphi^2 = 0$.

In this study, the fact that individual association parameters will be tested implies $\varphi_0 = \theta_{r\ell}$ with $r, \ell = 1, 2, 3$ therefore $q = 1$. Hence, $W^* \sim \chi^2_1$ and the normal distribution on the square root can be used to produce $p$-values.

### 6.4.2 Pseudo-score Test Statistics

This test is constructed by fitting the null model and it has the advantage over the Wald test that it is invariant to reparametrisation. Let us call $\boldsymbol{U}(\boldsymbol{\phi})$ the pseudo-score vector, specifically the derivative of the log of the pseudo-likelihood; and $\boldsymbol{U}_\varphi(\boldsymbol{\phi})$ the $q$-dimensional subvector. An empirically corrected version of this pseudo-score can be defined as

$$S^*(e.c) = \frac{1}{N}(\boldsymbol{U}_\varphi(\varphi_0, \hat{\beta}(\varphi_0))' J^{\varphi\varphi} \Sigma^{-1}_{\varphi\varphi} J^{\varphi\varphi} \boldsymbol{U}_\varphi(\varphi_0, \hat{\beta}(\varphi_0))),$$

where $\hat{\beta}(\varphi_0)$ is the maximum pseudo-likelihood estimator of $\beta$ when $\varphi$ is fixed to be $\varphi_0$, $J^{\varphi\varphi}$ is the $q \times q$ submatrix of the inverse of $J$, and $J^{\varphi\varphi}\Sigma^{-1}_{\varphi\varphi}J^{\varphi\varphi}$ is evaluated under $H_0$. We will use the fact that under mild regularity conditions $S^*(e.c.) \sim \chi^2_r$. However, computational problems were observed by Rotnitzky and Jewell (1990) in the context of generalised estimating equations, therefore an alternative model based version is proposed as follows

$$S^*(m.b) = \frac{1}{N}(\boldsymbol{U}_\varphi(\varphi_0, \hat{\beta}(\varphi_0))' J^{\varphi\varphi} \boldsymbol{U}_\varphi(\varphi_0, \hat{\beta}(\varphi_0))).$$

Its asymptotic distribution under $H_0$ is given by $\sum_{j=1}^{q} \eta_j \chi^2_{1(j)}$ where $\chi^2_{1(j)}$ are all independent random variables with $\chi^2_1$ distribution and $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_q$ are the eigenvalues of $(J^{\varphi\varphi})^{-1}\Sigma_{\varphi\varphi}$ under $H_0$.

To simplify calculations and to have a $\chi^2_q$ distribution we propose to compute the adjusted pseudo-score statistic, similar to Rotnizky and Jewell (1990), as follows:

$$S^*_a(m.b) = S^*(m.b)/\bar{\eta},$$

with

$$\bar{\eta} = \sum_{j=1}^{q} \eta_j/q.$$

Note that several adjustments have been proposed in literature by Rao and Scott (1987) and Robert, Rao, and Kumar (1987). One interesting feature of all tests is that in the maximum likelihood context all eigenvalues are equal to one and therefore all three statistics coincide. In our scalar case, $S^*(mb) = S^*_a(mb)$ holds because $q = 1$.

### 6.4.3 Pseudo-likelihood Ratio Test Statistics

Another proposal for testing $H_0$ is based on likelihood ratio ideas:

$$G^{*2} = 2[p\ell(\widehat{\boldsymbol{\phi}}) - p\ell(\varphi_0, \widehat{\beta}(\varphi_0))]$$

and is termed pseudo-likelihood ratio test statistic. The asymptotic distribution of $G^*$ can be written as $\sum_{j=1}^{q} \eta_j \chi^2_{1(j)}$, with $\chi^2_{1(j)}$ independently distributed according to $\chi^2_1$ and $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_r$ the eigenvalues of $(J^{\varphi\varphi})^{-1}\Sigma_{\varphi\varphi}$ under $H_0$ as before.

Similarly, we can define an adjusted pseudo-score statistic:

$$G^{*2}_a = G^{*2}/\bar{\eta},$$

that can be approximated by $\chi^2_q$. It can be shown that $G^{*2}$ can be expressed as an approximation to a Wald test. Note that, when applying the pseudo-likelihood ratio test, the model needs to be fitted twice, for the full and the reduced models, potentially making the procedure more time consuming. It is well known from (pseudo-)likelihood theory that the Wald test is the one with lowest power. However, from a practical point of view it is the more convenient one. All test statistics have been implemented using the SAS IML procedure.

## 6.5 Analysis of Moerzeke Data

The proposed methodology is used to analyze the Moerzeke data, making it the first application of this particular model to data of a familial type. To proceed we first fit

a multivariate Plackett-Dale model. Second, inferences are made by using the tests proposed in Section 6.4. Before we turn to the results, we give a short description of the model and the group selection procedure.

For reasons outlined in Section 6.1, we direct the analysis towards the questions of scientific interest. First, we carry out the analysis on the persons who were born and died in Moerzeke (the so-called "stable population"). In the next section, we will introduce a strategy to use censored lifetimes in order to incorporate some information from those individuals whose demographic data were for some reasons (mainly migration) not completely recorded. However, this does not provide much extra information. Second, we restrict the analysis to a subgroup of families having at least one child. For families having more than one child we will investigate the association of the parents and the oldest child where families having no children were excluded. Other approaches using the complete cluster of all family members can of course also be considered and are a topic of future research. Third, we restrict the analysis to families whose fathers were born between 1750 and 1830. From earlier studies we know that for this group, familial transmittance of longevity to daughters is relatively large (Matthijs *et al.* 2002). In this study, we address whether this association is mainly maternally or paternally transmitted.

The model contains three "longevity" variables: for the father, mother, and child, respectively. The longevity of a family member is measured by number of days lived, even though we use another scale in this analysis for numerical reasons (without any impact on the analysis' results). The year-of-birth of each family member and the gender (0 for females, 1 for males) of the child are included as covariates into model (6.1). Note that, of course, each family member's outcome is affected by a different year of birth, while a common gender-of-child effect is assumed for all members of the same family. This leads to the following parameters: $\beta_{\text{YB1}}$ year of birth of the mother, $\beta_{\text{YB2}}$ year of birth of the father, and $\beta_{\text{YB3}}$ year of birth of the child. The gender of the child, $\beta_{\text{G}}$, was included in all marginal Weibull distributions.

It is known that differences in longevity between males and females are not the same in all mortality groups. Whether or not we consider a cut-off point at 50 years for the age at death, differences are clearly seen. In Figure 6.2, we plotted the crude (i.e., unadjusted for year of birth) estimated survival curves for sons and daughters in three different age groups.

The selected group contains all families in which both father and mother reach the age of 50, while the child reaches at least the age of 10. We perform several analyses

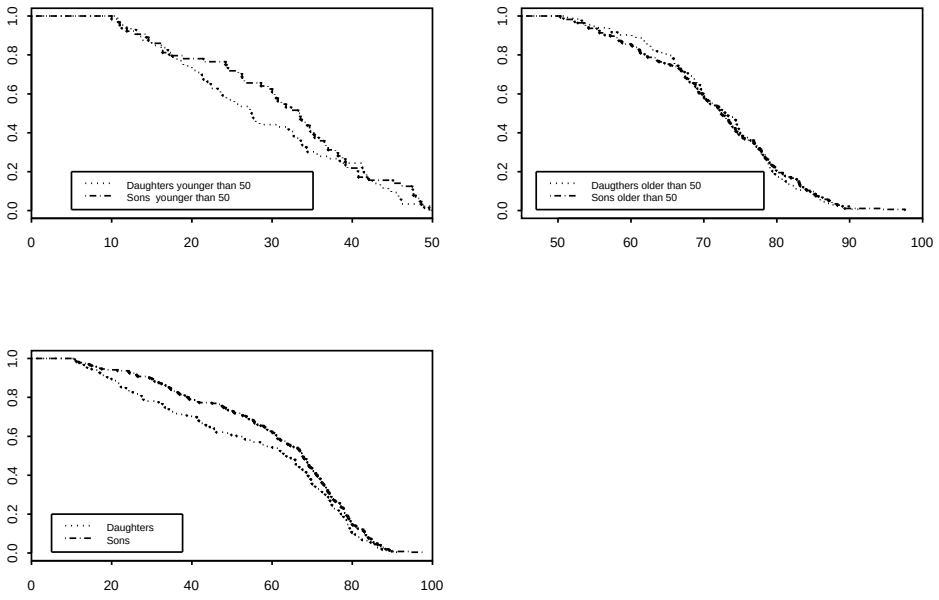


Figure 6.2: *Moerzeke Study: Survival curves for sons and daughters with a cutoff point of 50 years.*

with this model and selection of groups. First, we fit a multivariate Plackett-Dale model and estimate its parameters. The Weibull parameters $p_1$, $p_2$, $p_3$, $\lambda_1$, $\lambda_2$, and $\lambda_3$ also result from the fitting procedure. Table 6.1 shows the pseudo-likelihood estimates of the parameters. The estimated association parameter between mother and child is 1.349 (95% confidence interval $[1.002; 1.696]$), indicating a positive association between them. However, for father-child the value seems to be lower (0.983; not statistically significant). In addition, values of Spearman's and Kendall's coefficients, together with their confidence intervals, are presented in Table 6.2.

Second, inferences are made using the tests defined in Section 6.4. The null hypothesis of no association was tested in each case via the Wald, score, and pseudo-likelihood ratio tests and the results are displayed in Table 6.3.

From Table 6.3, similar conclusions are obtained irrespective of the test applied, while the Wald statistics gives the least significant $p$-value. Observe that the null hypotheses of no association between father's and mother's longevities on the one

Table 6.1: *Moerzeke Study: Model for father, mother, and child (son or daughter). Pseudo-likelihood estimates (empirically corrected standard errors) of the survival times.*

| Effect | Parameter | Estimates (s.e.) |
|---|---|---|
| Association (1,2) | $\theta_{12}$ | 1.136 (0.160) |
| Association (1,3) | $\theta_{13}$ | 1.349 (0.177) |
| Association (2,3) | $\theta_{23}$ | 0.983 (0.133) |
| Gender | $\beta_{\mathrm{G}}$ | -0.113 (0.041) |
| Year of birth of mother | $\beta_{\mathrm{YB1}}$ | -1.067 (1.313) |
| Year of birth of father | $\beta_{\mathrm{YB2}}$ | -0.899 (1.523) |
| Year of birth of child | $\beta_{\mathrm{YB3}}$ | -3.800 (1.293) |
| Shape parameter mother | $p_1$ | 4.799 (0.167) |
| Shape parameter father | $p_2$ | 5.770 (0.194) |
| Shape parameter child | $p_3$ | 3.002 (0.134) |
| Scale parameter mother | $\lambda_1$ | 0.215 (0.491) |
| Scale parameter father | $\lambda_2$ | 0.185 (0.473) |
| Scale parameter child | $\lambda_3$ | 1.587 (0.782) |

hand and father's and child's longevities on the other hand ($\theta_{12} = 1$ and $\theta_{23} = 1$) cannot be rejected, but the situation is different for mother and child. Indeed, we reject $\theta_{13} = 1$. The latter was already reflected in the fact that the 95% confidence intervals for Kendall's and Spearman's coefficients contain the zero value for the first and third hypotheses but not for the second one.

To explore this topic in more detail, we apply the model to different subsets. First, we make a distinction between sons and daughters. It is known that differences in mortality between males and females are not the same for all age groups. Figure 6.2 plots the estimated survival unadjusted curves for sons and daughters in three different groups for our data (same selection as outlined above). One possible way of tackling this problem in a simple way is by performing similar analysis in each group, i.e., sons and daughters separately. This is not the most efficient strategy but it will give us some idea about the hypotheses that the association depends on gender of the offspring.

Table 6.4 displays gender specific parameter estimates. We also performed the test for the association parameters and we can see from Table 6.5 that for sons there

Table 6.2: *Moerzeke Study: Model for father, mother, and child (son or daughter). Pseudo-likelihood estimates and inference for the association parameters $\theta$, Kendall and Spearman coefficients (95 % confidence intervals).*

| $(i,j)$ | $\theta_{ij}$ | $\log(\theta_{ij})$ | Kendall's $\tau_{ij}$ | Spearman's $\rho_{ij}$ |
|---|---|---|---|---|
| $(1,2)$ | 1.136 (0.823;1.449) | 0.127 (-0.013;0.268) | 0.028 (-0.013;0.044) | 0.042 (-0.049;0.134) |
| $(1,3)$ | 1.349 (1.002;1.696) | 0.299 (0.168;0.430) | 0.066 (0.052;0.081) | 0.099 (0.015;0.184) |
| $(2,3)$ | 0.983 (0.723;1.243) | -0.017 (-0.152;0.117) | -0.004 (-0.019;0.011) | -0.006 (-0.094;0.083) |

Table 6.3: *Moerzeke Study: Model for father, mother, and child (son or daughter). Pseudo-likelihood tests and theirs p-values*

| $H_0$ | Wald | $p$-value | $G^{*2}$ | $p$-value | $S^*(mb)$ | $p$-value | $S^*(ec)$ | $p$-value |
|---|---|---|---|---|---|---|---|---|
| $\theta_{12}=1$ | 0.854 | 0.395 | 0.922 | 0.337 | 1.224 | 0.269 | 0.821 | 0.365 |
| $\theta_{13}=1$ | 1.969 | 0.048 | 5.637 | 0.018 | 4.434 | 0.035 | 5.361 | 0.021 |
| $\theta_{23}=1$ | -0.128 | 0.898 | 0.017 | 0.896 | 0.025 | 0.874 | 0.016 | 0.899 |

are no significant differences at the 10% level, while for daughters there seems to be a stronger association in case of mothers and daughters than for the rest of the association parameters ($\theta_{13}$). The $p$-values corresponding to the Wald statistics are greater than 0.05 but the rest of the pseudo-likelihood tests gave significant results at the 10% level.

Second, we want to explore whether associations between families can also depend on the age at death of the offsprings (see Section 6.3). Therefore, we propose to fit the Plackett-Dale model in six different groups, i.e., we consider models for overlapping sets of offspring, reaching at least the age of 10, 20, 30, 40, 50, and 60 years, respectively. On the other hand this produces a decreasing number of observations in column $n$ from Table 6.6.

The results in Table 6.6 show the parameter estimates per subgroup. We need to interpret these results carefully, because of low (and different) numbers of observations. However, it seems clear that the association between mother and daughter is not gradually becoming stronger when stepwise excluding those daughters who have died at an early age. On the other hand, parameters for those who have reached the

Table 6.4: *Moerzeke Study: Model for father, mother, and child (son and daughter separately). Pseudo-likelihood estimates of the survival times.*

| | | Estimates (s.e.) | |
|---|---|---|---|
| Effect | Parameter | Sons | Daughters |
| Association (1,2) | $\theta_{12}$ | 0.979 (0.200) | 1.350 (0.266) |
| Association (1,3) | $\theta_{13}$ | 1.262 (0.244) | 1.413 (0.248) |
| Association (2,3) | $\theta_{23}$ | 0.953 (0.194) | 1.033 (0.188) |
| Year of birth of mother | $\beta_{\text{YB1}}$ | -2.732 (1.721) | 0.997 (2.069) |
| Year of birth of father | $\beta_{\text{YB2}}$ | 0.492 (2.179) | -2.360 (2.121) |
| Year of birth of child | $\beta_{\text{YB3}}$ | -5.149 (1.905) | -2.671 (1.827) |
| Shape parameter mother | $p_1$ | 4.778 (0.242) | 4.829 (0.234) |
| Shape parameter father | $p_2$ | 5.853 (0.267) | 5.689 (0.281) |
| Shape parameter child | $p_3$ | 3.498 (0.232) | 2.612 (0.152) |
| Scale parameter mother | $\lambda_1$ | 0.395 (0.645) | 0.100 (0.768) |
| Scale parameter father | $\lambda_2$ | 0.119 (0.667) | 0.295 (0.668) |
| Scale parameter child | $\lambda_3$ | 2.181 (0.977) | 1.044 (1.280) |

age of 50 are generally lower. This finding is somewhat surprising in the light of, for example, the findings of Korpelainen (1999; see Section 6.1). Nevertheless, the same findings are recovered for a set of French agricultural villages (Cournil *et al.*, 2000). It must be added that the fact that the association is visible for daughters at an early age, is not against a social explanation of the transmittance of mortality that focuses on the position of women in intra-household resource competition. It might not be excluded that the weak position of women within some households has effects at almost every age.

A graphical summary of the log of the association values (and their 95 % confidence intervals) is given in Figure 6.3, where we plot all three $\log(\theta)$ parameters for each group using as cut-off-point the age of mortality of the offspring.

Each set of parameters represents, from left to right, $\log(\theta_{12})$, $\log(\theta_{13})$, and $\log(\theta_{23})$. Therefore, the second $\log(\theta)$ estimate in each group of three corresponds to the mother–child relationship; this particular relationship typically exhibits a stronger association than for the other pairs. In addition to that for the group containing all offsprings dying older than 10 years, the confidence interval hardly contains the value of one, in agreement with the conclusions drawn from the tests before.

Table 6.5: *Moerzeke Study: Model for father, mother, and child (son and daughter separately). Pseudo-likelihood tests and theirs p-values for the association parameters for models from Table 6.4*

| Offspring | $H_0$ | Wald | $p$-value | $G^{*2}$ | $p$-value | $S^*(mb)$ | $p$-value | $S^*(ec)$ | $p$-value |
|---|---|---|---|---|---|---|---|---|---|
| Sons | $\theta_{12}=1$ | -0.104 | 0.916 | 0.013 | 0.910 | 0.022 | 0.883 | 0.011 | 0.917 |
| | $\theta_{13}=1$ | 1.074 | 0.282 | 1.656 | 0.198 | 1.373 | 0.241 | 1.502 | 0.220 |
| | $\theta_{23}=1$ | -0.244 | 0.807 | 0.066 | 0.798 | 0.121 | 0.728 | 0.057 | 0.812 |
| Daughters | $\theta_{12}=1$ | 1.316 | 0.188 | 2.473 | 0.116 | 2.521 | 0.112 | 2.283 | 0.131 |
| | $\theta_{13}=1$ | 1.663 | 0.091 | 3.884 | 0.049 | 2.952 | 0.086 | 3.928 | 0.048 |
| | $\theta_{23}=1$ | 0.177 | 0.859 | 0.033 | 0.8561 | 0.047 | 0.828 | 0.032 | 0.856 |

Moreover, in the second panel, associations between parents and daughters were plotted and there clearly is a different structure as opposed to the other two. Larger values of $\log(\theta_{13})$ are observed in almost every group. The latter result implies, once more, higher associations between longevity of mother and daughters. As we indicated before, some care has to be taken, given that the groups overlap. The differences observed in the length of the confidence intervals are due to the progressively decreasing sample sizes.

Finally, we will explore the influence of the age-at-death of the parents. Figure 6.4 gives a graphical display of their log's to ease interpretation of theses values. These confirm the previous findings that the association between mother and daughters is strongest and visible at all ages of the daughter. Consequently, this suggests that adult mortality of female family members is connected in a very general way, leading to associations in longevity between mothers and daughters irrespective of age groups. We must however add that interpretation should be done cautiously due to low numbers of observations.

Once again, in this case, qualitatively the same association picture is obtained, even though the significance has been removed.

## 6.6 The Impact of Censoring

In the work by Matthijs *et al.* (2002) censoring problems were avoided by limiting the analysis to a subset of the population with complete data on lifetimes.

Table 6.6: *Moerzeke Study: Model for father, mother, and child (all children; son and daughter separately); overlapping age groups. Pseudo-likelihood estimates (s.e.) of the association parameters between the survival times.*

| Year | $\theta_{12}$ | $p$-value | $\theta_{13}$ | $p$-value | $\theta_{23}$ | $p$-value | $n$ |
|---|---|---|---|---|---|---|---|
| All Children | | | | | | | |
| 10 | 1.136 (0.160) | 0.395 | 1.349 (0.177) | 0.048 | 0.983 (0.133) | 0.898 | 457 |
| 20 | 1.129 (0.164) | 0.431 | 1.338 (0.192) | 0.078 | 1.007 (0.148) | 0.962 | 421 |
| 30 | 1.079 (0.167) | 0.636 | 1.079 (0.167) | 0.636 | 0.974 (0.149) | 0.861 | 385 |
| 40 | 1.107 (0.181) | 0.554 | 1.303 (0.206) | 0.141 | 1.066 (0.192) | 0.731 | 342 |
| 50 | 1.096 (0.192) | 0.617 | 1.094 (0.175) | 0.591 | 1.077 (0.185) | 0.677 | 307 |
| 60 | 0.973 (0.184) | 0.883 | 0.950 (0.153) | 0.743 | 1.132 (0.216) | 0.541 | 269 |
| Sons | | | | | | | |
| 10 | 0.979 (0.200) | 0.916 | 1.262 (0.244) | 0.282 | 0.953 (0.194) | 0.808 | 238 |
| 20 | 1.038 (0.212) | 0.857 | 1.217 (0.239) | 0.363 | 1.112 (0.238) | 0.637 | 224 |
| 30 | 1.015 (0.213) | 0.943 | 1.194 (0.231) | 0.401 | 1.019 (0.216) | 0.929 | 214 |
| 40 | 1.136 (0.248) | 0.583 | 1.107 (0.219) | 0.625 | 1.190 (0.265) | 0.473 | 188 |
| 50 | 1.143 (0.269) | 0.595 | 1.029 (0.212) | 0.891 | 1.174 (0.261) | 0.504 | 174 |
| 60 | 1.038 (0.274) | 0.889 | 0.795 (0.158) | 0.194 | 1.270 (0.327) | 0.408 | 149 |
| Daughters | | | | | | | |
| 10 | 1.350 (0.266) | 0.188 | 1.413 (0.248) | 0.091 | 1.033 (0.188) | 0.860 | 219 |
| 20 | 1.265 (0.269) | 0.324 | 1.439 (0.296) | 0.138 | 0.940 (0.186) | 0.747 | 197 |
| 30 | 1.182 (0.278) | 0.512 | 1.427 (0.338) | 0.206 | 0.940 (0.210) | 0.775 | 171 |
| 40 | 1.098 (0.282) | 0.728 | 1.601 (0.400) | 0.132 | 0.929 (0.268) | 0.791 | 154 |
| 50 | 1.055 (0.288) | 0.848 | 1.200 (0.301) | 0.506 | 0.955 (0.250) | 0.857 | 133 |
| 60 | 0.928 (0.263) | 0.784 | 1.199 (0.323) | 0.537 | 0.977 (0.266) | 0.931 | 120 |

Family members with incomplete records were excluded from the analysis. However, in some situations, some information can be extracted from incomplete records. Even if not directly specified in the data, there are sometimes indication that the person was alive, and hence at risk of dying.

Essentially, we used two sources of information to recover additional observations. First, the date of marriage, if present, can be used as a censored time for lifetime of both parents. Second, the date of birth of the last child of the family can be used to define censoring for mother's survival time. By combining all of these strategies, only

Table 6.7: *Moerzeke Study: Model for father, mother, and child (son and daughter separately); overlapping age groups; parents dying at ages above 50 years. Pseudo-likelihood estimates (s.e.) of the association parameters between the survival times.*

| Interval | $\theta_{12}$ | $p$-value | $\theta_{13}$ | $p$-value | $\theta_{23}$ | $p$-value | $n$ |
|---|---|---|---|---|---|---|---|
| Sons | | | | | | | |
| 10 | 0.917 (0.223) | 0.708 | 1.191 (0.305) | 0.533 | 0.867 (0.207) | 0.519 | 163 |
| 20 | 0.970 (0.236) | 0.899 | 1.081 (0.293) | 0.781 | 0.996 (0.271) | 0.989 | 154 |
| 30 | 0.967 (0.245) | 0.892 | 1.255 (0.348) | 0.464 | 0.927 (0.246) | 0.768 | 149 |
| 40 | 1.188 (0.307) | 0.540 | 0.982 (0.252) | 0.944 | 1.039 (0.284) | 0.891 | 129 |
| 50 | 1.142 (0.303) | 0.640 | 0.943 (0.250) | 0.820 | 1.045 (0.290) | 0.877 | 121 |
| 60 | 1.205 (0.339) | 0.546 | 0.827 (0.230) | 0.451 | 1.362 (0.411) | 0.379 | 107 |
| Daughters | | | | | | | |
| 10 | 1.201 (0.303) | 0.507 | 1.518 (0.312) | 0.098 | 0.784 (0.191) | 0.258 | 147 |
| 20 | 0.970 (0.236) | 0.989 | 1.081 (0.293) | 0.269 | 0.996 (0.271) | 0.058 | 132 |
| 30 | 1.053 (0.315) | 0.866 | 1.272 (0.372) | 0.465 | 0.774 (0.204) | 0.268 | 116 |
| 40 | 1.015 (0.315) | 0.962 | 1.482 (0.482) | 0.317 | 0.715 (0.231) | 0.218 | 103 |
| 50 | 1.358 (0.450) | 0.425 | 1.281 (0.398) | 0.480 | 0.928 (0.275) | 0.794 | 91 |
| 60 | 1.335 (0.459) | 0.465 | 1.218 (0.416) | 0.600 | 1.050 (0.310) | 0.873 | 81 |

17 new families were incorporated and the analyses were repeated not only due to an expanded basis of inference, but also by way of sensitivity analysis. Table 6.8 contains the results of all fits. The first column shows the results for sons and daughters together; the last two columns show the results for each of the genders separately.

Interestingly enough, the previously obtained mild but significant association between mothers and children in terms of longevity has disappeared when additionally taking censored lifetimes into account. Of course, the differences between the $p$-values is fairly small and it is fair to say that, in all cases, evidence for an association between mothers and children is modest. We want to point out that these techniques easily allow one to work with censored times. Whether or not to include censored observations depends, of course, on the context.

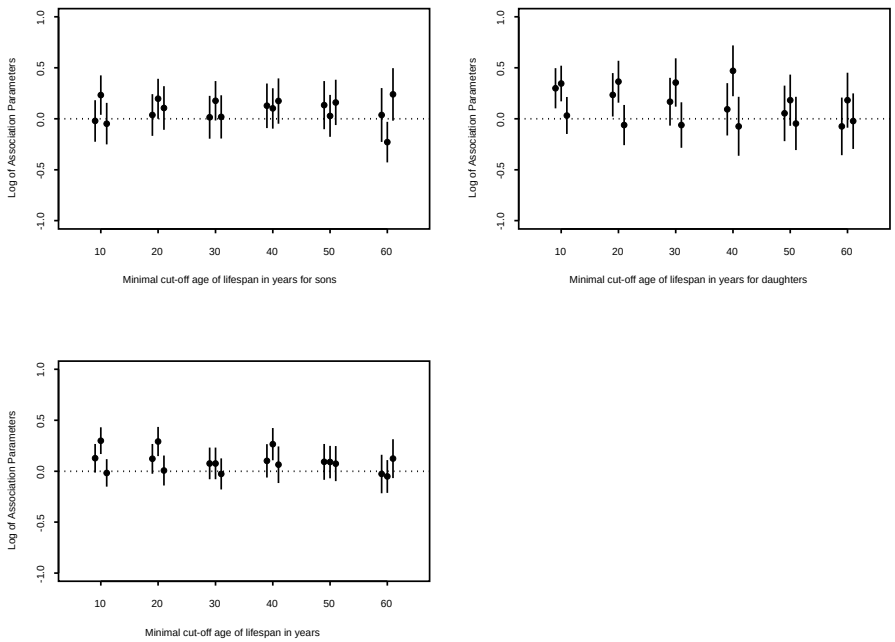


Figure 6.3: *Moerzeke Study: Log of association parameters $\theta_{12}$, $\theta_{13}$, and $\theta_{23}$ (from left to right) for offspring mortality group.*

## 6.7 Conclusions

In this chapter we have applied a multivariate Plackett-Dale model to study the inheritance of longevity in a Flemish village ($18^{\text{th}}$–$20^{\text{th}}$ century). The model was applied for the first time in a family study and, in particular, to a subset of the whole population with the characteristics explained in Section 6.5. The associations of the longevity between mother, father and the first child were estimated.

We proposed three different alternatives to perform inferences for the model parameters: Wald, pseudo-likelihood ratio, and score type tests. We illustrated how these test can be performed. Even if the Wald test is the one with less power, in this context we noticed that it is easily implemented from a computational point of view. Even though the pseudo-likelihood and pseudo-score tests are the most powerful, as was observed with other types of data (Geys *et al.* 1999), here it demands to fit

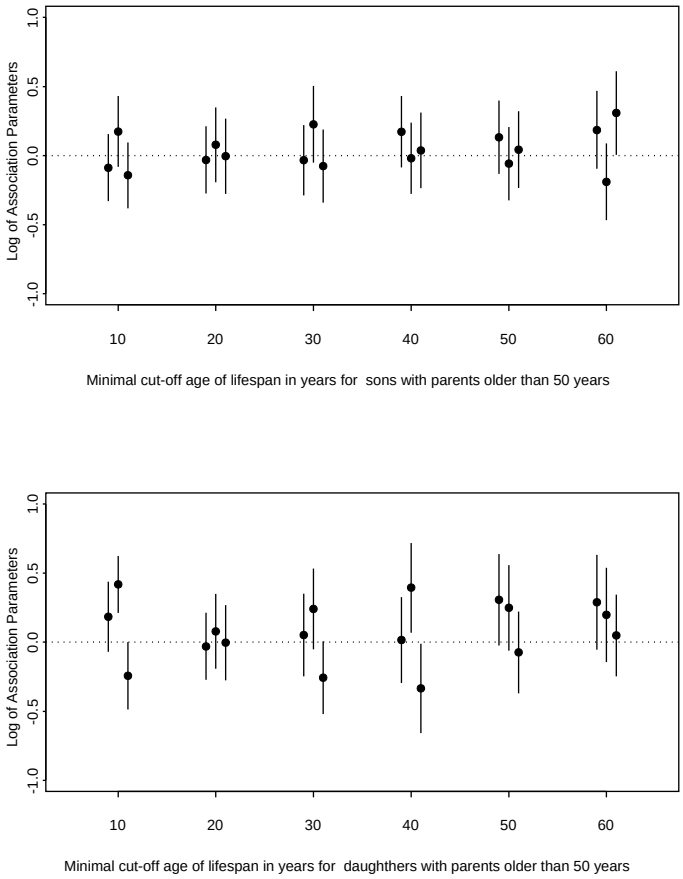


Figure 6.4: *Moerzeke Study: Log of associations for sons and daughters using intervals for mortality, considering only offspring with parents dying older than 50 years.*

different models. Given the complexity of these models it can be hard to obtain the building blocks needed to calculate these statistics.

Turning attention to the results regarding this study, we built on the analysis done in the original paper by Matthijs *et al.* (2002), using a different methodological approach. We defined a strategy for censoring to incorporate more observations into the analysis. However, the results did not change drastically and we opted to report

Table 6.8: *Moerzeke Study: Model for father, mother, and child (all children; son and daughter separately). Pseudo-likelihood estimates of the survival times. Censored observations are included*

| | | Estimates (s.e.) | | |
|---|---|---|---|---|
| Effect | Parameter | All children | Sons | Daughters |
| Association (1,2) | $\theta_{12}$ | 1.139 (0.160) | 0.981 (0.200) | 1.352 (0.266) |
| Association (1,3) | $\theta_{13}$ | 1.329 (0.171) | 1.282 (0.242) | 1.356 (0.235) |
| Association (2,3) | $\theta_{23}$ | 1.020 (0.137) | 0.982 (0.202) | 1.075 (0.194) |
| Gender | $\beta_{\mathrm{G}}$ | -0.117 (0.041) | – | – |
| Year of birth of mother | $\beta_{\mathrm{YB1}}$ | -0.713 (1.299) | -2.154 (1.687) | 1.115 (2.063) |
| Year of birth of father | $\beta_{\mathrm{YB2}}$ | -0.812 (1.511) | 0.315 (2.165) | -1.990 (2.103) |
| Year of birth of child | $\beta_{\mathrm{YB3}}$ | -3.760 (1.274) | -4.749 (1.892) | -3.010 (1.781) |
| Shape parameter mother | $p_1$ | 4.732 (0.162) | 4.698 (0.234) | 4.775 (0.225) |
| Shape parameter father | $p_2$ | 5.733 (0.191) | 5.799 (0.266) | 5.669 (0.274) |
| Shape parameter child | $p_3$ | 2.976 (0.129) | 3.455 (0.222) | 2.599 (0.147) |
| Scale parameter mother | $\lambda_1$ | 0.190 (0.493) | 0.323 (0.643) | 0.100 (0.775) |
| Scale parameter father | $\lambda_2$ | 0.182 (0.472) | 0.125 (0.669) | 0.264 (0.665) |
| Scale parameter child | $\lambda_3$ | 1.586 (0.777) | 1.831 (0.980) | 1.340 (1.256) |

results primarily for the complete cases. The main substantial conclusion is that significant associations were detected between mother and child. In a second step the associations were modelled within the group of daughters and sons separately and we observed significant associations between mother and daughter, but not between mother and sons.

One should make a careful distinction between the effect of a covariate, such as gender and year of birth, on the individual longevities on the one hand and on the association on the other hand. Our model has been conceived such that the longevities are allowed to depend on important covariates while the association between them is kept constant. Now, in line with Molenberghs and Lesaffre (1994), the association model can be extended to be covariate dependent rather than constant, but we consider this to be outside of the scope of this paper.

This finding confirms the role of the mother in the transmission of longevity. However, as these findings were present not only for mothers and daughters above the age of fifty, but also for mothers and daughters reaching at least the age of 10, this

finding does not support the view that familial association in adult mortality is only discernible at older ages. In our opinion, this finding is not in contradiction with social explanations that claim that females' weak position in intra-household competition leads to the association of mortality of females belonging to poor households. In other words, in some ages food competition placed the females in a weak position compared to the males within the family.

# Chapter 7

# Application of a Plackett-Dale Model to Study Associations in a Pilot Cancer Clinical Trial

## 7.1 Introduction

The work developed in this chapter was motivated by the need to find surrogate endpoints for survival of patients in oncology studies. In Chapter 2 we have introduced the basic ideas of surrogacy in a general framework with only one surrogate and one true endpoint. This methodology has been presented for normally distributed endpoints, in addition simplified strategies have been proposed within this setting.

Specifically, for endpoints of survival type, Burzykowski *et al.*, (2001) studied this situation by using a bivariate survival model with copula functions to model the associations between the margins. However, there are cases where more than one measurement is used as potential surrogate. Then, the need for multivariate models becomes clear. Even though most of the work in this area assumes that only one potential surrogate is going to be evaluated, Alonso *et al.*(2003) studied the validation of surrogates markers in multiple randomized clinical trials with repeated

measurements, where the concept of surrogacy was extended to settings with more than one surrogate marker. Further, studies involving survival times like in oncology were not extensively considered and it is a topic of ongoing research.

A starting point when studying surrogacy is to state a joint model for the true and the surrogate endpoints. Oftentimes, this task is far from trivial. For time-to-event responses with possibly censored outcomes the problem of finding such a model is even more complex, not many multivariate survival models are available.

In Chapter 5, a multivariate survival model to analyze non independent survival responses has been proposed, and to some extent, we have applied it to different settings with three survival outcomes. In this chapter, we intent to determine associations between five time-to-event outcomes, coming from three clinical trials for non small cell lung cancer. In particular, we use the multivariate Dale model for time-to-event and we fit the model to these data, using a pseudo-likelihood approach to estimate the model parameters.

We evaluate and compare the performance of different dimensional models and we relate the Dale model association parameter, i.e., the odds ratio, to well known quantities such as Kendall's $\tau$ and Spearman's $\rho$. Finally, the results are discussed with a perspective on surrogate marker validation. Some suggestions are made regarding further studies in this field.

Survival time of patients is one of the most common outcomes when assessing response to treatment in cancer clinical trials. While tumor response or percentage of tumor shrinkage has been used as a surrogate endpoint for cytotoxic drugs, it has been questioned at several occasions (Anderson *et al.*, 1983; Ellenberg *et al.*, 1989; Buyse *et al.*, 1998). There is a need to detect potential surrogate endpoints to decrease costs, time, and/or to improve the quality of life of cancer patients. Appropriate models, considering the type of response (continuous, binary, time-to-event, etc.) have to be proposed and applied to this effect.

Here, we use a multivariate survival model to estimate associations between time-to-event responses, to explore surrogacy of candidate markers, potentially after adjustment for other factors. The model used here has the advantage that its association parameter, the odds ratio, can be translated without difficulty into quantities that are considered easier to interpret, such as Spearman's rank correlation coefficient $\rho$ or Kendall's $\tau$. Appropriate hypothesis tests can be applied to assess the strength of the association.

Survival-type models using copulas were developed by Burzykowski *et al.* (2001)

and extended by Tibaldi *et al.*, (2003) to the multivariate case by using pseudo-likelihood estimation of the parameters.

In Section 7.2, a pilot study in cancer vaccination for non small cell lung cancer (NSCLC) patients is described. The focus is the assessment of the association between five time-to-event outcomes, one of which can be considered the true endpoint from a surrogate marker point of view. The statistical model and pseudo-likelihood estimation of its parameters are presented in Section 7.3. The analysis of the data is presented in Section 7.4.

## 7.2 Clinical Trials for Non Small Cell Lung Cancer

Three pilot clinical trials were performed, with the aim of testing safety, immunogenicity, and survival of a therapeutic vaccine based on the epidermal growth factor (EGF) molecule in patients with advanced non small cell lung cancer (NSCLC)(Gonzalez *et al.*, 1998; Gonzalez *et al.*, 2002). A first pilot study tested the vaccine in 20 patients with NSCLC, randomized to the EGF vaccine with two different adjuvants Alum and Montanide ISA-51. The vaccine was administered in a 5 doses schedule for 51 days. Immunogenicity data were collected weekly during the treatment period and monthly during follow-up. The second pilot trial studied the same vaccines in an additional group of 20 patients, but with common 3 days pre-treatment with cyclosphosphamide. In the third trial, 21 patients were assigned randomly to two different EGF vaccine doses.

In all three trials, the scope of patients is reduced to very advanced cancer patients at stages III, IIIb or IV without any other alternative of oncospecific treatment, with ECOG performance status less than 3. Survival time was considered from the day of random treatment assignment until the day of death, regardless of its cause. There were three participating hospitals. The mechanism of vaccine activity ought to induce an anti-tumoral immune response. Time to a good immune response could be an indicator of a possible clinical effect. The quality of the immune response is assessed by its titer and the titer ratio with respect to the baseline value.

The immune response was measured through an immuno-enzymatic experiment (ELISA). The titer was defined as the highest dilution for which the sample develops a significant intensity compared to the control sample. An immune response, 2X, is obtained when the sample achieves a titer value two times greater than the control sample. An immune response 1:2000 and 4X, is obtained when the sample achieves a

titer value four times greater than the control sample at least at a level 1:2000.

We consider five time-to-event outcomes. Time 1 is time to response immunogenecity 2X; Time 2 is time to response of immunogenecity 1:2000 and 4X; Time 3 is time to maximum titer; TTP is time to progression and TSV is overall survival time. The latter is the true endpoint whereas the earlier can be seen as four potential surrogate endpoints. All times are expressed in months. Available covariate information includes age (on a continuous scale), disease stage (categories III, IIIb, and IV), indicator for patient's previous treatment (e.g., chemotherapy), and, of course, treatment assignment.

In a previous analysis, a relationship between immunological response and survival time was detected (Torres *et al.*, 2001). For one of the trials, there was a clear advantage on survival for the group of high immunological responders (Torres *et al.*, 2002).

## 7.3 Statistical Model

Let us consider a trial involving $N$ subjects with $k$ time-to-event measurements. In our case study, $k = 5$ with times Time 1, Time 2, Time 3, TTP, and TSV defined in Section 7.2.

Suppose that we also observe a vector of covariates $Z$ and assume a Weibull distribution for each time $T_j$ with $\lambda_{T_j}$ and $p_{T_j}$ the scale and shape parameters, respectively. For any pair of survival times $(T_1, T_2)$ assume that $T_1$ and $T_2$ are correlated survival times, then the joint survival function can be written as

$$S_{T_1T_2}(t_1,\, t_2) = P(T_1 \geq t_1,\, T_2 \geq t_2) = C_{\theta_{12}}\{S_{T_1}(t_1),\, S_{T_2}(t_2)\}, \quad t_1,\, t_2 \geq 0, \tag{7.1}$$

where $S_{T_1}$ and $S_{T_2}$ denote marginal survival functions and $C_{\theta_{12}}$ is the Plackett copula (Chapter 3). To model the effect of specific covariates on the marginal distributions of $T_1$ and $T_2$ in (7.1) we propose to use the proportional hazard model:

$$S_{T_1}(t_1) \;=\; \exp\left\{-\int_0^{t_1} h_{T_1}(x)\exp(\boldsymbol{\beta}_{T_1}Z_1)dx\right\}, \tag{7.2}$$

$$S_{T_2}(t_2) \;=\; \exp\left\{-\int_0^{t_2} h_{T_2}(x)\exp(\boldsymbol{\beta}_{T_2}Z_2)dx\right\}, \tag{7.3}$$

where $h_{T_1}$ and $h_{T_2}$ are marginal baseline hazard functions and $\boldsymbol{\beta}_{T_1}$ and $\boldsymbol{\beta}_{T_2}$ are vectors of unknown regression parameters corresponding to the covariates $Z$. The classical

model proposed by Cox (1972) is used for the hazard functions. In this case the dependence can be defined using a *global cross-ratio* $\theta_{12}$ at defined in Section (3.6).

While we will focus on Weibull marginals, choosing different univariate marginal survival distributions will not induce additional complexities. The choice of marginal survivorship functions will, of course, impact the fit of the marginal outcomes but is expected to have less impact on the estimated values of the association parameters. Express the observed information on individual $i$ in the format: $(T_{i1}, \ldots, T_{ik}, \Delta_{i1}, \ldots, \Delta_{ik}, z_{i1}, \ldots, z_{in_k})$ so that $\boldsymbol{W}_{ij} = (T_{ij}, \Delta_{ij}, Z_i)$ are the values for a particular subject $i$ and time point $j$, with $j = 1, \ldots, k$.

As it was done before also pseudo-likelihood estimation is used here. It is well known that full maximum likelihood estimation can become prohibitive for many (marginal) models in particular in this study having five outcomes to be considered. The pseudo-likelihood function constructed for the estimation of the parameters of this model is based on considering all (in our case, ten) possible pairs of outcomes on an individual $(\boldsymbol{W}_{ir}, \boldsymbol{W}_{i\ell})$ with $1 \leq r < \ell \leq 5$. These pairs produce $f_{T_rT_\ell}(\boldsymbol{W}_{ir}, \boldsymbol{W}_{i\ell})$, where $f_{T_rT_\ell}$ is the density function of the Plackett-Dale distribution defined in Chapter 3.

Precisely, we can define the pseudo-likelihood function $PL$ through its logarithm, that in this case is expressed by

$$p\ell(\boldsymbol{\phi}) = \sum_{i=1}^{N} \sum_{(r,\ell)\in S} \ln f_{T_rT_\ell}(\boldsymbol{W}_{ir}, \boldsymbol{W}_{i\ell}, \boldsymbol{\phi}), \tag{7.4}$$

where $S = \{(1,2),(1,3),(1,4),(1,5),(2,3),(2,4),(2,5),(3,4),(3,5),(4,5)\}$ is the set of all ten possible pairs of outcomes, $f_{T_rT_\ell}$ is the value of the function defined earlier and evaluated in the corresponding outcomes for subject $i$, and $\boldsymbol{\phi}$ is the vector of parameters, $\boldsymbol{\phi}' = (\boldsymbol{\theta}', \boldsymbol{\beta}_T', \boldsymbol{\lambda}_T', \boldsymbol{p}_T')$, with $\boldsymbol{\theta}$ the subvector of association parameters, $\boldsymbol{\beta}_T$ the subvector of coefficients corresponding to the covariates $Z$ and, $\boldsymbol{\lambda}_T$ and $\boldsymbol{p}_T$ subvector of parameters from the Weibull distribution.

The pseudo-likelihood estimator $\widehat{\boldsymbol{\phi}}$ is defined as the maximizer of (7.4). The Plackett-Dale model allows us to estimate and interpret the strength of the association between a pair of survival times via global cross ratios, the $\theta$ parameters in the model. We consider a transformation of $\theta$ that has the interpretational properties of a correlation coefficient, this is the case of Spearman's $\rho$ or Kendall's $\tau$.

Kendall's $\tau$ measures the association between both time points after adjustment for the covariates used in the model; and it was defined in Section 5.4. Estimates and

confidence intervals, using the delta method, are accordingly easily obtained. Since there is no closed form for Kendall's $\tau$ in the Plackett-Dale case, an estimate has to be obtained from (5.11).

Spearman's $\rho$ is also independent of the margins, and belongs to the unit interval. An estimate can be obtained from $\widehat{\rho} = \rho(\widehat{\theta})$, with variance estimated using a straightforward application of the delta method.

This allows estimation of the associations between the five outcomes by fitting a multivariate model and adjusting for other variables: age of the patients, previous treatment status, stage of the diseases, etc., as we will see in Section 7.4.

Pseudo-likelihood estimates were obtained using Newton-Raphson with analytical first derivatives and numerical second derivatives, as implemented in SAS IML 8.02 and using routine NLPNRR (SAS Institute Inc. 1999–2001). Standard errors of the parameters were calculated using the inverse of the observed matrix of second derivatives.

This model has important implications in the assessment of surrogacy. In previous studies (Burzykowski *et al.*, 2001), the validation of a new variable as surrogate was performed on only one surrogate and only one true endpoint. In our case, the model allows to study several surrogates and several true endpoints at the same time. It gives also the possibility of developing new strategies not only to validate already identified candidates, but also to identify new variables that have potential regarding surrogacy.

Both Kendall's $\tau$ and Spearman's $\rho$ quantities can serve as an indication of individual level surrogacy in the sense of Buyse *et al.* (2000a). In case data are available from a sufficiently large number of trials and/or centers, these authors' meta-analytic perspectives can be adopted as well.

## 7.4 Analysis of the Data

In this section we will analyse the data introduced in Section 7.2. The overall median survival time achieved by the patients was 8.13 months with a range of 25 months. The median time to immune response given by time to response immunogenicity 2X (Time 1), time to response immunogenicity 1:2000 and 4X (Time 2) and time to achieve the maximum titer (Time 3) are less than 3 months. In addition, the time to progression endpoint has a median of 3 months. The average difference of these endpoints with respect to survival time is 5 months, giving an opportunity to look for

a possible surrogacy of the immune response. The age of the patients was on average 59 years, 57.38% of the patients were in stage IV and 54.1% were previously treated with other oncospecific treatments. The objective response in these advanced non-small cell lung cancer patients is hardly seen during their evaluation period, which is their lifetime. Hence, it would be of interest to assess the immune response endpoints and the time to progression as possible surrogates of the survival time.

We will now fit the proposed model to the data described in Section 7.2. Even when the association between outcomes is of primary scientific interest, as is the case here, it is mandatory to appropriately adjust the marginal survival regressions for covariate effects. We have included patients' characteristics: age (as a continuous variable), disease stage (three categories labelled III, IIIb, and IV), whether or not a patient received previous treatment (e.g., chemotherapy), and treatment arm. The time unit for the outcomes was months.

We will use the indices 1, 2, 3, 4 and 5 to identify the outcomes Time 1, Time 2, Time 3, TTP and TSV, respectively. Thus, for example, $\theta_{15}$ denotes the association between outcomes Time 1 and TSV. Note that the models of primary interest are those containing the variable TSV, considered to be the true endpoint in this study. Nevertheless, the other models are useful to further insight into the association structure.

In the first part of the analysis we explored the importance of hospital and trial to estimate the pairwise associations. We fitted all possible bivariate models using as covariates *age*, *stage*, *prevtrt* and *group* in four different situations. Firstly, we fitted models with the variables *hospital* and *trial*; secondly, with the variable *hospital* only; thirdly, with the variable *trial* only and fourthly with neither of these variables. The results, not shown here, reveal that no large differences were observed between the association parameters across the four choices, so it was decided to retain the simplest model and both *trial* and *hospital* were dropped from further consideration.

We first considered all possible bivariate models (1B to 10B) and all different trivariate models (1T to 10T). The association parameters obtained from these models, as well as those from the five-variate model (1F), are presented in Table 7.1. The primary use of the bivariate and trivariate models lies in their comparison with the full 5-variate model. Indeed, given the marginal nature of the models, corresponding associations have the same meaning. While each association occurs only once in the collection of bivariate models, they do so several times in the collection of trivariate models, disallowing their easy use. Similarly, each association is used only once in

the full 5-variate model. The most obvious advantage is that all associations feature within a single, integrated model. They are also estimated with increased precision as opposed to their bivariate and trivariate counterparts. The bivariate models are also useful to provide starting values for the 5-variate model. Indeed, the model is not easy to bring to convergence in the absence of reasonable starting values.

Let us zoom in on the comparison of association parameters across models. For example, the association between TTP and TSV, $\theta_{45}$, can be found from Models 10B, 1T, 2T, 3T, and 1F. The results are very similar, as can be seen in most other rows in Table 7.1, with somewhat exceptional behavior for $\theta_{12}$ and $\theta_{13}$. Such behavior is not uncommon for relatively large odds ratios, and the difference is less prominent on the log odds ratio scale.

Full details of the parameter estimates from the 5-variate model are given in Tables 7.2 and 7.3. Table 7.2 described the association parameters. Apart from the original odds-ratio scale ($\theta$ parameters), the easier-to-interpret Kendall's $\tau$ and the Spearman's $\rho$ coefficient are included, together with asymptotic 95% confidence intervals. The $\theta$-confidence intervals not containing one provide evidence for association between the corresponding pair of times, after correction for the covariates. Note that the covariates and other marginal regression parameters are displayed in Table 7.3. A corresponding association assessment based on Kendall's $\tau$ and Spearman's $\rho$ requires exclusion of the zero value from the corresponding confidence intervals.

Several substantive conclusions can be drawn from the model fits. From Model 1F we see that the highest association is observed between TTP and TSV. We further observe a significant association between Time 1 and Time 2. While the first two of these three associations are of direct interest, and may lead to reconsideration of Model 2T (containing Time 2, TTP, and TSV), it is of interest to consider a 4-variate model as well, i.e., a model with outcomes 1, 2, 4, and 5 (i.e., Time 1, Time 2, TTP, and TSV). Indeed, through its association with Time 2, Time 1 may indirectly contribute useful information. In any case, Time 3 appears to have no association with any of the other outcomes. Thus, a 4-variate model as presented in Table 7.4 will be considered our final model.

In summary, we have some evidence that TTP and Time 2 can be used as surrogates for TSV, with some auxiliary information coming from Time 1. Of course, the evidence apported here is just from three relatively small trials, and is based on an assessment of the association between responses only. Clearly, more exhaustive studies need to be designed in order to evaluate the surrogacy in a more authorita-

tive fashion, preferably in a meta-analytic setting such as the one proposed by Buyse *et al.* (2000a) or Burzykowski *et al.* (2001).

## 7.5 Conclusions

We have proposed the use of a multivariate Plackett-Dale model for estimating associations between, possibly censored, time-to-event outcomes. Specifically, we showed how this methodology can be useful in the context of surrogate marker validation.

Given the difficulties of manipulating the likelihood function in this case, a pseudo-likelihood approach was undertaken as a viable and attractive alternative to maximum likelihood. The computational complexity of the algorithm used for the estimation of the model parameters was overcome by using initial values obtained from the bivariate fitted models. Good numerical results were obtained in most cases.

Kendall's $\tau$ and Spearman's $\rho$ coefficients can be used as measures of individual level surrogacy (Burzykowski *et al.*, 2001). In spite of the multivariate flavor of this model, the pairwise pseudo-likelihood approach provides only bivariate association measures. Valid confidence intervals for such quantities were constructed using the delta method.

One of the primary purposes of this study was to detect or identify possible new surrogate endpoints for survival time. We are particularly interested in the validation of four different surrogate variables (Time 1, Time 2, Time 3 and TTP). This implies the need of a multivariate model considering all of these surrogates and the true endpoint.

We want to point out that the methodology we applied here focuses only on the individual level surrogacy but similar ideas as in Buyse *et al.* (2000a) for the meta-analytic framework need to be developed further.

Using the selected 4-variate model a high association between Time 1 and Time 2 can be observed. This evidenced that the time of reaching the double baseline titer, for most of the patients, had a strong relationship with the time to achieve a high titer value (1:2000 and 4X).

None of the times to reach a good immune response seem to have a high association with TTP, or with survival time. It seems that, with the accumulated evidences in this patient population, time to a good immune response is not strongly associated with the time to reach a good immune response. Other immune information seems to be more important and this should be the objective of further research. In addition,

Table 7.1: *Pilot Clinical Trial Study: Comparison of the association parameters $\theta$, obtained from bivariate (B), trivariate (T), and five-variate(F) models. Estimates of the association parameters and standard errors.*

| | Model | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Par. | 1T | 2T | 3T | 4T | 5T | 6T | 7T | 8T | 9T | 10T | 1F | 1B-10B |
| $\theta_{12}$ | – | – | – | 11.10 | – | – | 10.96 | – | – | 11.09 | **8.82** | 17.52 |
| | | | | (6.89) | | | (0.46) | | | (6.98) | **(2.79)** | (8.85) |
| $\theta_{13}$ | – | – | – | – | – | 5.58 | – | 5.61 | – | 5.81 | **4.79** | 8.93 |
| | | | | | | (5.04) | | (5.03) | | (5.12) | **(2.28)** | (5.14) |
| $\theta_{14}$ | 0.84 | – | – | – | – | – | 0.86 | 0.85 | – | – | **0.86** | 0.83 |
| | (0.48) | | | | | | (0.46) | (0.46) | | | **(0.26)** | (0.41) |
| $\theta_{15}$ | 0.71 | – | – | 0.71 | – | 0.72 | – | – | – | – | **0.72** | 0.68 |
| | (0.30) | | | (0.28) | | (0.29) | | | | | **(0.19)** | (0.31) |
| $\theta_{23}$ | – | – | – | – | 1.55 | – | – | – | 1.56 | 1.64 | **1.57** | 1.58 |
| | | | | | (0.76) | | | | (0.76) | (0.90) | **(0.41)** | (0.68) |
| $\theta_{24}$ | – | 1.04 | – | – | – | – | 1.06 | – | 1.07 | – | **1.05** | 1.07 |
| | | (0.60) | | | | | (0.46) | | (0.61) | | **(0.37)** | (0.50) |
| $\theta_{25}$ | – | 0.55 | – | 0.51 | 0.53 | – | – | – | – | – | **0.55** | 0.51 |
| | | (0.29) | | (0.28) | (0.28) | | | | | | **(0.17)** | (0.24) |
| $\theta_{34}$ | – | – | 1.06 | – | – | – | – | 1.06 | 1.05 | – | **1.06** | 1.05 |
| | | | (0.42) | | | | | (0.43) | (0.41) | | **(0.28)** | (0.39) |
| $\theta_{35}$ | – | – | 1.86 | – | 1.90 | 1.94 | – | – | – | – | **1.90** | 1.91 |
| | | | (0.73) | | (0.75) | (0.80) | | | | | **(0.52)** | (0.76) |
| $\theta_{45}$ | 11.22 | 11.17 | 11.16 | – | – | – | – | – | – | – | **10.56** | 11.93 |
| | (5.09) | (5.10) | (5.06) | | | | | | | | **(3.31)** | (4.67) |

if individual-level associations are weak, then time to immune response are unlikely to be good surrogates for survival, but this is a topic that deserves further analysis.

However, there is evidence that TTP is highly associated with survival time. In practice, this variable is not very convienient given its closeness to the actual survival time. The marginal gain does not justify its use as a surrogate.

Table 7.2: *Pilot Clinical Trial Study: Pseudo-likelihood estimates of the association parameters (95% confidence intervals) of the five-variate model, with outcomes Time1, Time2, Time3, TTP, and TSV. Apart from the original odds ratio scale, Kendall's $\tau$ and Spearman's $\rho$ are presented.*

| $(i,j)$ | $\theta_{ij}$ | Kendall's $\tau_{ij}$ | Spearman's $\rho_{ij}$ |
|---|---|---|---|
| $(1,2)$ | 8.821 (3.363;14.280) | 0.454 (0.426;0.482) | 0.628 (0.497;0.759) |
| $(1,3)$ | 4.790 (0.325;9.255) | 0.337 (0.290;0.384) | 0.483 (0.238;0.727) |
| $(1,4)$ | 0.857 (0.356;1.358) | -0.034 (-0.067;-0.002) | -0.051 (-0.246;0.143) |
| $(1,5)$ | 0.716 (0.348;1.083) | -0.074 (-0.103;-0.046) | -0.111 (-0.280;0.058) |
| $(2,3)$ | 1.565 (0.766;2.363) | 0.099 (0.071;0.127) | 0.148 (-0.019;0.315) |
| $(2,4)$ | 1.045 (0.311;1.779) | 0.010 (-0.029;0.049) | 0.015 (-0.219;0.249) |
| $(2,5)$ | 0.545 (0.209;0.881) | -0.134 (-0.168;-0.100) | -0.200 (-0.398;-0.002) |
| $(3,4)$ | 1.060 (0.521;1.599) | 0.013 (-0.015;0.041) | 0.019 (-0.150;0.189) |
| $(3,5)$ | 1.896 (0.882;2.910) | 0.141 (0.112;0.171) | 0.210 (0.039;0.381) |
| $(4,5)$ | 10.567 (4.088;17.046) | 0.487 (0.460;0.514) | 0.665 (0.544;0.785) |

Table 7.3: *Pilot Clinical Trial Study: Pseudo-likelihood estimates (standard errors) of the survival regression parameters in the five-variate model with outcomes Time1, Time2, Time3, TTP, and TSV.*

| Parameters | $k$ | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| $age_k$ | 0.404 (0.077) | 0.143 (0.060) | -0.103 (0.072) | -0.106 (0.057) | -0.220 (0.097) |
| $stage1_k$ | 0.746 (0.209) | -0.196 (0.216) | 0.235 (0.230) | -0.220 (0.146) | -0.143 (0.194) |
| $stage2_k$ | -0.789 (0.252) | -0.903 (0.241) | -0.472 (0.270) | 0.122 (0.180) | -0.007 (0.241) |
| $prvtrt_k$ | 0.001 (0.158) | -0.065 (0.137) | -0.326 (0.124) | -0.420 (0.124) | 0.004 (0.156) |
| $trt_k$ | 0.538 (0.165) | 1.251 (0.162) | 0.310 (0.142) | -0.208 (0.118) | -0.039 (0.141) |
| $p_k$ | 1.230 (0.053) | 0.903 (0.039) | 1.184 (0.039) | 1.085 (0.041) | 1.638 (0.066) |
| $\lambda_k$ | -2.659 (0.438) | -2.901 (0.551) | -0.665 (0.412) | -0.599 (0.284) | -1.539 (0.335) |

Table 7.4: *Pilot Clinical Trial Study: Pseudo-likelihood estimates (standard errors) of the survival regression and association parameters in four-variate model with outcomes Time1, Time2, TTP, and TSV.*

| Par. | Time1-Time2-TTP-SVT |
|---|---|
| $\theta_{12}$ | 9.441 (5.417) |
| $\theta_{14}$ | 0.856 (0.245) |
| $\theta_{15}$ | 0.712 (0.184) |
| $\theta_{24}$ | 1.041 (0.558) |
| $\theta_{25}$ | 0.543 (0.189) |
| $\theta_{45}$ | 10.820 (3.559) |
| $age_1$ | 0.424 (0.112) |
| $age_2$ | 0.141 (0.118) |
| $age_4$ | -0.109 (0.071) |
| $age_5$ | -0.209 (0.100) |
| $stage1_1$ | 0.826 (0.261) |
| $stage1_2$ | -0.164 (0.341) |
| $stage1_4$ | -0.216 (0.201) |
| $stage1_5$ | -0.140 (0.196) |
| $stage2_1$ | -0.758 (0.287) |
| $stage2_2$ | -0.904 (0.333) |
| $stage2_4$ | 0.133 (0.216) |
| $stage2_5$ | -0.022 (0.247) |
| $prvtrt_1$ | 0.001 (0.199) |
| $prvtrt_2$ | -0.072 (0.227) |
| $prvtrt_4$ | -0.422 (0.153) |
| $prvtrt_5$ | 0.008 (0.171) |
| $trt_1$ | 0.531 (0.185) |
| $trt_2$ | 1.240 (0.272) |
| $trt_4$ | -0.203 (0.143) |
| $trt_5$ | -0.057 (0.155) |
| $p_1$ | 1.231 (0.053) |
| $p_2$ | 0.881 (0.057) |
| $p_4$ | 1.081 (0.055) |
| $p_5$ | 1.630 (0.074) |
| $\lambda_1$ | -2.784 (0.697) |
| $\lambda_2$ | -2.928 (1.119) |
| $\lambda_4$ | -0.593 (0.391) |
| $\lambda_5$ | -1.561 (0.336) |

# Chapter 8

# Conditional Linear Mixed Models with Crossed Random-Effects

## 8.1 Introduction

The analysis of continuous hierarchical data such as, for example, repeated measures or data from meta-analyses can be done by means of the linear mixed-effects model (Verbeke and Molenberghs, 2000). However, in some situations this model, in its standard form, does pose computational problems. For example, when dealing with crossed random-effects models the estimation of the variance components becomes a non-trivial task if only one observation is available for each cross-classified level. Pseudo-likelihood ideas were used by Renard *et al.* (2002) in the context of binary data with standard generalized linear multilevel models. Also in this case the problem of the estimation of the variance remains non-trivial. We propose a method to fit a crossed random-effects model with two levels and continuous outcomes, by borrowing ideas from the conditional linear mixed effects models theory. We apply this method to a case study with data coming from the field of psychometry and study a series of items (responses) crossed with subjects. A simulation study assesses the operating characteristics of the method.

Models where population units are hierarchically structured have been studied

extensively for several decades. However, in many cases, units at the same level of a hierarchy are simultaneously classified by more than one factor. To fix ideas, let us consider the case where school pupils are classified by the school they attend as well as by the neighborhood they live in. Since schools usually attract pupils from several neighborhoods and children from the same neighborhood usually attend several schools, these two factors are crossed, i.e., one factor is not nested within the levels of the other. While crossed random-effects models extend classical hierarchical multilevel models, they can be fitted using procedures designed for purely hierarchical or multilevel structures. Of course, they are technically and computationally more demanding and can become prohibitive since the data cannot be grouped into independent blocks. For binary data, Renard *et al.* (2002) considered crossed random-effects models and used pseudo-likelihood for parameter estimation. Here, in the context of continuous outcomes, we apply conditional linear mixed model ideas (Verbeke, Spiessens, and Lesaffre 2001) to conveniently estimate parameters as well as precision. The major advantage of the approach is that, by appropriate conditioning, the original model maps into two hierarchical ones, for which conventional and hence computationally efficient and fast techniques can be used.

We organize this chapter as follows. In Section 8.2 we introduce the cross classification multilevel models in a psychometric context. The case study we use to illustrate our method is described in Section 8.3. A full description of the methodology is given in Section 8.4 and the case study is analyzed in Section 8.5. To conclude, a simulation study is carried out in Section 8.6, followed by a discussion in Section 8.7.

## 8.2 Cross-classification Multilevel Models in Psychometry

Suppose a set of items has been offered to a group of pupils and for each pupil the correctness of the responses has been recorded. These responses have random and systematic components, due to a difference in the ability of pupils to correctly respond to an item, as well as due to a difference in the difficulty of the items. Although one often encounters dichotomous items (e.g., correct/incorrect), item responses may also be categorical (e.g., yes/perhaps/no), and an appropriate aggregation can convert them into quasi-continuous responses. Let us refer to such aggregations as *targets*. We will focus on such targets and hence on continuous outcomes. In the analysis of

the case study, we will comment on issues stemming from targets that are made up of variable numbers of items. Extensions of the methodology from the continuous case to the binary case is the subject of further research.

If both targets as well as persons could be regarded as random samples from a population of targets and a population of persons, one can define a random residual error for both persons and targets. Because persons and targets are in a non-hierarchical relationship, classical methodology for hierarchical models is in need of extension. In the multilevel literature, such a model therefore is often referred to as a cross-classification model or crossed random-effects model (Goldstein 1987; Raudenbush 1993). Note that there usually is only one observation for each target $\times$ person combination.

## 8.3 Psychometric Study

The Flemish Community in Belgium issued a set of attainment targets that specify the basic competences that are expected from pupils leaving primary education. De Boeck *et al.* (1997) explored the assessment of the attainment targets of reading comprehension in Dutch. These attainment targets can be characterized by the text type and by the level of processing. In the example, we use the data of one of the tests that were developed by De Boeck *et al.* (1997) and studied by Van den Noortgate *et al.* (2003). These data were analyzed before by Janssen *et al.* (2000). The data consist of the responses of a group of 539 pupils from 15 schools who answered 57 items assumed to measure 9 attainment targets. In Table 8.1, the 9 attainment targets are described by the type of text and by the level of processing. In addition, we indicate the number of items ($I_k$) that were used to measure each one of the targets.

As our response variables, we will use the sum of all positive answers within a category and we will assume them to be continuous. The methodological aspect will be discussed in the next sections.

## 8.4 Methodology

Let us consider a continuous random variable $Y_{ij}$ and a general model with two random factors: $\alpha_i$ for subjects or persons ($i = 1, \ldots, I$) and $\beta_j$ ($j = 1, \ldots, J$) for targets, assumed to be crossed with each other. Further, we assume $\alpha_i$ and $\beta_j$ to follow $N(0, \sigma_\alpha^2)$ and $N(0, \sigma_\beta^2)$ distributions, respectively. The residual errors are assumed to

Table 8.1: *Psychometric Study: Text type, Level of Processing, and Number of Items for the Attainment Targets of the Text. From Janssen* et al. *(2000), used with the permission of the authors.*

| $k$ | Text Type | Level of Processing | $I_k$ |
|---|---|---|---|
| 1 | Instructions | Retrieving | 4 |
| 2 | Articles in magazine | Retrieving | 6 |
| 3 | Study material | Structuring | 8 |
| 4 | Tasks in textbook | Structuring | 5 |
| 5 | Comics | Structuring | 9 |
| 6 | Stories, novels | Structuring | 6 |
| 7 | Poems | Structuring | 8 |
| 8 | Newspapers for children, textbooks encyclopedias | Evaluating | 6 |
| 9 | Advertising material | Evaluating | 5 |

follow a $N(0, \sigma^2)$ distribution. All random terms are assumed to be independent of each other. The model can be written as:

$$Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}. \tag{8.1}$$

The parameters of interest are the variances $\sigma_\alpha^2$ and $\sigma_\beta^2$ of the random effects and the residual variance $\sigma^2$.

We will now briefly describe the general conditional linear mixed model (Section 8.4.1) and then focus on the particular situation of crossed random-effects (Section 8.4.2).

### 8.4.1 Conditional Linear Mixed Models

Conditional linear mixed models were used by Verbeke, Spiessens and Lesaffre (2001) (see also Verbeke and Molenberghs 2000) to analyze longitudinal data without the need to specify time-independent effects such as, for example, the effect of baseline covariates that are assumed to have a constant effect over time.

Consider the general linear mixed-effects model, of which (8.1) is a special case:

$$\boldsymbol{Y} = X\boldsymbol{\beta} + Z\boldsymbol{b} + \boldsymbol{\varepsilon}, \tag{8.2}$$

where $\boldsymbol{\beta}$ corresponds to the fixed part of the model, $\boldsymbol{b}$ to the random part and the errors $\boldsymbol{\varepsilon}$ are assumed to be normally distributed with zero mean and variance matrix equal to $\sigma^2 I$. Typically, the random effects $\boldsymbol{b}$ are assumed to be zero-mean normally distributed.

Verbeke, Spiessens, and Lesaffre (2001) conceived conditional linear mixed-effects models to consist of two steps. In the first step, they conditioned on sufficient statistics for the cross-sectional component of the model. In order to proceed, they rewrite the general model

$$\boldsymbol{Y_i} = \mathbf{1}_{n_i} b_i^* + X_i \boldsymbol{\beta} + Z_i \boldsymbol{b_i} + \boldsymbol{\varepsilon_i}, \tag{8.3}$$

where the matrices $X_i$ and $Z_i$ and the vectors $\boldsymbol{\beta}$ and $\boldsymbol{b_i}$ are those submatrices and subvectors of their original counterparts $X_i$, $Z_i$, $\boldsymbol{\beta}$, and $\boldsymbol{b_i}$ obtained from deleting the elements which correspond to the cross-sectional component. The component $b_i^*$ groups all cross-sectional components, considered to be of nuisance in this approach, and combining, for example, random intercepts and time-invariant effects of baseline covariates. Conditional linear mixed models now proceed in two steps. In a first step, we condition on sufficient statistics for the nuisance parameters $b_i^*$. In a second step, classical estimation procedures for nested random effects are used to estimate the remaining parameters in the conditional distribution of the $\boldsymbol{Y_i}$ given these sufficient statistics.

Conditional on the subject-specific parameters $b_i^*$ and $\boldsymbol{b_i}$ in (8.3), we have that $\boldsymbol{Y_i}$ is normally distributed with mean vector $\mathbf{1}_{n_i} b_i^* + X_i \boldsymbol{\beta} + Z_i \boldsymbol{b_i}$ and covariance matrix $\sigma^2 I_{n_i}$, from which it readily follows that $\overline{y}_i = \sum_j y_{ij}/n_i$ is sufficient for $b_i^*$. Further, the distribution of $\boldsymbol{Y_i}$, conditional on $\overline{y}_i$ and on the remaining subject-specific effects $\boldsymbol{b_i}$, is given by

$$\begin{aligned} f_i(\boldsymbol{y_i}|\overline{y}_i, \boldsymbol{b_i}) &= \frac{f_i(\boldsymbol{y_i}|b_i^*, \boldsymbol{b_i})}{f_i(\overline{y}_i|b_i^*, \boldsymbol{b_i})} \\ &= \left(2\pi\sigma^2\right)^{-(n_i-1)/2} \sqrt{n_i} \exp\left\{ -\frac{1}{2\sigma^2} \left(\boldsymbol{y_i} - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b_i}\right)' \right. \\ &\quad \left. \times \left( I_{n_i} - \mathbf{1}_{n_i} \left(\mathbf{1}_{n_i}' \mathbf{1}_{n_i}\right)^{-1} \mathbf{1}_{n_i}' \right) \left(\boldsymbol{y_i} - X_i\boldsymbol{\beta} - Z_i\boldsymbol{b_i}\right) \right\}. \end{aligned} \tag{8.4}$$

It now follows directly from some matrix algebra (Seber 1984, property B3.5, p. 536), that (8.4) is proportional to

$$\begin{aligned} \left(2\pi\sigma^2\right)^{-(n_i-1)/2} \exp\{ -\frac{1}{2\sigma^2} \left(A_i'\boldsymbol{y_i} - A_i'X_i\boldsymbol{\beta} - A_i'Z_i\boldsymbol{b_i}\right)' \left(A_i'A_i\right)^{-1} \\ \left(A_i'\boldsymbol{y_i} - A_i'X_i\boldsymbol{\beta} - A_i'Z_i\boldsymbol{b_i}\right) \} \end{aligned} \tag{8.5}$$

for any set of $n_i \times (n_i - 1)$ matrices $A_i$ of rank $n_i - 1$ which satisfy $A_i' \mathbf{1}_{n_i} = 0$. This shows that the conditional approach is equivalent to transforming each vector $\boldsymbol{Y_i}$ orthogonal to $\mathbf{1}_{n_i}$. If we now also require the $A_i$ to satisfy $A_i' A_i = I_{(n_i-1)}$, we have that the transformed vectors $A_i' \boldsymbol{Y_i}$ satisfy

$$\begin{aligned} \boldsymbol{Y_i}^* \equiv A_i' \boldsymbol{Y_i} &= A_i' X_i \boldsymbol{\beta} + A_i' Z_i \boldsymbol{b_i} + A_i' \boldsymbol{\varepsilon_i} \\ &= X_i^* \boldsymbol{\beta} + Z_i^* \boldsymbol{b_i} + \boldsymbol{\varepsilon_i^*}, \end{aligned} \tag{8.6}$$

where $X_i^* = A_i' X_i$ and $Z_i^* = A_i' Z_i$ and where the $\boldsymbol{\varepsilon_i^*} = A_i' \boldsymbol{\varepsilon_i}$ are normally distributed with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I_{n_i - 1}$.

Model (8.6) is now again a linear mixed model, but with transformed data and covariates, and such that the only parameters still in the model are the longitudinal effects and the residual variance. Hence, the second step in fitting conditional linear mixed models is to fit model (8.6) using maximum likelihood or restricted maximum likelihood methods. Note that once the transformed responses and covariates have been calculated, standard software for fitting linear mixed models (e.g., SAS procedure MIXED) can be used for the estimation of all parameters in model (8.6).

Note that the conditional linear mixed model is, in spirit, very similar to REML estimation in the classical linear mixed model, where the variance components are estimated after transforming the data such that the fixed effects vanish from the model. As shown by Harville (1974) and by Patterson and Thompson (1971), and as discussed in Verbeke and Molenberghs (2000, Sec. 5.3.4), the REML estimates for the variance components do not depend on the selected transformation, and no information on the variance components is lost in the absence of information on the fixed effects. It has been shown by Verbeke, Spiessens and Lesaffre (2001) that similar properties hold for inferences obtained from conditional linear mixed models; that is, it was shown that results do not depend on the selected transformation $\boldsymbol{Y_i} \rightarrow A_i' \boldsymbol{Y_i}$ and that no information is lost on the average, nor on the subject-specific longitudinal effects, from conditioning on sufficient statistics for the cross-sectional components $b_i^*$ in the original model.

### 8.4.2 Models for Crossed Random-Effects

Let us return to model (8.1) and apply conditional linear mixed model ideas to it. It is convenient to rewrite our model in vectorized form. To this end, group the outcomes into a vector $\boldsymbol{Y}$, with the second index varying more rapidly than the first one, and

write (8.1) as

$$\boldsymbol{Y} = \mu\mathbf{1} + Z(\alpha)\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_I\end{pmatrix} + Z(\beta)\begin{pmatrix}\beta_1\\ \vdots\\ \beta_J\end{pmatrix} + \boldsymbol{\varepsilon}, \tag{8.7}$$

where the design matrices, using Kronecker products, are $Z(\alpha) = I_I \otimes \mathbf{1}_J$ and $Z(\beta) = \mathbf{1}_I \otimes I_J$, respectively. It means that,

$$Z(\alpha) = \begin{pmatrix}1&0&\dots&0\\ \vdots&\vdots&&\vdots\\ 1&0&\dots&0\\ 0&1&\dots&0\\ \vdots&\vdots&&\vdots\\ 0&1&\dots&0\\ \vdots&\vdots&&\vdots\\ 0&0&\dots&1\\ \vdots&\vdots&&\vdots\\ 0&0&\dots&1\end{pmatrix} \text{ and } \quad Z(\beta) = \begin{pmatrix}1&0&\dots&\dots&0\\ 0&1&0&\dots&0\\ 0&0&1&\dots&0\\ \vdots&\vdots&\vdots&\vdots&\vdots\\ 0&0&0&\dots&1\\ \vdots&\vdots&\vdots&\vdots&\vdots\\ \vdots&\vdots&\vdots&\vdots&\vdots\\ 1&0&\dots&\dots&0\\ 0&1&0&\dots&0\\ 0&0&1&\dots&0\\ \vdots&\vdots&\vdots&\vdots&\vdots\\ 0&0&0&\dots&1\end{pmatrix}.$$

The dimensions of these are $IJ \times I$ and $IJ \times J$, respectively. We will now apply the conditional linear mixed model idea twice, once to remove the $\alpha_i$ and once to remove the $\beta_j$.

First, focusing on removal of the $\alpha_i$ effects, we need to construct a matrix $A$ with dimensions $IJ \times J(I-1)$ such that $A'Z(\alpha) = \mathbf{0}$ and $A'A = I$. Assuming such a matrix has been constructed, we obtain a transformed response vector

$$\begin{aligned}\boldsymbol{Y}^* &= A'\boldsymbol{Y}\\ &= A'\mu\mathbf{1} + A'Z(\alpha)\begin{pmatrix}\alpha_1\\ \vdots\\ \alpha_I\end{pmatrix} + A'Z(\beta)\begin{pmatrix}\beta_1\\ \vdots\\ \beta_J\end{pmatrix} + A'\boldsymbol{\varepsilon} \qquad (8.8)\\ &= Z^*(\beta)\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*,\end{aligned}$$

where $\boldsymbol{\varepsilon}^*$ are normally distributed with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I_{J-1}$.

Let us now find a matrix $A$ such that $Z^*(\alpha) = A'Z(\alpha) = \mathbf{0}$ or, equivalently, $A'(I_I \otimes \mathbf{1}_J) = \mathbf{0}$. In this case, due to the specific format of $Z(\alpha)$, the matrix $A'$ can be written as $A_1' \otimes A_2'$. By using the fact that $(A_1' \otimes A_2')(I_I \otimes \mathbf{1}_J) = (A_1' I_I) \otimes (A_2' \mathbf{1}_J)$, we just need to find a matrix $A_2'$ such as $A_2' A_2 = I_{J-1}$.

For a $p$-dimensional matrix, we define

$$A = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{12} & \dots & 1/\sqrt{p+p^2} \\ -1/\sqrt{2} & 1/\sqrt{6} & 1/\sqrt{12} & \dots & 1/\sqrt{p+p^2} \\ 0 & -2/\sqrt{6} & 1/\sqrt{12} & \dots & 1/\sqrt{p+p^2} \\ 0 & 0 & -3/\sqrt{12} & \dots & 1/\sqrt{p+p^2} \\ 0 & 0 & 0 & \dots & 1/\sqrt{p+p^2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -p/\sqrt{p+p^2} \end{pmatrix}.$$

Therefore, if we define

$$A_{2,ij} = \begin{cases} 0, & 1+j < i, \\ -j/\sqrt{j+j^2}, & 1+j = i, \\ 1/\sqrt{j+j^2}, & 1+j > i, \end{cases}$$

and using $A_1' = I_I$, we get

$$Z^*(\alpha) = A'Z(\alpha) = (A_1' \otimes A_2')(I_I \otimes \mathbf{1}_J) = (A_1' I_I) \otimes (A_2' \mathbf{1}_J) = I_I \otimes \mathbf{0}_{(J-1)\times 1} = \mathbf{0}_{I(J-1)\times I}$$

and the resulting model contains only one random factor $\beta$.

Second, with entirely similar logic, we need to apply a different transformation, $B$ say, to (8.7) to eliminate $\beta_j$, and details are omitted. We obtain a second conditional linear mixed model:

$$\boldsymbol{Y}^{**} = Z^{**}(\alpha)\boldsymbol{\alpha} + \boldsymbol{\varepsilon}^{**}. \tag{8.9}$$

Note that, just as $Z(\alpha)$ has been removed by the transformation $A$, now $Z(\beta)$ has been removed. Further, $\boldsymbol{\varepsilon}^{**} \sim N(0, \sigma^2 I_{I-1})$, and hence the residual error variance component occurs in both (8.8) and (8.9), while the random effects occur in just one of these models.

## 8.5 Analysis of Data from Psychometric Study

To fit model (8.1) to the data described in Section 8.3, we will pass by conditionally derived models (8.8) and (8.9). The models were fitted to the data of our case study

Table 8.2: *Psychometric Study: Parameters estimates (standard errors) for the conditional linear mixed effects model.*

| Effect | Parameter | Model 1 | Model 2 | Combined | Standard Error |
|---|---|---|---|---|---|
| Person | $\hat{\sigma}^2_\alpha$ | 1.3634 | – | 1.3634 | 0.0017 |
| Item Group | $\hat{\sigma}^2_\beta$ | – | 2.2155 | 2.2155 | 0.0040 |
| Residual | $\hat{\sigma}^2$ | 0.6704 | 0.7477 | 0.7090 | 0.0030 |

by means of the SAS procedure MIXED. Macros constructing the matrices $A$ and $B$, applying the orthogonal transformations before applying the SAS procedure MIXED, are given in the Appendix of this chapter. In Table 8.2, the estimated values are displayed. The parameters, obtained with (8.9) and (8.8) are labeled 'Model 1' and 'Model 2', respectively.

Note that $\widehat{\sigma}^2_\alpha$ and $\widehat{\sigma}^2_\beta$ are obtained from just one of the models, while the residual variance is estimated twice.

In order to obtain a unique estimated value for the residual variance we propose to combine these two numbers into a single one using the following strategy. Define $\widehat{\sigma}^2 = (\widehat{\sigma}^2_{(1)} + \widehat{\sigma}^2_{(2)})/2$, producing an overall estimate of the residual variance. To obtain its standard error, let us proceed as follow. Construct

$$\widehat{B} = \sum_{k=1}^{2} (\widehat{\sigma}^2_{(k)} - \widehat{\sigma}^2)^2,$$

which reduces to $\widehat{B} = (\widehat{\sigma}^2_{(1)} - \widehat{\sigma}^2_{(2)})^2/2$, and $\widehat{W} = (v_{(1),11} + v_{(2),11})/2$, with $v_{(k),11}$ the element of the covariance matrix of the covariance parameters for model $k = 1, 2$. The covariance matrix of the covariance parameters can then be written as $\widehat{V} = \widehat{W} + \widehat{B}$. Table 8.2 shows the final numerical results of the estimation process in the last two columns.

### 8.5.1 Discussion and Scope of Results

Table 8.2 shows that the expected score of a pupil on a target varies over persons and especially over targets. Both variance components are highly significant ($p < 0.0001$) using a Wald test, whether or not one corrects for the fact that the null hypothesis lies on the boundary of the parameter space (Verbeke and Molenberghs, 2000 Section 6.3).

Although pupils and targets explain a very large part of the total variability, 32% and 52%, respectively, the residual variability is substantial and statistically highly significant as well ($p < 0.0001$). This indicates that the score for a pupil on a target may deviate from the score that is expected based on the person ability and the target difficulty. It implies that there is some interaction between persons and targets.

Our model contains random effects only; no fixed effects are included apart from an intercept that is removed in the conditioning process. This means that both person abilities as well as target difficulties are assumed to be independently and identically drawn from a normal distribution with variances as in Table 8.2. Further exploration of these person abilities and target difficulties can be done by studying the empirical Bayes estimates of the $\alpha_i$ and $\beta_j$. In many practical situations, however, one may want to explain, for example, differences in target difficulties based on a priori grounds, say, using target-level covariates. The same might be true at the level of the pupil. Such person-level covariates may be continuous (e.g., age) or categorical (e.g., sex). Similarly, target characteristics (e.g., number of subtasks, as discussed below, or the type of problem) can often be assumed to influence the target difficulty. It is also possible that person-by-target characteristics have an effect. This is a topic of further research.

Of course, model (8.7) can easily be extended by including such person, target, or interaction effects as covariates. This is conveniently done by replacing $\mu\mathbf{1}$ in (8.7) by a full fixed-effects design, $X\boldsymbol{\beta}$. The inclusion of covariates can be based on prior beliefs of the researcher about effects of these characteristics, but may also be a tool to explore possible relations. The use of the cross-classification model complemented with target and person predictors yields a flexible predictive and explanatory approach, as it includes error terms at both sides. For instance, the targets in our example are characterized by a combination of text type and the level of processing. One could include one or both variables in the model to explore if they explain part of the variance between targets. Since this would side track from our methodological development, we have chosen not to discuss this further.

A person covariate of particular interest is a person group. For instance, pupils can be grouped into schools. When the groups are seen as randomly drawn from a population of groups, the group effects can be modeled as random rather than as fixed effects. For the cross-classification model discussed above, this would result in a model with crossed as well as nested random effects. Although the conditional linear mixed model approach could still be used, the discussion of this extension of

the simple cross-classification model is beyond the scope of this chapter.

### 8.5.2 Unequal Number of Items per Target

From Table 8.1 we see that the number of items per target varies. This is bound to occur whenever a compounded score, such as the targets used here, are subject of analysis. In cases, unlike here, the number of items per targets varies widely and one insists on retaining stability of variances at the item level, rather than at the target level, then one can proceed as follows. Let us extend (8.1) by denoting $Y_{ijk}$ the response by person $i = 1, \ldots, I$ on item $k = 1, \ldots, K_j$ contributing to target $j = 1, \ldots, J$ and write the item level model:

$$Y_{ijk} = \widetilde{\mu} + \alpha_i + \beta_j + \widetilde{\varepsilon}_{ijk}, \tag{8.10}$$

with distributions $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\beta_j \sim N(0, \sigma_\beta^2)$, and $\widetilde{\varepsilon}_{ijk} \sim N(0, \sigma^2)$. Then, the derived target level model is

$$Y_{ij} = \frac{\sum_{k=1}^{K_j} Y_{ijk}}{K_j} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \tag{8.11}$$

with unaltered distributions for the random effects and

$$\varepsilon_{ij} \sim N\left(0, \sigma^2/K_j\right). \tag{8.12}$$

In other words, this model is similar to (8.1), except for a heteroscedatstic measurement error variance. Next, we need to apply transformations $A$ and $B$ as in Section 8.4.2, but with a slightly different requirement. For example, we would transform still $\boldsymbol{Y_i} \rightarrow \widetilde{A}_i' \boldsymbol{Y_i}$, requiring $\widetilde{A}_i$ to be orthogonal onto the unit vector, but with condition $\widetilde{A}_i' \Lambda_i \widetilde{A}_i = I_{(n_i - 1)}$, where $\Lambda_i$ is a diagonal matrix with $j$th diagonal element equal to $1/K_j$. Then, after transformation, the residual errors are zero-mean normally distributed with covariance $\sigma^2 I_{(n_i-1)}$ and the only programming required is the implementation of this alternative transformation. For the $B$ transformation, where the data are grouped by targets $j$ rather than by persons $i$, the only modification required is to multiply all elements of $B$, as described in Section 8.4.2, by $\sqrt{K_j}$.

Doing so, produces $\hat{\sigma}_\beta^2 = 29.11$ (s.e. 0.42), $\hat{\sigma}_\alpha^2 = 13.08$ (s.e. 0.08), and $\hat{\sigma}^2 = 0.9506$ (s.e. 0.0006). The difference in random-effects variances is, of course, due to the fact that the variability is now at item level rather than at target level, and hence the inclusion of the factor $K_j$ in (8.12) changes the balance between the random-effects

variances and the measurement error variance. A further reason why the results differ here, in comparison to those obtained at the target level is that we now properly account for the different number of items per target, whereas before this aspect was ignored.

## 8.6 Simulation Study

We conducted a simulation study to evaluate the performance of the conditional linear mixed model for crossed random-effects. The design of the simulation study was carried out under different settings to investigate the impact of number of subjects and number of items on the performance.

First, we generated data where the true parameters were set equal to the estimates obtained from the analysis done in Section 8.5. Five hundred simulation data sets were generated. Other, additional settings were also used to study changes on the variances of the random effects. They were defined in the following way:

**Setting 1.**

$\sigma^2_\alpha = 1.3634$; $\sigma^2_\beta = 2.2155$; $\sigma^2 = 0.7090$; $I = 20, 50, 100$; $J = 5, 10, 20, 30$.

**Setting 2.**

$\sigma^2_\alpha = 0.50$ to $8.50$ by $0.5$; $\sigma^2_\beta = 2.2155$; $\sigma^2 = 0.7090$; $I = 10, 20, 50, 100$; $J = 10$.

**Setting 3.**

$\sigma^2_\alpha = 1.3634$; $\sigma^2_\beta = 0.50$ to $8.50$ by $0.5$; $\sigma^2 = 0.7090$; $I = 20, 50, 100$; $J = 10$.

We report the results of Setting 1 in Table 8.3. Bias and relative bias was calculated by taking the average of $\widehat{\sigma}^2 - \sigma^2$ from 500 replicates, Mean(SE) denotes the average of the estimated standard errors of the estimates. The 95% confidence interval coverage probabilities are included as well.

From this table it can be seen that the method performs well in most cases. The relative bias regarding the estimation of $\sigma^2_\alpha$ decreases when the number of subjects increases, however it is always smaller than 0.06. An identical situation can be observed for $\sigma^2_\beta$ where the relative bias decrease when the number of items increase being always smaller than 0.04. This similarity is, of course, to be expected.

The 95% coverage probabilities are all rather high and even in the most unfavorable situation, i.e., a low number of subjects as well as of items, the coverage probabilities are 98.6% and 90.2% for the variance of the subjects and items, respectively. It is

Table 8.3: *Results of the simulation study.*

| | Subjects | 20 | | | 50 | | | 100 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Items | Parameters | $\sigma^2_\alpha$ | $\sigma^2_\beta$ | $\sigma^2_\varepsilon$ | $\sigma^2_\alpha$ | $\sigma^2_\beta$ | $\sigma^2_\varepsilon$ | $\sigma^2_\alpha$ | $\sigma^2_\beta$ | $\sigma^2_\varepsilon$ |
| 5 | Estimate | 1.281 | 2.125 | 0.769 | 1.287 | 2.126 | 0.771 | 1.289 | 2.132 | 0.772 |
| | Bias | 0.081 | 0.089 | 0.060 | 0.076 | 0.088 | 0.062 | 0.073 | 0.083 | 0.064 |
| | Rel. Bias | 0.060 | 0.040 | 0.084 | 0.055 | 0.040 | 0.088 | 0.054 | 0.037 | 0.089 |
| | Mean(SE) | 0.847 | 0.339 | 0.018 | 0.536 | 0.213 | 0.011 | 0.381 | 0.150 | 0.008 |
| | 95% coverage | 98.6 | 90.2 | | 99.2 | 94.0 | | 99.8 | 95.8 | |
| 10 | Estimate | 1.290 | 2.133 | 0.772 | 1.287 | 2.135 | 0.772 | 1.286 | 2.134 | 0.772 |
| | Bias | 0.072 | 0.082 | 0.063 | 0.075 | 0.079 | 0.063 | 0.076 | 0.081 | 0.063 |
| | Rel. Bias | 0.053 | 0.037 | 0.088 | 0.055 | 0.035 | 0.089 | 0.056 | 0.036 | 0.090 |
| | Mean(SE) | 0.568 | 0.242 | 0.012 | 0.360 | 0.150 | 0.008 | 0.254 | 0.105 | 0.006 |
| | 95% coverage | 99.4 | 91.2 | | 99.8 | 94.2 | | 100 | 96.4 | |
| 20 | Estimate | 1.289 | 2.131 | 0.772 | 1.287 | 2.132 | 0.772 | 1.289 | 2.134 | 0.773 |
| | Bias | 0.074 | 0.084 | 0.063 | 0.076 | 0.082 | 0.064 | 0.077 | 0.081 | 0.064 |
| | Rel. Bias | 0.054 | 0.038 | 0.089 | 0.055 | 0.037 | 0.089 | 0.056 | 0.036 | 0.090 |
| | Mean(SE) | 0.390 | 0.171 | 0.008 | 0.247 | 0.106 | 0.006 | 0.175 | 0.074 | 0.004 |
| | 95% coverage | 99.8 | 95.4 | | 100 | 96.8 | | 100 | 97.4 | |
| 30 | Estimate | 1.288 | 2.133 | 0.772 | 1.287 | 2.134 | 0.773 | 1.286 | 2.135 | 0.773 |
| | Bias | 0.074 | 0.081 | 0.064 | 0.075 | 0.080 | 0.064 | 0.076 | 0.079 | 0.064 |
| | Rel. Bias | 0.054 | 0.036 | 0.089 | 0.055 | 0.036 | 0.090 | 0.056 | 0.035 | 0.090 |
| | Mean(SE) | 0.316 | 0.140 | 0.007 | 0.200 | 0.087 | 0.004 | 0.141 | 0.061 | 0.003 |
| | 95% coverage | 99.8 | 97.0 | | 100 | 96.6 | | 100 | 96.2 | |

important to point out that our case study contains 9 items and 500 subjects being the scenario with 10 items and 100 subjects the closest. In this setting we obtained a coverage of 100% and 96.4% for each of the crossed random-effects.

To explore the impact of the magnitude of the variances on the estimation, we performed simulations by fixing one of the variances and varying the other, between 0.5 and 8.5. The variances were fixed to be 1.3634 and 2.2155, according to the values obtained from the data set. We conducted the simulation study for a fixed number

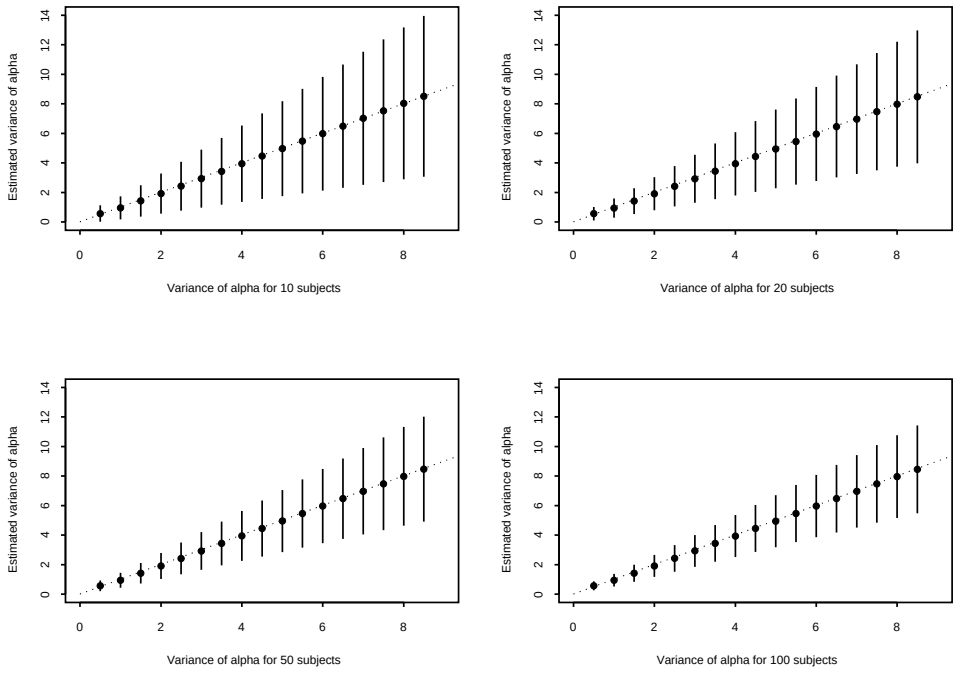


Figure 8.1: *Simulation results for variance of beta equal to 2.2155 and variance of alpha varying between 0.5 and 8.5. Each panel corresponds to different numbers of subjects. The segments indicate the size of the 95% confidence intervals.*

10 of items and for four different choices for the number of subjects, i.e., 10, 20, 50, and 100.

The numerical results are graphically displayed to facilitate interpretation. Figure 8.1 contains the results of the simulations with fixed variance of beta and Figure 8.2 displays equivalent results with fixed variance of alpha. Each panel corresponds to a different number of subjects. True values are plotted against true values, together with their confidence intervals. The dotted lines indicate the estimated values of alpha.

These figures show that almost all points virtually fall on the dotted line, indicating a high agreement between true and estimated values. As it can be expected, the size of the confidence intervals increases with variance and decreases with number of subjects.

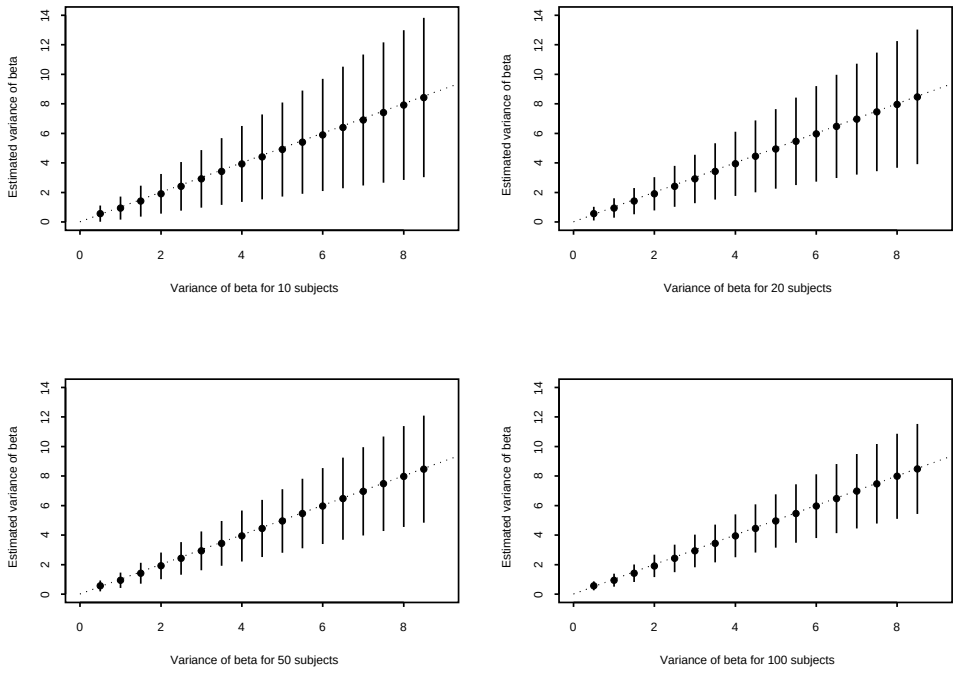


Figure 8.2: *Simulation results for variance of alpha equal to 1.3634 and variance of beta varying between 0.5 and 8.5. Each panel corresponds to different numbers of subjects. The segments indicate the size of the 95% confidence intervals.*

## 8.7 Conclusions

We have proposed an estimation method for the variance components of a crossed random-effects model when only one observation is available in each cross classified level. We have illustrated the methodology using data coming from a psychometric study. Of course, there is no restriction for our methodology to be applied in different scientific fields. To estimate the variance of the error and the standard deviation of these variances, we proposed an approach to combine the values obtained from the fit of two simple models after an appropriate transformation of the response and the effects. We implemented our approach by means of a SAS macro.

We have conducted several simulation studies of the proposed method under a

variety of circumstances, e.g., different number of subjects, items and a range of values for the variance of alpha and beta. Our method performs quite well for the estimation of the variance components in most of the cases.

Some extensions of this method can easily be carried out, for example, fixed covariates can be included in the model without major changes to the general structure of the programs. However, it would be interesting to investigate the possibilities of applying this strategy with other type of responses, such as binary or general categorical outcomes.

# Appendix: SAS Macros

The macro `condlin1` finds the transformation for the model that conditions over subjects and fits the linear model with items. The macro `condlin2` reverses the roles of subjects and items.

- id: random factor corresponding to subject
- target: random factor corresponding to item
- I: Number of subjects
- J: Number of items

## First Conditional Linear Mixed Model

```
%macro condlin1(I,J);

proc sort data=<dataset>;
   by <factor1> <factor2>;
run;
data help;
   set <dataset>;
   by <factor1> <factor2>;
   if first.<factor2>;
run;

proc iml;

a=J(&J,&J-1,0);
  do i= 1 to &J;
    do j=1 to (&J-1);
       a[i,j]=( (1+j=i)*(-j)+(1+j>i)*1 ) / ( sqrt(j+j*j) );
    end;
end;

 A2t=t(a);
 Za=I(&I)@J(&J,1,1);
 Zb=J(&I,1,1)@I(&J);
 A1t=I(&I);
 At=A1t@A2t ;
 Zbstar=(At*Zb);
```

```
use help;
 labelx = {scorec target};
 labelid = {id};
 read all var labelid into id;
 read all var labelx into x;
 close help;
 do s=1 to &I;
   do ss=1 to (&J-1);
     id2=id2//s;
   end;
 end;

 ytt=At*x[,1];
 btt=zbstar*x[1:&J,2];
 hulp=id2||ytt||btt;
 name = labelid||labelx;
 create outdata1 var name;
 append from hulp;
quit;

%mend;


proc sort data=outdata1;
   by <factor1> <factor2>;
run;

proc mixed data=outdata1 method=reml  asycov covtest;
   class <factor1> <factor2>;
   model <response> = / solution noint;
   random <factor2> /  type=vc subject=<factor1> V solution;
run;
```

## Second Conditional Linear Mixed Model

```
%macro condlin2(I,J);

data help;
   set <dataset>;
   by <factor1> <factor2>;
   if first.<factor2>;
run;
```

```
proc sort data=help;
   by <factor2> <factor1>;
run;

proc iml;

a=J(&I,&I-1,0);
 do i= 1 to &I;
    do j=1 to (&I-1);
       a[i,j]=( (1+j=i)*(-j)+(1+j>i)*1 ) / ( sqrt(j+j*j) );
    end;

 end;

 A2t=t(a);
 Za=I(&J)@J(&I,1,1);
 Zb=J(&J,1,1)@I(&I);
 A1t=I(&J);
 At=A1t@A2t ;
 Zbstar=(At*Zb);

use help;
 labelx = {scorec id};
 labelid = {target};
 read all var labelid into target;
 read all var labelx into x;
 close crossre;
 do s=1 to &J;
   do ss=1 to (&I-1);
    id3=id3//s;
   end;
 end;

ytt=At*x[,1];
btt=zbstar*x[1:&I,2];
hulp=id3||ytt||btt;
name = labelid||labelx;
create outdata2 var name;
append from hulp;
quit;

%mend;
```

```
proc mixed data=outdata2 asycov covtest;
    class <factor1> <factor2>;
    model <response> = / solution noint;
    random <factor1> / type=vc subject=<factor2> solution;
run;
```

# Chapter 9

# Conditional Linear Mixed Models with Crossed Random-Effects for Binary Data

## 9.1 Introduction

In the previous chapter we have shown how conditional linear mixed models can be used in presence of crossed random-effects. In particular, if the response variable is continuous, conditioning on sufficient statistics allow us to produce estimates of the variances of the random effects. We used data from a psychometric study and we have proposed an estimation strategy in case of dealing with continuous or quasi-continuous responses.

The first challenge we faced was the fact that not all responses (the so-called targets) were based on averaging an equal number of items. To tackle this, we proposed an alternative method properly adjusting for this unequal number of items per target. The main idea is to modify the original matrices in order to take into account the extra variability introduced by the different number of items within each target.

Thus, our modelling approach still assumes response variables (targets) to be continuous. However, the nature of this experiment produces binary responses because

the correctness of each item is recorded as 0 and 1.

In this chapter we intend to present a solution to this problem and we develop a method based on conditional logistic regression and pseudo-likelihood methods for the estimation of the model parameters.

## 9.2 Methodology

We will start by introducing the model that will be used to analyze the data of the psychometric study, described in Section 8.3.

Given that we have only two random effects, let us consider

$$\text{logit}(\Pr(Y_{ij} = 1|a_i, b_j)) = \mu + a_i + b_j,$$

where $a_i$ and $b_j$ represent the random effects corresponding to person $i$ and item $j$, respectively. Of course, the effect of other covariates can be explored by adding an extra term in the model. However, we will focus only on models with two random effects.

For example, a logistic model for binary data is useful in designs with matched pairs. In such a case one expresses by $(Y_{i1}, Y_{i2})$ pairs of matched observations where $i = 1, \ldots, n$. In addition, $Y_{ij}$ is a binary response with two possible outcomes: 1 for success and 0 for failure.

One often considers a model which permits separate response distributions for each pair (Agresti, 1990). A common effect is then considered as follows

$$\begin{aligned} \text{logit}(\Pr(Y_{i1} = 1)) &= a_i, \\ \text{logit}(\Pr(Y_{i2} = 1)) &= a_i + \beta. \end{aligned}$$

It immediately follows that the marginal probabilities can be written as

$$\begin{aligned} \Pr(Y_{i1} = 1) &= \exp(a_i)/[1 + \exp(a_i)], \\ \Pr(Y_{i2} = 1) &= \exp(a_i + \beta)/[1 + \exp(a_i + \beta)]. \end{aligned}$$

Using ideas of sufficiency, the parameters $a_i$'s can be eliminated by conditioning on $S_i = y_{i1} + y_{i2}$. Of course,

$$\begin{aligned} \Pr(Y_{i1} = Y_{i2} = 0|S_i = 0) = 1, \\ \Pr(Y_{i1} = Y_{i2} = 1|S_i = 2) = 1. \end{aligned}$$

Thus, we conclude that such probabilities depend on $\beta$ only when $S_i = 1$ because

$$\Pr(Y_{i1}=y_{i1},Y_{i2}=y_{i2}|S_i=1)$$

$$=\frac{\Pr(Y_{i1}=y_{i1},Y_{i2}=y_{i2})}{\Pr(Y_{i1}=1,Y_{i2}=0)+\Pr(Y_{i1}=0,Y_{i2}=1)}$$

$$=\frac{\left(\frac{e^{a_i}}{1+e^{a_i}}\right)^{y_{i1}}\left(\frac{1}{1+e^{a_i}}\right)^{1-y_{i1}}\left(\frac{e^{a_i+\beta}}{1+e^{a_i+\beta}}\right)^{y_{i2}}\left(\frac{1}{1+e^{a_i+\beta}}\right)^{1-y_{i2}}}{\left(\frac{e^{a_i}}{1+e^{a_i}}\right)\left(\frac{1}{1+e^{a_i+\beta}}\right)+\left(\frac{1}{1+e^{a_i}}\right)\left(\frac{e^{a_i+\beta}}{1+e^{a_i+\beta}}\right)}.$$

There are only two possibilities for this last expression depending on the values of $y_{ij}$ correspond to $S_i=1$.
If $y_{i1}=1$ and $y_{i2}=0$ then

$$\begin{aligned}&\Pr(Y_{i1}=1,Y_{i2}=0|S_i=1)\\&=\frac{\left(\frac{e^{a_i}}{1+e^{a_i}}\right)\left(\frac{1}{1+e^{a_i+\beta}}\right)}{\left(\frac{e^{a_i}}{1+e^{a_i}}\right)\left(\frac{1}{1+e^{a_i+\beta}}\right)+\left(\frac{1}{1+e^{a_i}}\right)\left(\frac{e^{a_i+\beta}}{1+e^{a_i+\beta}}\right)}\\&=1/(1+e^{\beta}).\end{aligned}$$

If $y_{i1}=0$ and $y_{i2}=1$ then

$$\begin{aligned}&\Pr(Y_{i1}=0,Y_{i2}=1|S_i=1)\\&=\frac{\left(\frac{1}{1+e^{a_i}}\right)\left(\frac{e^{a_i+\beta}}{1+e^{a_i+\beta}}\right)}{\left(\frac{e^{a_i}}{1+e^{a_i}}\right)\left(\frac{1}{1+e^{a_i+\beta}}\right)+\left(\frac{1}{1+e^{a_i}}\right)\left(\frac{e^{a_i+\beta}}{1+e^{a_i+\beta}}\right)}\\&=e^{\beta}/(1+e^{\beta}).\end{aligned}$$

Hence, conditional on $S_i=1$, the joint distribution of the matched pairs is

$$\prod_{S_i=1}\left(\frac{1}{1+e^{\beta}}\right)^{y_{i1}}\left(\frac{e^{\beta}}{1+e^{\beta}}\right)^{y_{i2}}=e^{\beta\sum y_{i2}}(1+e^{\beta})^{-\sum S_i}$$

and the maximum likelihood estimates are obtained differentiating the log of this so-called conditional likelihood.

In classical conditional logistic regression models one has that

$$\text{logit}(\Pr(Y_{ij=1}|x_{ij},a_i))=a_i+x'_{ij}\beta,$$

and the likelihood is constructed by conditioning upon sufficient statistics for $a_i$. In our case, the response $Y_{ij}$ is binary. Therefore $S_i = \sum_{j=1}^{n_i} Y_{ij}$, representing the number of successes, is a sufficient statistics.

In this case the likelihood function can be written as

$$\frac{\prod_{j=1}^{n_i} f(y_{ij}|x_{ij})}{\sum_{p\in\wp}\prod_{j=1}^{n_i} f(y_{ij}|x_{ip(j)})}, \tag{9.1}$$

where $\wp$ is the set of all permutation of $(y_{i1},\ldots,y_{in_i})$.

Following ideas similar to the matched pairs situation in the case of $n_i = 2$, (9.1) can be expressed as

$$\begin{aligned}\frac{f(y_{i1}|x_{i1})f(y_{i2}|x_{i2})}{f(y_{i1}|x_{i1})f(y_{i2}|x_{i2}) + f(y_{i1}|x_{i2})f(y_{i2}|x_{i1})} &= \frac{e^{\beta(x_{i1}y_{i1}+x_{i2}y_{i2})}}{e^{\beta(x_{i1}y_{i1}+x_{i2}y_{i2})} + e^{\beta(x_{i2}y_{i1}+x_{i1}y_{i2})}}\\ &= \frac{e^{\{\beta(y_{i1}(x_{i1}-x_{i2})+y_{i2}(x_{i2}-x_{i1})\}}}{1+e^{\{\beta(y_{i1}(x_{i1}-x_{i2})+y_{i2}(x_{i2}-x_{i1})\}}}\end{aligned}$$

Therefore, it follows for the case of matched pairs that this expression is the same with $y_{i1} \equiv 0$ and $y_{i2} \equiv 1$.

The likelihood can be easily written if we have 2 factors only in each crossed level. In other words, let us assume that we have two persons and two items, hence we can write

$$\begin{aligned}\frac{f(y_{i1}|b_1)f(y_{i2}|b_2)}{f(y_{i1}|b_1)f(y_{i2}|b_2) + f(y_{i1}|b_2)f(y_{i2}|b_1)} &= \frac{e^{b_1y_{i1}+b_2y_{i2}}}{e^{b_1y_{i1}+b_2y_{i2}} + e^{b_1y_{i2}+b_2y_{i1}}}\\ &= \frac{e^{\{(b_1-b_2)y_{i1}+(b_2-b_1)y_{i2}\}}}{1+e^{\{(b_1-b_2)y_{i1}+(b_2-b_1)y_{i2}\}}}\\ &= \frac{e^{\widetilde{b}(y_{i1}-y_{i2})}}{1+e^{\widetilde{b}(y_{i1}-y_{i2})}},\end{aligned}$$

where $\widetilde{b} \sim N(0, 2\sigma^2)$.

However, in our case study, we have more than two items. Then the aforementioned approach fails to be applicable in this way and needs to be generalized. Let us suppose that we have three items and three subjects. Then, the previous expression

becomes

$$\frac{f(y_{i1}|b_1)f(y_{i2}|b_2)f(y_{i3}|b_3)}{\sum\limits_{t_k\leq 3,\, k=1,\ldots,6} f(y_{it_1}|b_{t_4})f(y_{it_2}|b_{t_5})f(y_{it_3}|b_{t_6})}.$$

It is clear that for more than 2 items (subjects), this expression gets cumbersome. To overcome this, we propose to use the case of two items and two subjects nevertheless and combine it using pseudolikelihood ideas.

This procedure has to be implemented in two stages. First, consider all possible pairs of items $(j, j')$. Therefore, there will be $\binom{J}{2}$ of such pairs within person $i$ and where $J$ is the total number of items. If this strategy is repeated for all $I$ subjects we will then have a new dataset containing $\binom{J}{2} \times I$ observations. Second, conditional logistic regression is applied by considering each pair as a unit. The resulting variance is $2\sigma_b^2$, and the variance of this last estimator can be obtained by means of the sandwich estimator $var(\sigma^2)(J-1)$.

Finally, in order to estimate the variance and its standard deviation corresponding to the factor item, a complete symmetric situation is considered where $\binom{I}{2}$ pairs of observations are constructed by using $(i, i')$ all possible subjects. Then, $var(\sigma_a^2)(I-1)$ gives an estimator of the variance of the point estimator.

In the next section, we will illustrate this strategy using the case example introduced in Section 8.3.

## 9.3 Application to the Psychometric Study

To fit the model described in Section 9.2 to the data from the study introduced in Section 8.3 we will start by structuring the data in an appropriate way. To this end, we construct two new datasets. The first one contains $I \times 2 \times J$ observations where each row corresponds to a pair, being a combination of two different items, and $I$ is the total number of persons in the original dataset. Similarly, the second dataset contains $J \times 2 \times I$ observations where now each row corresponds to a determined item and contains pairs of observation corresponding to different subjects.

We developed a SAS macro to this effect. After having arranged the data in a more convienient way, we fit two conditional logistic models by means of the SAS NLMIXED procedure (the code is presented in the appendix of this chapter).

In Table 9.1 the estimated values of $\sigma_a^2$ and $\sigma_b^2$ and their standard errors are presented using the fact that SAS NLMIXED procedure in our case gives an estimate of $2\sigma_a^2$ and $2\sigma_b^2$. Thus the estimates in the table were obtained merely by dividing

Table 9.1: *Psychometric Study: Parameters estimates (standard errors) for the conditional logistic mixed effects model, fitted to the psychometric data.*

| Effect | Parameter | Estimates | Standard Error |
|---|---|---|---|
| Person | $\hat{\sigma}_a^2$ | 0.0795 | 0.927 |
| Item Group | $\hat{\sigma}_b^2$ | 0.0405 | 0.212 |

the obtained values by 2. We want to point out that there exists a way of obtaining these values directly from the SAS procedure.

An important issue when using the SAS NLMIXED procedure is to reach convergency. First, different initial values of the parameters can be tried if convergency fails with the default option. Second, different optimization techniques can be chosen and some of them are faster to reach convergency. In our case, we have tried different choices of initial values of the parameters and different optimization techniques before obtaining convergency. However, once the convergency is reached it seems that the algorithm is quite stable with respect to different initial values.

The numerical results obtained via this approach cannot be compared with the ones presented in the previous chapter. An obvious reason is that the nature of the outcome is different, in other words, while before the response was assumed to be continuous and a linear model used, in the present approach the response was treated as binary and a conditional logistic model applied.

From the numerical values obtained after fitting the models, it can be seen that the variability between subjects is larger than the variability between items, but this conclusion should be used carefully given that the Wald test will fail to reject the null hypotheses due to the large standard errors.

## 9.4 Conclusions

We have proposed an alternative estimation method for the variance component of a crossed random-effect model when the response is binary and in addition to that only one observation is available in each crossed classified level.

We have illustrated this method with data coming from a psychometric study. The estimation of the variance of the random effects has been carried out by using conditional logistic regression together with pseudo-likelihood ideas. For more details about pseudo-likelihood methods, see Chapter 4. Throughout the text we have

established the relationship between our approach and the framework of matched pairs.

Extensions can be carried out for example including covariates and the study of the performance of this method as it was done in the continuous case via simulations is a topic of further research.

# Appendix: SAS Procedures

We display here the SAS code for fitting both conditional logistic models. We assumed the data are structured in the way we mentioned in Section 9.2. The first procedure is for fitting the model with random effect person and the second analogous for item.

- id: random factor corresponding to subject.
- item: random factor corresponding to item.
- y: difference between the values corresponding to each pair of item of persons depending on which model we are fitting.

## First Conditional Logistic Model

```
data model1;set dataset1;
  one=1;
  y=x1-x2;
run;

proc nlmixed data = model1 tech=nrridg;
  parms   s2b=<start>;
  eta = b*Y ;
  expeta = exp(eta);
  p = expeta/(1+expeta);
  model one ~ binary(p);
  random b ~ normal(0,s2b) subject = id ;
run;
```

## Second Conditional Logistic Model

```
data model2;set dataset2;
  one=1;
  y=x1-x2;
run;

proc nlmixed data = model2 tech=nrridg;
  parms   s2b=<start>;
  eta = b*Y;
  expeta = exp(eta);
  p = expeta/(1+expeta);
  model one ~ binary(p);
  random b ~ normal(0,s2b) subject = item;
run;
```

# Chapter 10

# Conclusions and Topics for Further Research

The main goal of this thesis has been the development of models for different types of complex structures, all of which involving dependent outcomes. The first chapters were devoted to the study and modelling of continuous data within the framework of surrogate endpoints. We contributed to the modelling of correlated survival data, using a Plackett copula and a set of inferential tools. Towards the end of this thesis, we studied models with crossed random effects in different settings and we proposed an alternative method for the estimation of the effects in particular situations.

## 10.1 Methodology for the Evaluation of Surrogate Endpoints

In Chapter 2 we have presented simplified approaches to surrogate endpoint validation in a metanalytic framework. Following the ideas of Buyse *et al.* (2000a), we have investigated several strategies to deal with the computational problems when using hierarchical models. The methodology has been developed in the context of surrogate marker validation, but in principle it could be applied to any other settings were hierarchies are present.

As a result of our research in Chapter 2 we have recommended to use these simplified methods primarily because they are faster and easier to implement. We support our conclusions with some simulation results for the case when both endpoints are

normally distributed. A natural extension of this method could be to situations where the responses cannot be considered normally distributed. This is the case for binary-binary, survival-survival or other settings where both responses are of different type. This last issue was not studied yet and it may be interesting to explore these situations in order to analyze further the performance of our method.

## 10.2 Multivariate Survival Model with Pseudo-likelihood Estimation

An important part of this work has been devoted to constructing and testing in a new multivariate survival model.

Like in any modelling problem, checking and diagnostics are important issues that need to be explored. In particular, given the large number of copula functions, goodness-of-fit is one of the points of attention in this area. An extension of the method proposed by Wang and Wells (2000) could be a possible solution and it is a topic of further research.

The case studies used in this research contain no censored or right censored variables. However, in Chapter 3, we displayed the bivariate loglikelihood function in a very general way, and it can be used to construct the pseudolikelihood function in presence of other kinds of censoring mechanisms. In the same spirit other marginal distribution functions can be considered, for example, when dealing with longevity data, Gavrilova *et al.* (1991) suggested that the Gompertz-Makeham distribution could be more appropriate in some cases. Obviously, this problem can be solved by using other approaches such as semiparametric or non parametric marginal functions, as suggested by Shih and Louis (1995).

It was our choice to use a Plackett-Dale model to estimate the association between any pair of survival outcomes. This choice was motivated by the fact that the parameter obtained using this model is easy to interpret. However, the methodology we have developed is not restricted to this particular copula expression. Hence, other copulas (see Chapter 3) as Clayton, Frank, etc. can be used combined with pseudolikelihood ideas in analogous way as it was in case of the Plackett copula.

In a broader extension, we believe that the idea of using pseudolikelihood concepts to fit marginal multivariate models can be extended to other settings apart from the survival framework. To fix ideas, let us suppose we have three variables: one categorical and two time-to-event variables where the goal is to estimate associations in a multivariate way. By using the bivariate distribution functions, i.e., the one

corresponding to the categorical-survival pair, following ideas of Geys *et al.* (1999) and the one corresponding to the survival-survival pair, as it was done already in Chapter 5, the construction of the PL function is straightforward and all parameters can be estimated at once. This issue was not tackled in this work but we want to set up some lines of further research.

Another area where these methods could be an important help is in the surrogate validation field. Our model allows us to easily estimate associations but it is well known that this is not enough to assess surrogacy and measures, such as $R^2$ at trial and at individual-level need to be extended to this case. It will be interesting to construct a global measure of surrogacy based on the copula model. The latter will help us to evaluate surrogate endpoints and to identify new potential surrogate variables as well.

## 10.3 Crossed Random-Effect Models

We have considered models where random-effects are assumed to be crossed and in addition to that only one observation per crossed level is available. We used data from a psychometric study and as a consequence of this specific design only two factors were studied in Chapters 8 and 9. We developed a method for two different situations assuming continuous responses and binary responses as well. Two important issues should be explored in the future. First, the inclusion of covariates in the model, similarly as it was done by Verbeke *et al.* (2001) within the framework of longitudinal data. Second, to explore situations with a larger number of crossed factors.

# References


Aerts, M., Geys, H., Molenberghs, G., and Ryan, L. (2002). *Topics in Modelling of Clustered Data*. London: Chapman & Hall.

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley & Sons.

Alonso, A., Geys, H., Molenberghs, G., and Vangeneugden, T. (2001). Investigating the criterion validity of psychiatric symptom scales using surrogate marker validation methodology. *Submitted for publication*.

Alonso, A., Geys, H., Molenberghs, G., Kenward, M. and Vangeneugden, T. (2003). Validation of surrogate markers in multiple randomized clinical trials with repeated measurements. *Biometrical Journal*, **45**, 1–15.

Anderson, J.R., Cain, K.C., and Gelber, R.D. (1983). Analysis of survival by tumour response. *Journal of Clinical Oncology*, **1**, 710–719.

Arnold, B.C. and Strauss, D. (1991). Pseudolikelihood estimation: some examples. *Sankhya B*, **53**, 233–243.

Bagdonavicius, V., Malov, S. and Nikulin, M. (1999). Characterizations and semiparametric regression estimation in Archimedean copula. *Journal of Applied Statistical Science*, **8**, 137–153.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Society*, **88**, 9–25.

Breslow, N.E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika*, **82**, 81–91.

Burzykowski, T. (2001). *Validation of Surrogate Endpoints from Multiple Randomized Clinical Trials with a Failure-Time True Endpoint*. Unpublished Ph.D. dissertation, Limburgs Universitair Centrum.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H., and Renard, D. (2001). Validation of surrogate endpoints in multiple randomized clinical trials with failure-time endpoints. *Applied Statistics*, **50**, 405–422.

Burzykowski, T., Molenberghs, G., Tafforeau, J., Van Oyen, H., Demarest, S., and Bellamammer, L. (1999). Missing data in the Health Interview Survey 1997 in Belgium. *Archives of Public Health*, **57**, 107–130.

Buyse, M. and Molenberghs, G. (1998). The validation of surrogate endpoints in randomized experiments. *Biometrics*, **54**, 186–201.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, **1**, 49–67.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000a). The validation of surrogate endpoints in meta-analyses of randomized experiments. *Biostatistics*, **1**, 1–19.

Buyse, M., Thirion, P., Carlson, R.W., Burzykowski, T., Molenberghs, G., and Piedbois, P. (2000b). Tumour response to first line chemotherapy improves the survival of patients with advanced colorectal cancer. *Lancet*, **356**, 373–378.

Clayton, D. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65(1)**, 141–151.

Corfu-A Study Group (1995). Phase III randomized study of two fluorouracil combinations with either interferon alfa-2a or leucovorin for advanced colorectal cancer. *Journal of Clinical Oncology*, **13**, 921–928.

Cournil, A., Legay, J., and Schächter, F. (2000). Evidence of sex-linked effects onthe inheritance of human longevity: a population-based study in the Valserine valley (French Jura) $18^{th}$-$20^{th}$ centuries. *Proceedings of the Royal Society of London B*, **267**, 1021–1025

Cox, D.R. (1972). Regression models in life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, **34**, 187–220.

Cox, D.R. and Snell, E.J. (1989). *Analysis of Binary Data*, 2nd edition. London: Chapman and Hall.

Dale, J. R. (1986). Global cross–ratio models for bivariate, discrete, ordered responses. *Biometrics*, **42**, 909–917.

Daniels, M.J. and Hughes, M.D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Statistics in Medicine*, **16**, 1965–1982.

De Beule, O. (1962). *Moerzeke. Religieus-sociale monografie van een plattelandsgemeente.* K.U.Leuven. Leuven

De Boeck, P., Daems, F. Meulders, M., and Rymenams, R. (1997). Onwikkeling van een toets voor de eindtermen begrijpend lezen [Construction of a test for the educational target of reading comprehension]. Leuven/Antwerpen (Belgium): University of Leuven/University of Antwerpen.

De Gruttola, V., Fleming, T.R., Lin, D.Y., and Coombs, R. (1997). Validating surrogate markers – Are we being naive? *Journal of Infectious Diseases*, **175**, 237–246.

Deheuvels, P. (1978). Caractérisation complète des lois extremês multivariées et de a convergence des types extremês. *Publication de l'Institute de Statistique de l'Université de Paris*, **23(3-4)**, 1–36.

De Ridder, J. (1984). *Moerzeke 1710–1796. Een historisch-demografische analyse van een plattelandsparochie in Oost-Vlaanderen.* Oudheidkundigen Kring van het Land van Dendermonde. Dendermonde.

Dolin, R., Amato, D., Fischl, M.A., Pettinelli, C., Beltangady, M., Liou, S., Brown, M.J., Cross, A.P., Hirsch, M.S., Hardy, W.D., Mildvan, D., Blair, D.C., Powderly, W.G., Para, M.F., Fife, K.H., Steigbigel, R.T., Smaldone, L., and the National Institute of Allergy and Infectious Diseases Clinical Trials Group (1995). Zidovudine compared with didanosine in patients with advanced human immunodeficiency virus type I infection and little or no previous experience with zidovudine. *Archives of Internal Medicine*, **155**, 961–974.

Ellenberg, S.S. and Hamilton, J.M. (1989). Surrogate endpoints in clinical trials: cancer. *Statistics in Medicine*, **8**, 405–413.

Fears, T.R., Benichou, J. and Gail, M.H. (1996). A Reminder of the Fallibility of the Wald Statistics. *The American Statistician*, **50**, 226–227.

Finkelstein, D.M., Williams, P.L., Molenberghs, G., Feinberg, J., Powderly, W., Kahn, J., Dolins, R. and Cotton, D. (1996). Patterns of opportunistic infections in patients with HIV infection. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, **12**, 38–45.

Fleming, T.R. and DeMets, D.L. (1996). Surrogate endpoints in clinical trials: Are we being misled? *Annals of Internal Medicine*, **125**, 605–613.

Fleming, T.R., Prentice, R.L., Pepe, M.S., and Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials, with potential applications in cancer and AIDS research. *Statistics in Medicine*, **13**, 955–968.

Franck, M. (1979). On the simultaneous associativity of $F(x, y)$ and $x + y - F(x, y)$ *Aequationes Mathematicae*, **19**, 194–226.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Annals Université de Lyon, Section A, Series 3*, **14**, 143–153.

Freedman, L.S., Graubard, B.I., and Schatzkin, A. (1992). Statistical validation of intermediate endpoints for chronic diseases. *Statistics in Medicine*, **11**, 167–178.

Gail, M.H., Pfeiffer, R., Van Houwelingen, H.C., Carroll, R. (2000). On meta-analytic assessment of surrogate outcomes. *Biostatistics*, **1**, 231–246.

Gavrilova, N.S. and Gavrilov, L.A. (1991) *Biology of Life Span: A Quantitative Approach*. Harwood: Taylor & Francis.

Gavrilova, N., Gavrilov, L., Evdokushkina, G., and Semyonova, V. (1998). Demarcation of the bounderies for human longevity. *The Gerontologist*, **39**, 146–147.

Gavrilova N.S., Gavrilov L.A., Evdokushkina G., Semyonova V.G., Gavrilova A.L., Evdokushkina N.N., Kushnareva T.E., Kroutko V.N., and Andreyew A.Y. (1998). Evolution, mutations and human longevity: European royal and noble families. *Human Biology*, **70**, 799–804.

Gavrilova N.S., Gavrilov L.A., Evdokushkina G., Semyonova V.G. (1998). Mechanisms of familial transmission of human longevity: comparison of maternal and paternal contributions into offspring lifespan. *Paper presented at the Population Association of America 1998 Annual Meeting*, Chicago, Illinois.

Genest, C. and McKay, J. (1986). The joy of copulas: bivariate distributions with uniform marginals. *American Statistician*, **40**, 280–283.

Genest, C. Ghoudi K. and Rivest, L.P. (1995). A semiparametric estimation procedure for dependence parameters in multivariate families of distributions. *Biometrika*, **82(3)**, 543–552.

Geys, H., Molenberghs, G. and Ryan, L. (1999). Pseudo-likelihood modelling of multivariate outcomes in developmental toxicology. *Journal of the American Statistical Association*, **94**, 34–745.

Geys, H. (1999). *Pseudo-likelihood Methods and Generalized Estimating Equations: Efficient Estimation Techniques for the Analysis of Correlated Multivariate Data*. Unpublished Ph.D. dissertation, Limburgs Universitair Centrum.

Geys, H., Molenberghs, G., and Lipsitz, S. (1998). A note on the comparison of pseudo-likelihood and generalized estimating equations for marginally specified odds ratio models with exchangeable association structure. *Journal of Computational Statistics and Simulations*, **62**, 45–71.

Geys, H., Molenberghs, G., and Ryan, L. (1997). Pseudo-likelihood inference for clustered binary data. *Communications in Statistics: Theory and Methods*, **26**, 2743–2767.

Goldstein, H. (1987). Multilevel covariance components models. *Biometrika*, **74**, 430–431.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A*, **159**, 505–513.

Graubard, B.I. and Korn, E.L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, **5**, 263–281.

González, G., Crombet, T., Catalá, M., Mirabal, V., Hernández, J., González, Y., Marinello, P., Guillén, G., and Lage, A. (1998). A novel cancer vaccine composed of human-recombinant epidermal growth factor linked to a carrier protein: Report of a pilot clinical trial. *Annals of Oncology*, **9**, 431–435.

González, G., Crombet, T., Torres, F., Catalá, M., Alfonso, L., Osorio, M., Neninger, E., García, B., Mulet, A., Pérez, R., Lage, A. (2002). Epidermal growth factor based cancer vaccine for non small cell lung cancer therapy. *Annals of Oncology*, 000–000.

Greco, F.A., Figlin, R., York, M., Einhorn, L., Schilsky, R., Marshall, E.M., *et al.* (1996). Phase III randomized study to compare interferon alfa-2a in combination with fluorouracil versus fluorouracil alone in patients with advanced colorectal cancer. *Journal of Clinical Oncology*, **14**, 2674–2681.

Gudmundsson, H., Gudbjartsson, D., Kong, A., Gudbjartsson, H., Frigge, M., Gulcher, J., and Stefánsson, K. (2000). Inheritance of human longevity in Iceland. *European Journal of Human Genetics*, **8**, 743–749.

Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, **61**, 383–385.

Heagerty, P.J. and Lele, S.R. (1998). A composite likelihood approach to binary spatial data. *Journal of the American Statistical Association*, **93**, 1099–1111.

Hjort, N.L. (1993). A quasi-likelihood method for estimating parameters in spatial covariance functions. Technical Report SAND/93, Norwegian Computing Centre, Oslo.

Hougaard, P. (1986). Suvival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73**, 387–396.

Hougaard, P. (2000). Analysis of multivariate survival data. *Statistics for Biology and Health*, Springer-Verlag, New York.

Janssen, R., Tuerlinckx, F., Meulders, M., and De Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal of Educational and Behavioral Statistics*, **25**, 285–306.

Joe, H. (1997). *Multivariates Models and Dependence Concepts.* London: Chapman and Hall.

Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis.* 3rd ed. Englewood Cliffs: Prentice-Hall.

Kahn, J.O., Lagakos, S.W., Richman, D.D., Cross, A., Petinelli, C., Liou, S., Brown, M., Volberding, P.A., Crumpacker, C.S., Beall, G., Sacks, H.S., Merigan, T.C., Beltangady, M., Smaldone, L., Dolin, R., and the NIAID AIDS Clinical Trials Group (1992). A controlled trial comparing continued zidovudine with didanosine in human immunodeficiency virus infection. *New England Journal of Medicine*, **327**, 581–587.

Kaplan, E.L., and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

Kimberling C. (1974). A probabilistic interpretation of complete monotonicity. *Aequationes Math.*, **10**, 152–164.

Korpelainen, H. (1999). Genetic maternal effects on human life span through the inheritance of mitocondrila DNA. *Human Heredity*, **49**, 183–185.

Kuk, A.Y.C. and Nott D.J. (2000). A pairwise likelihood approach to analyzing correlated binary data. *Statistics and Probability Letters*, **47**, 329–335.

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Le Cessie, S. and Van Houwelingen, J.C. (1994). Logistic regression for correlated binary data. *Applied Statistics*, **43**, 95–108.

Lehmann, E.L. (1983). *Theory of Point Estimation.* New York: John Wiley & Sons.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics*, **50**, 325–335.

Lesaffre, E. and Verbeke, G. (1998). Local influence in linear mixed models. *Biometrics*, **54**, 570–582.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Liang, K., Zeger, S. and Qaqish, B. (1992). Multivariate regression analysis for categorical data. *Journal of the Royal Statistical Society, Series B*, **54**, 3–40.

Lindeboom, M. and Van Den Berg, G. (1994). Heterogeneity in Models for Bivariate Survival: the Importance of Mixing Distribution. *Journal of the Royal Statistical Society, Series B*, **56**, 49–60.

Lipsitz, S.R., Laird, N.M., and Harrington, D.P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, **78**, 153–160.

Littell, R.C., Milliken, G.A., Stroup, W.W., and Wolfinger, R.D. (1996). *SAS System for Mixed Models.* Cary, NC: SAS Institute Inc.

Little, R.J.A. and Rubin, D.B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley & Sons.

Mardia, K.V. (1970). *Families of Bivariate Distributions.* London: Griffin.

Marshall, A. and Olkin (1988). Families of multivariate distributions. *Journal of the American Statistical Association*, **83**, 834–841.

Matthijs, K., Van de Putte, B., and Vlietinck, R. (2002). The inheritance of longevity in a Flemish village. *European Journal of Population*, **18**, 59–81.

Molenberghs, G. (1992). *A Full Maximum Likelihood Method for the Analysis of Multivariate Ordered Categorical Data*. Unpublished Ph.D. dissertation, University of Antwerp.

Molenberghs, G., Williams, P.L., and Lipsitz, S.R. (2002). Prediction of survival and opportunistic infections in HIV infected patients: a comparison of imputation methods of incomplete CD4 counts. *Statistics in Medicine*, **21**, 1387–1408.

Molenberghs, G. and Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *Journal of the American Statistical Association*, **89**, 633–644.

Molenberghs, G. and Lesaffre, E. (1999). Marginal modelling of multivariate categorical data. *Statistics in Medicine*, **18**, 2237–2255.

Molenberghs, G., Buyse, M., Burzykowski, T., Renard, D., and Geys, H. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Submitted for publication.*

Molenberghs, G., Geys, H., and Buyse, M. (2001). Evaluation of surrogate endpoints in randomized experiments with mixed discrete and continuous outcomes. *Statistics in Medicine*, **20**, 3023–3038.

Molenberghs, G. and Ryan, L. (1999). An exponential family model for clustered multivariate binary data. *Environmetrics*, **10**, 279–300.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T., and Alonso, A. (2002). Statistical challenges in the evaluation of surrogate endpoints in randomized trials. *Controlled Clinical Trials*, **00**, 000–000.

Nelsen, R.G. (1999). An introduction to copulas. *Lecture Notes in Statistics,* **139**. New York: Springer-Verlag.

Nielsen, G.G., Gill, R.D., Andersen, P.K., and Sørensen, T.I.A. (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics*, **19**, 25–44.

Oakes, D. (1989). Bivariate survival models induced by frailities. *Journal of the American Statistical Association*, **84**, 487–493.

Parner, E.T. (2001). A composite likelihood approach to multivariate survival data. *Scandinavian Journal of Statistics*, **28**, 295–302.

Patterson, H., Thompson, R. (1971). Recovery the inter-block information when blocks sizes are unequal. *Biometrika*, **78**, 609–619.

Plackett, R. L. (1965). A class of bivariate distributions. *Journal of the American Statistical Association*, **60**, 516–522.

Prentice, R.L. (1989). Surrogate endpoints in clinical trials: definitions and operational criteria. *Statistics in Medicine*, **8**, 431–440.

Rao, J.N.K, and Scott, A.J. (1987). On simple adjustments to Chi-square tests with sample survey data. *The Annals of Statistics*, **15**, 385–397.

Raudenbush, S.W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, **18**, 321–349.

Renard, D. (2002). *Topics in modeling multilevel and longitudinal data.* Unpublished Ph.D. dissertation, Limburgs Universitair Centrum.

Rodríguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the American Statistical Society*, **158**, 73–89.

Robins, J.M., Rotnitzky, A., and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.

Rotnitzky, A. and Jewell, P. (1990). Hypothesis testing of regression parameters in semiparametric generalized linear models for clustered correlated data. *Biometrika*, **77**, 485–497.

Royall, R.M. (1986). Model robust confidence intervals using maximum likelihood estimators. *International Statistical Review*, **54**, 221–226.

Saah, A.J., Hoover, D.R., He, Y., Kingsley, L.A., Phair, J.P., and the Multicenter AIDS Cohort Study (1994). Factors influencing survival after AIDS: Report from the Multicenter AIDS Cohort Study (MACS). *Journal of Acquired Immune Deficiency Syndromes*, **7**, 287–295.

SAS Institute Inc. (1995). *SAS/IML Software: Changes and Enhancements Through Release 6.11.* Cary, NC: SAS Institute Inc.

Schweizer, B. and Wolff, E.F. (1981). On nonparametric measures of dependence for random variables. *Annals of Statistics*, **9** 879–885.

Seber, G.A.F. (1984) *Multivariate Observations*. New York: John Wiley.

Shih, J.H.and Louis, T.A. (1995). Inferences on association parameter in copula models for bivariate survival data. *Biometrics*, **51**, 1384–1399.

Smith, D.C., Dunn, R.L., Stawderman, M.S., and Pienta, K.J. (1998). Change in serum prostate-specific antigen as a marker of response to cytotoxic therapy for hormone-refractory prostate cancer. *Journal of Clinical Oncology*, **16**, 1835–1843.

Sørensen, T.I.A., Nielsen, G.G., Andersen, P.K., and Teasdale, T.W. (1988). Genetic and familial environmental influence on premature death of adults adoptees. *New England Journal of Medicine*, **318**, 727–732.

Spiegelhalter D.J., Thomas, A., Best, N.G., and Gilks, W.R. (1995). *BUGS Manual and Examples: Version 0.50.* Cambridge: MRC Biostatistics Unit, Institute of Public Health, University of Cambridge.

Tanaka, M. and Gong, J.S. (1998). Mitochondrial genotype associated with longevity. *Lancet*, **351**, 9097, p.185.

Tibaldi, F., Molenberghs, G., Burzykowski, T., Geys, H. (2003). Pseudo-likelihood estimation for a marginal multivariate survival model. *Statistics in Medicine*, **00**, 000–000.

Tibaldi, F., Cortiñas Abrahantes, J., Molenberghs, G., Renard, D., Burzykowski, T., Buyse, M., Parmar, M., Stijnen, T., and Wolfinger, R. (2003). Simplified hierarchical linear models for the evaluation of surrogate endpoints. *Journal of Statitical Computation and Simulation*, **73**, 643–658.

Tibaldi, F., Verbeke, G., Molenberghs, G., Renard, D., Van de Noortgate, W., De Boeck, P. (2003). Conditional linear mixed models with crossed-random effects. *Submitted for publication.*

Tibaldi, F., Molenberghs, G., Verbeke, G., Renard, D., Van de Noortgate, W., De Boeck, P. (2003). Conditional linear mixed models with crossed-random effects for binary data. *In preparation.*

Tibaldi, F., Van de Putte, B., Geys, H., Molenberghs, G., Matthijs, K., Vlietinck, R.(2003). Multivariate Plackett-Dale Inference to Study the Inheritance of Longevity in a Belgian Village ($18^{th}$-$20^{th}$ Century) *Submitted for publication.*

Tibaldi, F., Van de Putte, B., Geys, H., Molenberghs, G., Matthijs, K., Vlietinck, R.(2003). Multivariate Plackett-Dale Inference to Study the Inheritance of Longevity in a Belgian Village ($18^{th}$-$20^{th}$ Century)In: *Proceedings of the 18th Intyernational Workshop on Statistical Modelling, Verbeke, G., Molenberghs, G., Aers, M. and Fieuws, S.* (Eds.). Leuven: Katholieke Universiteit Leuven, pp. 415–420.

Tibaldi, F., Torres Barbosa, F., Molenberghs, G.(2002). Modelling associations between time-to-event responses in pilot cancer clinical trials using a Plackett-Dale model. *Statistics in Medicine*, **00**, 000–000.

Tibaldi F., Bruckers L., Van Oyen H., Van der Heyden J., and Molenberghs G. (2003). Statistical software for calculating properly weighted estimates from health interview survey data. *International Journal of Public Health, Hints and Kinks*, **48**(4).

Torres, F., Gonzàlez, G., Crombet, T., and Lage, A. (2001). Empirical Bayes analysis of early clinical trials for cancer vaccines. *Submitted for publication.*

Torres, F., Tibaldi, F., Cortiñas Abrahantes, J., Geys, H., González, G., and Molenberghs, G. (2002). Surrogacy Evaluation of Immunological Parameters in Pilot Cancer Clinical Trials. *Submitted for publication.*

Tsiatis, A.A., De Gruttola, V., and Wulfsohn, M.S. (1995). Modeling the relationship of survival to longitudinal data measured with error. Applications to survival and CD4 counts in patients with AIDS. *Journal of the American Statistical Association*, **90**, 27–37.

Van den Noortgate, W., De Boeck, P., and Meulders, M. (2003). Cross-classification multilevel logistic models in psychometry. *Journal of Educational and Behavioral Statistics*, **00**, 000–000.

Van Houwelingen, J.C., Arends, L.A., Stijnen, T. (2002). Advanced methods for meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*, **21**, 589–624.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects distribution. *Journal of the American Statistical Association*, **91**, 217–221.

Verbeke, G., Spiessens, B., and Lesaffre, E. (2001). Conditional linear mixed models. *The American Statistician*, **55**, 25–34.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*. New York: Springer-Verlag.

Wang, W. and Wells, M.T. (2000). Model selection and semiparametric inferences for bivariate failure-time data. *Journal of the American Statistical Association*, **95**, 62–76.

Wolfinger, R.D. and O'Connell, M. (1993). Generalized linear mixed models: a pseudo-likelihood appoach. *Journal of Statistical Computation and Simulation*, **48**, 233–243.

Zhao, L. and Prentice, R. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–648.

Zeger, L. and Liang, K. (1986). Longitudinal data analysis for discrete and continouos outcomes. *Biometrics*, **42**, 121–130.

Zeger, L., Liang, K. and Albert, P. (1986). Models for longitudinal data: a generalized estimated equation approach. *Biometrics*, **44**, 1049–1060.

# Samenvatting

## Statistische Modellen en Gecorreleerde Gegevens

In deze sectie geven we een overzicht van het statistisch modelleren in de context van gecorreleerde gegevens, zowel van overlevingstijden, als van continue en categorische gegevens. Het hoofddoel in heel wat gebieden van de statistiek is een model opbouwen om een afhankelijke of responsvariabele te verklaren. Dit kan al of niet gebeuren in aanwezigheid van covariaten of verklarende variabelen. In een aantal gevallen gaat de interesse verder dan dit objectief en is het doel de studie van correlatie tussen responsvariabelen onderling. Juist in situaties waar gegevens in groepen of clusters worden verzameld, is de associatie vaak doel van wetenschappelijke vraagstelling. Naast de mogelijkheid interessante wetenschappelijke vragen te beantwoorden, wordt er ook een veelheid aan modelleringsvragen opgeroepen.

Inderdaad, het analyseren van dergelijke gegevensstructuren brengt een zekere complexiteit met zich mee. Onderzoek naar het modelleren van gecorreleerde gegevens heeft enorm aan populariteit gewonnen gedurende de laatste jaren. Deze modellen behoren tot het vitale en zeker interessante hart van het statistisch modelleren. *Estimating equations* (schattingsvergelijkingen) en modellen met *random effects* (toevalseffecten) zijn ongetwijfeld de twee populairste aanpakken om gegevens van een dergelijk type te analyseren, in een groot aantal toepassingsgebieden. Longitudinale toepassingen zijn heel belangrijk, maar ook ruimtelijke gegevens komen veel voor, in de medische wetenschappen, genetische studies, klinische studies, volksgezondheid, economie, politieke wetenschappen en sociologie.

Aanvankelijk werd veel onderzoek verricht voor continue gegevens, maar een deel van de complexiteit ontstaat omdat men gegevens van verscheidene types ontmoet, soms zelfs in één en dezelfde studie. Gegevens kunnen, naast continu (normaal verdeeld), ook binair zijn, categorisch, of overlevingstijden. Random-effects modellen, zoals het lineair gemengd model, werden ontwikkeld voor continue gegevens,

maar aanpassingen zijn nodig voor elk ander type van responsvariabele.

Aangepaste modellen werden ontwikkeld in de context van herhaalde metingen (longitudinale gegevens, Verbeke en Molenberghs, 2000) en van cluster gegevens (Aerts *et al.*, 2002). Nochtans werden deze methoden in hoofdzaak ontwikkeld voor klassieke gegevensstructuren. Plausibele en flexibele alternatieven zijn nodig voor complexere structuren, en precies daaraan draagt deze thesis bij.

Uiteraard werd er reeds heel wat werk verricht. Toen Liang en Zeger (1986), Zeger en Liang (1986) en Zeger, Liang en Albert (1988) hun *generalized estimating equations* (GEE, veralgemeende schattingsvergelijkingen) voorstelden voor gecorreleerde binaire gegevens, werd het een onmiddellijke succes. Als de wetenschappelijke vraag beperkt blijft tot eerste momenten, kan men gebruik maken van GEE1. Toch zijn er situaties waar ook de tweede momenten (associaties) van belang zijn, en daartoe stelden Zhao en Prentice (1990) en Liang, Zeger en Qaqish (1992) gepaste uitbreidingen voor: de GEE2 methode. Uitgaande van deze modellen werden heel wat alternatieve marginale modellen voorgesteld, waarbij de klemtoon ligt op het efficiënt schatten van de effecten van covariaten op marginale kansen, op de verwachte waarde van een aantal, enz.

Een aantrekkelijk alternatief voor GEE, ontwikkeld in de context van multivariate gecorreleeerde binaire gegevens, bestaat erin van likelihood te vervangen door pseudo-likelihood (Geys 1999), waarbij de echte multivariate dichtheidsfunctie vervangen wordt door een veel makkelijker te manipuleren product van marginale of conditionele dichtheden. Reeds bij herhaalde meetreeksen van een matige lengte blijkt de normaliseringsconstante van bepaalde types van loglineaire modellen zo goed als onoverkomelijke computationele vereisten te stellen. Een gepaste keuze van de pseudo-likelihoodfunctie vermijdt het voorkomen van een dergelijke constante. Sterke punten van deze methode zijn, naast het vereenvoudigen van de berekeningen: (1) de interpretatie van de modelparameters wijzigt niet en (2) de efficiëntie van de methode is in veel realistische situaties niet noemenswaardig lager dan wanneer likelihood gebruikt wordt. Uit onderzoek van Geys (1999) blijkt dat deze beweringen breed geldig zijn.

## Gegevensstructuren

We geven een overzicht van de diverse hiërarchische gegevensstructuren die zich voordoen, naast in de studies beschouwd in deze thesis, ook meer algemeen. Dit sluit aan bij de modelleringsconcepten van de vorige sectie.

Naast gegevens afkomstig van observationele studies, verzameld in humane en biologische experimenten, zijn er nog heel wat voorbeelden te noemen van gecorreleerde, cluster, of hiërarchische structuren. Bijvoorbeeld, erfelijkheidsstudies in mens en dier

leiden bijna altijd tot een hiërarchie, waarbij nakomelingen familie-gerelateerde clusters vormen. Uiteraard hebben nakomelingen van dezelfde ouders de neiging om gelijkaardig te scoren in hun lichamelijke en geestelijke kenmerken, dan wanneer men individuen vergelijkt die uit totaal verschillende families afkomstig zijn. Een eenvoudig voorbeeld bestaat in de vaststelling dat kinderen binnen eenzelfde familie dezelfde neiging hebben voor een grote, gemiddelde, of kleine lichaamsbouw. Naast genetische factoren spelen hierbij typisch ook omgevingsfactoren een belangrijke rol.

Ook in survey gegevens ontmoet men op zeer regelmatige basis hiërarchische gegevens, omdat het studie-opzet daar in vele gevallen expliciet in voorziet. Populaire manieren om een dergelijke situatie aan te pakken omvatten (1) het gebruik van het opzet om te corrigeren voor aspecten zoals clustering, het voorkomen van hiërarchische gegevens, e.d. en (2) het expliciet modelleren van de correlatie bij hiërarchische gegevens (Tibaldi *et al.* 2003).

Andere voorbeelden omvatten klinische studies, uitgevoerd in verschillende groepen van individuen, zoals in verschillende centra. Recent is hieromtrent veel onderzoek verricht en dit werk beschouwt dergelijke studies met als oogmerk de evaluatie van surrogaatrespons. Als de primaire respons duur is (zoals in het geval van bepaalde laboratoriumwaarden) of als het verzamelen ervan veel tijd in beslag neemt (zoals overlevingstijd), wensen onderzoekers over te schakelen op een zogenaamde surrogaatrespons. Natuurlijk kan dit niet zomaar gebeuren en dient men na te gaan of de potentiële surrogaatrespons voldoende goed is, d.w.z. een voldoende betrouwbare vervanger is voor de primaire respons. Al gedurende meer dan een decennium wordt onderzoek gedaan naar dergelijke gegevens en deze thesis richt zich specifiek op continue, normaal verdeelde surrogaat- en primaire respons. Daarnaast zullen we aandacht schenken aan het geval van overlevingstijden.

Aan de andere kant vinden we gecorreleerde gegevens in familiestudies, waar correlatie afkomstig is van karakteristieken, gemeenschappelijk aan alle familieleden. Eerst introduceren we een adoptiestudie, waar de associatie tussen overlevingstijden van biologische en adoptieve familie wordt gemodelleerd. Vervolgens presenteren we een zogenaamde *longevity* studie, waar gelijkaardige karakteristieken gevonden worden in een grote dataset. In heel wat andere gebieden komt associatie tussen overlevingstijden voor. Ter illustratie bestuderen we correlatie tussen overlevingstijden van dezelfde patiënt, in een klinische studie.

Aan het einde van dit werk concentreren we ons op gegevens afkomstig van een psychometrische studie. Meer in het bijzonder ligt de klemtoon op zogenaamde gekruiste random effecten. Er worden een aantal vragen (items) gesteld aan een aantal studenten, en voor beide, gekruiste, niveaus worden random effecten voorzien. Daar

waar dergelijke gegevens, mits gepaste keuzes, nog vrij vlot kunnen gemodelleerd en geanalyzeerd worden in het geval van continue gegevens, doen zich bijkomende complicaties voor wanneer de respons binair of categorisch is. In het bijzonder besteden we aandacht aan het binaire geval, waar pseudo-likelihood wordt gebruikt, samen met concepten uit conditionele logistische regressie, om dit probleem aan te pakken.

Alle studies hebben een gemeenschappelijk kenmerk: **complexe, gecorreleerde gegevensstructuren**. Likelihood methodologie voor complexe hiërarchische modellen wordt aangevuld met concepten uit pseudo-likelihood, conditionele regressie, en vereenvoudigde methoden, om een grote klasse van dergelijke problemen aan te pakken. Uit de toepassing van deze methoden vloeit nieuw inzicht voor op basis van bestaande en reeds voorheen geanalyzeerde gegevens. Natuurlijk kunnen technieken maar toegepast worden indien betrouwbare en gebruiksvriendelijke software aanwezig is. Aan de hand van SAS routines wordt de toepassing van de voorgestelde methoden mogelijk.

Naast gecorreleerde overlevingstijden concentreren we ons ook op continue en op binaire gegevens. Toepassingen worden gevonden in de klinische studies, in het bijzonder met het oog op de validering van surrogaatrespons, de populatie-genetische studies en de psychometrische studies.